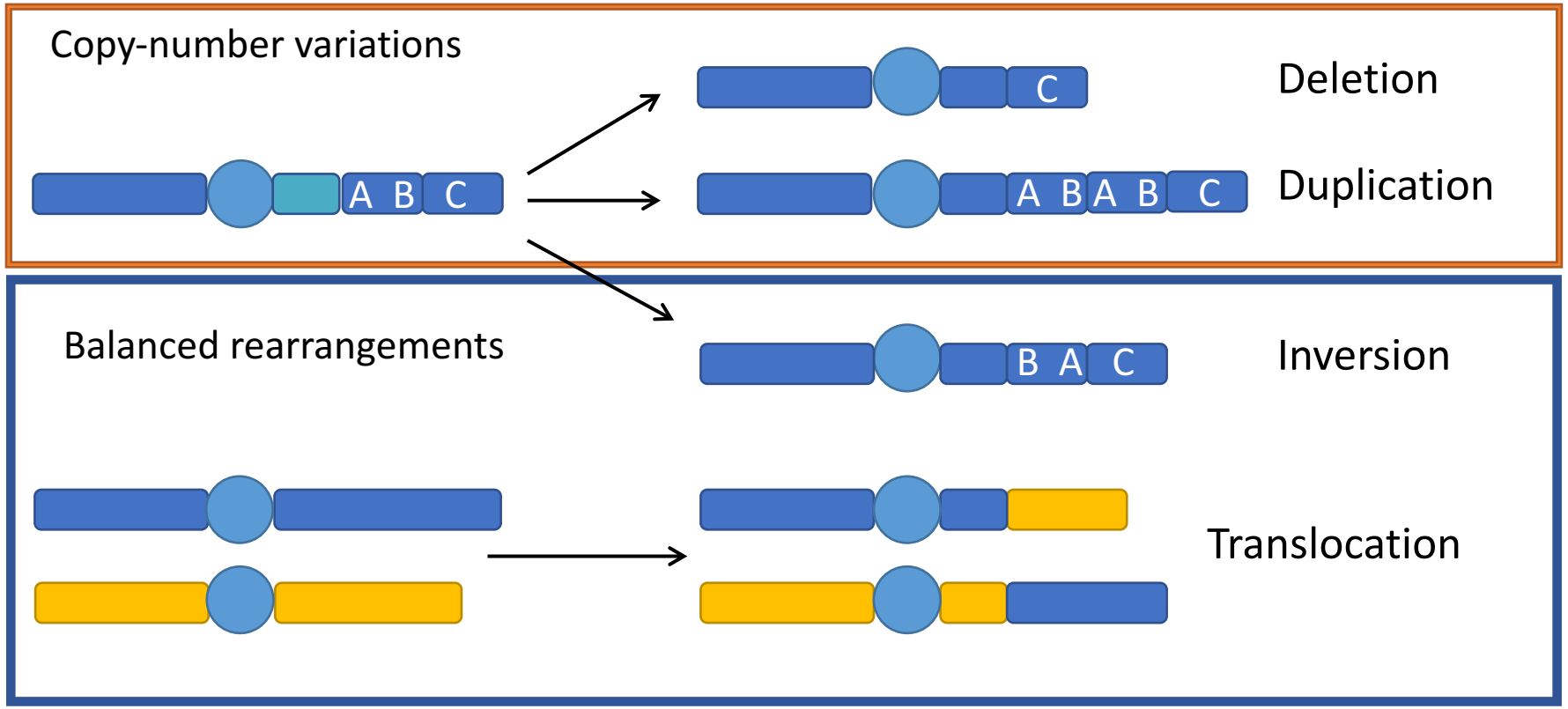
Genome-wide copy-number calling (CNAs not CNVs!)

Dr Geoff Macintyre

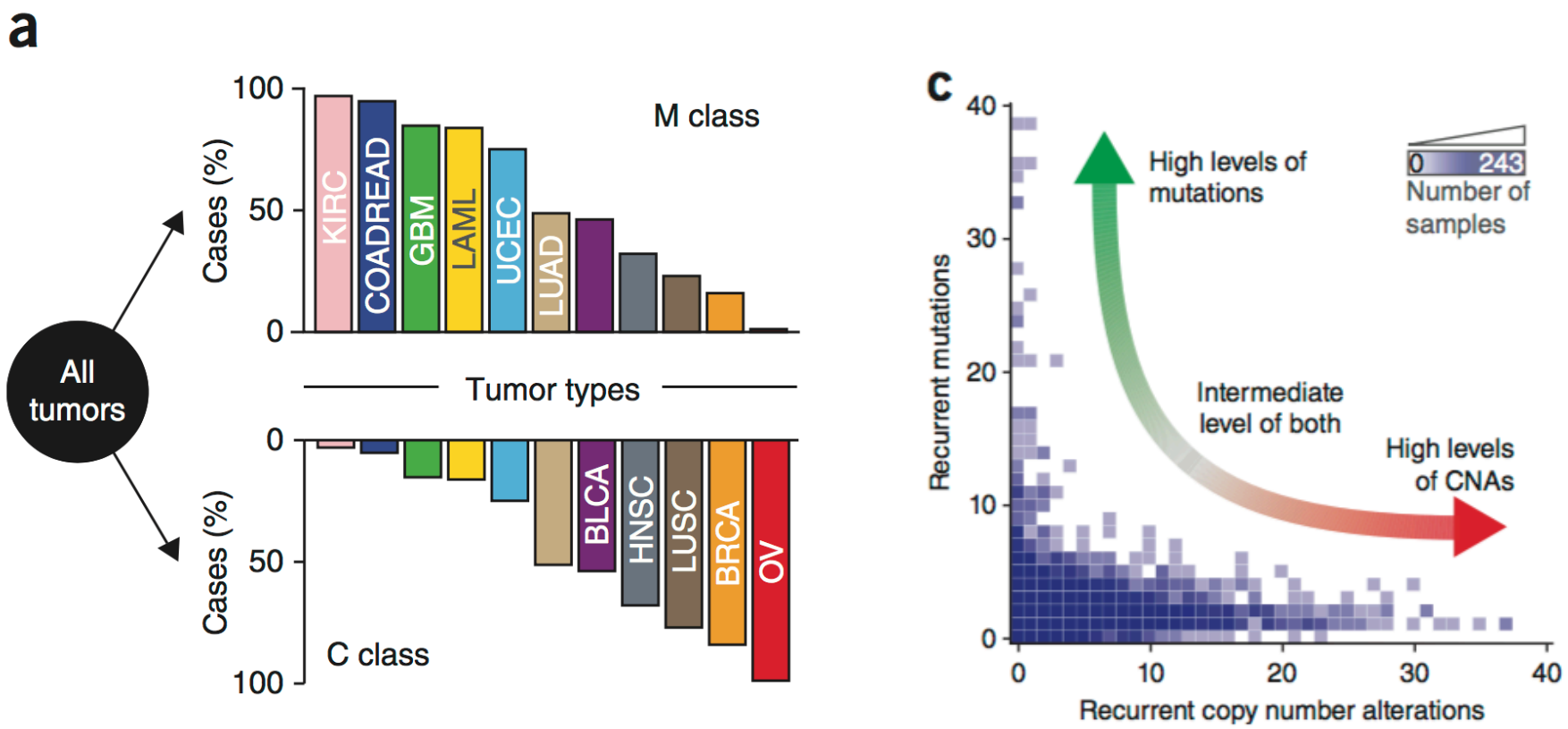
Structural variation (SVs)



Causes

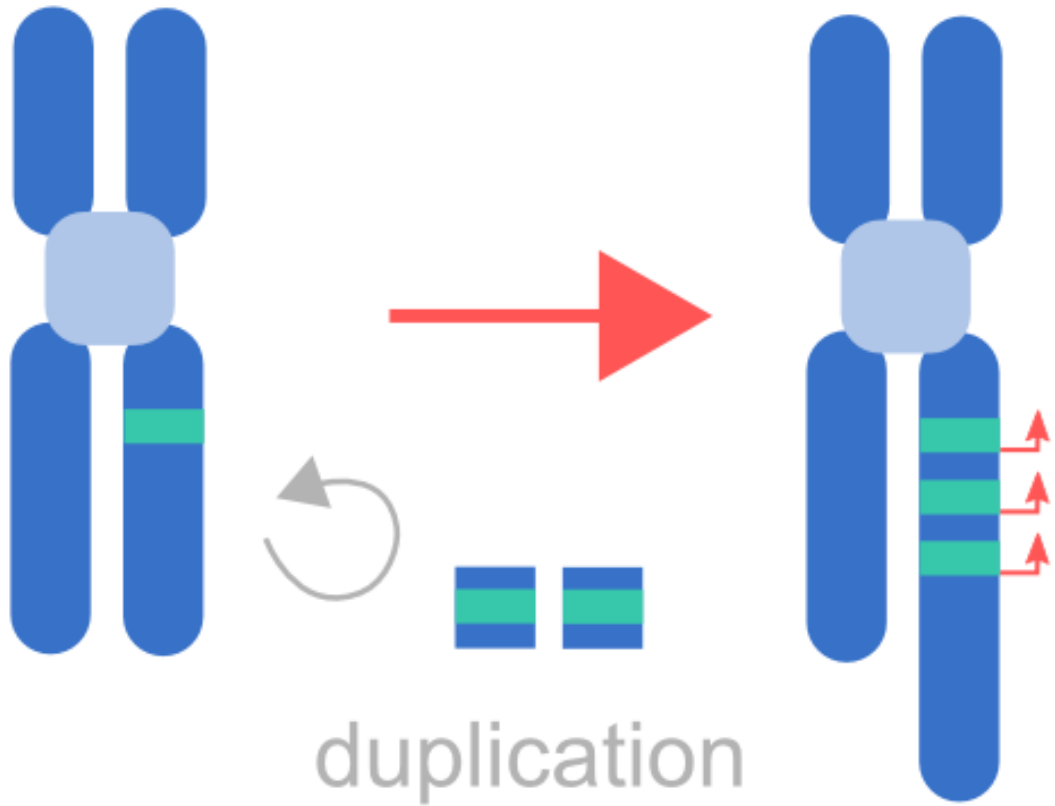
- Replication errors
- Retrotransposition
- Repair errors
- Recombination errors

Why is copy-number important?

