UNIVERSITY OF CAMBRIDGE

MPhil in Computational Biology 2016/17

Course: **Functional Genomics**

Lecture 14 (Wednesday 23rd November 2016)

Genotyping/Copy Number

Geoff Macintyre

geoff.macintyre@cruk.cam.ac.uk

(Contributions by Oscar Rueda and Andy Lynch)

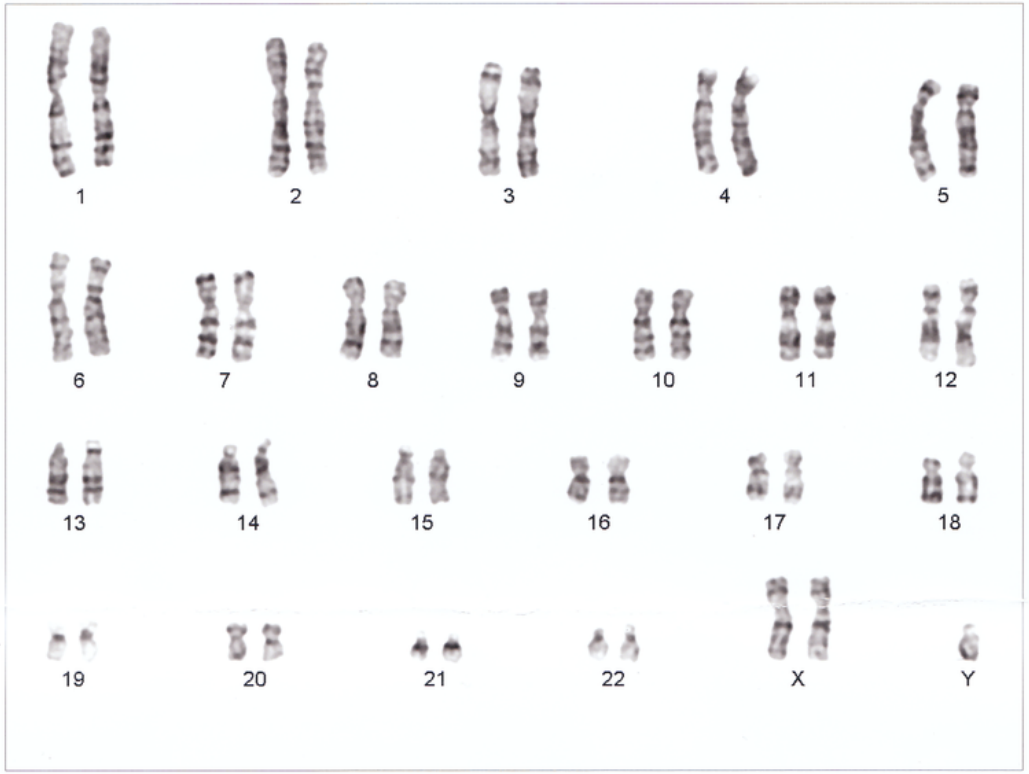
Background:
Types of genetic variation

Germline variation

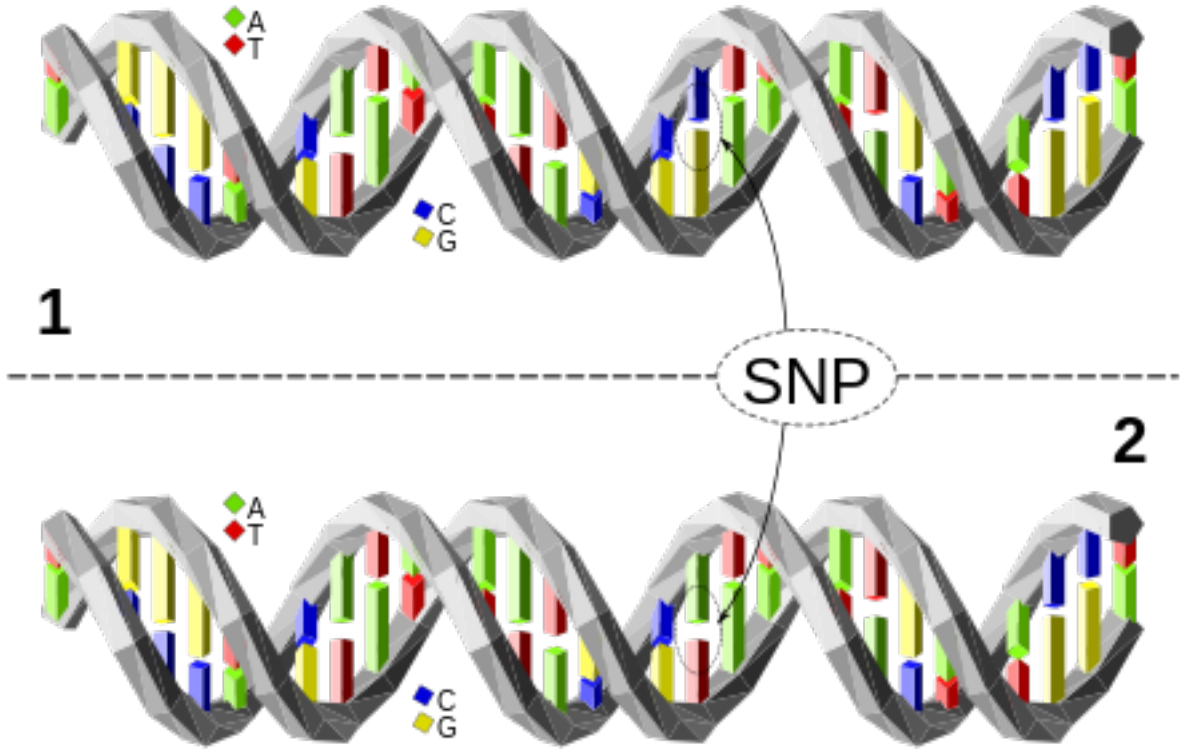
Inherited mutations present in *germ* cells

In humans:

- We have 23 chromosome pairs
- Our genome is therefore *diploid*
- You differ from me by about 3 million mutations
- There are many different types of germline mutation



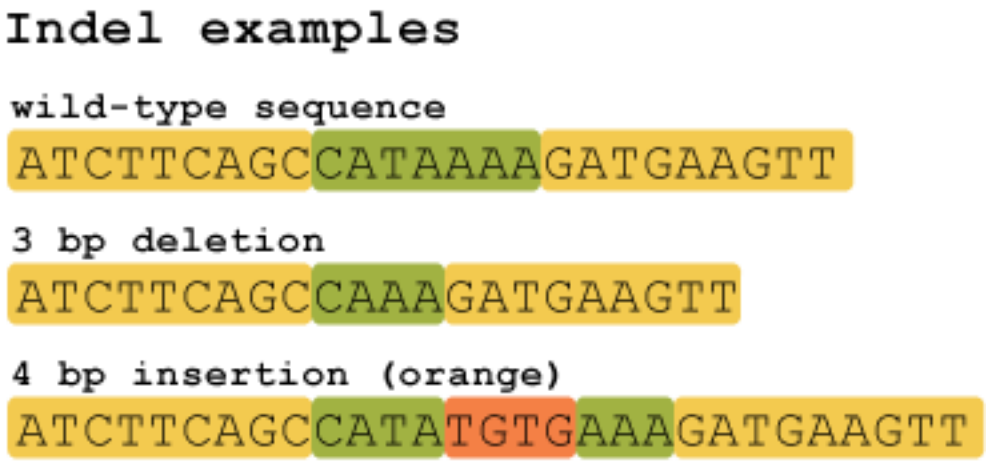
Single nucleotide polymorphisms - SNPs



Causes

- Replication errors
- Repair error
- Mutagens
- Spontaneous

Insertions and deletions - INDELs

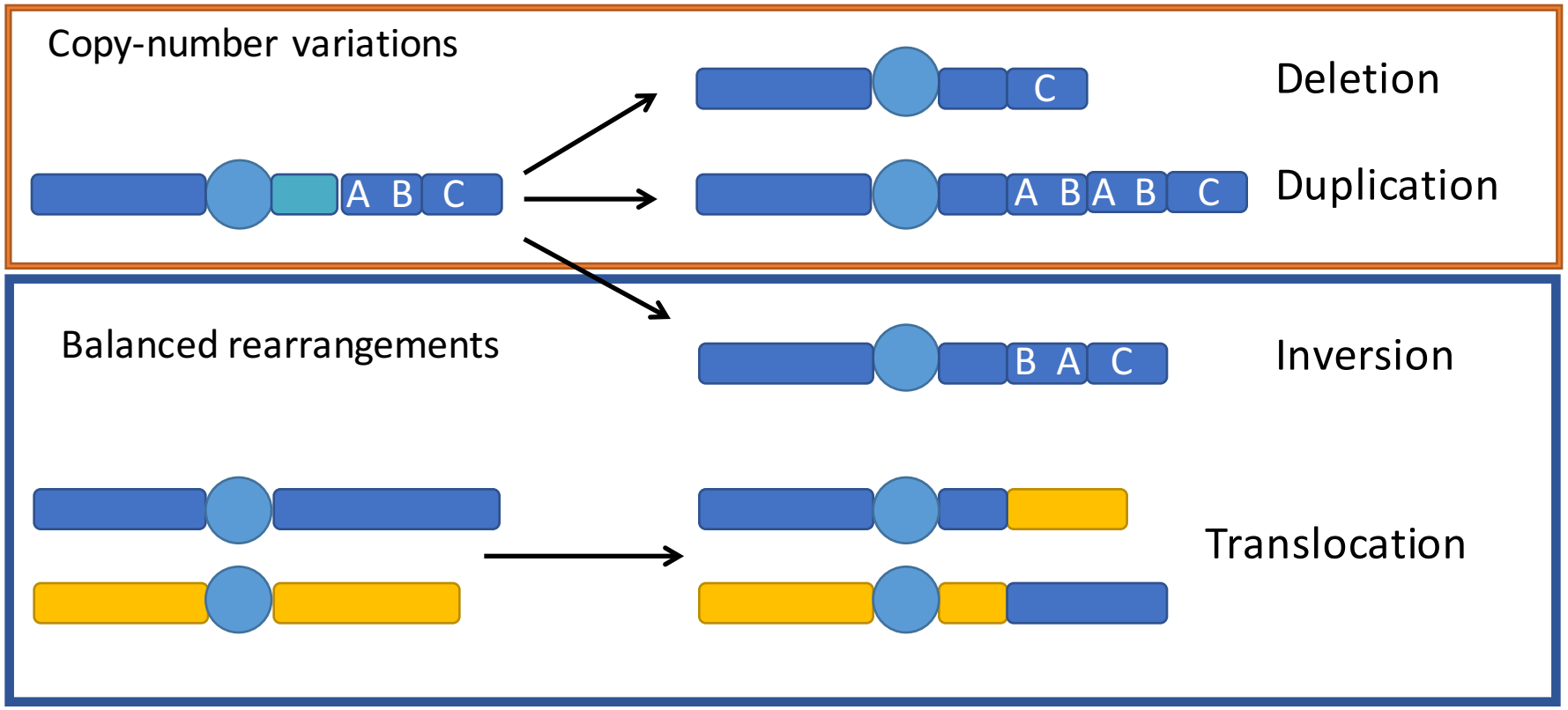


Causes

- Strand slippage
- Aberrant repair
- Retrotransposons

https://hackbrightacademy.com/blog/indel-finder-how-the-python-version-of-this-program-works/

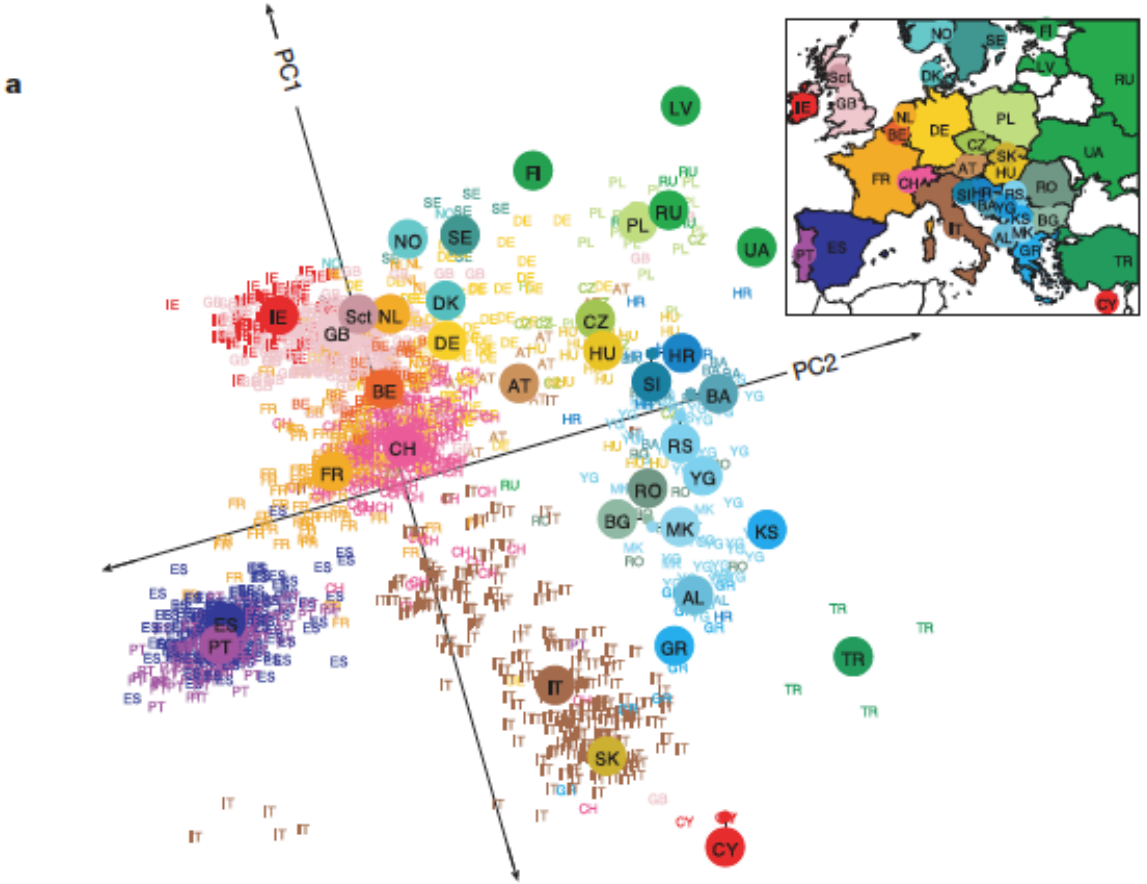
Structural variation (SVs)



