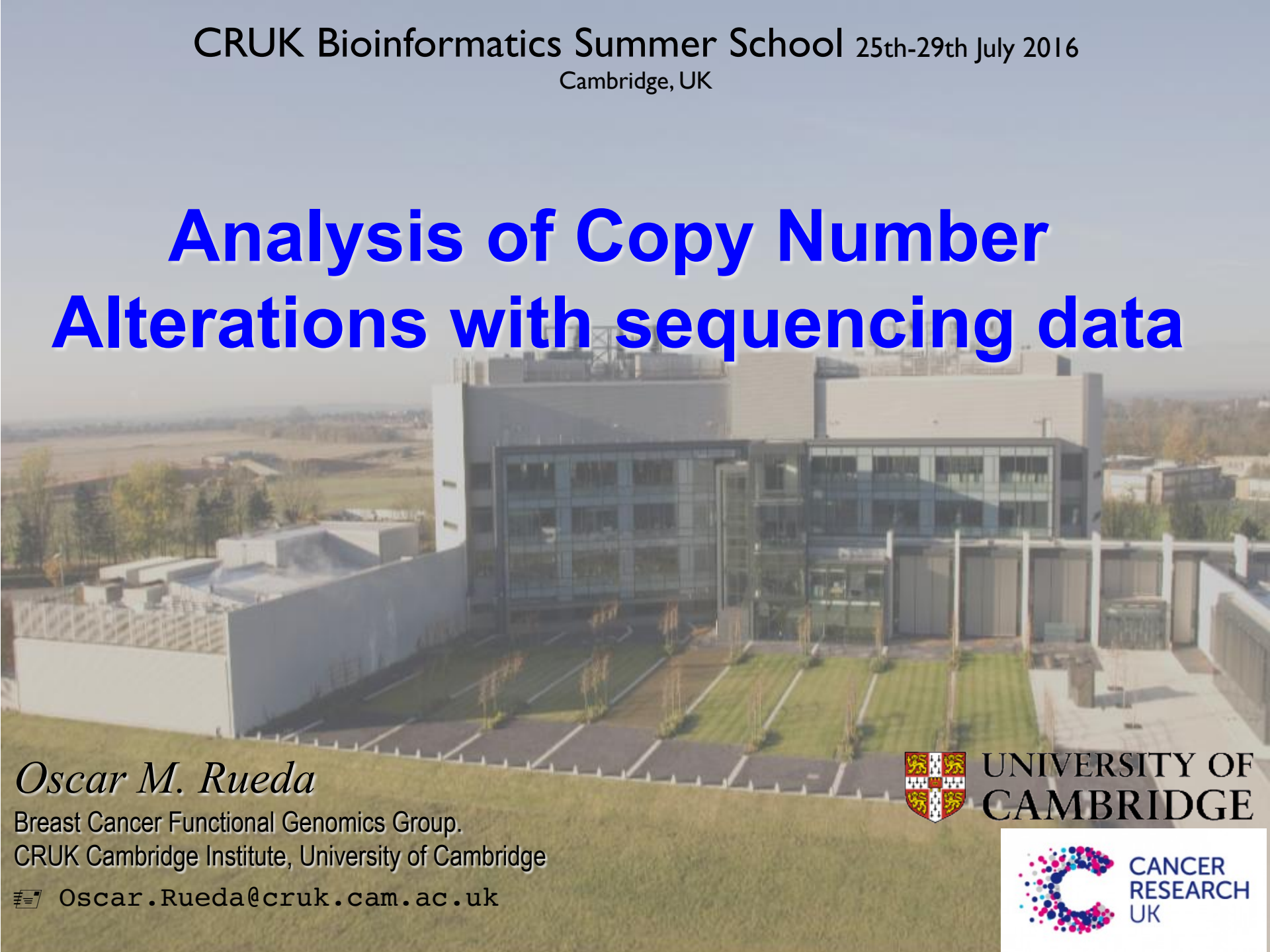
CRUK Bioinformatics Summer School 25th-29th July 2016

Cambridge, UK

Analysis of Copy Number Alterations with sequencing data



Oscar M. Rueda

Breast Cancer Functional Genomics Group.

CRUK Cambridge Institute, University of Cambridge

Oscar.Rueda@cruk.cam.ac.uk

UNIVERSITY OF CAMBRIDGE

CANCER RESEARCH UK

Overview

- Introduction.
- Methods for summarisation and normalisation.
- Methods for CN based on read depth.
- Methods for CN based on read depth and minor allele frequency.