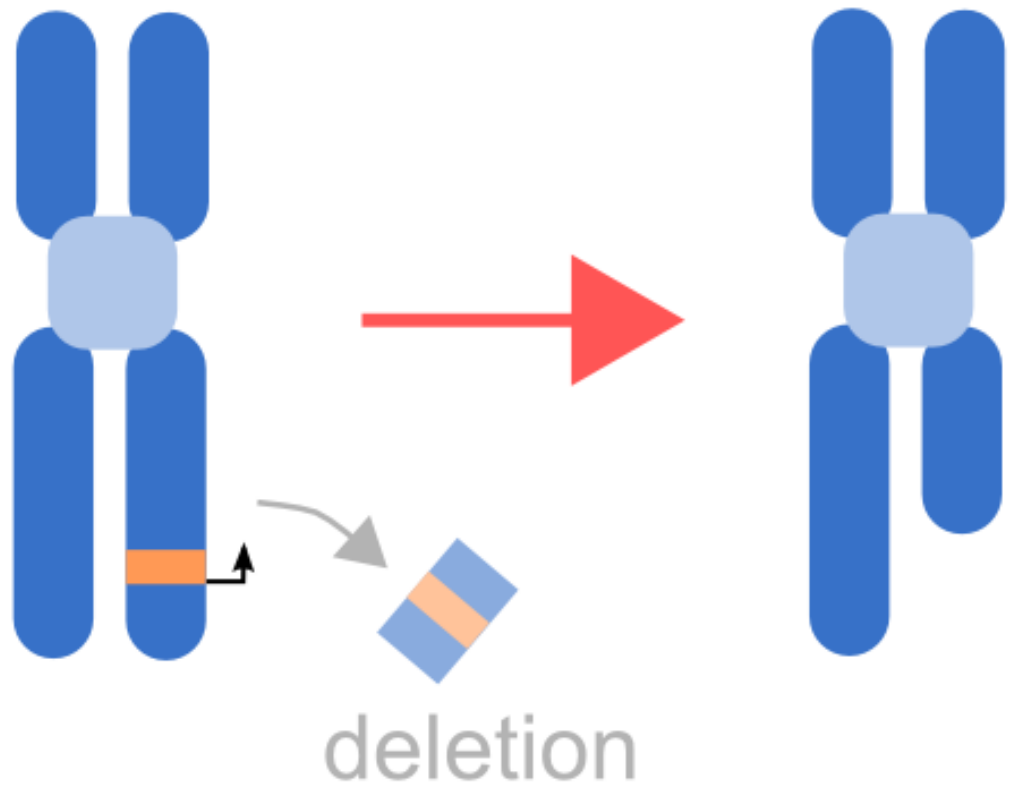


Ciriello et al (2015). Nature Genetics

Oncogene amplification



Tumour suppressor deletion



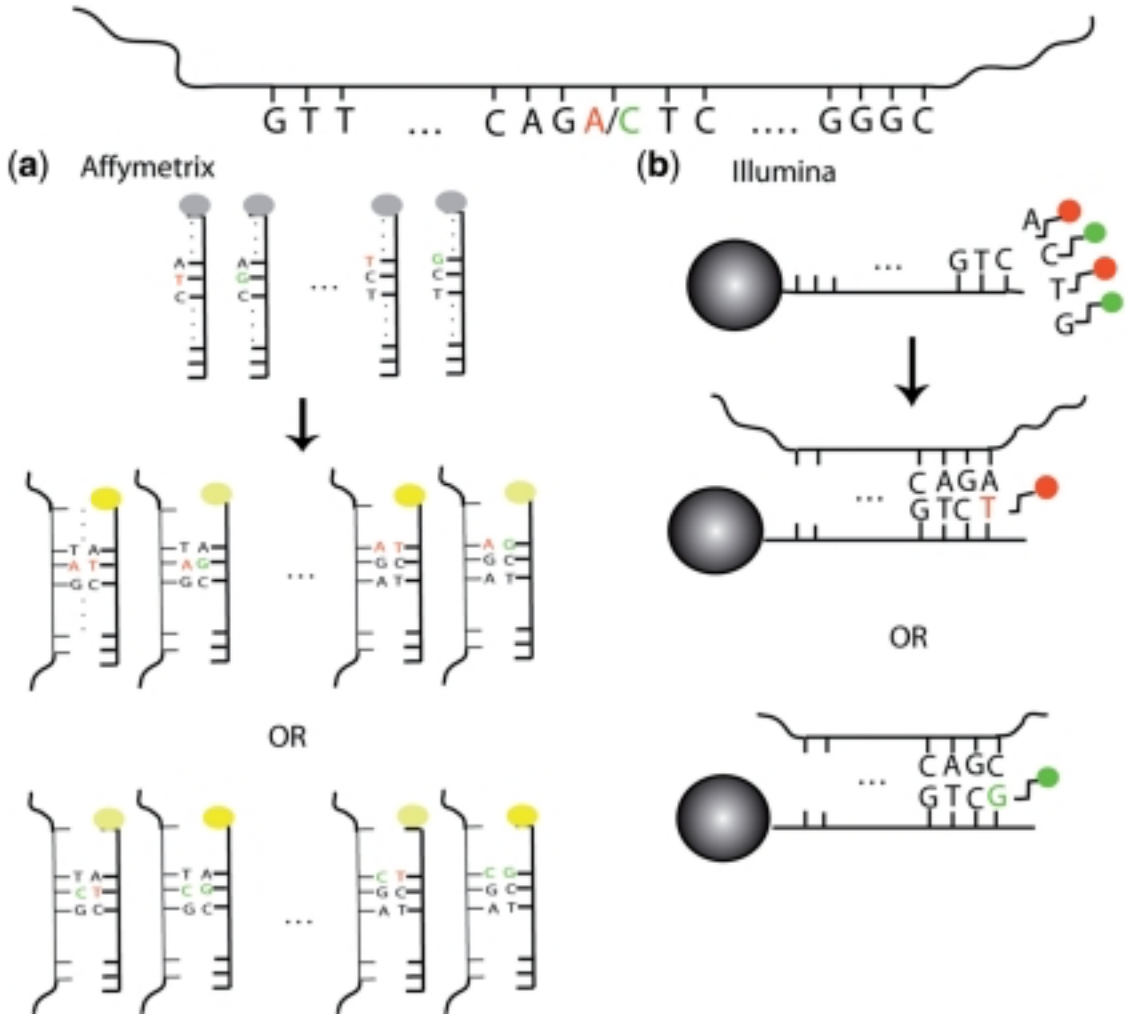
The data

Genome-wide SNP allele frequencies

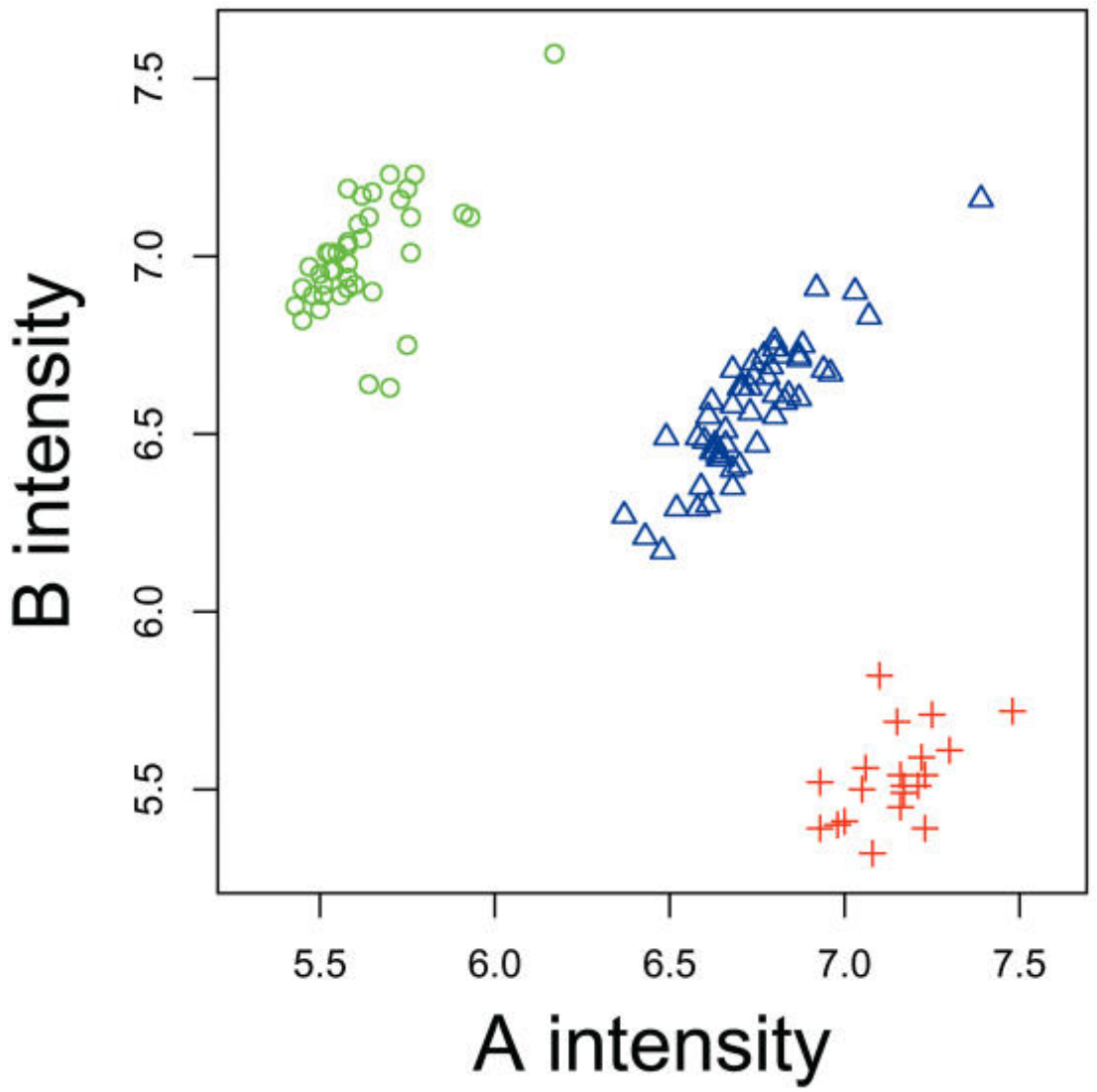
Some measure of the amount of DNA for a given locus