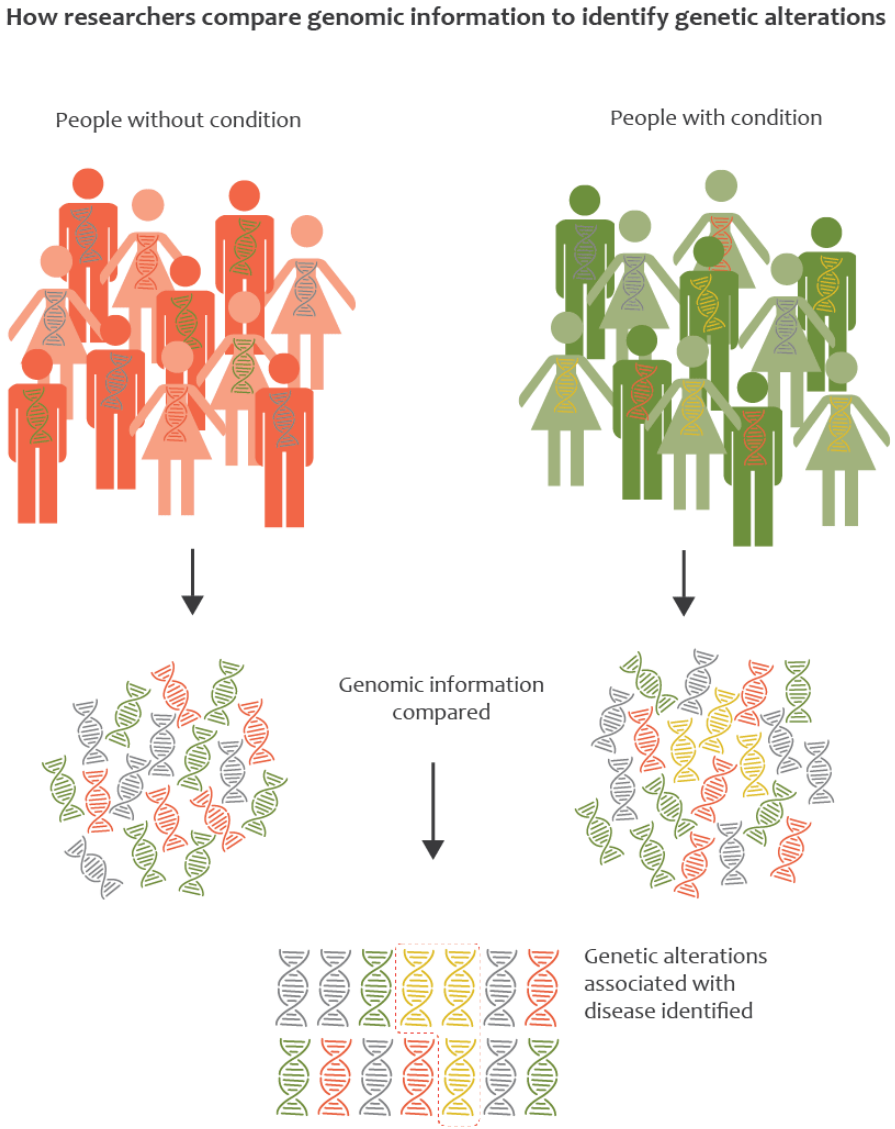
Causes

- Replication errors
- Retrotransposition
- Repair errors
- Recombination errors

Why do we look at germline variation?



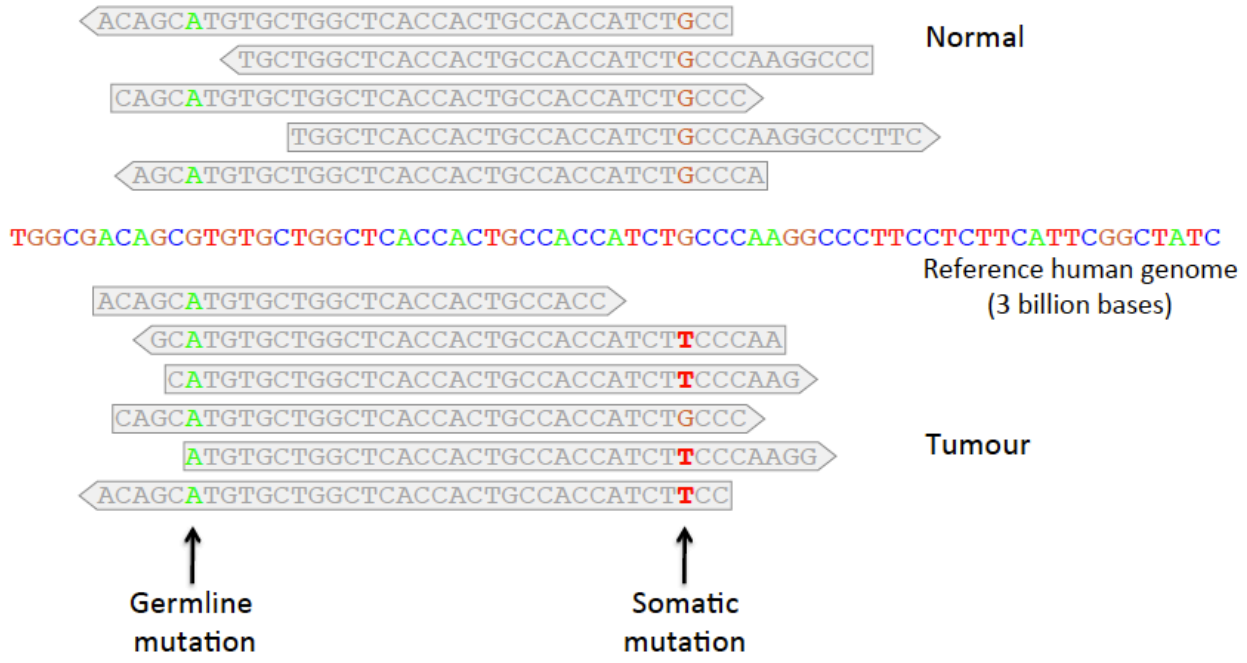
Nature Vol 456|6 November 2008| doi:10.1038/nature07331



Somatic variation

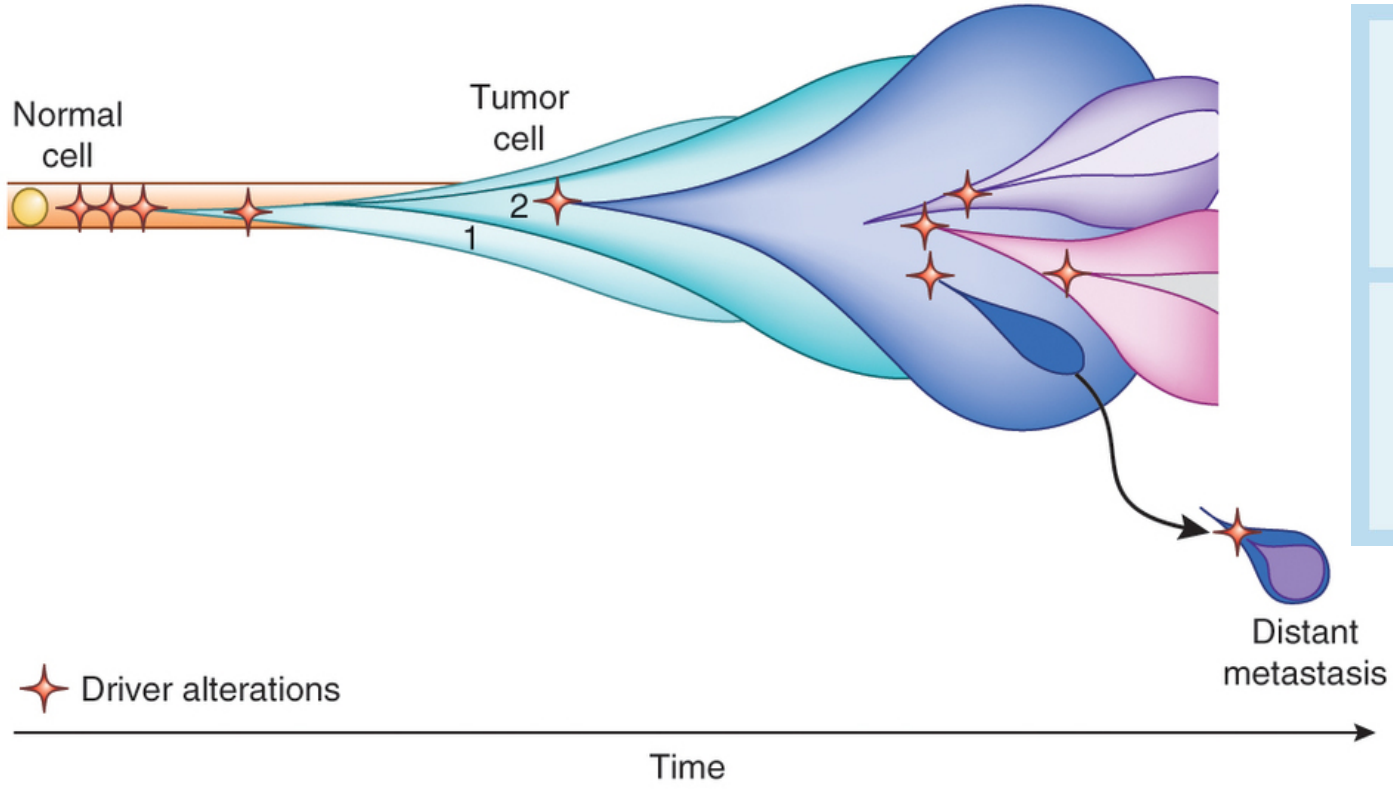
Mutations acquired post conception

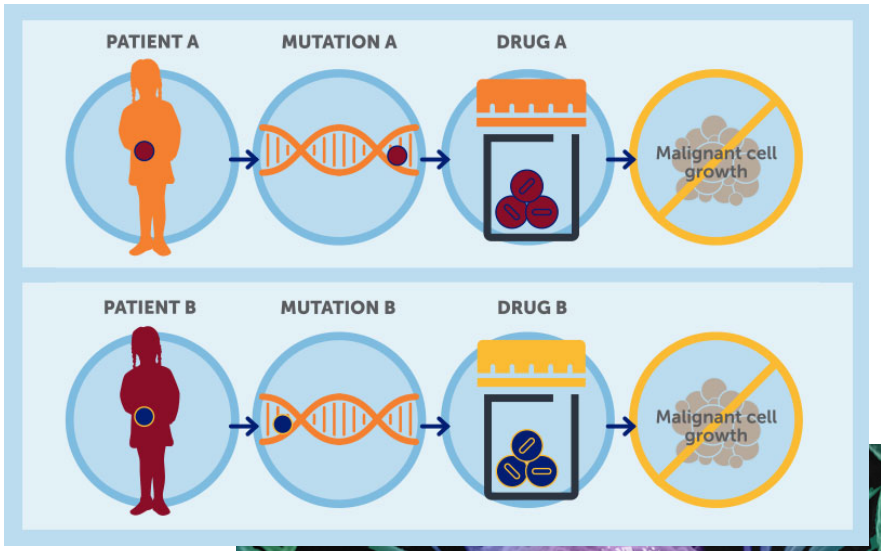
- Single nucleotide variation
  SNVs not SNPs!
- Copy-number aberration
  CNAs not CNVs!
- INDELs
- SVs