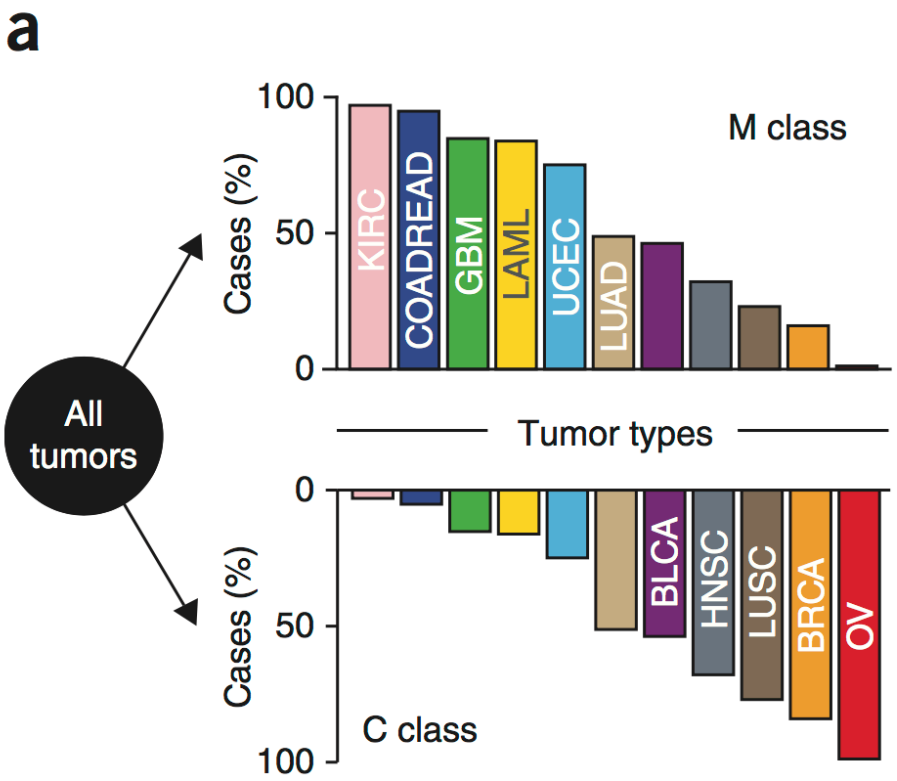
Introduction

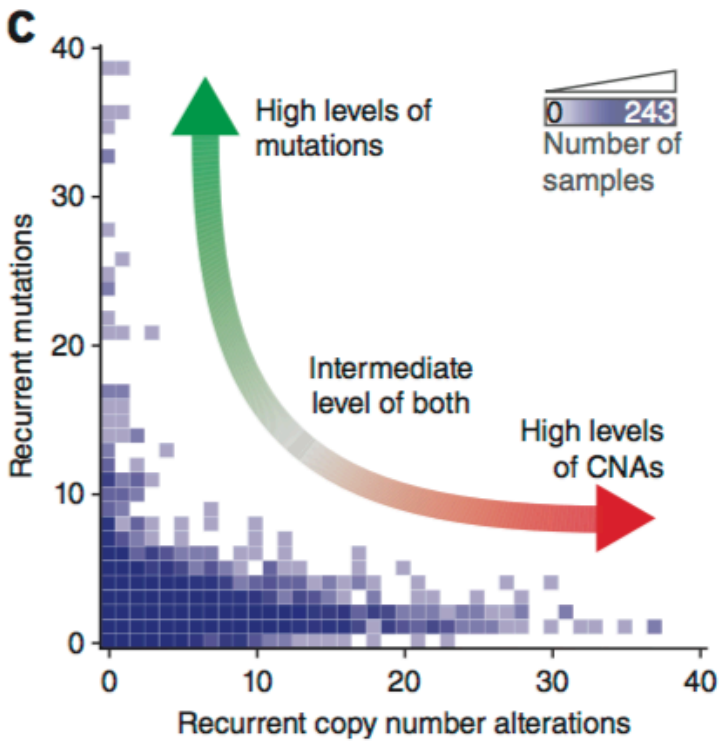
Copy number alterations

- 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 chemicals or viruses…

CNAs are very common in cancer

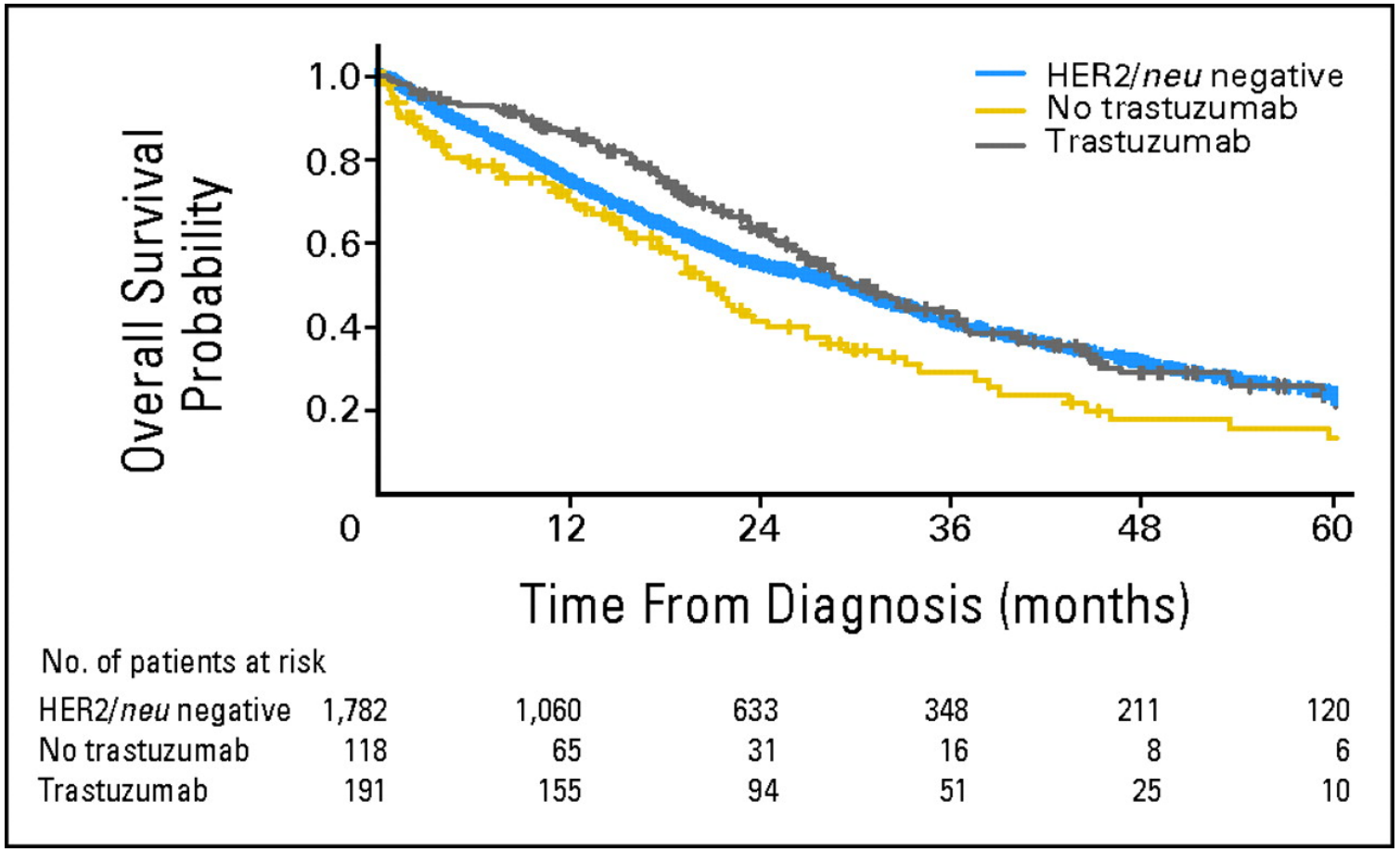




Ciriello et al (2015). Nature Genetics

CNAs are important for treatment

Overall survival by trastuzumab treatment group.



