Analysis of High-throughput sequencing data with Bioconductor
3rd-5th, September 2014
University of Cambridge, Cambridge, UK

Analysis of Copy Number Alterations with HT-seq 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 normalization.
- Methods for CN based on read depth.
- Methods for CNbased 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.**

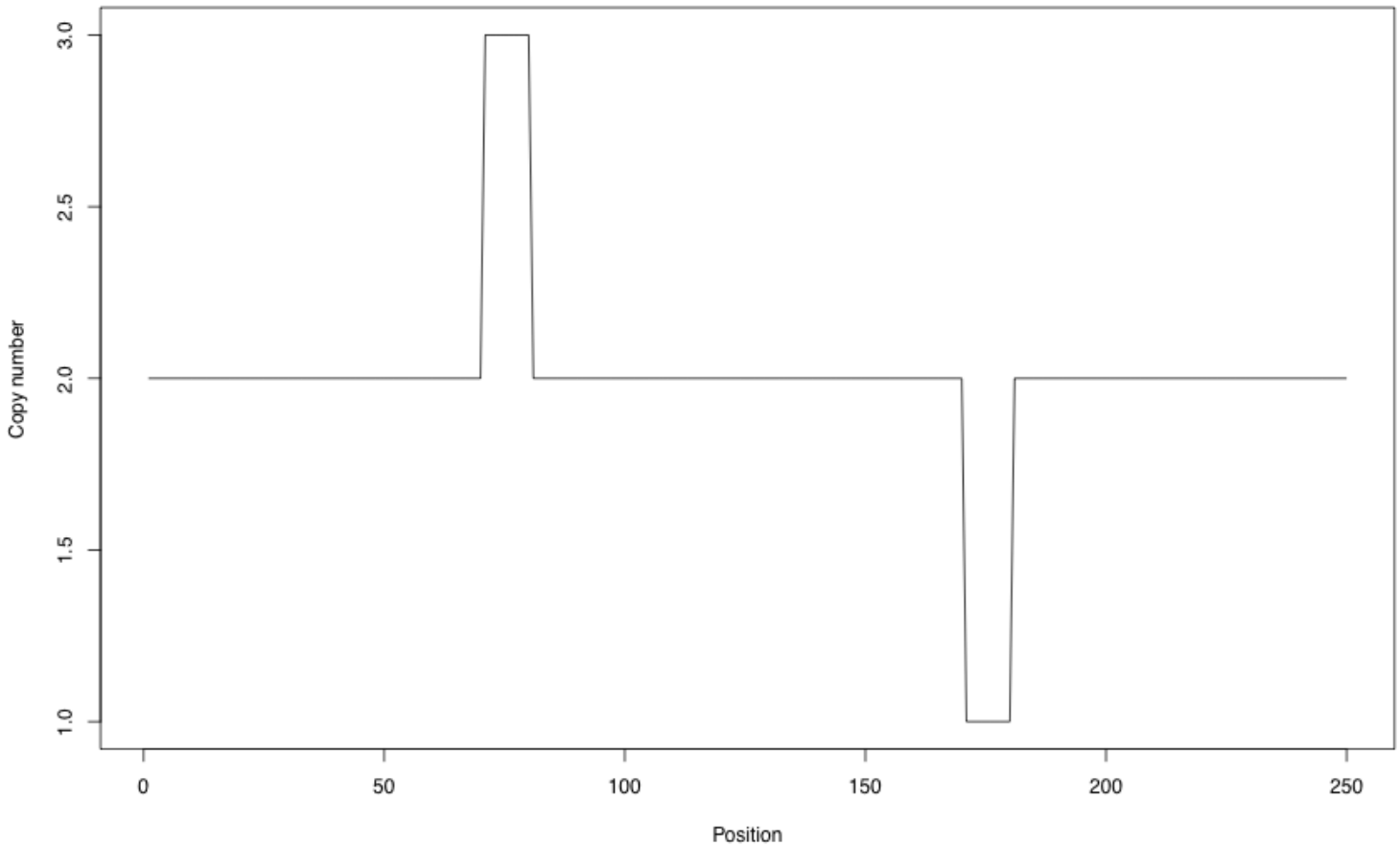
* Other external agents can also produce mutations, like exposure to radiation, certain chemical or viruses…

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.

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

Copy number alterations



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

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:

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

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

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

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.