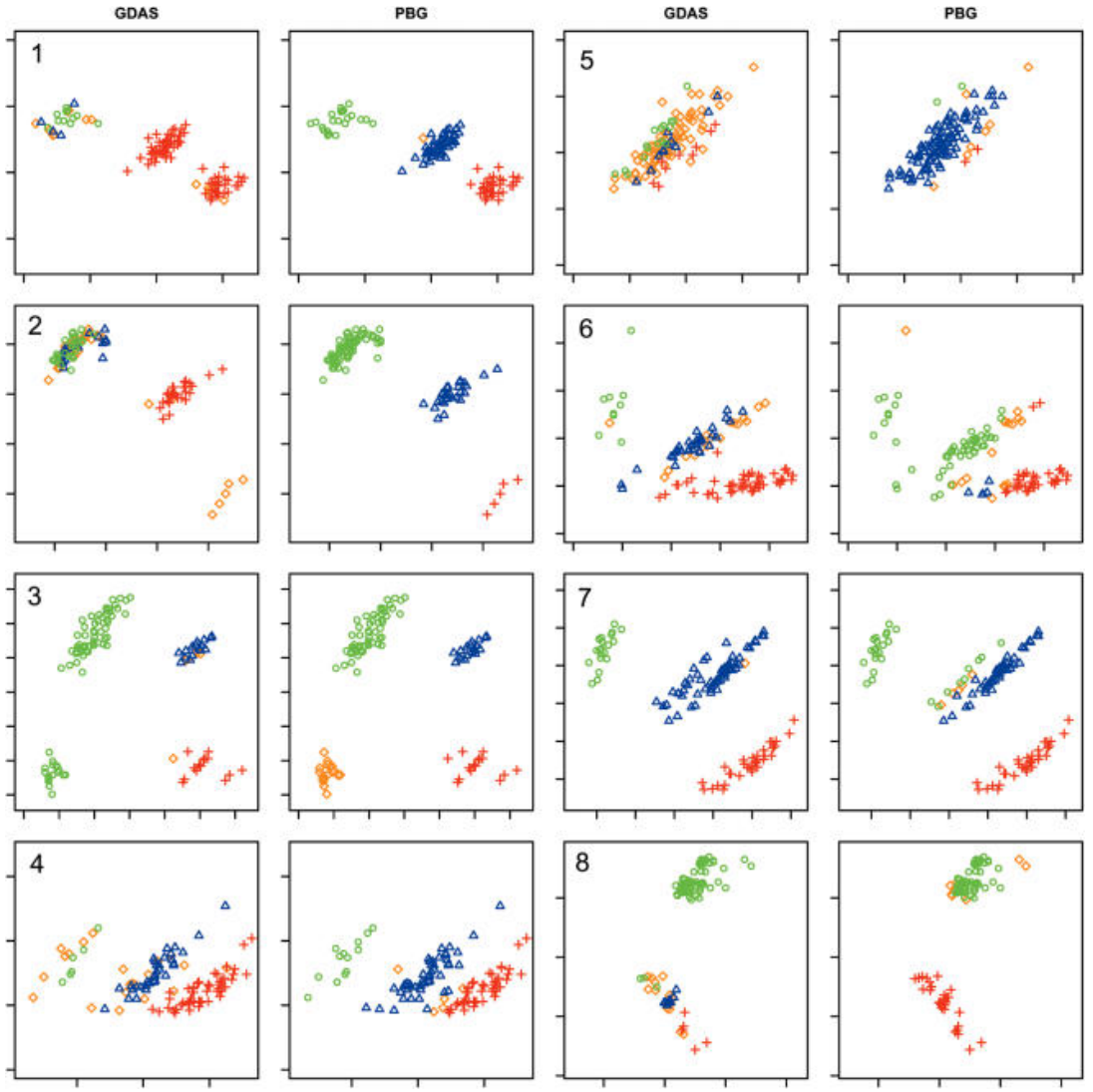
(e.g. sequencing depth)

Array based detection of SNPs (hybridisation)



SNP calling using affymetrix arrays



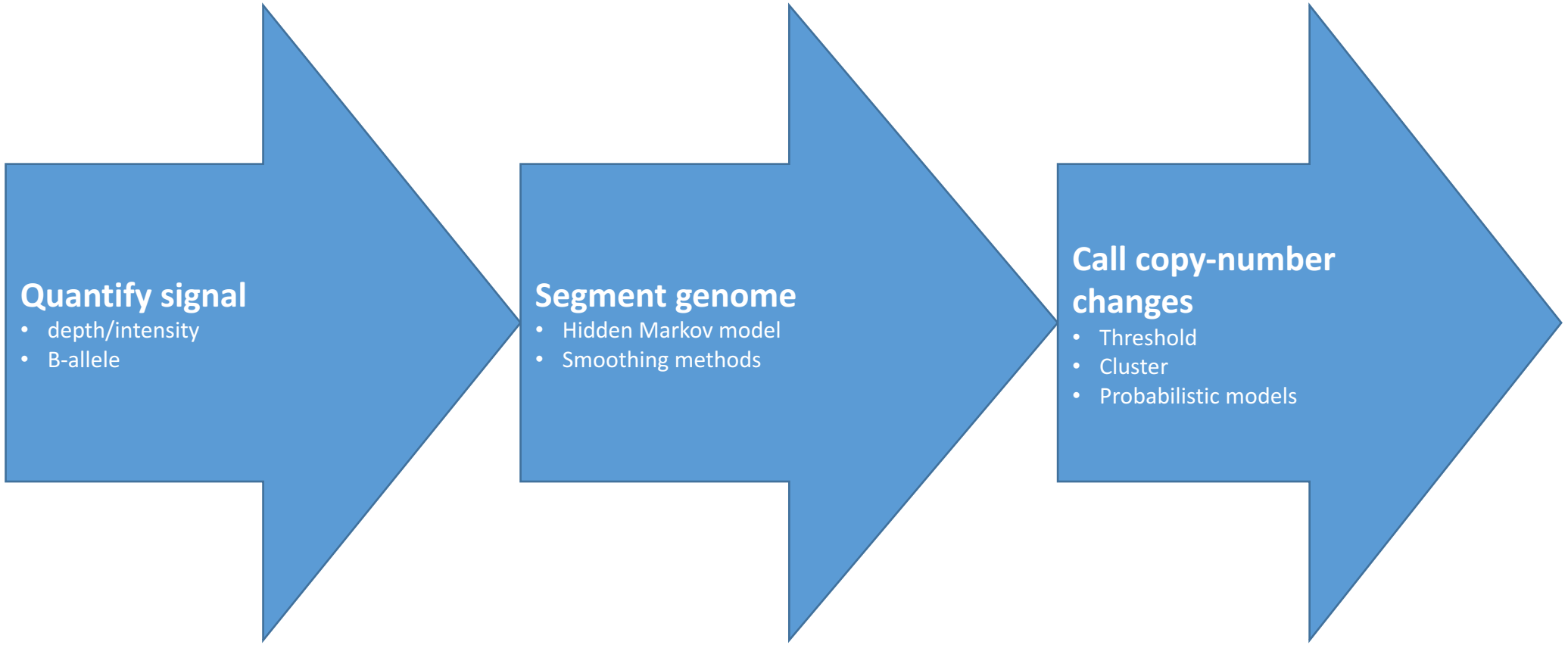


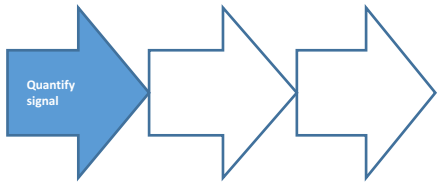
Nucleic Acids Res. 2006; 34(14): e100 2006 doi: 10.1093/nar/gkl475