No. of patients at risk

HER2/*neu* negative	1,782	1,060	633	348	211	120
No trastuzumab	118	65	31	16	8	6
Trastuzumab	191	155	94	51	25	10

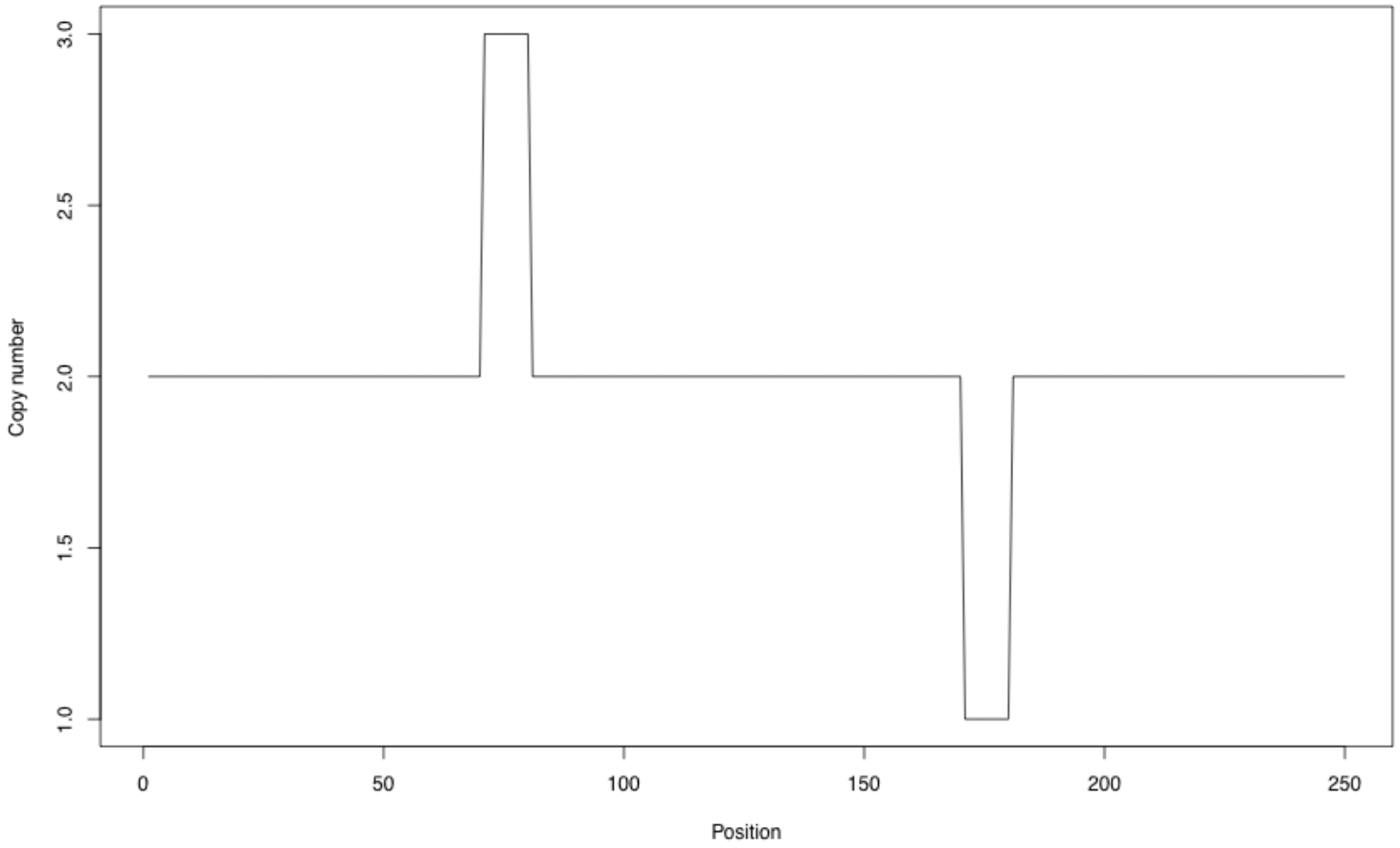
Shaheenah Dawood et al. JCO 2010;28:92-98

CNVs and CNAs

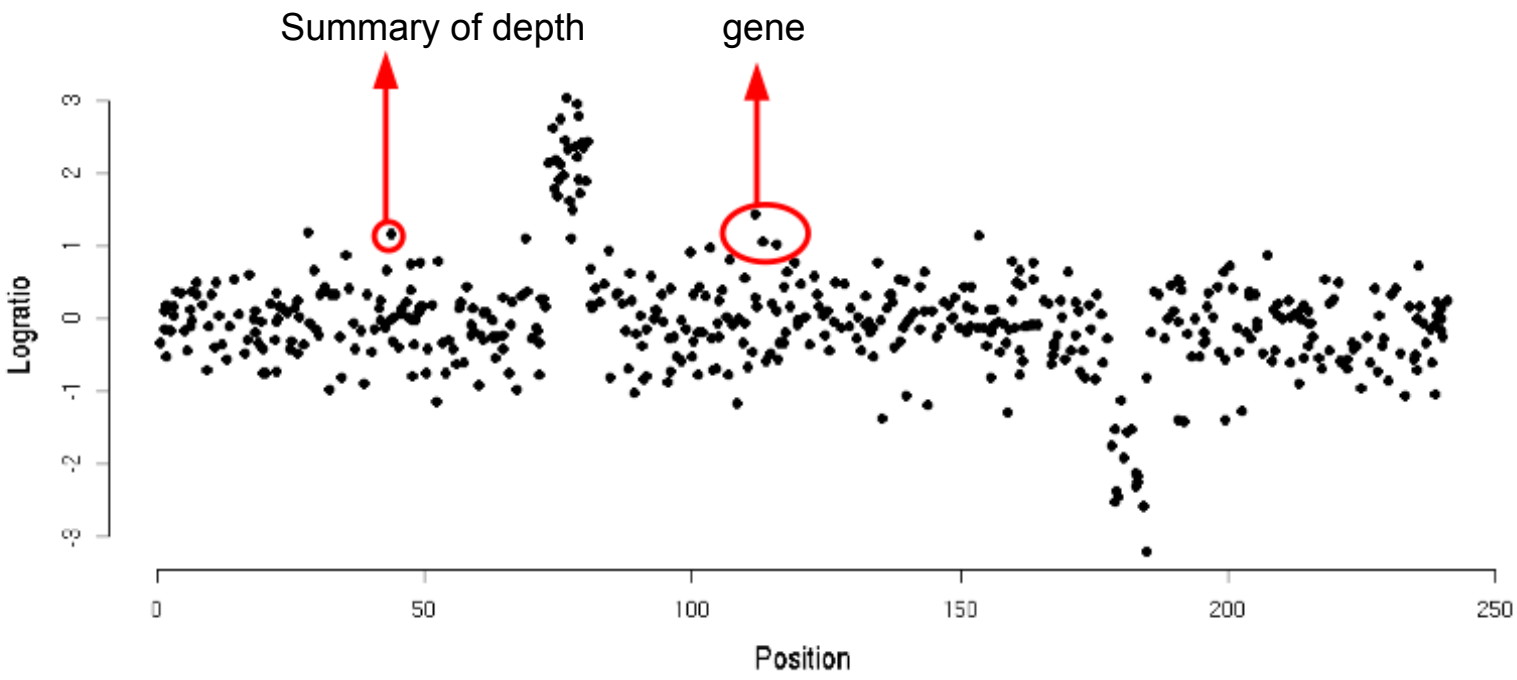
- Copy Number Alterations is a generic name for Copy Number Variations and Copy Number Aberrations.
- **Copy Number Variations (CNVs):** Germline alterations, individual and not disease related.
- **Copy Number Aberrations (CNAs):** Somatic alterations, disease related.