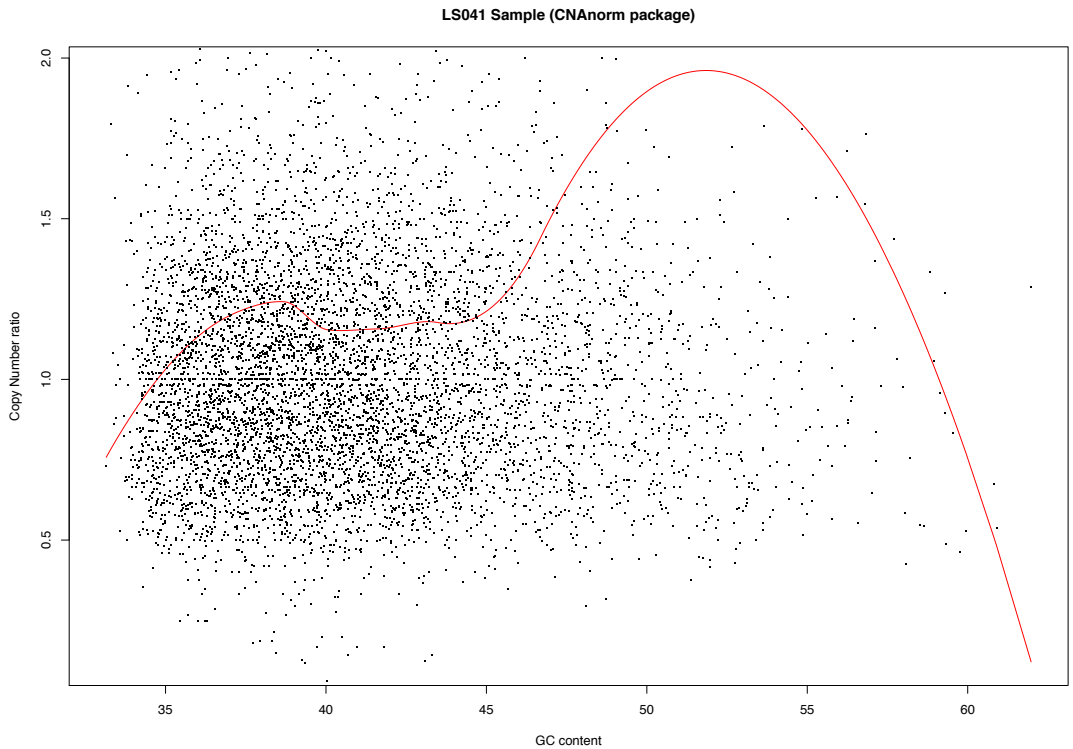
Methods for normalization

Background normalization

- 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).

GC content normalization

- 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.



Target normalization

- 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
- Bioconductor package exomeCopy performs a comprehensive normalization.

Methods for CN based on read depth.