Further reading on genotyping

- Birdseed: http://www.nature.com/ng/journal/v40/n10/full/ng.237.html
- CRLMM: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3329223/
- HaplotypeCaller and UnifiedGenotyper: http://www.nature.com/ng/journal/v43/n5/full/ng.806.html
- Varscan: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2734323/
- GWAS primer: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4181332/

Basic workflow





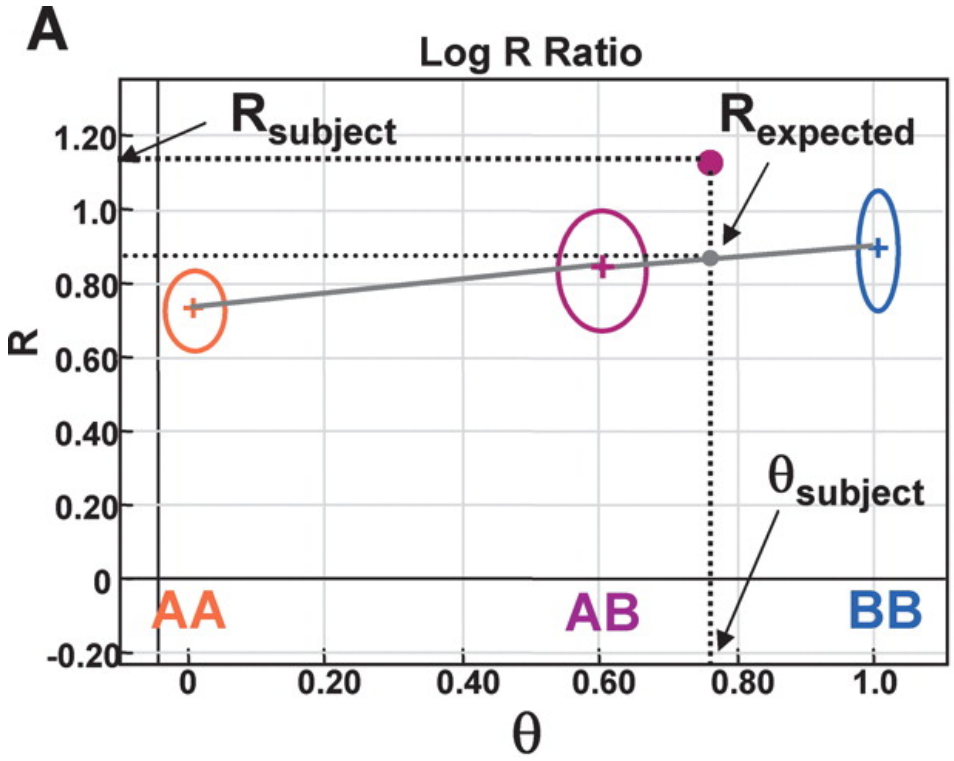
The data: logR

Sequencing:
$\log_2(\text{depth})$

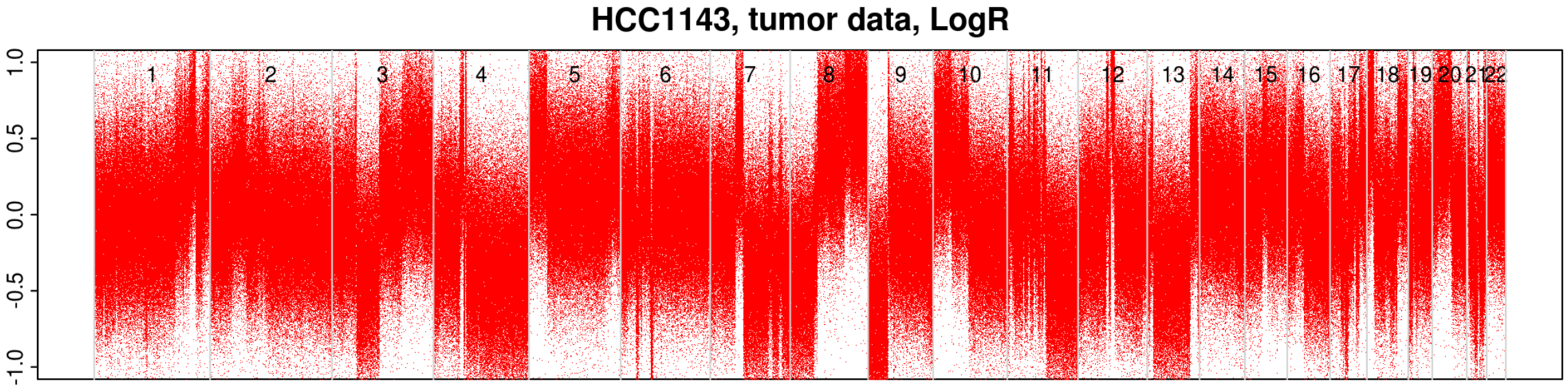
Array:
$R(\theta)_{subject}$ = normalised intensity of probes from sample

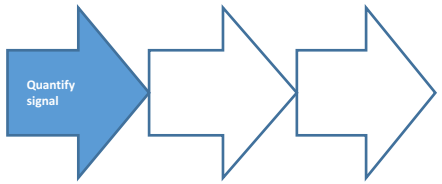
$R(\theta)_{expected}$ = normalised intensity of probes from control

$\log R = \log_2(R(\theta)_{observed}/R(\theta)_{expected})$



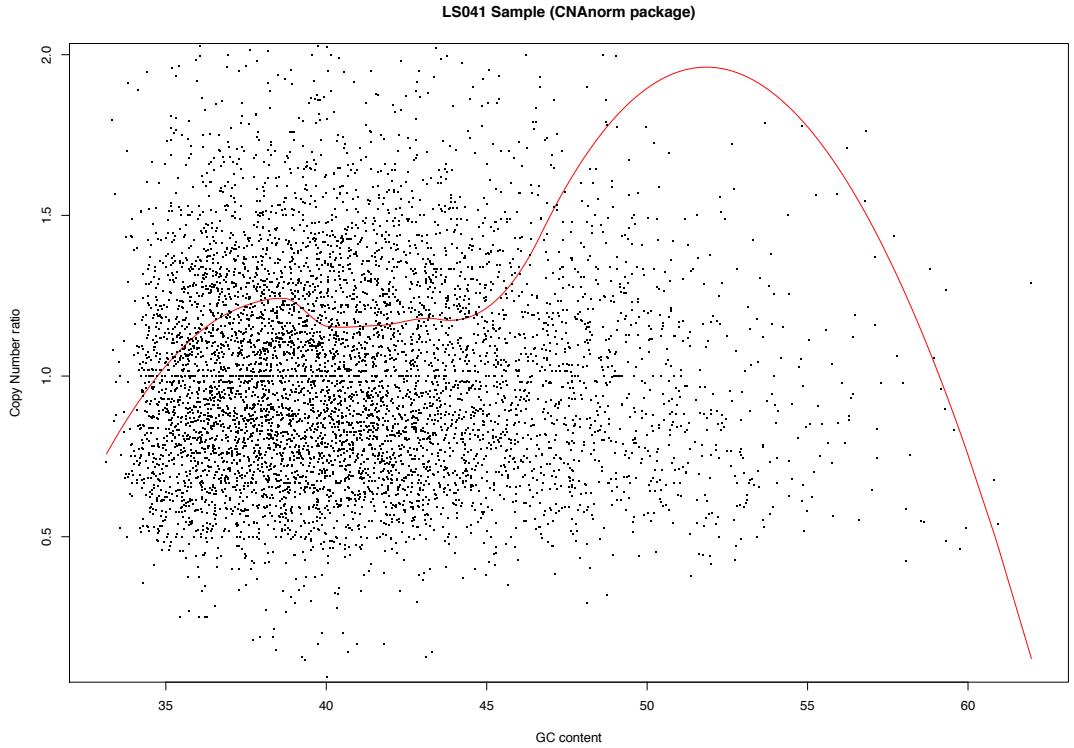
logR of HCC1143 cell-line using affy SNP6

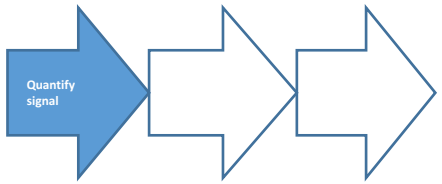




logR/depth normalisation

- 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.