We need the pair to distinguish germline from somatic!!!

Copy number alterations



Data obtained from sequencing reads



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 regions are
- Some regions may present **specific effects** due to GC content, target enrichment, etc that may correlate across different samples.

Different approaches to sequencing

- **Whole genome sequencing:** reads from the complete DNA sequencing of the sample. WGS with low coverage is sometimes called "**shallow sequencing**"
- **Exome sequencing**: reads from the protein-coding genes in the genome
- **Target sequencing**: reads from a subset of genes in the genome.

Methods for summarisation and normalisation

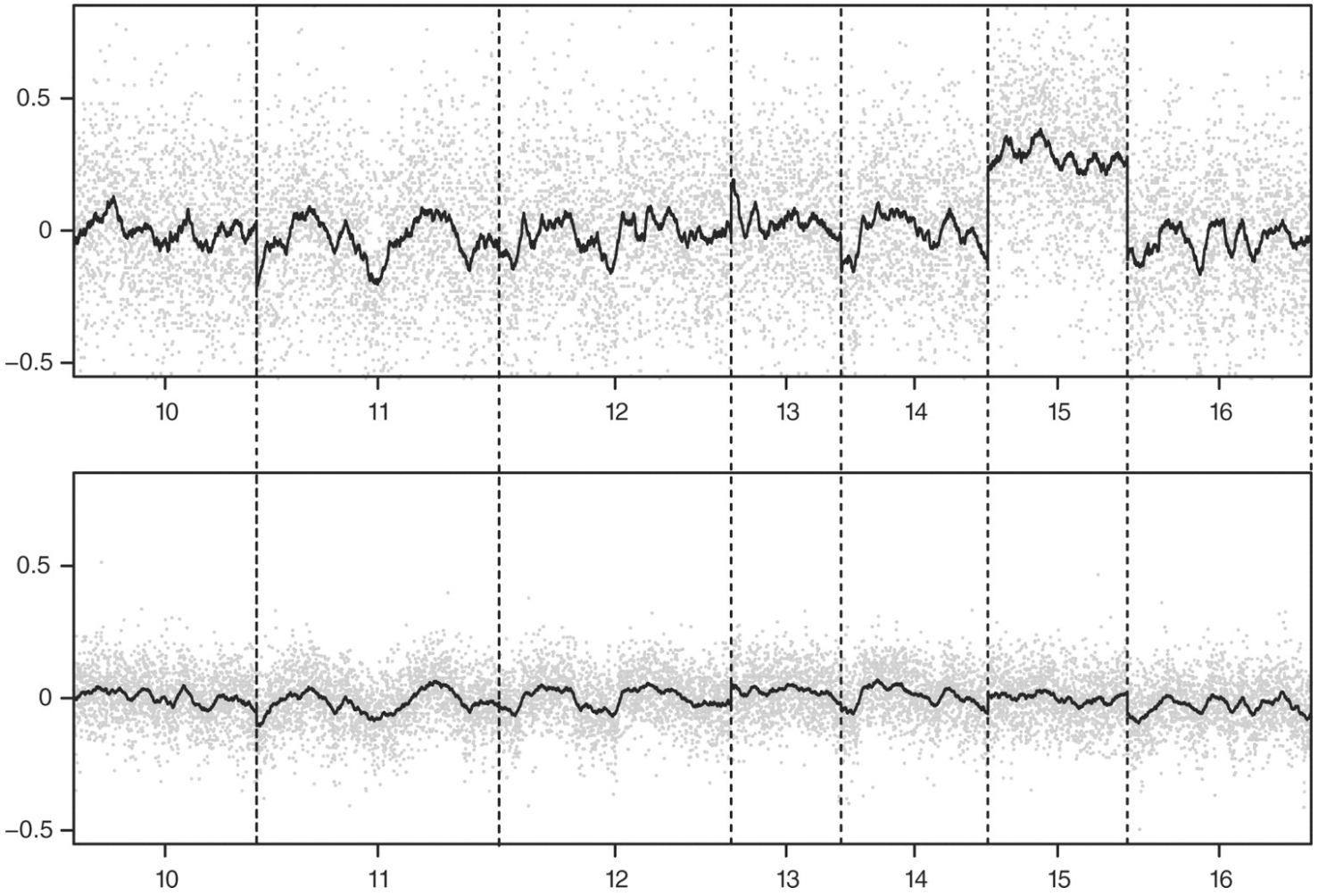
Basic premise

- Single nucleotide depths can be very noisy
- We can reduce that noise taking bins of a given length across the genome and adding the total read depth
- The length of the bins can depend on the sequencing method used (for example, the baits in exome/target), our coverage depth, or the resolution needed in copy number estimation.

Filtering genomic regions

- **Uncharacterized bases**
- **Repetitive regions**
- **Unmappable regions**

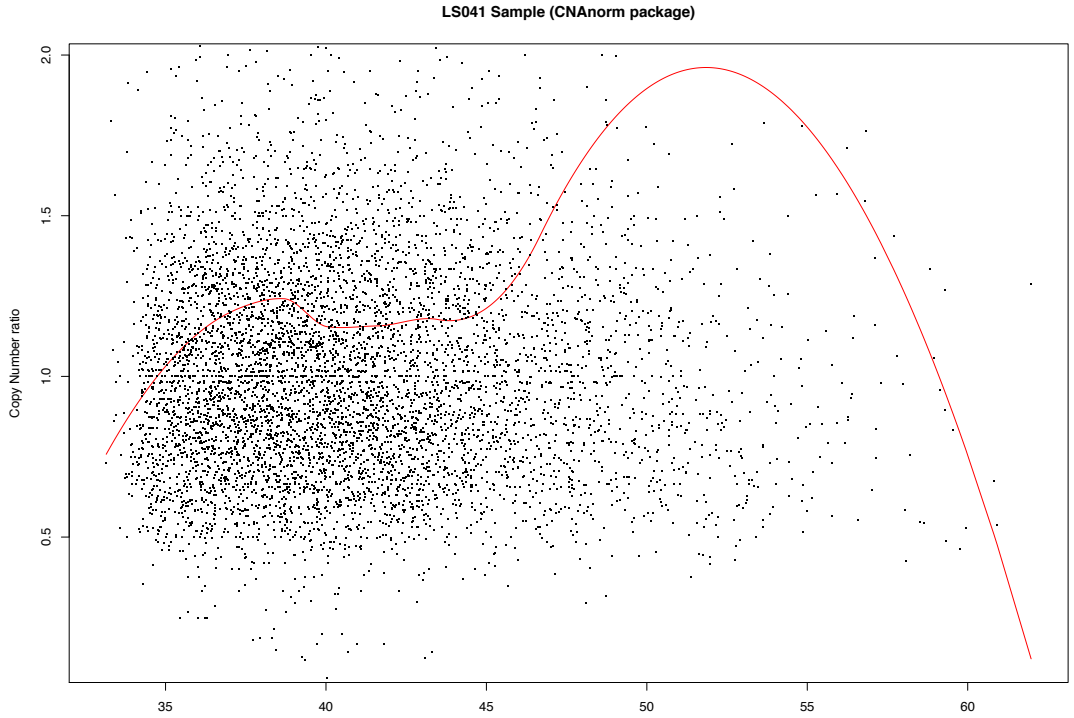
Wave effect



Mark A. van de Wiel et al. Bioinformatics 2009;25:1099-1104

GC content normalisation

- Different proportions of GC in each region can produce a bias in the read depth (wave artifact)
- We can fit a loess model and remove the effect.