Segmentation methods

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

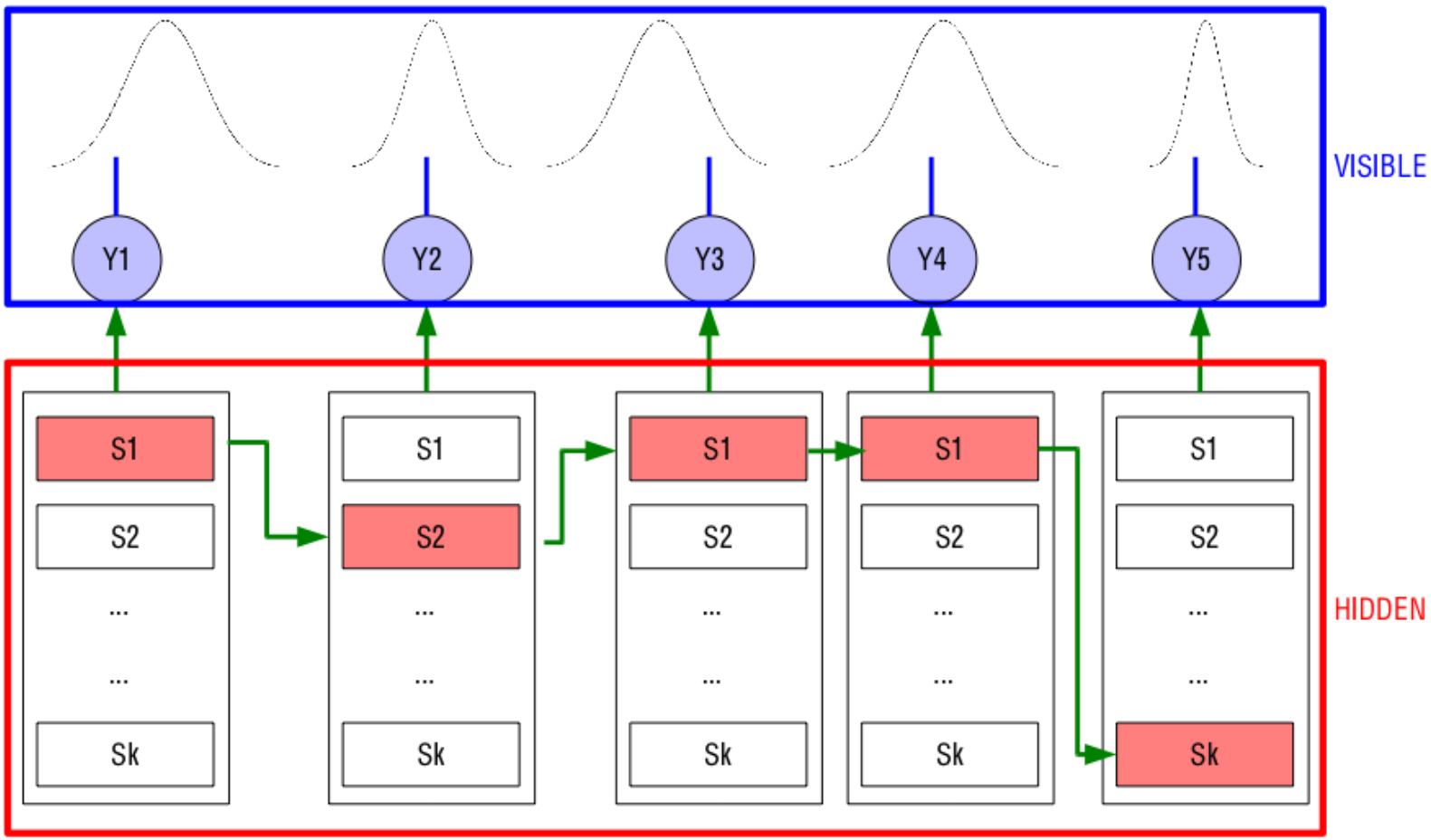
From ratios or $\log_2$ ratios to segmented means: $y_t \Rightarrow m_t$

- **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)

DNAcopy

- **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)



Methods:

- **CNAnorm**
  - Gusnanto et al., 2012.
  - Divides genome in windows of the same size
  - Performs tumour content and ploidy estimation
  - Appropriate for whole genome sequencing
  - Bioconductor package CNAnorm

- **exomeCopy**
  - Love et al., 2011.
  - Fits a Hidden Markov Model
  - Suitable for CNVs (normal samples)
  - Appropriate for exome sequencing
  - Bioconductor package exomeCopy

Methods for CN based on read depth and minor allele frequency.

Minor allele frequency

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

  A: common allele
  B: minor allele

  AA: sample is homozygous for that SNP
  AB: sample is heterozygous for that SNP
  AA: sample is homozygous for that SNP

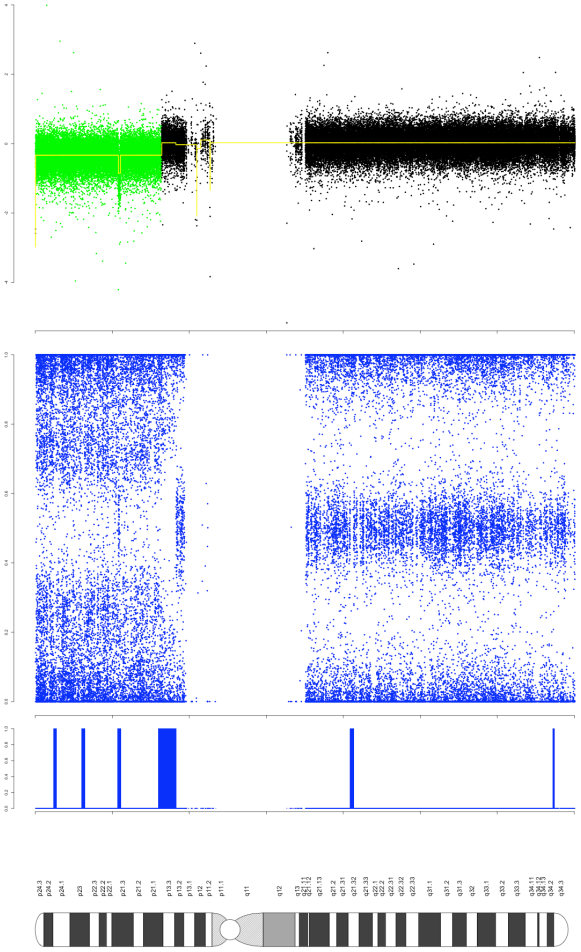
  `maf = #reads(B)/(#reads(A) + #reads(B))`