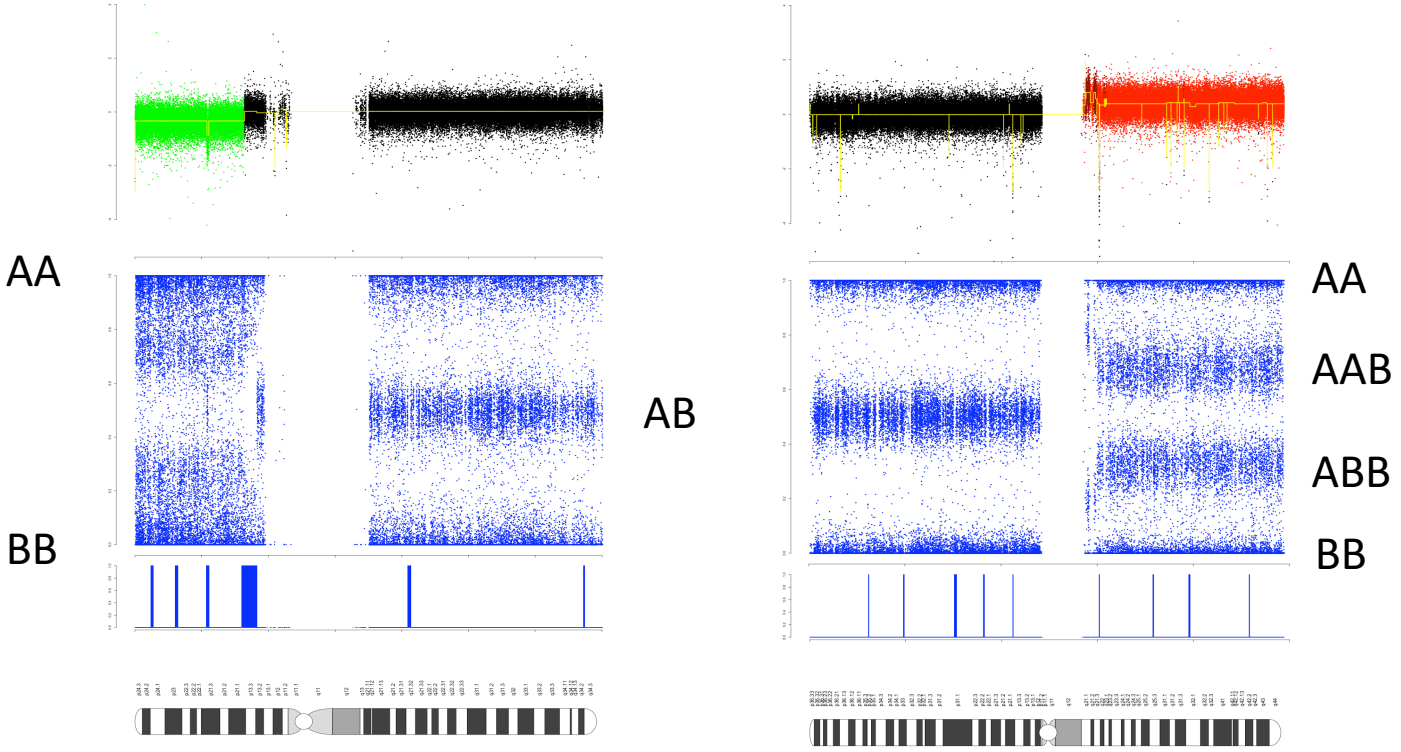


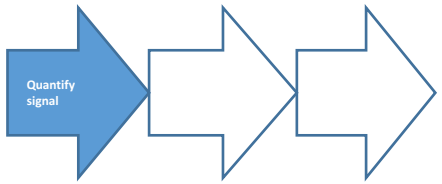


The data: B-allele frequency

θ_A= intensity of probe for allele A

$BAF = \theta_A/(\theta_A + \theta_B)$

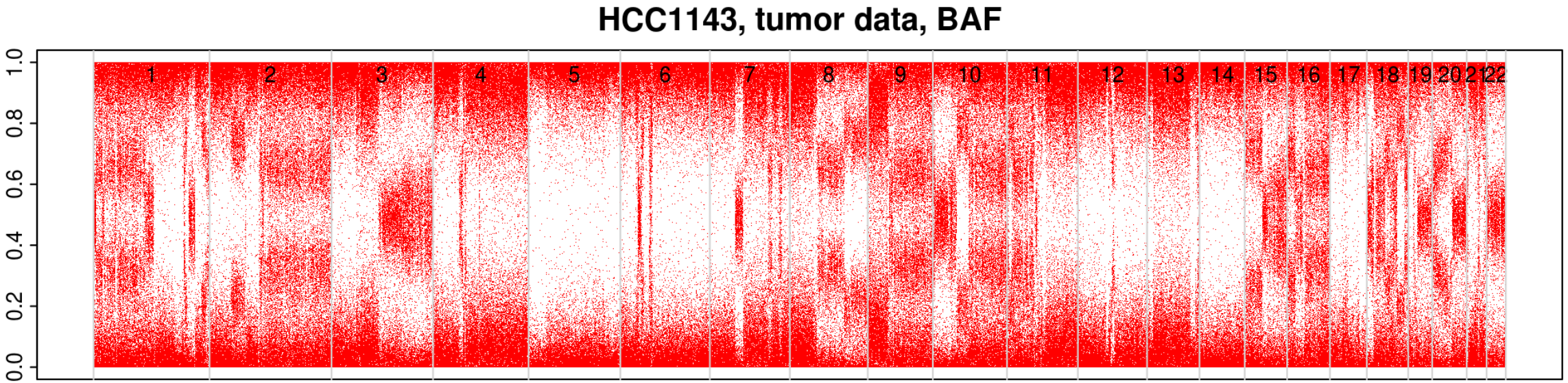


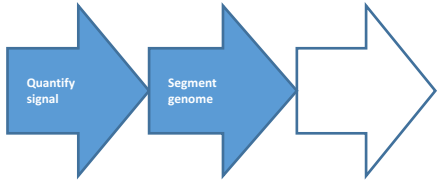


BAF banding

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

BAF of HCC1143 cell-line using affy SNP6

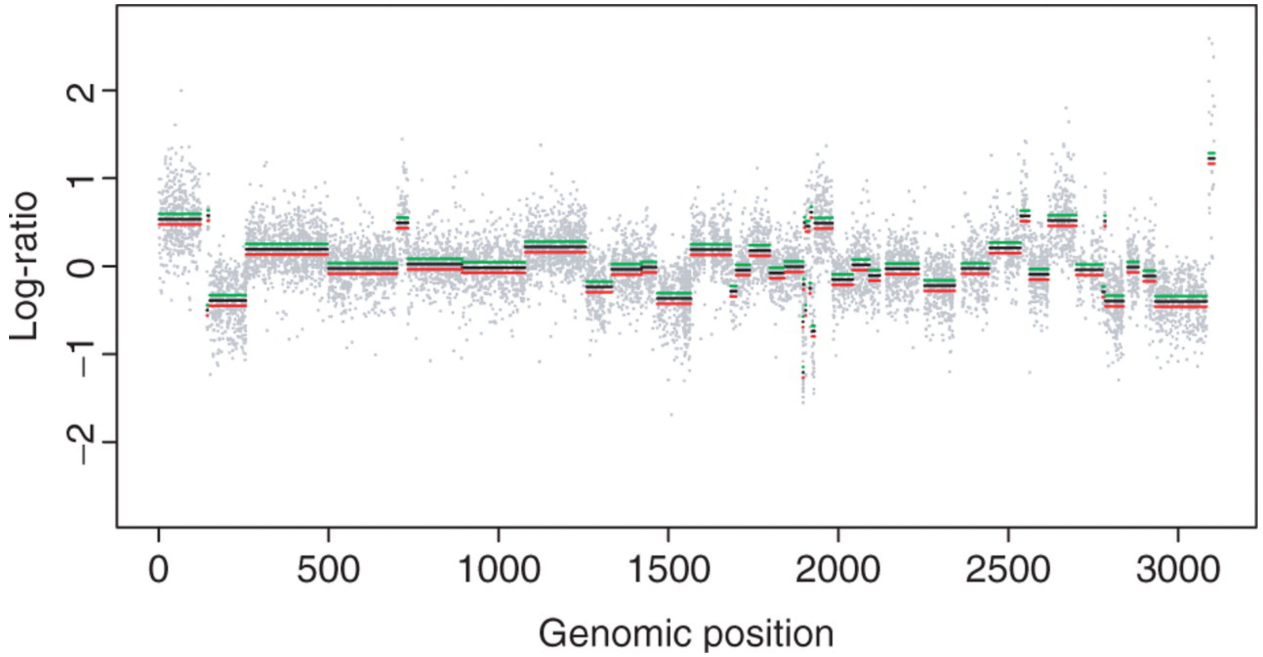




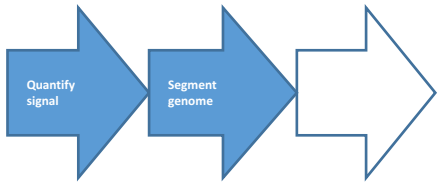
Segmentation: Circular binary segmentation

Olshen et al., 2004.