Centering

Common practices:

- Median centering around zero
- Data is transformed to log2 ratios to reflect comparison against a diploid reference
- The assumption in some normalisation methods that the proportion of altered probes is the same for each sample is **NOT** true.

Target normalisation

- In exome/target sequencing different targets can have non-uniform read-depth
- We expect that these enrichment effects are correlated across samples, therefore we can estimate these effects
- A background normalisation can help mitigate these effects

Background normalisation

- We need a sample or a set of samples that represent the expected profile of a diploid genome
- It can be a matched normal sample from the same tissue or from blood in the case of a tumour sample, or a pool of normal samples
- We compute the ratios between the sample and the control (or sometimes the log2 ratios).

Methods for copy number analysis based on read depth

Two steps:

1. Segmentation
2. Calling

Copy Number Segmentation

Segmentation methods

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

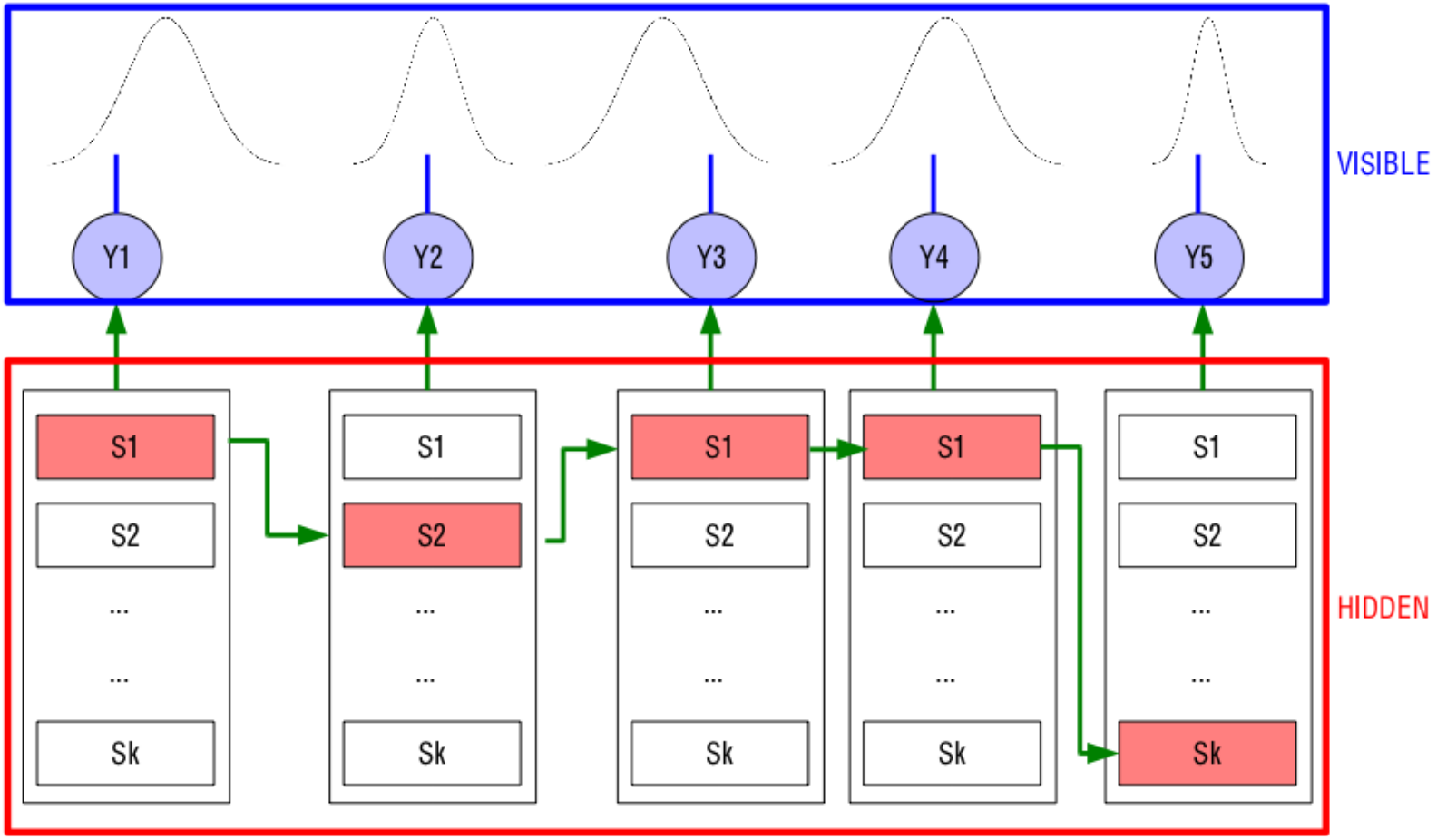
From ratios or $\log_2$ ratios to segmented means: $y_t \Rightarrow m_t$

- **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)