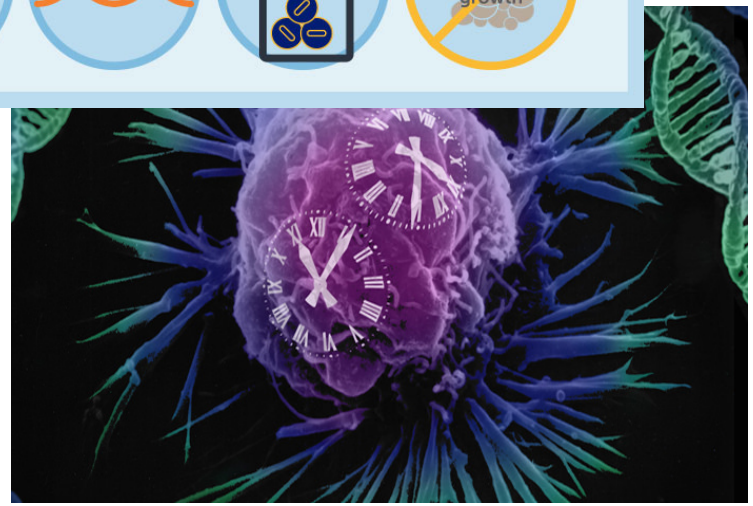


Why do we look at somatic variation?

https://vector.childrenshospital.org/2016/02/precision-cancer-medicine-in-pediatric-oncology/







Nature Medicine **21**, 846–853 (2015) doi:10.1038/nm.3915

Nature Genetics **47**, 1402–1407 (2015) doi:10.1038/ng.3441

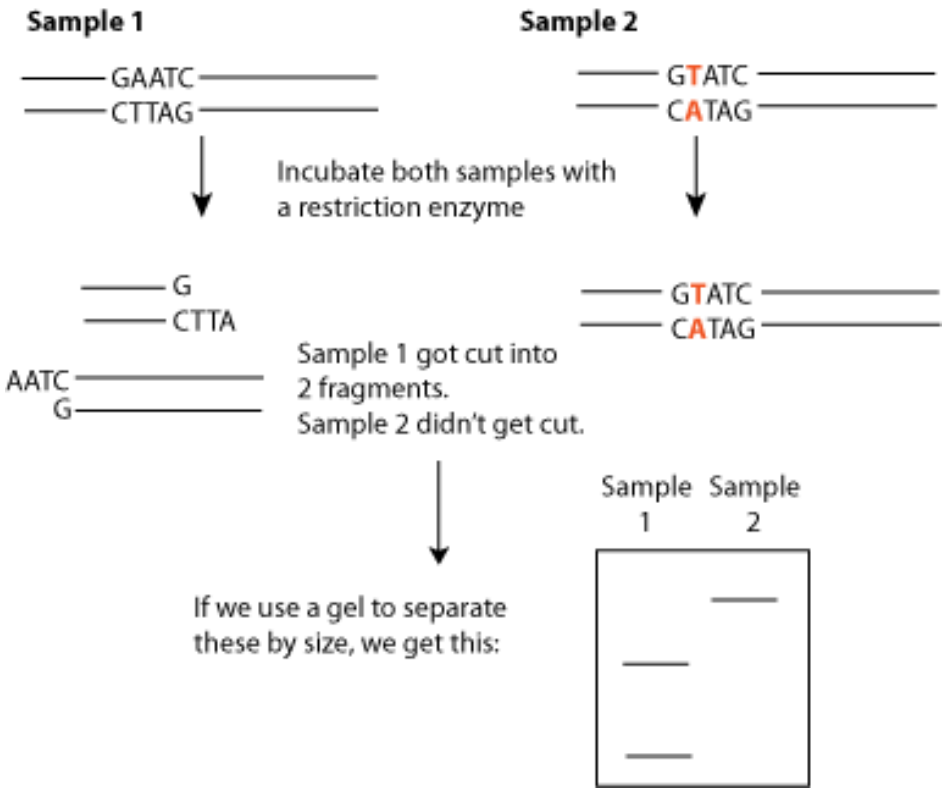
Genotyping

Determining the genetic makeup of individuals

Genotyping

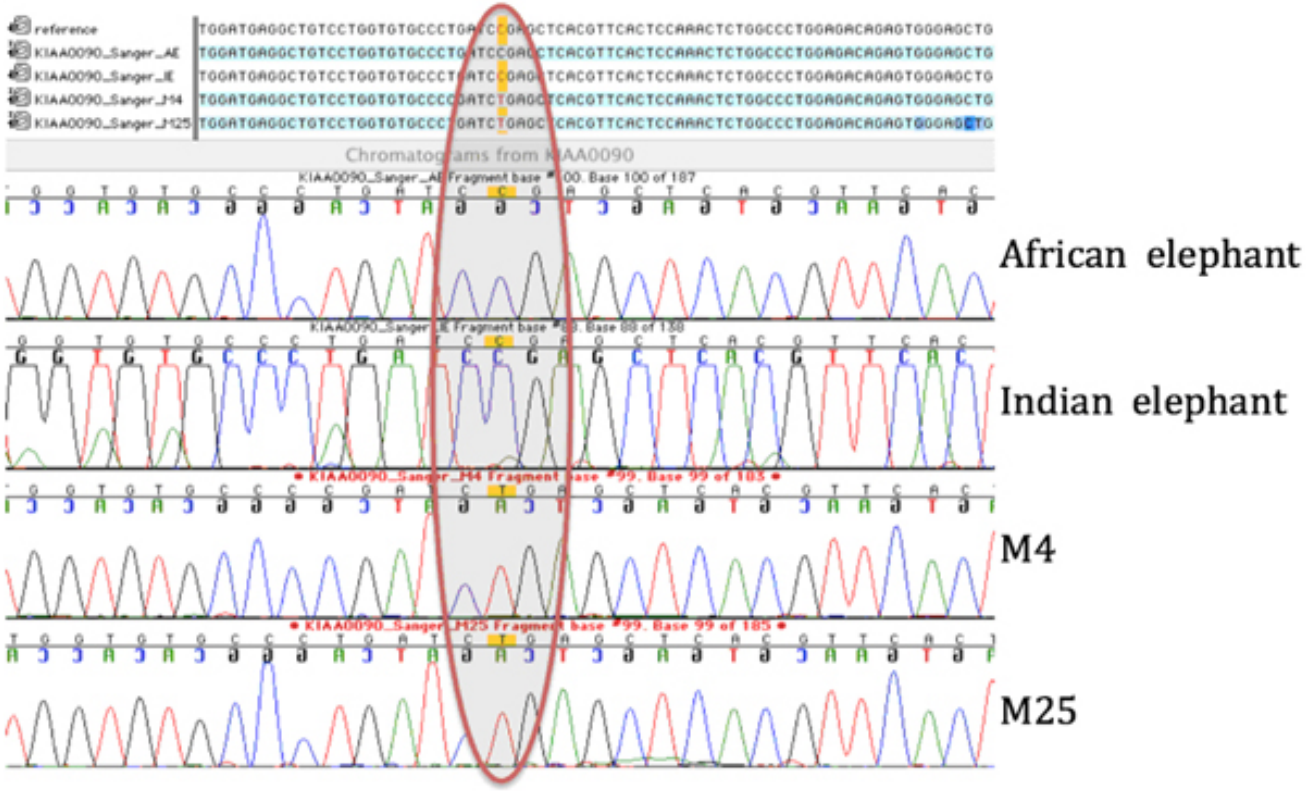
Requires comparison to a reference genome, or a comparison amongst individuals

RFLP – Restriction fragment length polymorphism



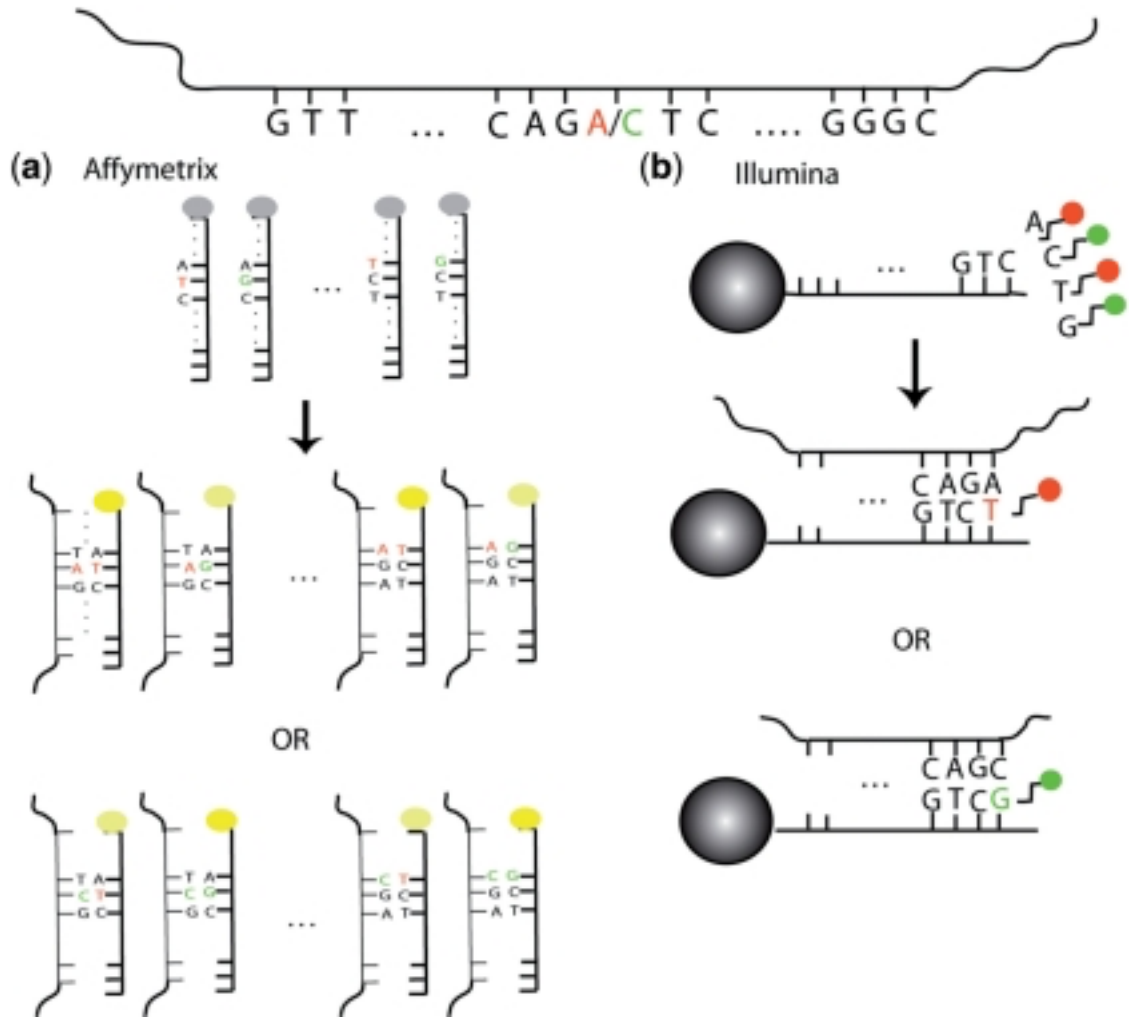
http://scienceblogs.com/digitalbio/2008/05/29/genetic-variation-ii-what-is-a/