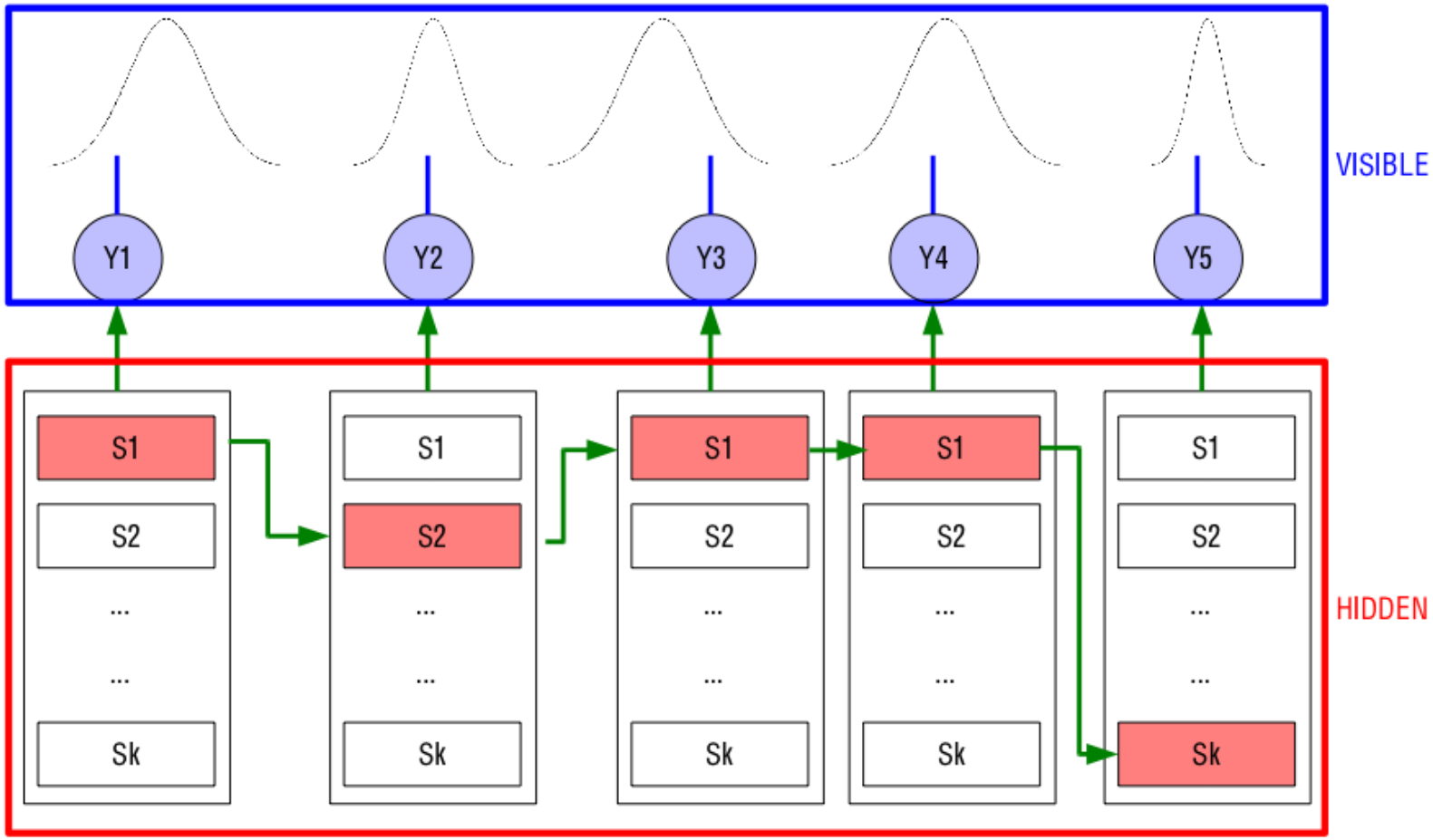
- It can be used with array and sequencing data
- Finds change points using a t-test under a permutation model.
- Bioconductor package DNAcopy.

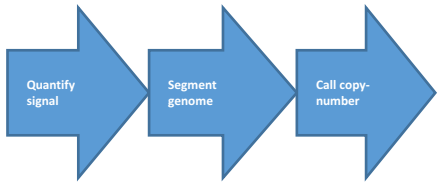


https://academic.oup.com/biostatistics/article/5/4/557/275197/Circular-binary-segmentation-for-the-analysis-of?searchresult=1



Segmentation: Hidden markov models



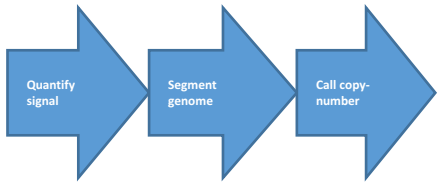


Copy-number calling: threshold based

Individual thresholds based on the variability of each sample:

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

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



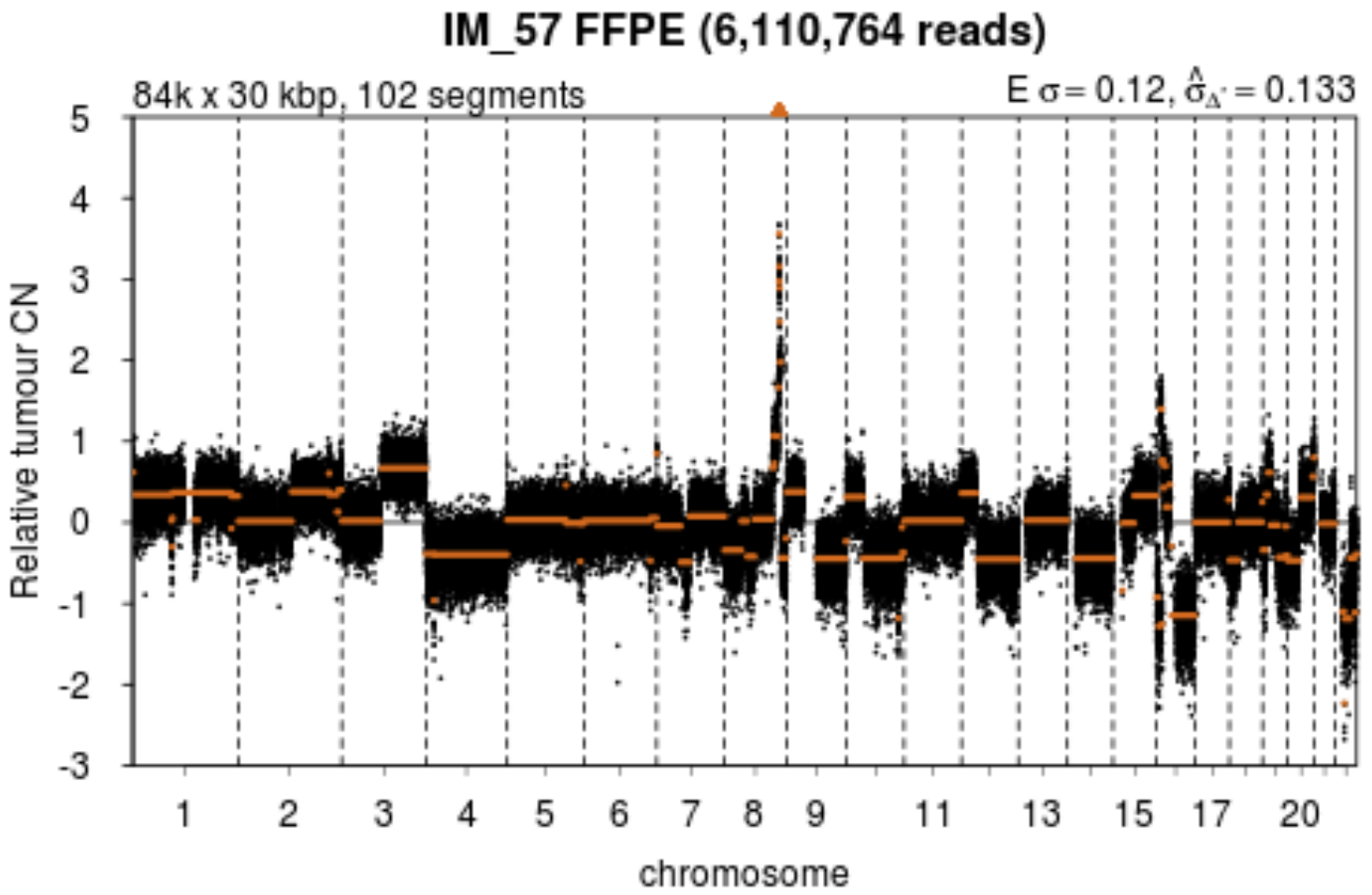
Copy-number calling: cluster based

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

- The segmented means come from a mixture of six normal populations.
- The model is fit by EM algorithm.
- Classification reduced to 3 or 4 states. (Usually loss, gain, normal)