Algorithms for segmentation

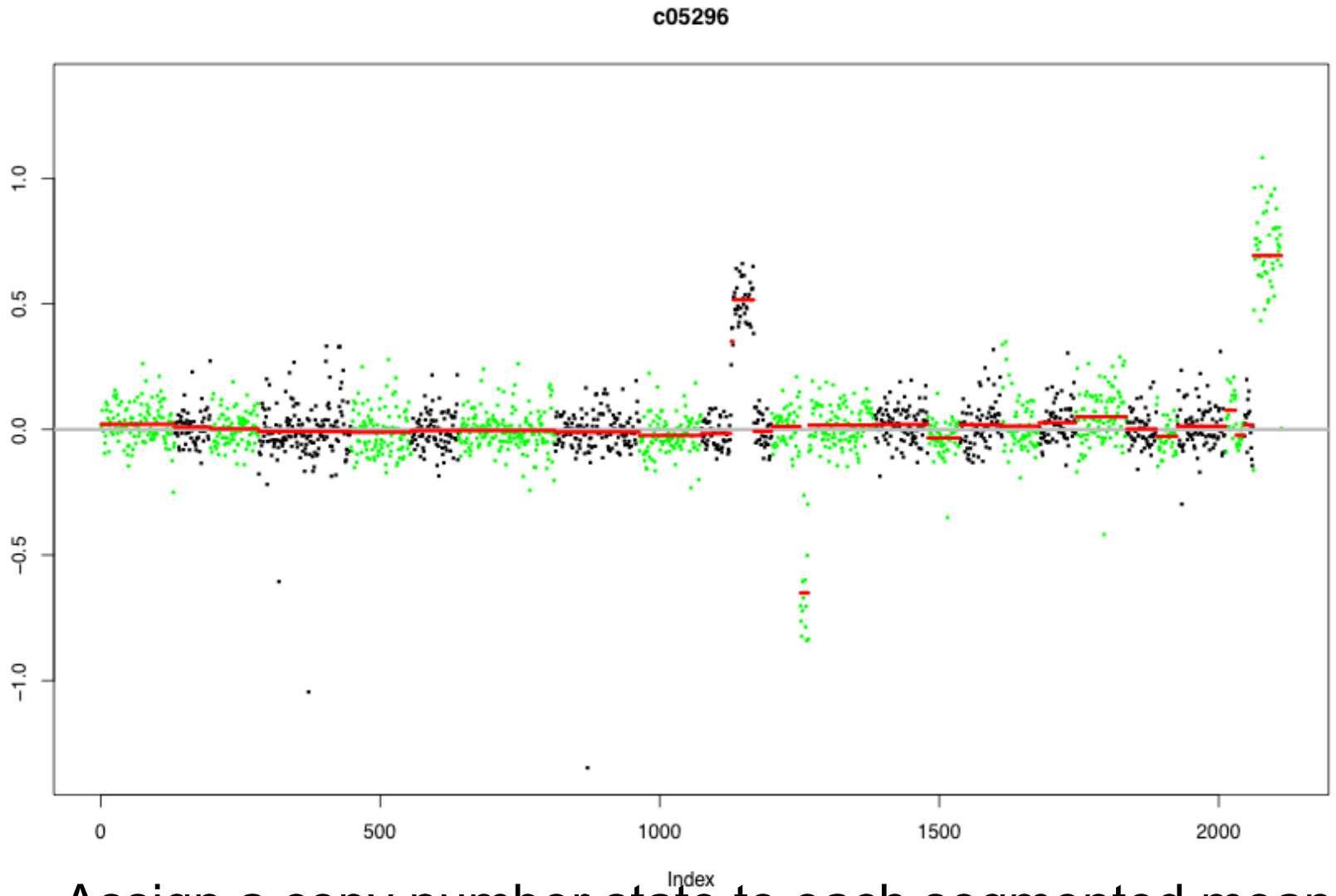
- **Circular Binary Segmentation (CBS)**
  - Olshen et al., 2004.
  - It can be used with array and sequencing data
  - Finds change points using a t-test under a permutation model.
  - Bioconductor package DNAcopy.

Hidden Markov Models (HMMs)



Copy Number Calling

Calling of gains and losses



Assign a copy number state to each segmented mean.

Threshold-based methods

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

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

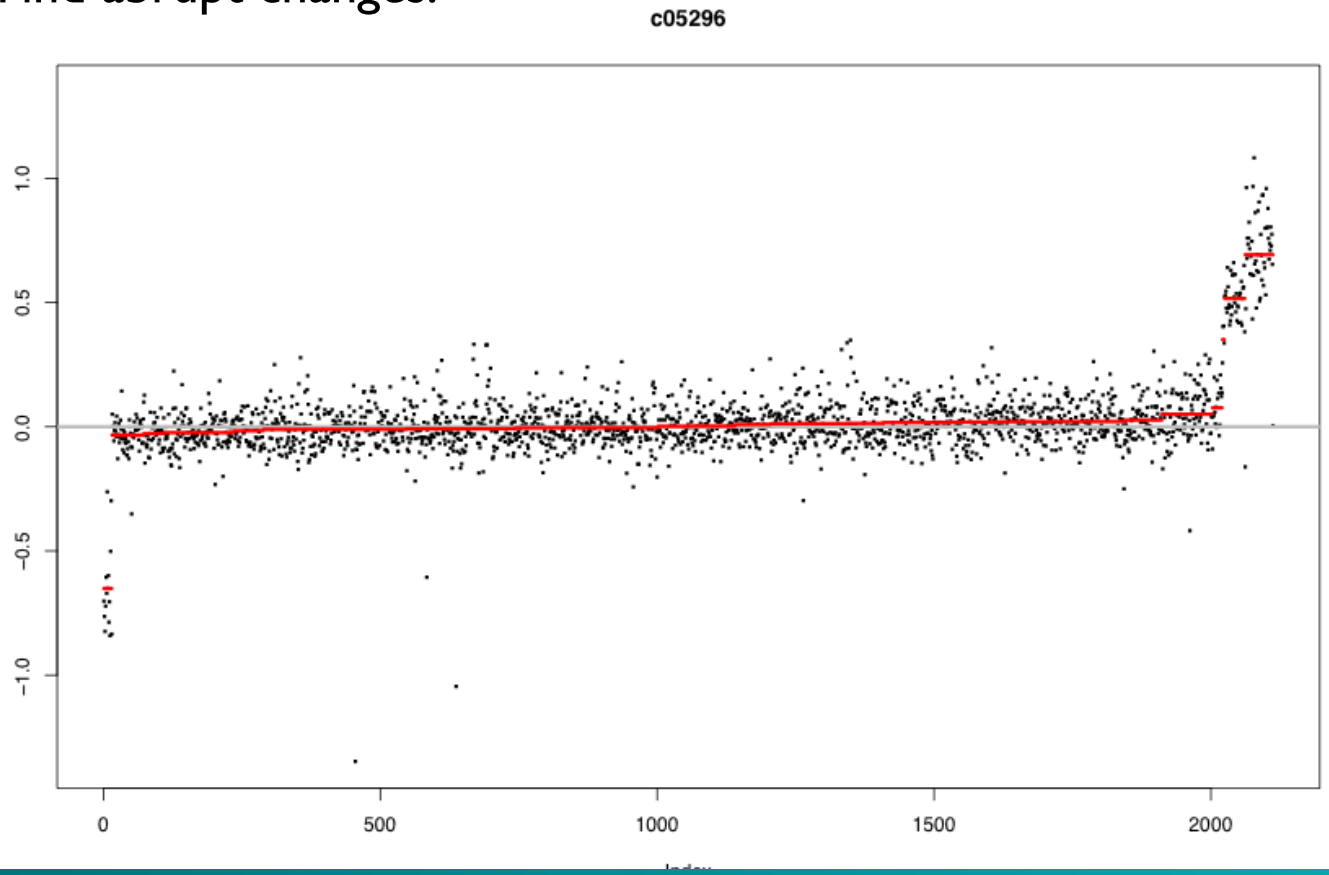
$$t / m_t \leq \bar{y} - k_L \sigma_Y \rightarrow LOSS$$

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

Plateau plots

Olshen and Venkatraman, 2005 (DNAcopy R package).