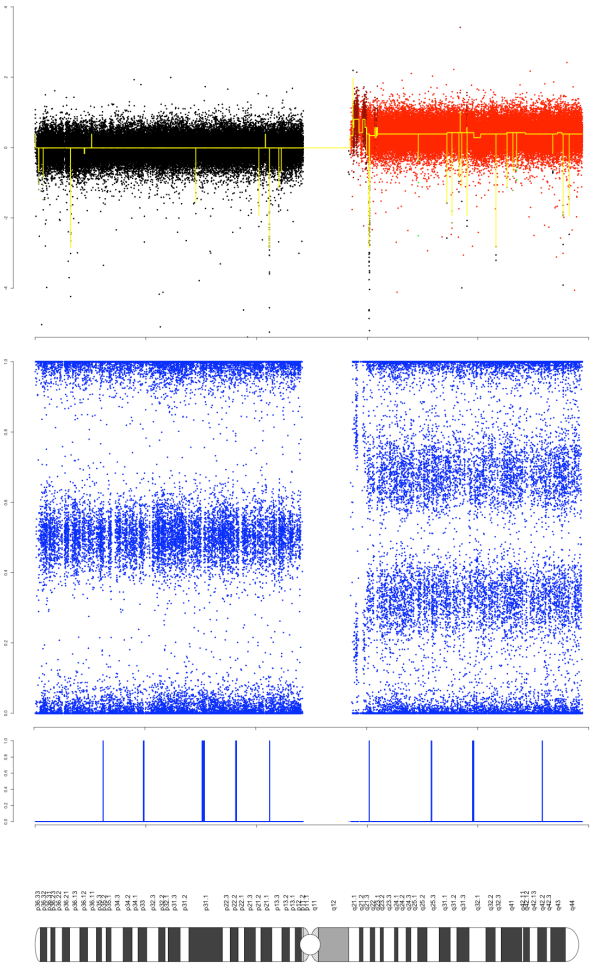
- Now we have two sets of data (similar to SNP arrays):
  - ratios
  - mafs

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





Methods:

- **SomatiCA**
  - Chen et al., 2013
  - Adjusts for tumour content and subclonal heterogeneity
  - Fits a Bayesian Finite Mixture Model
  - Appropriate for whole genome sequencing
  - Bioconductor package somatiCA

- **ExomeCNV**
  - Sathirapongsasuti et al., 2011
  - Uses segmentation on ratios and minor allele frequencies
  - Detects LOH
  - Appropriate for exome sequencing
  - Bioconductor package exomeCNV.

References

- 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
- A. Alkodsi, R. Louhim and S. Hautaniemi. ***Comparatove analysis of methods for identifying somatic copy number alterations from deep sequencing data***. Breif.Bioinform, 2014.
- M. Chen, M. Gunel and H. Zhao. ***SomatiCA: identifying, characterizing and quantifying somatic copy number alterations from cancer genome sequecncing data***. PLoS One, 8(11),2013
- JF Sathirapongsasuti, H Lee, BAJ Horst, G Brunner AJ Cochran, S Binder, J Quackenbush, SF Nelson. ***Exome sequencing-based ocopy number variation and loss of heterozygosity detection: ExomeCNV***. Bioniformatics, 27(19), 2011.
- M I Love, A Mysickova, R Sun, V Kalscheuer, M Vingron, S A Haas. ***Modeling Read Counts for CNV detection in exome sequencing data***. Stat Appl Genet Mol Biol, 10(1), 2011
- A Gusnanto, H M Wood, Y Pawitan, P Rabbits, S berri. ***Correcting for cancer genome size and tumour cell content enables better estimation of copy number alterations from next generation sequencing data***. Bioinformatics, 28(1), 2012.