http://bioconductor.org/packages/release/bioc/html/CGHcall.html

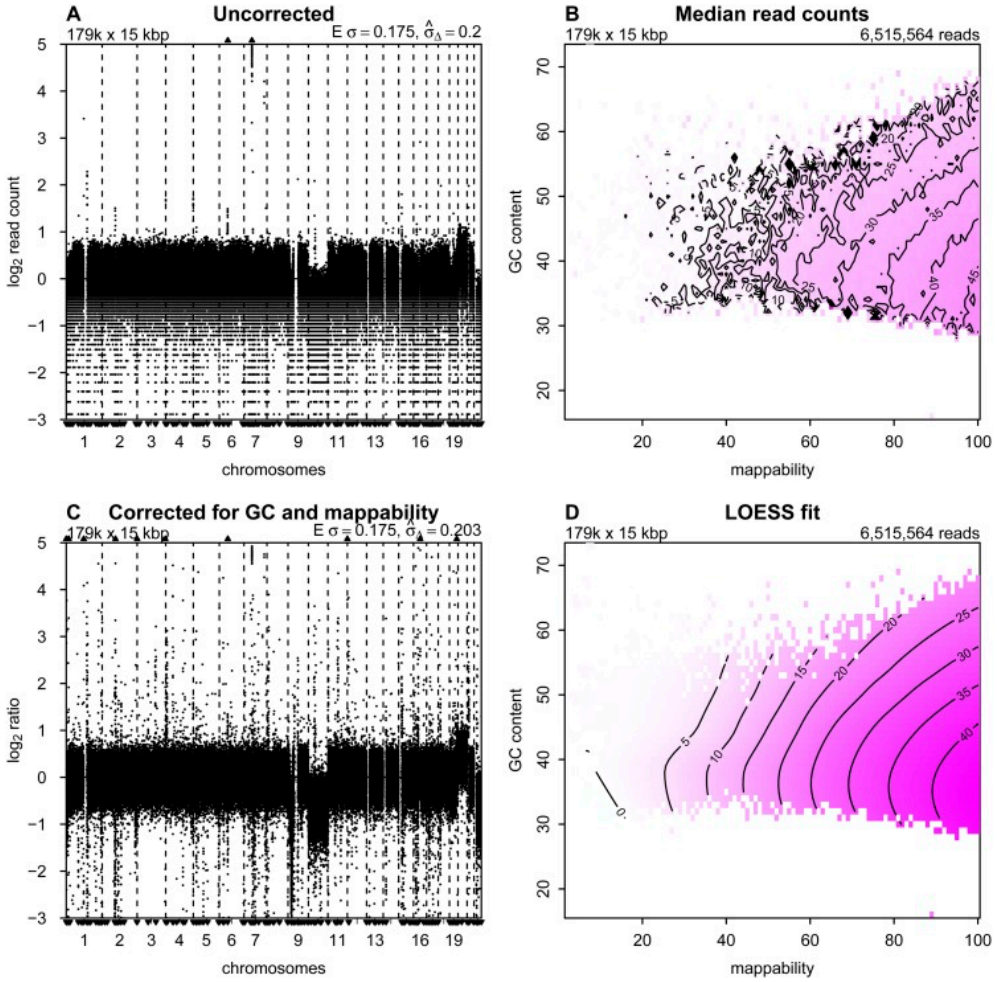
Relative copy-number profile (ovarian cancer)



Method: QDNAseq

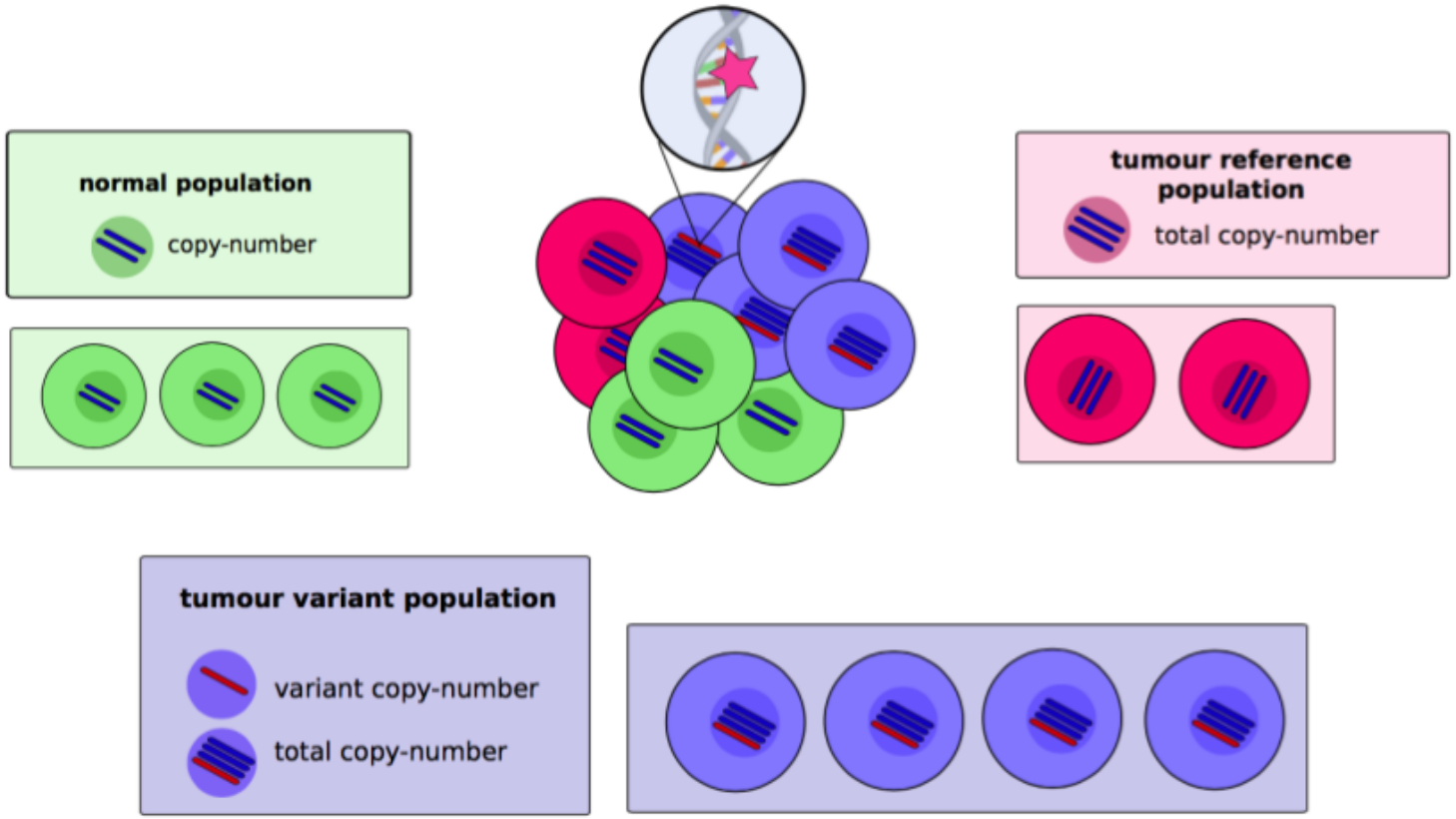
Scheinin I et al., 2014 (QDNAseq Bioconductor package).