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



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.

Methods specific for sequencing data

QDNAseq

Scheinin I et al., 2014 (QDNAseq Bioconductor package).

- Divides genome into bins of equal size.
- Normalisation based on blacklisted regions, GC content,....
- Segmentation with DNAcopy.
- Optional calling with CGHcall.

CopywriteR

Kuilman et al., 2016 (CopywriteR Bioconductor package).

- Appropriate for exome/targeted sequencing: it uses the off-target reads
- Peak calling, removal of reads in peaks
- Divides genome into bins of equal size.
- Normalisation based on blacklisted regions, GC content,....
- Segmentation with DNAcopy.

Methods for copy number analysis based on read depth and variant allele frequency

Variant allele frequency

- We can gain information about the copy number of sample if we incorporate the variant (minor) allele frequency of a list of SNPs:

  A: common allele (reference)
  B: minor allele (alternate)

  AA: sample is (reference) homozygous for that SNP
  AB: sample is heterozygous for that SNP
  BB: sample is (alternate) homozygous for that SNP

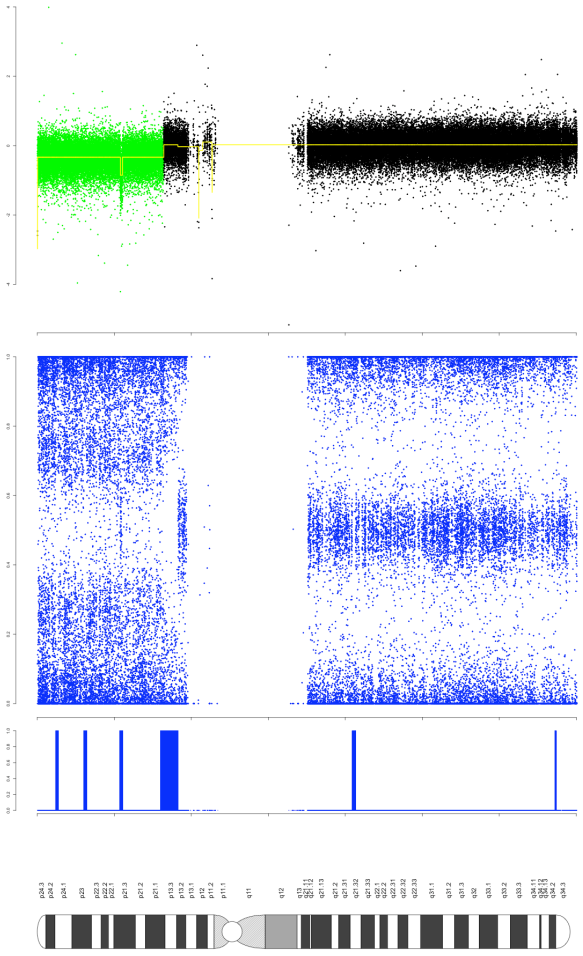
  $$vaf = \#reads(B)/(\#reads(A) + \#reads(B))$$

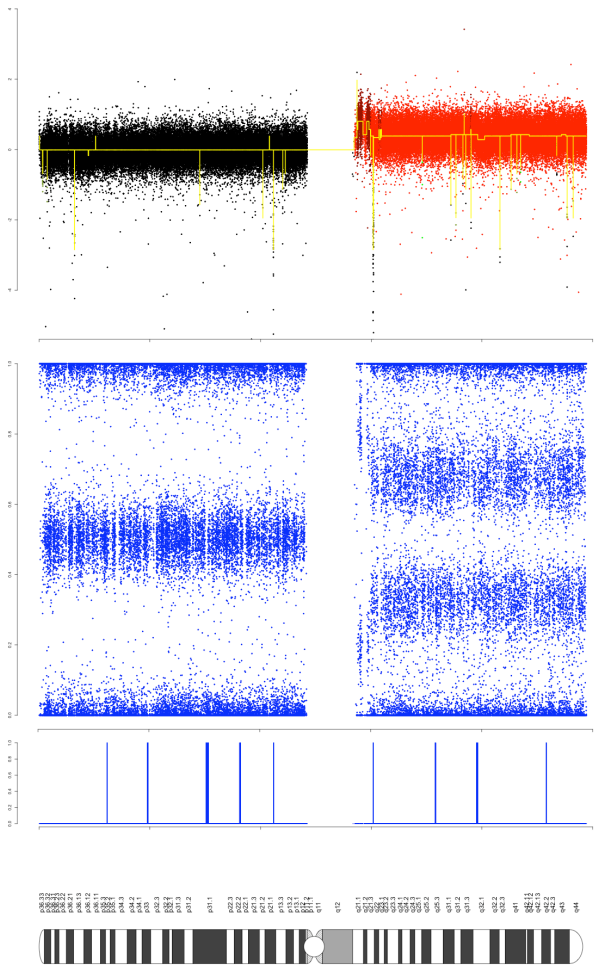
- Now we have two sets of data (similar to SNP arrays):
  - Ratios
  - vafs