Sanger sequencing

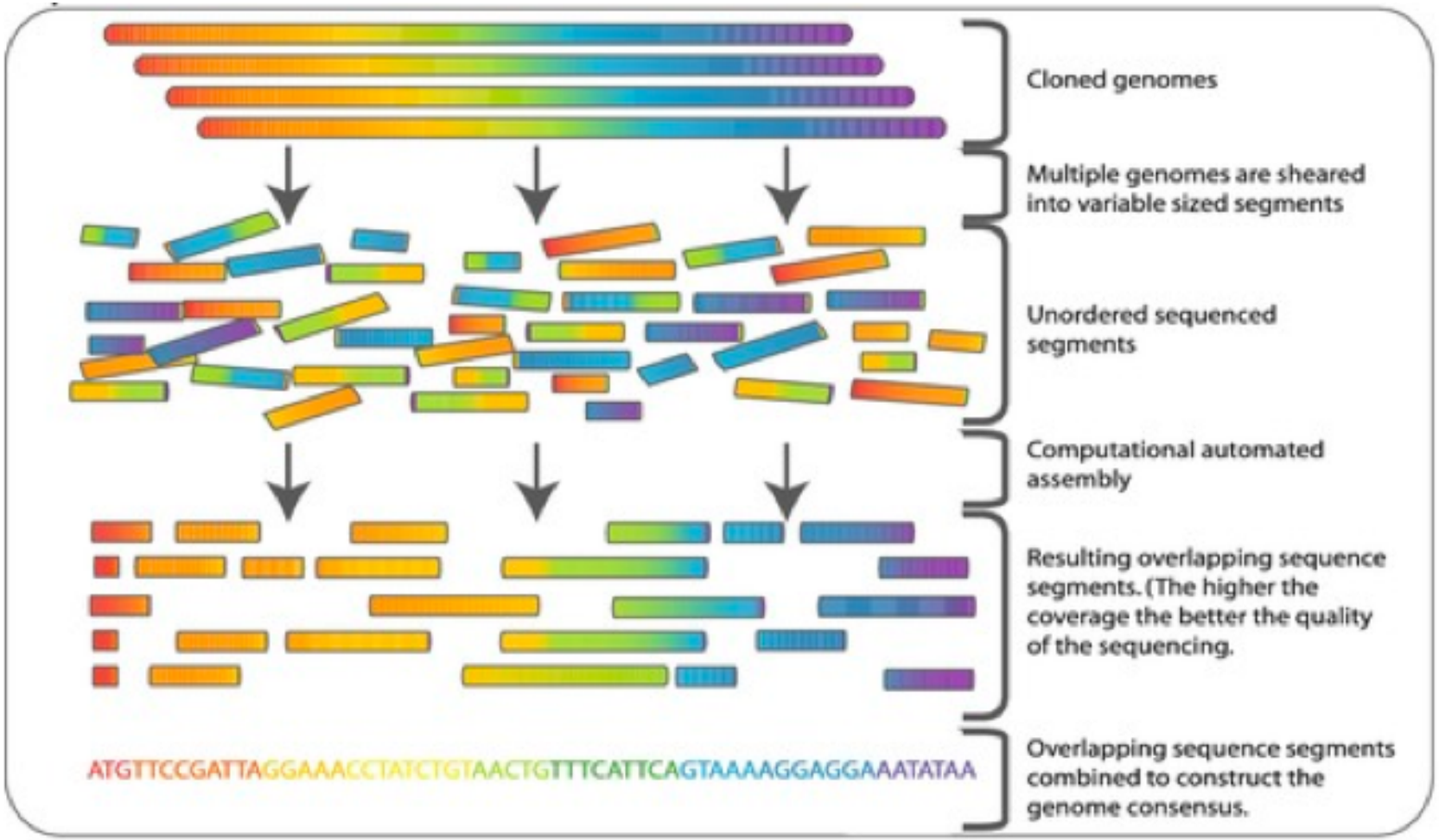


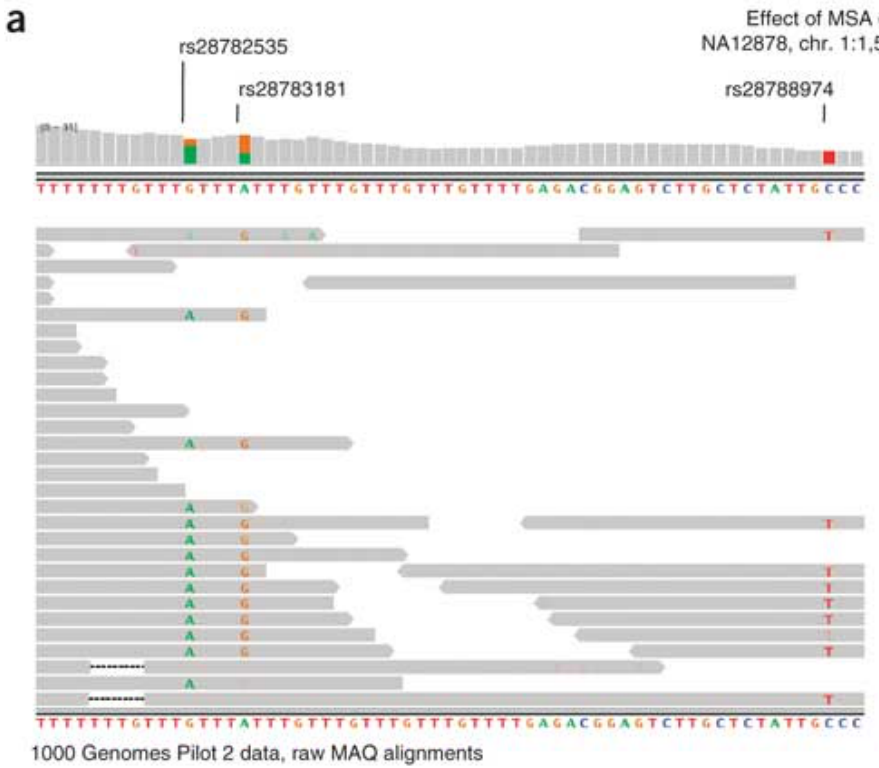
http://mammoth.psu.edu/ultraConserved.html

Array based detection of SNPs (hybridisation)



(Whole-genome) sequencing

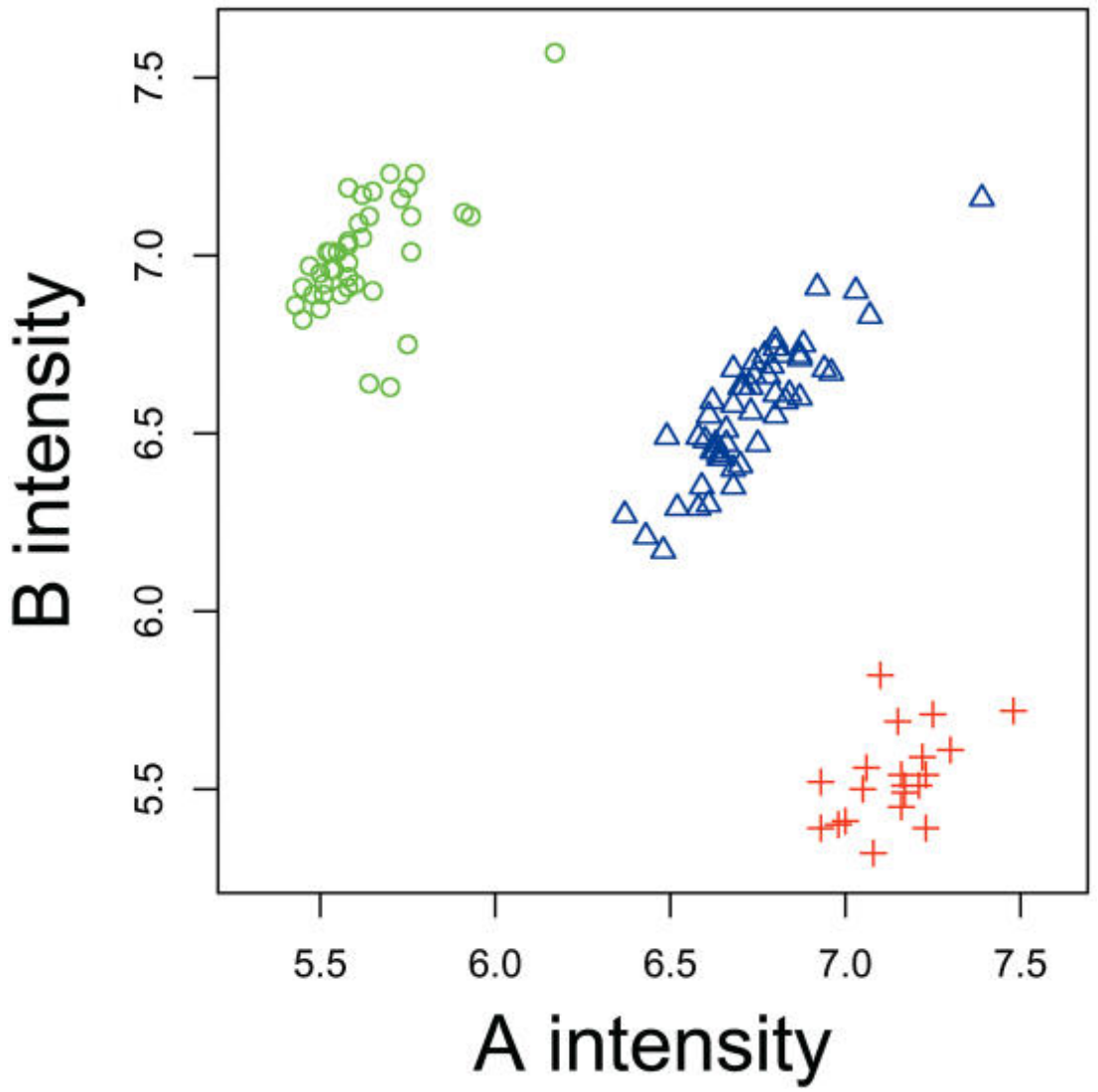


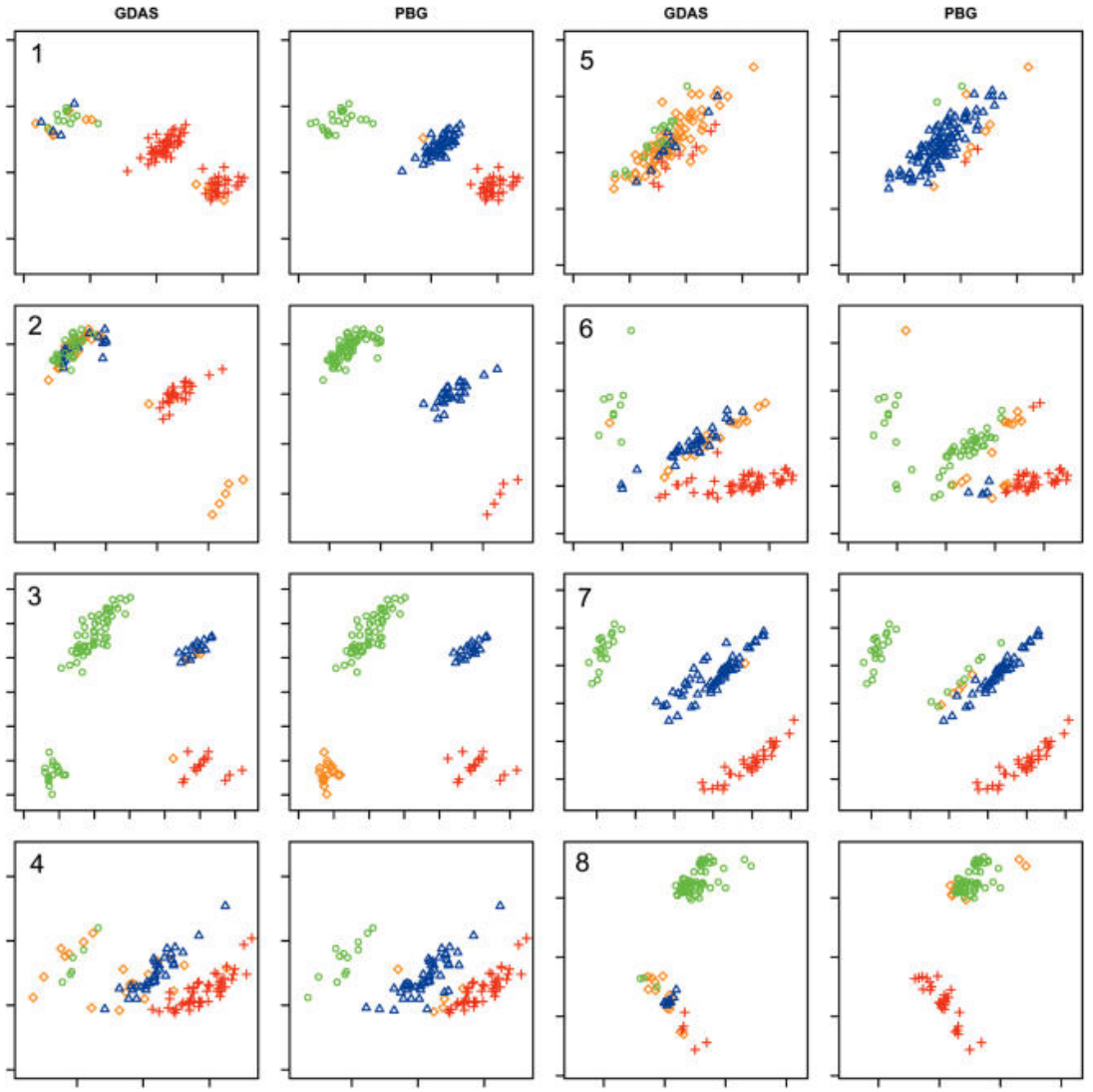


Nature Genetics **43**, 491–498 (2011) doi:10.1038/ng.806

https://commons.wikimedia.org/wiki/File:Whole_genome_shotgun_sequencing_versus_Hierarchical_shotgun_sequencing.png

SNP calling using affymetrix arrays

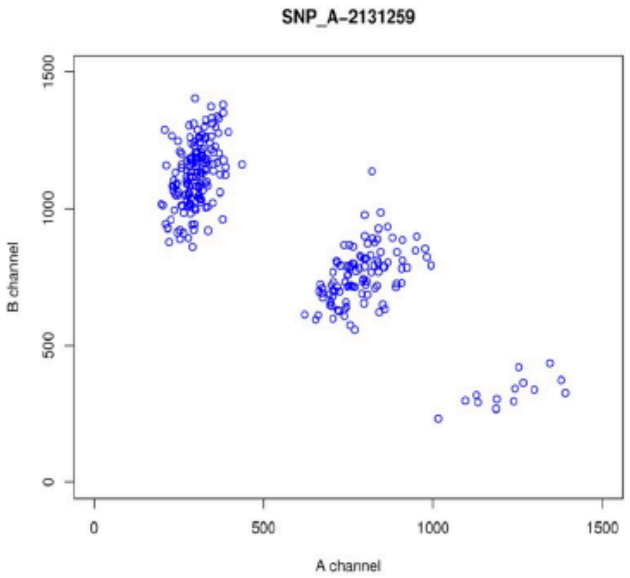


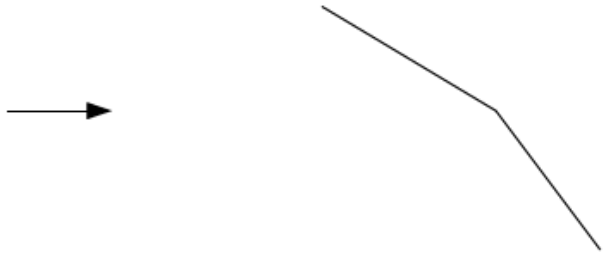


Nucleic Acids Res. 2006; 34(14): e100 2006 doi: 10.1093/nar/gkl475

Birdseed – SNP Caller

In Phase I, Birdseed builds models of all SNPs by using a training data set (Hapmap)