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

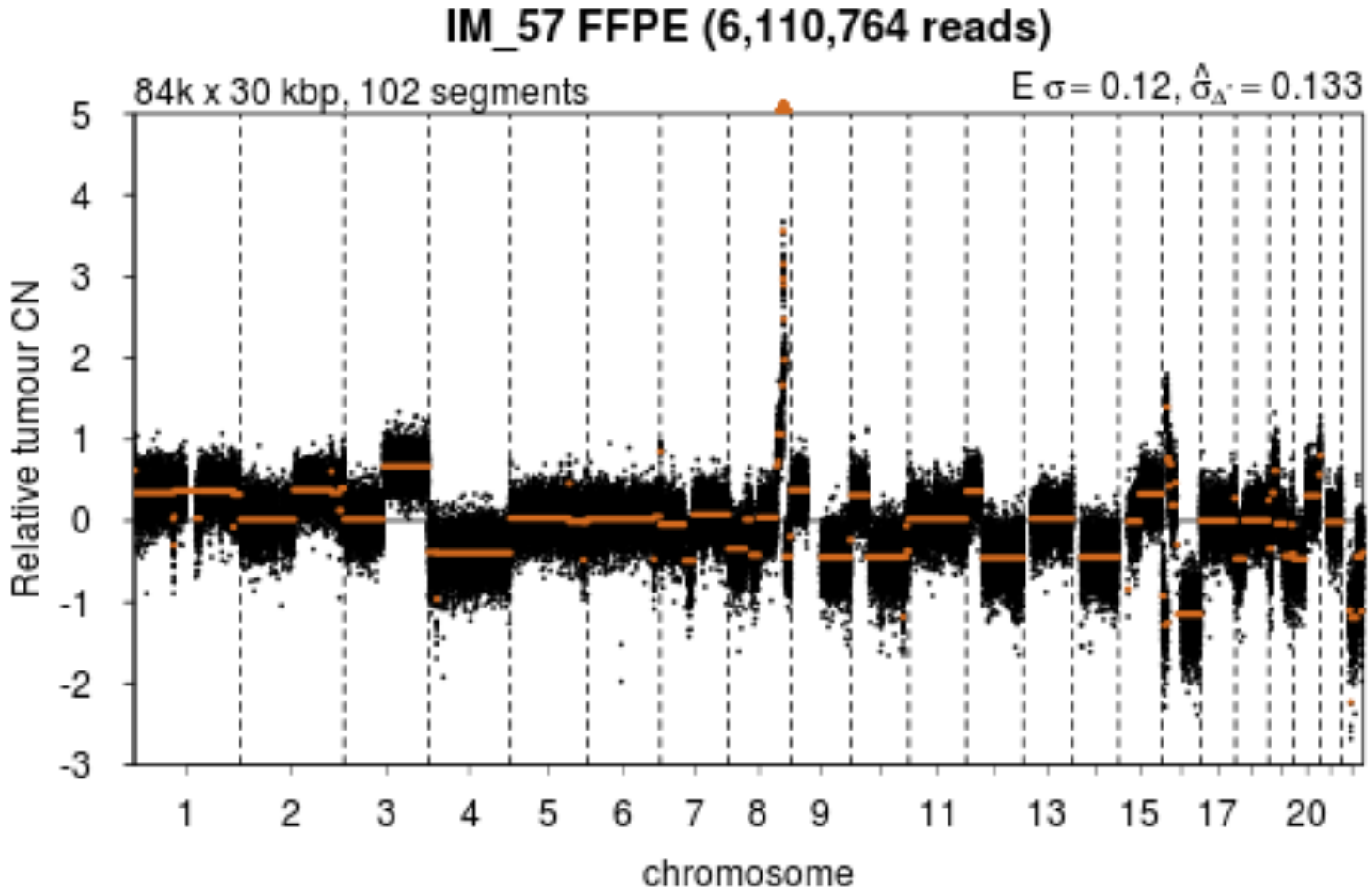


https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4248318/

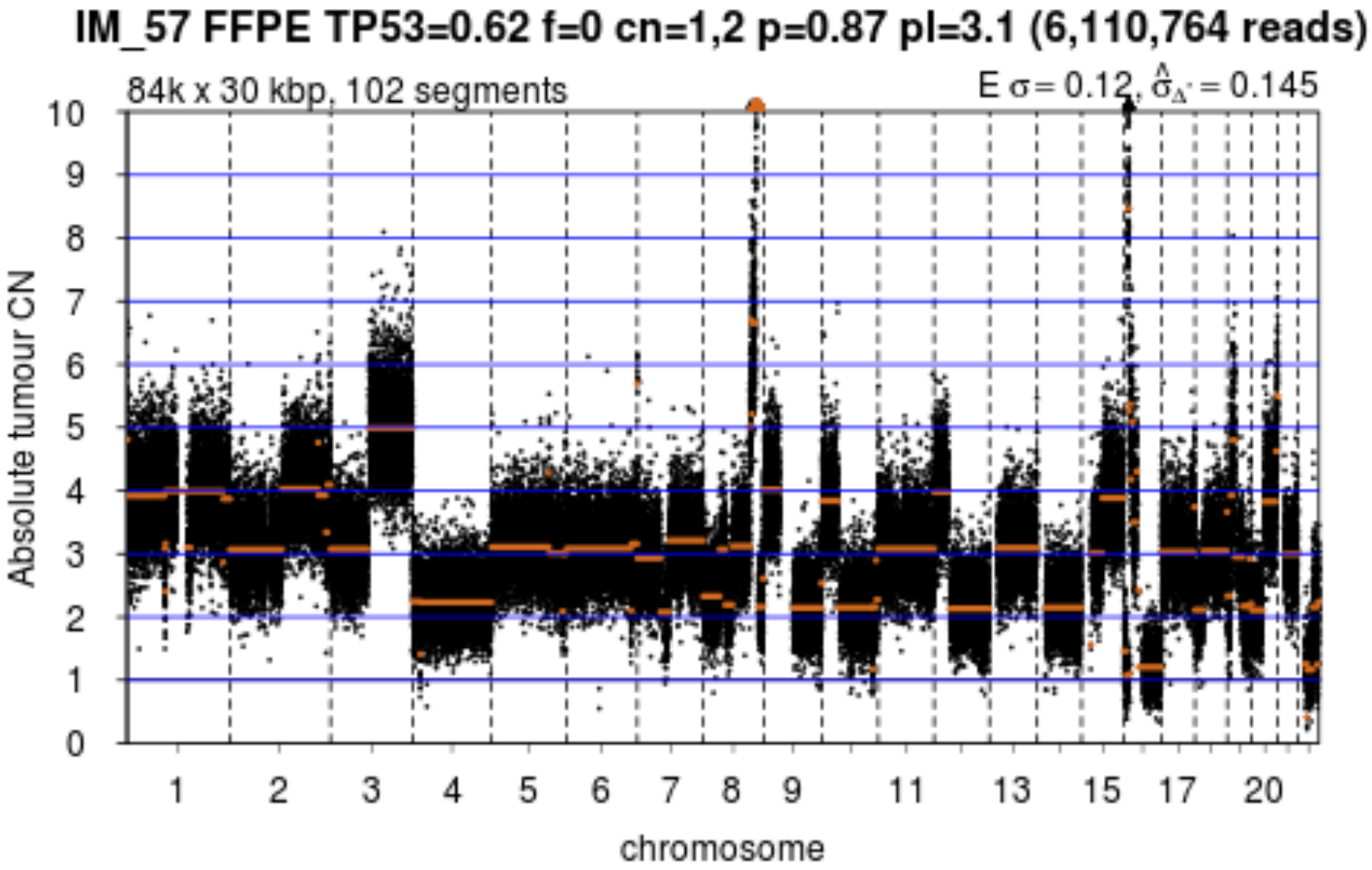
Problems: purity and heterogeneity



What are the effects on relative copy-number?



Absolute copy-number profile (ovarian cancer)



Method: Allele-Specific Copy number Analysis of Tumours (ASCAT)

A-allele copy-number

tumour fraction

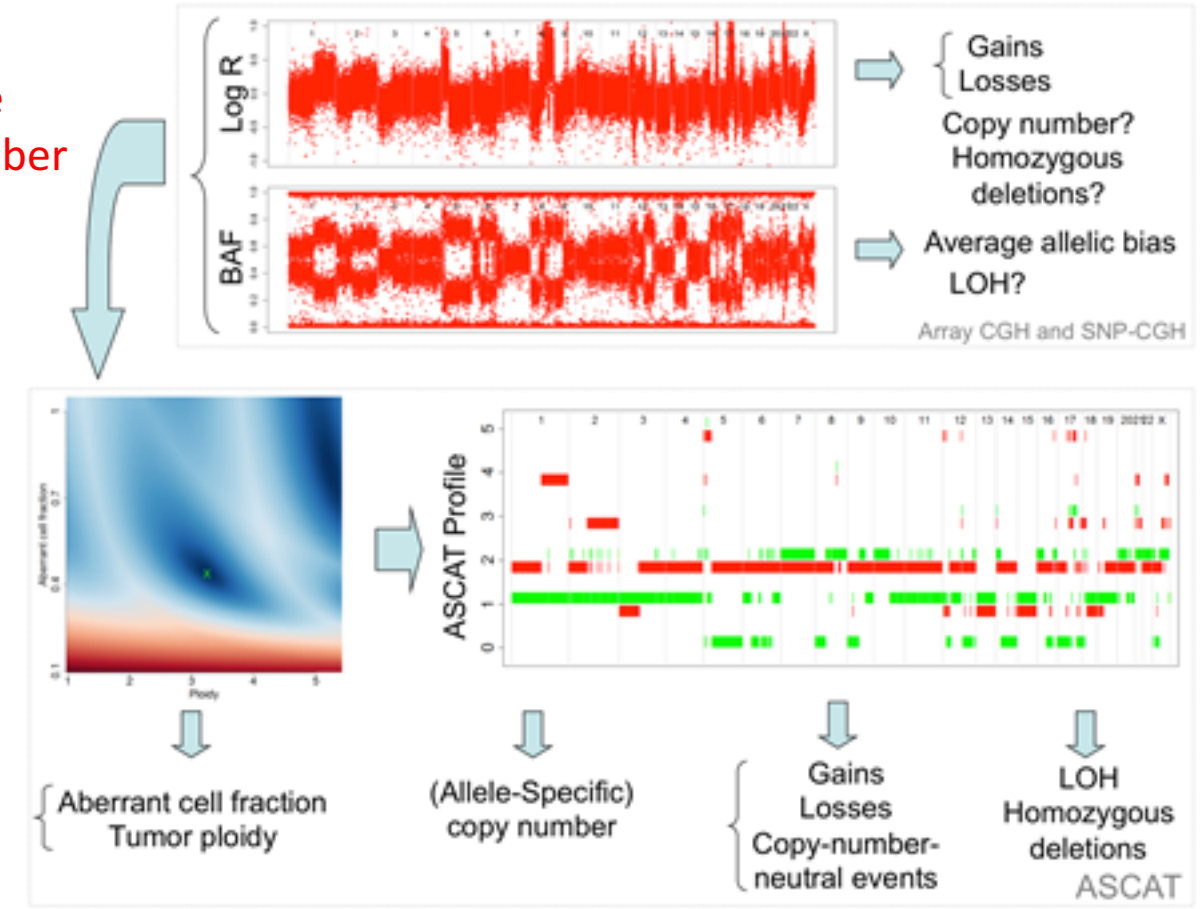
B-allele copy-number

constant

$$r_i = \gamma \log_2 \left(\frac{2(1-\rho) + \rho(n_{A,i} + n_{B,i})}{\psi} \right)$$

ploidy

$$b_i = \frac{1 - \rho + \rho n_{B,i}}{2 - 2\rho + \rho(n_{A,i} + n_{B,i})}$$



Further reading on copy-number

- Methods for CN detection (array data): https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2697494/
- Tools for CN detection (sequence data): http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-14-S11-S1
- PennCNV, a package for CNV calling: http://penncnv.openbioinformatics.org/en/latest/
- Large scale analysis of CNAs in cancer: http://www.nature.com/ng/journal/v45/n10/full/ng.2760.html