VAF 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)/ $i < j$

VAF patterns help with copy number calling





Realistic scenarios

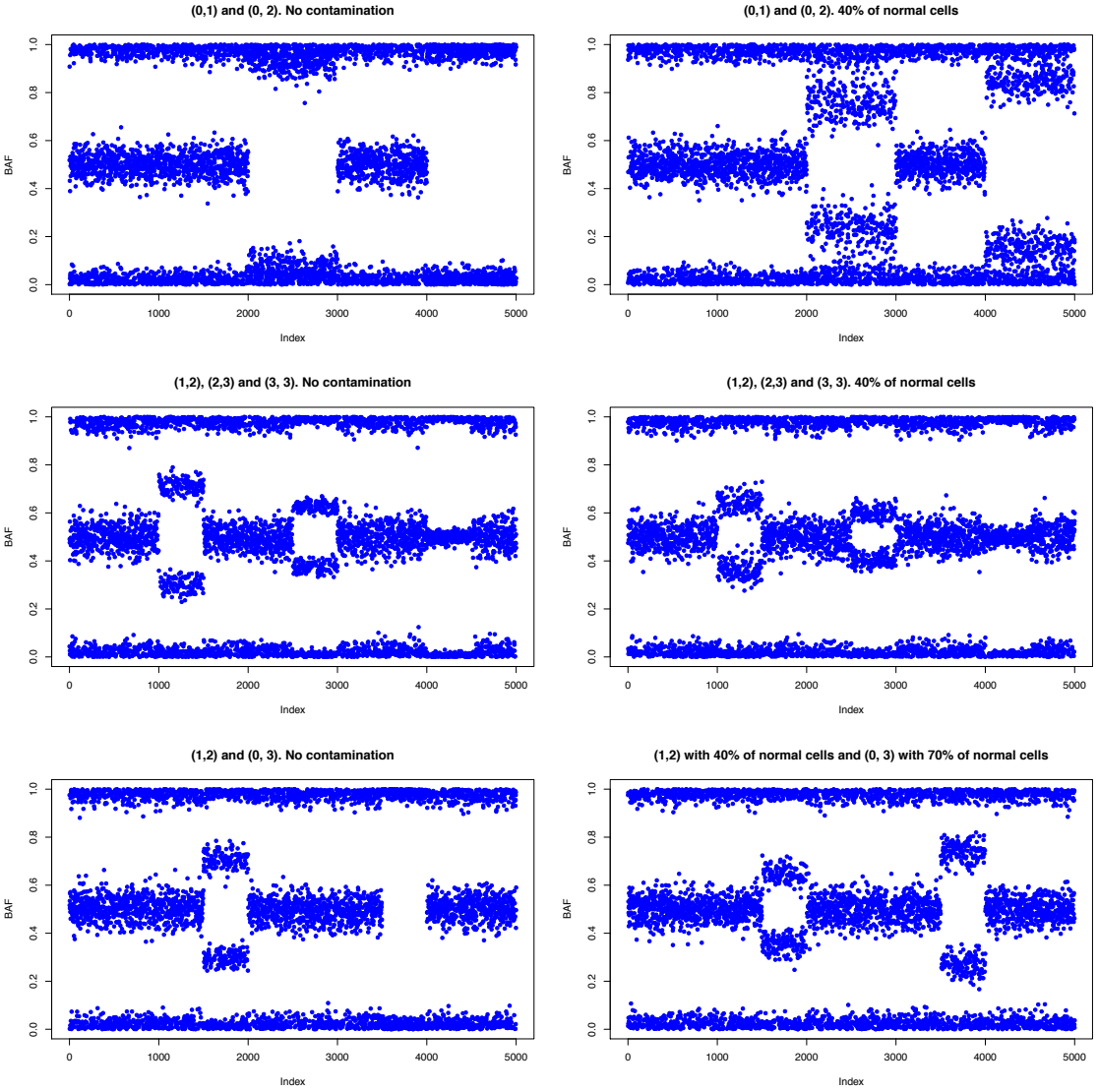
- **Aneuploidy**
  - The baseline of a sample is not 2 copies.
- **Normal contamination**
  - Only a given percentage of the cells in our sample are tumor cells:

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

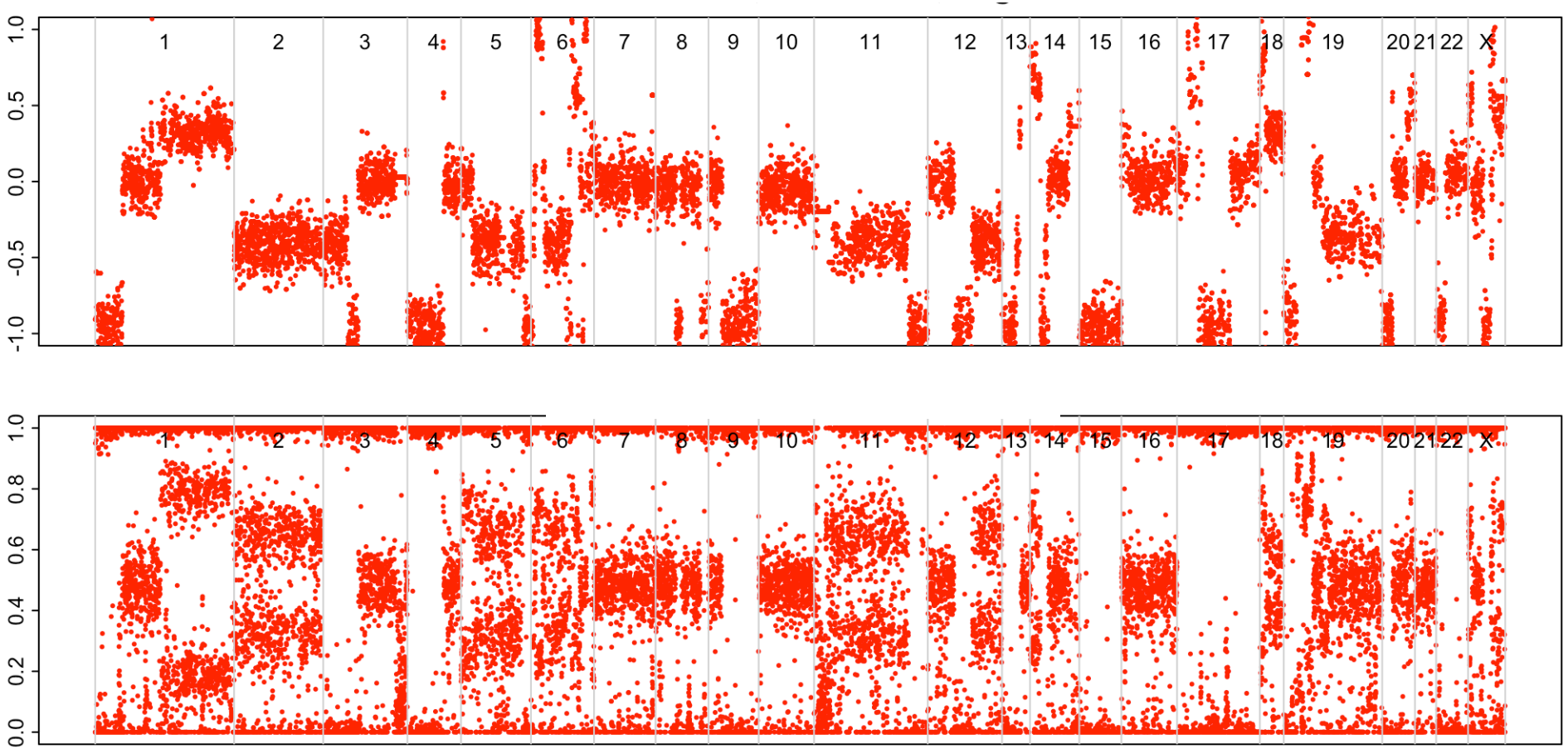
- **Intra-tumoral heterogeneity**
  - Alterations are shared by different proportions of tumor cells.

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

VAF Plots help detecting contamination



The combination of log-ratios and VAF Plots help detecting aneuploidy



Methods:

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…