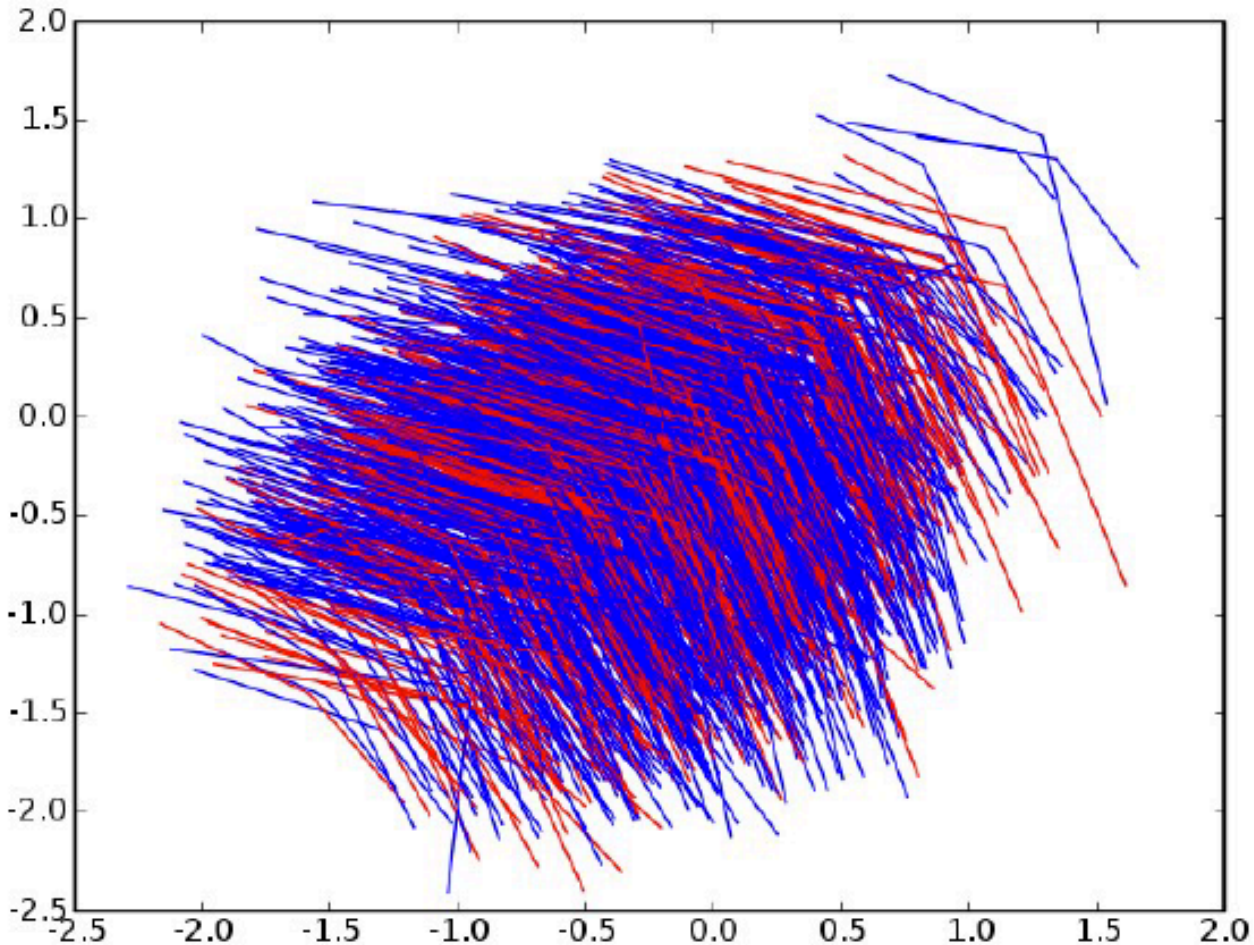


Each SNP can be thought of as a bird. The wingtips are AA and BB, the body is AB. Birds are computed for all SNPs.

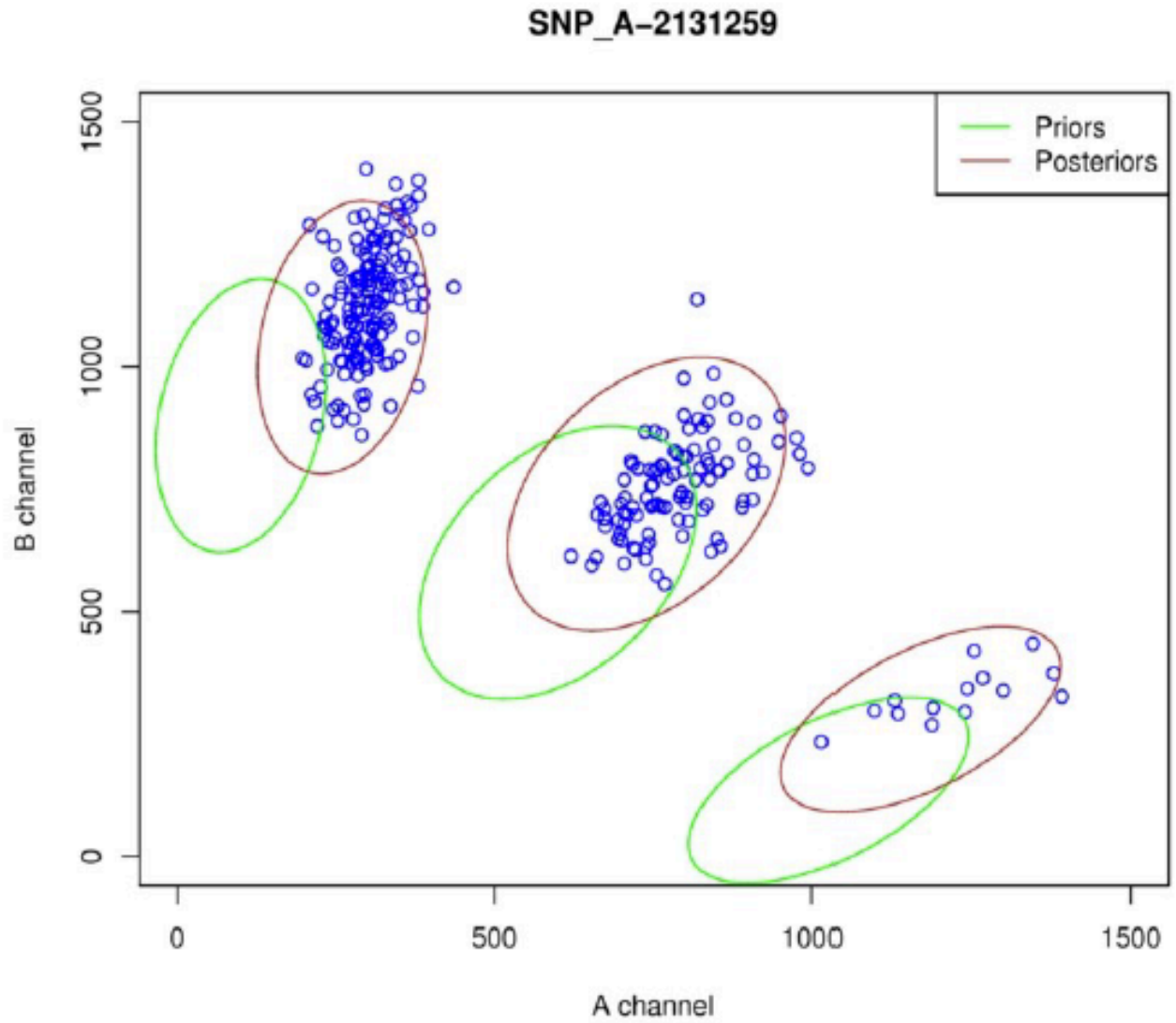
AA: 1.1671 0.3133 0.0108 0.0039 0.0028 14
AB: 0.7499 0.7224 0.0056 0.0034 0.0089 102
BB: 0.2852 1.0713 0.0018 0.0019 0.0125 154

https://www.genome.gov/pages/about/od/opg/gain-workshopii/gain-ii-kuruvillabirdseed.pdf

Birdseed can make highly accurate predictions because it has learned cluster morphology patterns by studying flocks of birds



In Phase II Birdseed uses a highly customized EM algorithm using the SNP-specific bird as the “seed” (hence the name) & as cluster anchors

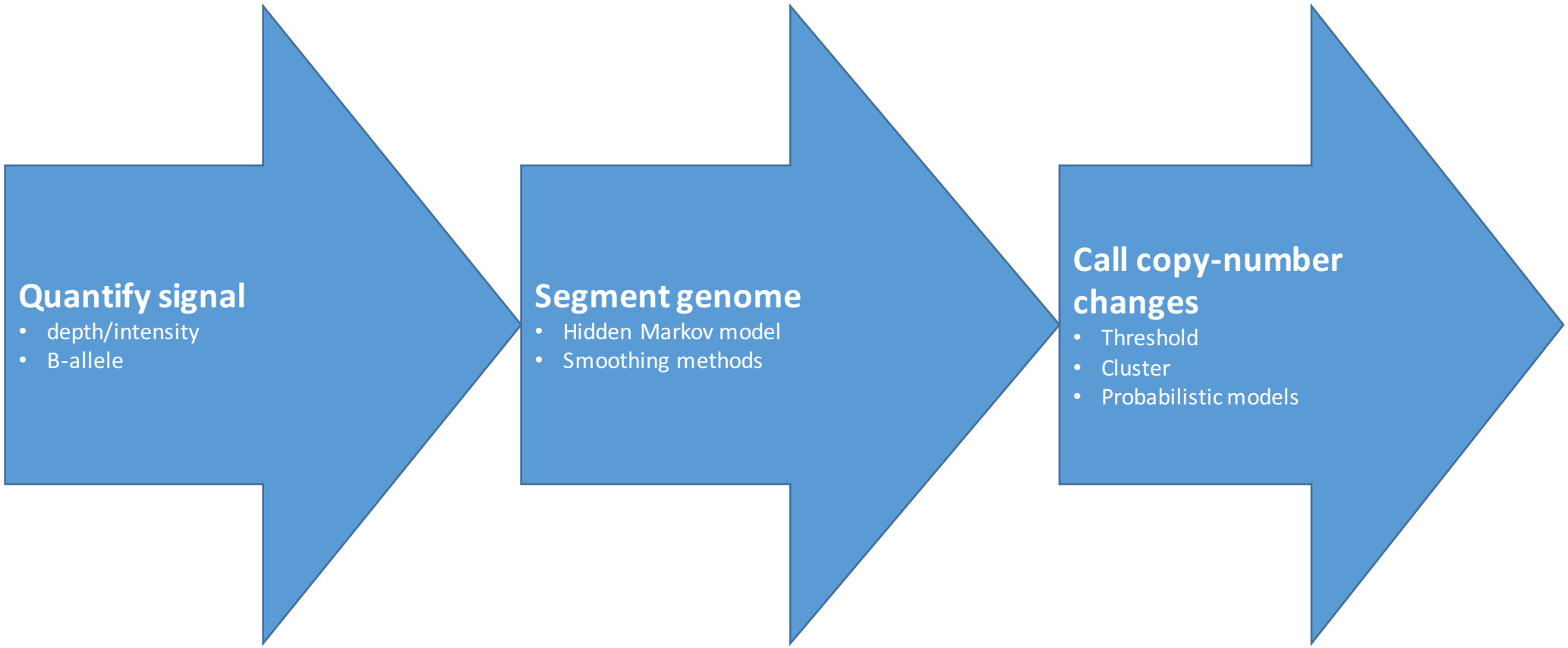


Further reading on genotyping

- 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/

Copy-number calling

Basic workflow



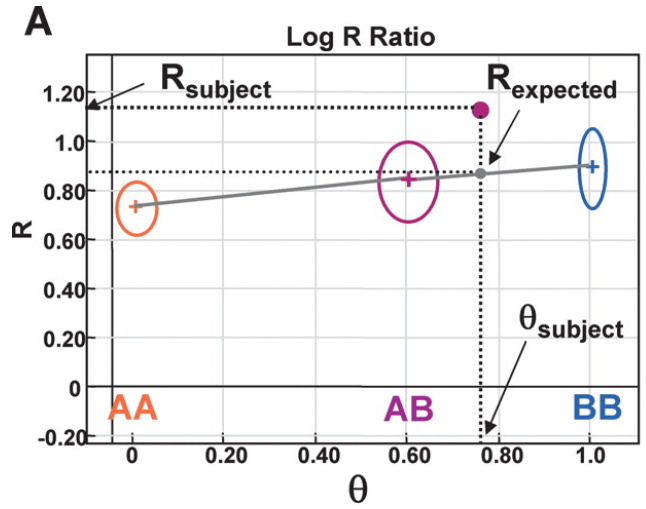
Signal: depth of coverage/intensity

Array based methods use logR:

$R_{subject}$ = normalised intensity of probes from sample

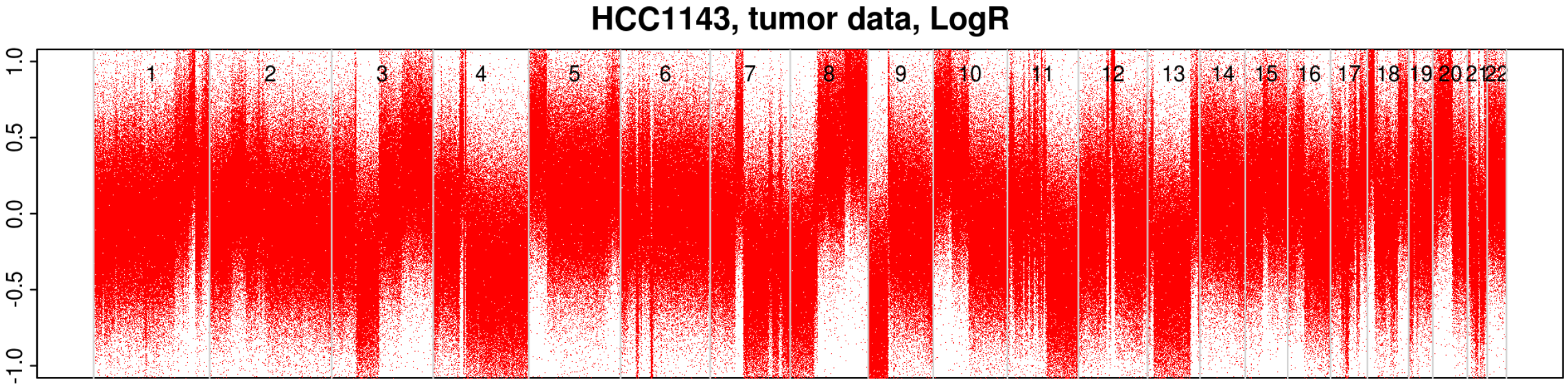
$R_{expected}$ = normalised intensity of probes from control

$$logR = \log_2(\theta_{observed}/\theta_{expected})$$



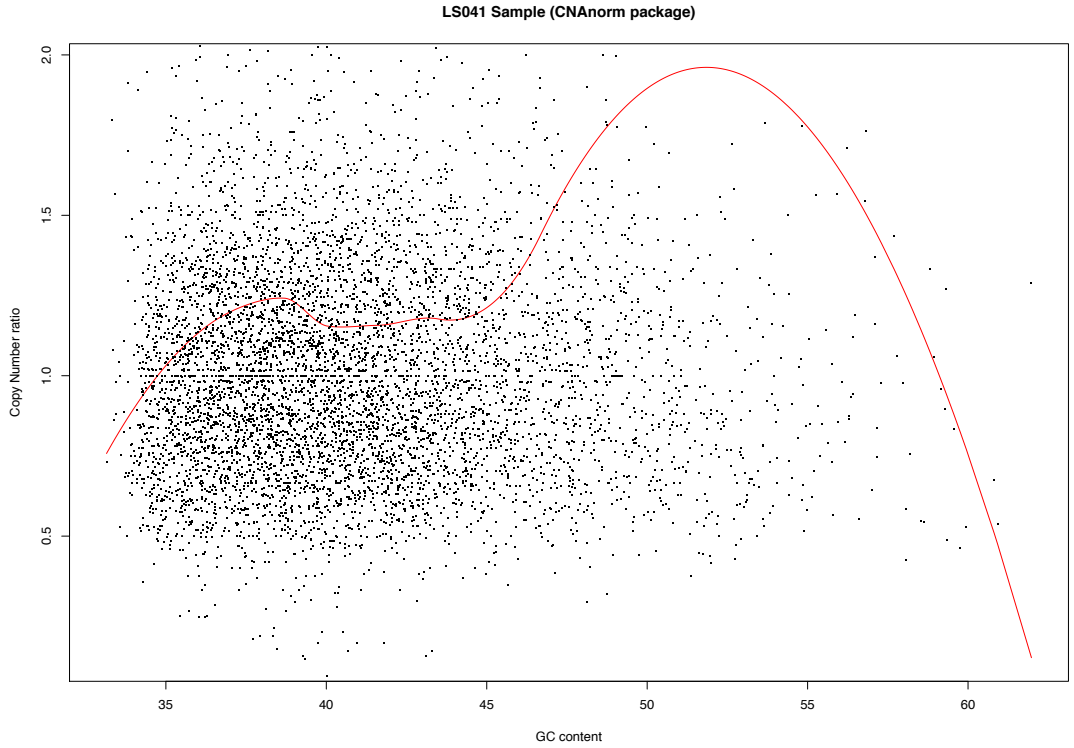
Sequence based methods use read depth (also called logR).

logR of HCC1143 cell-line using affy SNP6



logR/depth normalisation

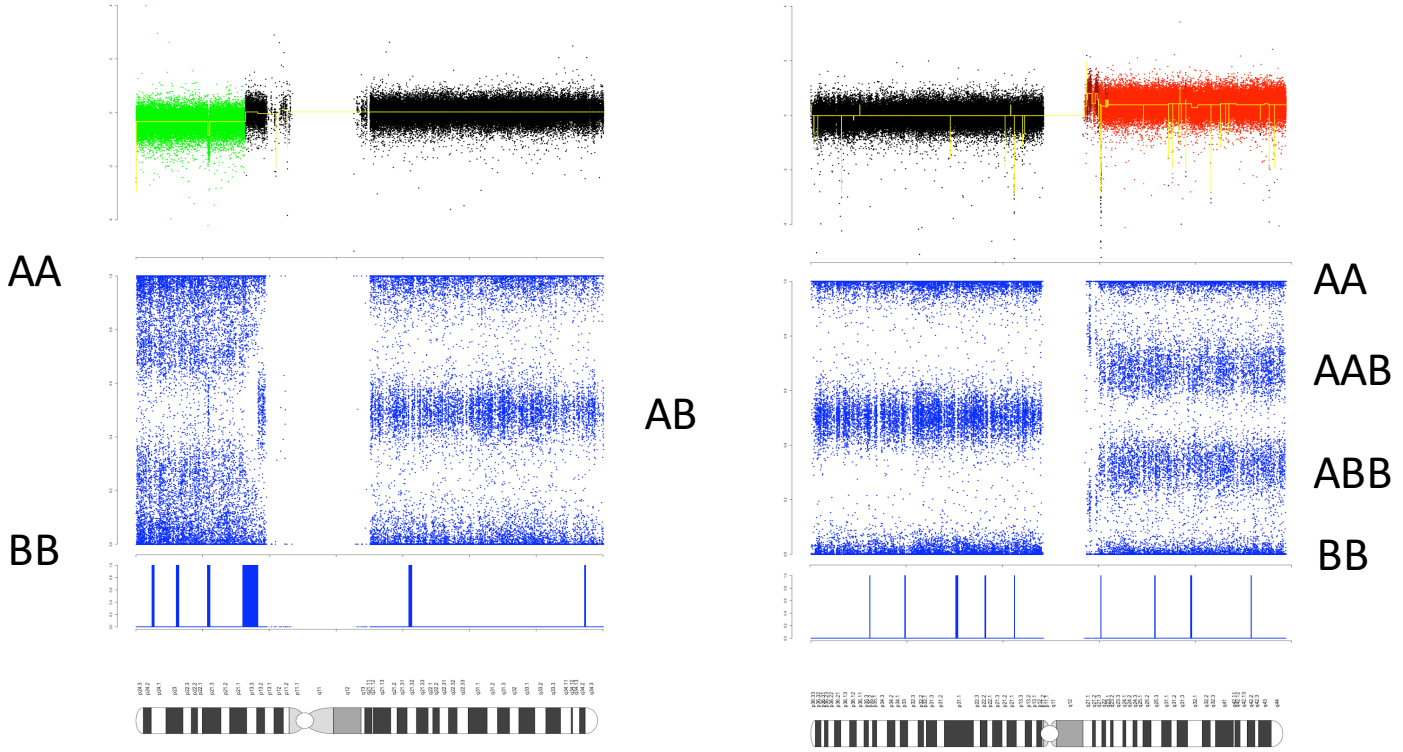
- Different proportions of GC in each region can produce a bias in the read depth (wave artifact)
- We can fit a loess model and remove the effect.



Signal: B-allele frequency

θ_A= intensity of probe for allele A

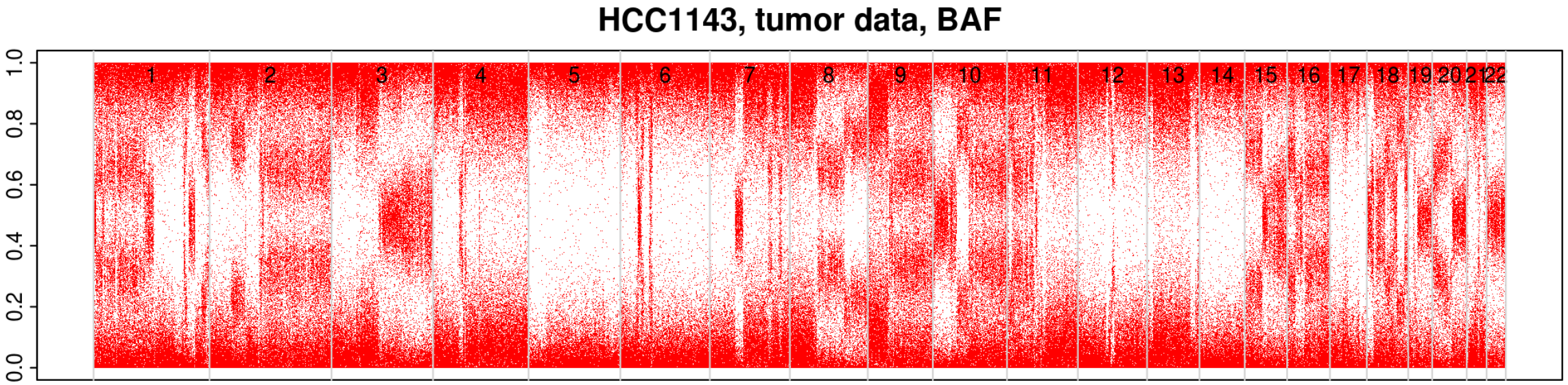
$BAF = \theta_A/(\theta_A + \theta_B)$



BAF banding

- **1 band:**
  - Background noise (0 copies).
- **2 bands:**
  - {A,B}, {AA,BB}, or {AAA,BBB},… Copy numbers (0, 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 of HCC1143 cell-line using affy SNP6



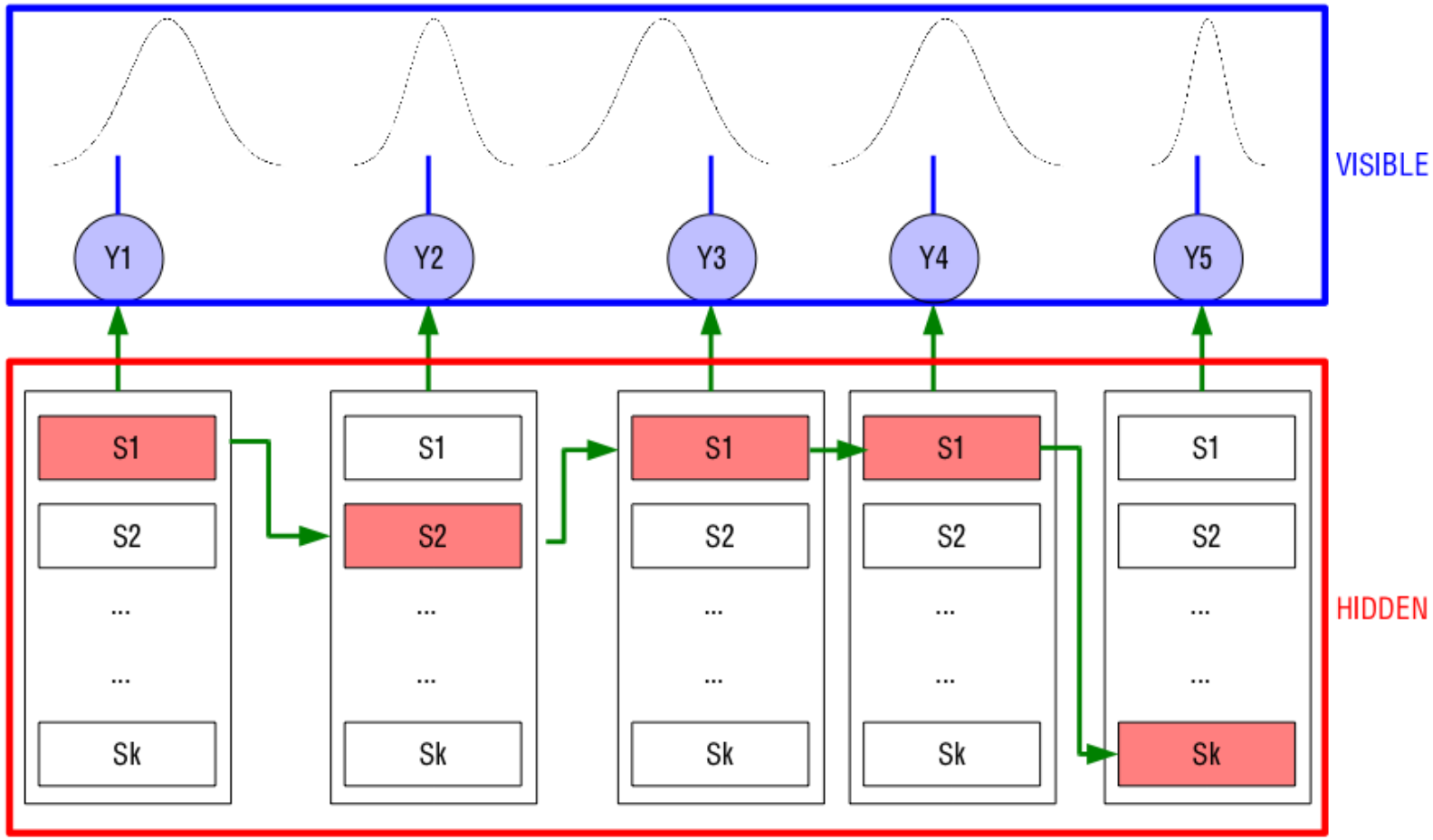
Segmentation: Circular binary segmentation

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).

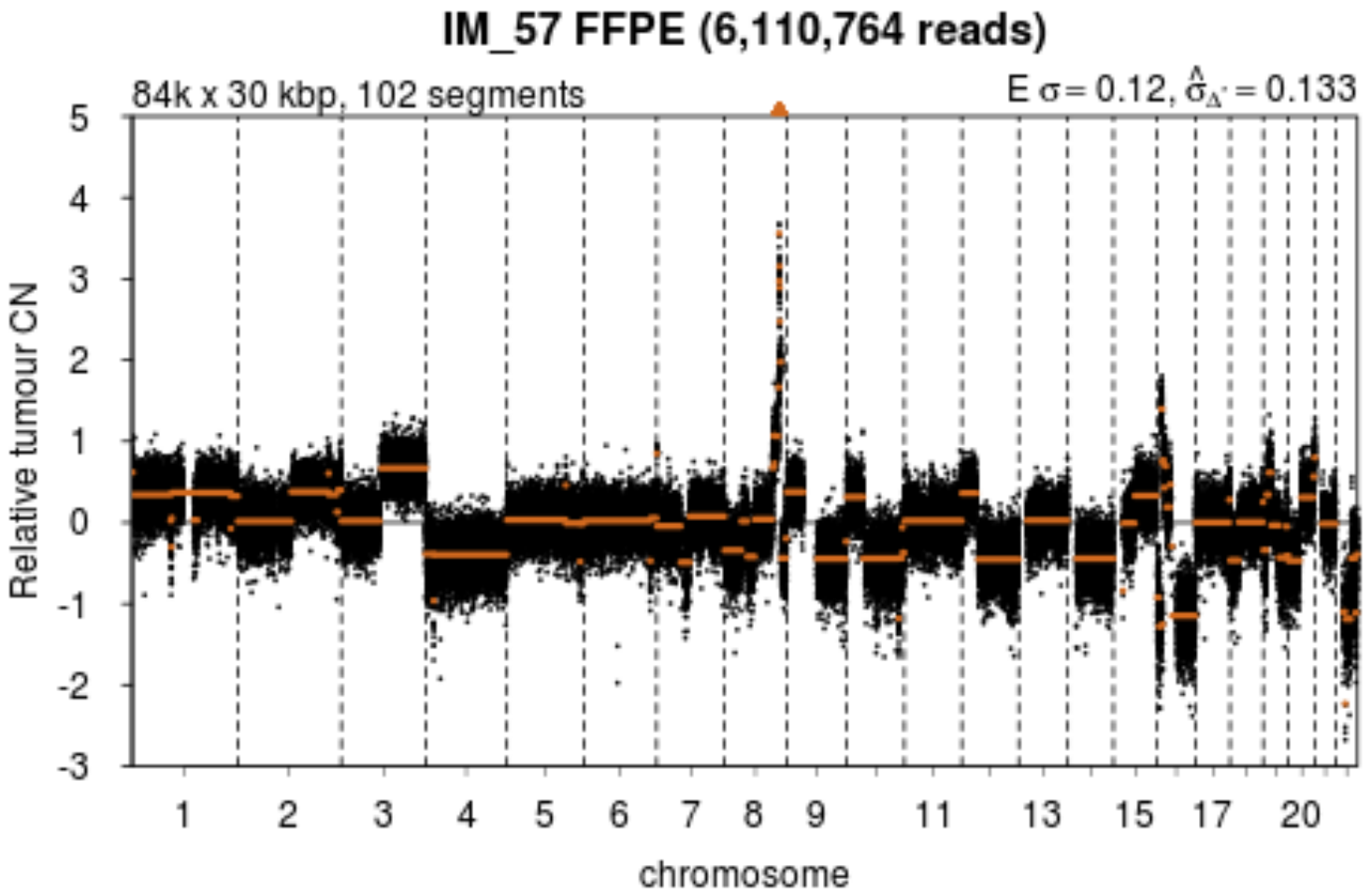
- The segmented means come from a mixture of six normal populations.
- The model is fit by EM algorithm.
- Classification reduced to 3 or 4 states. (Usually loss, gain, normal)

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

Profile types and methods

Relative versus absolute copy-number

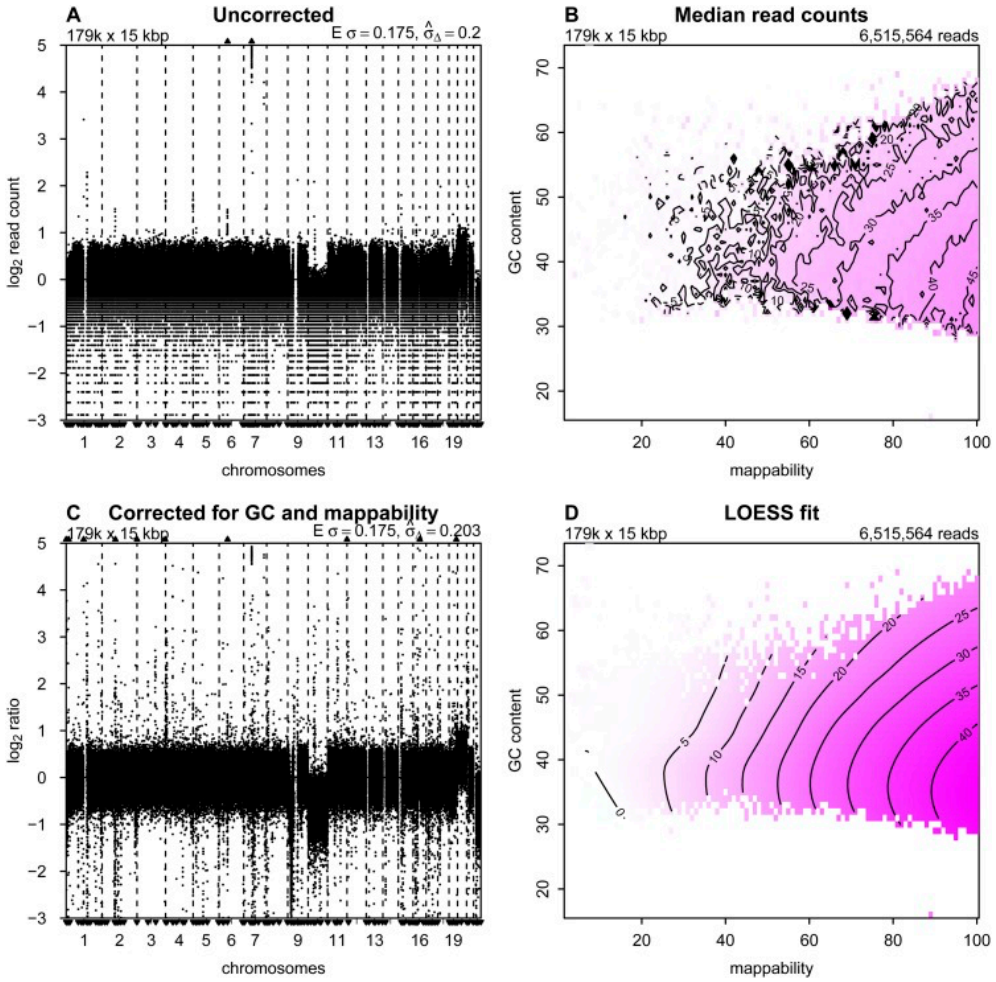
Relative copy-number profile (ovarian cancer)



Method: QDNAseq

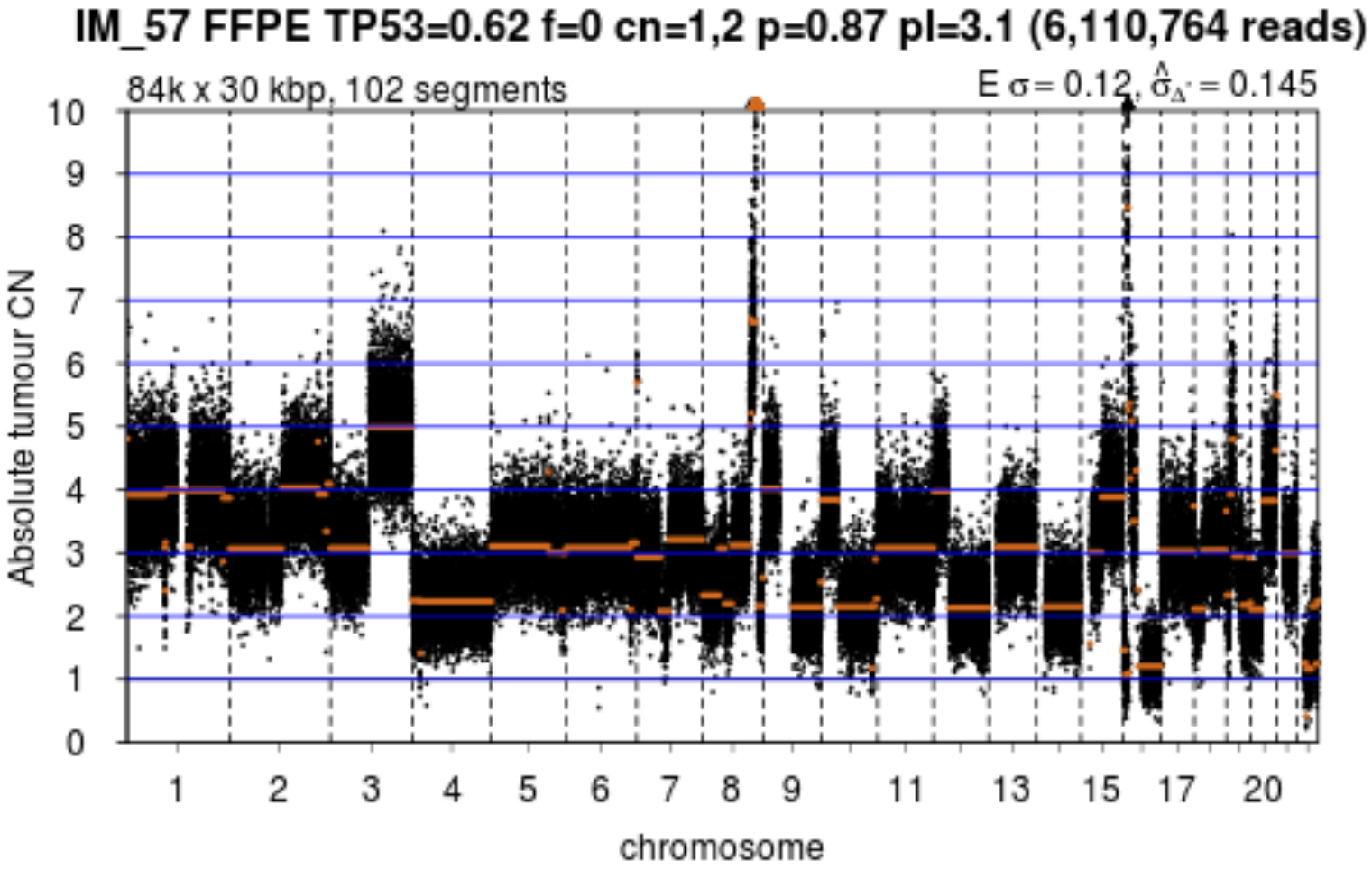
Scheinin I et al., 2014 (QDNAseq Bioconductor package).

- 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/

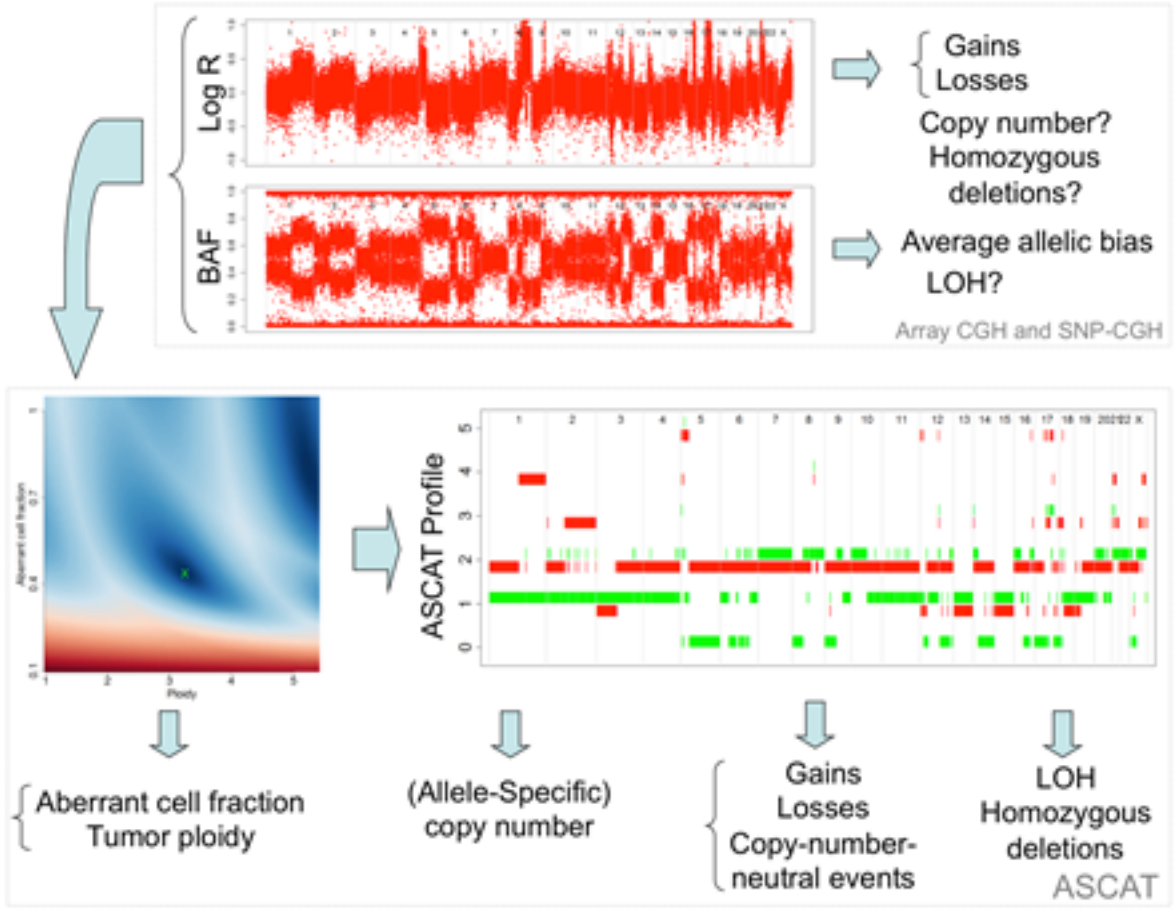
Absolute copy-number profile (ovarian cancer)



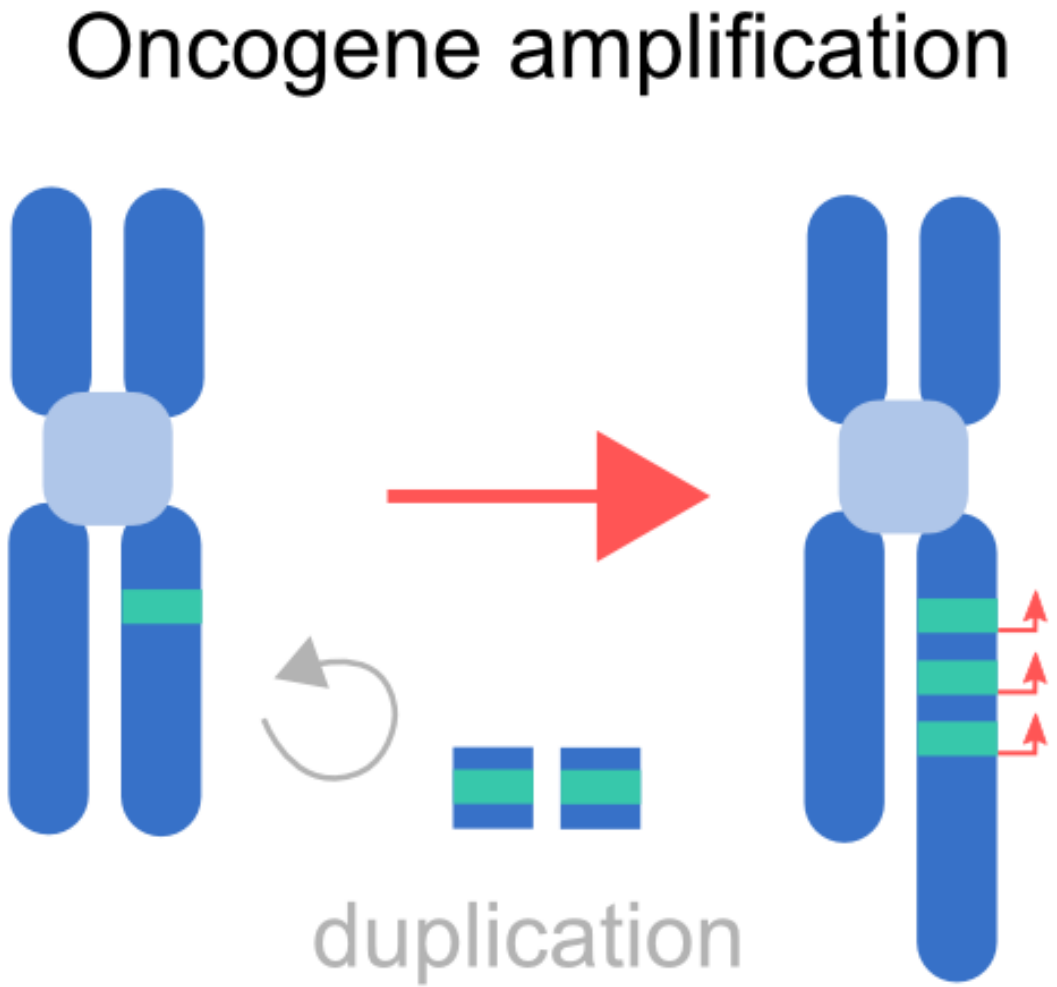
Method: Allele-Specific Copy number Analysis of Tumours (ASCAT)

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

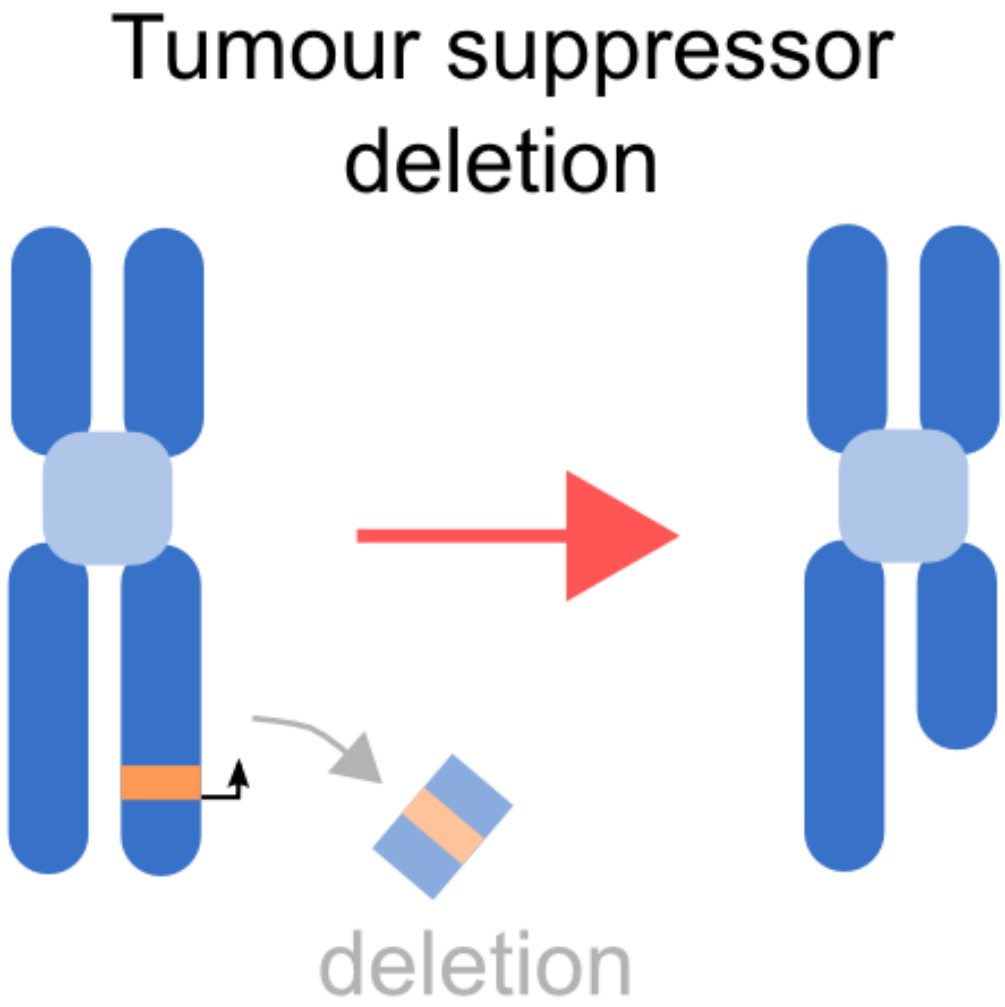
$$b_i = \frac{1 - \rho + \rho n_{B,i}}{2 - 2\rho + \rho\left(n_{A,i} + n_{B,i}\right)}$$



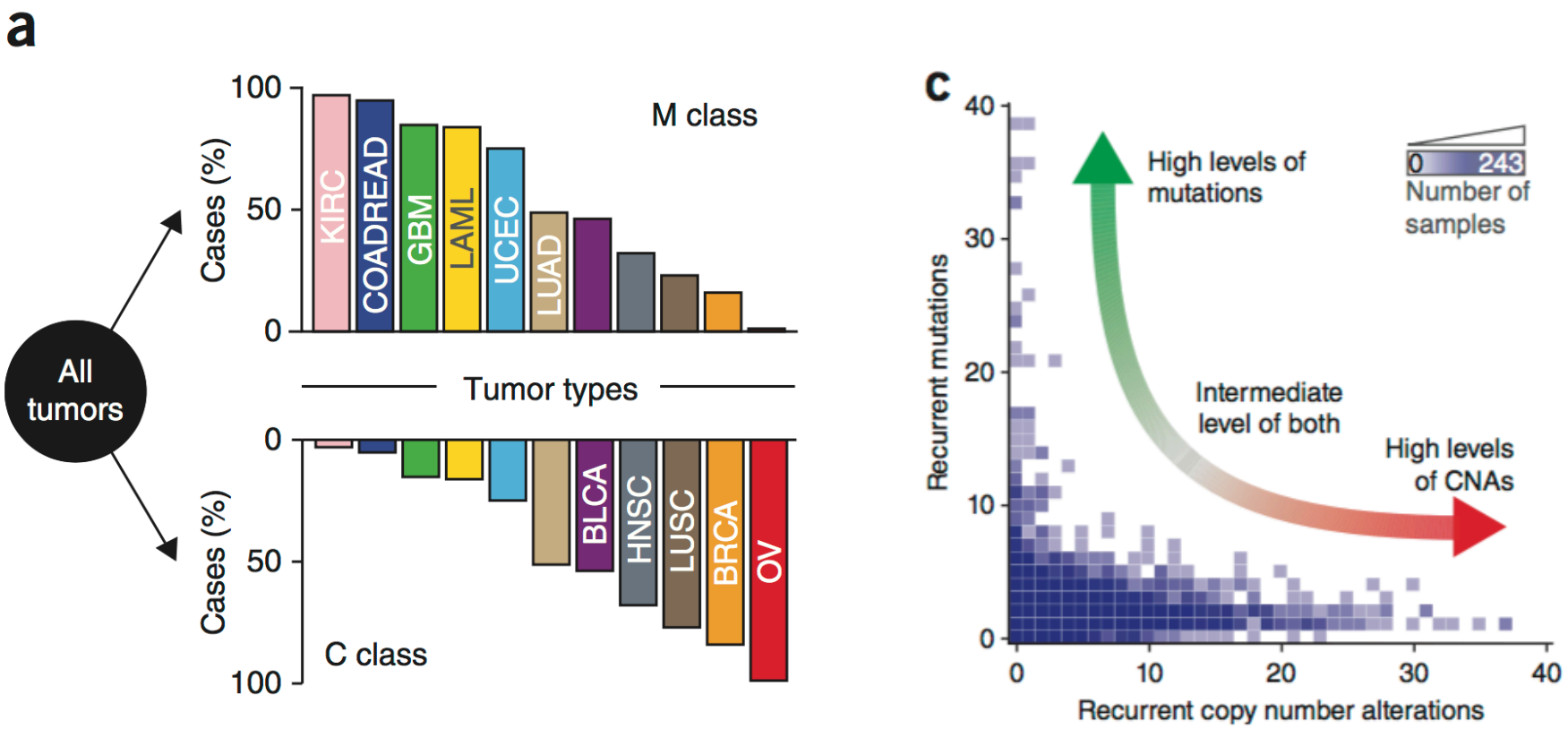
Why is copy-number important?



Why is this important?



Why is copy-number important?



Ciriello et al (2015). Nature Genetics

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