QDNAseq: Quantitative DNA sequencing for chromosomal aberrations

Scheinin I, Sie D, Bengtsson H, van de Wiel MA, Olshen AB, van Thuijl HF, van Essen HF, Eijk PP, Rustenburg F, Meijer GA, Reijneveld JC, Wesseling P, Pinkel D, Albertson DG and Ylstra B.
Bioconductor package QDNAseq
Genome Res., 2014, doi: 10.1101/gr.175141.114

Presentation by:
Christian Rausch
MSc Biotechnology, PhD Bioinformatics
Netherlands Cancer Institute (NKI)
Amsterdam
Netherlands

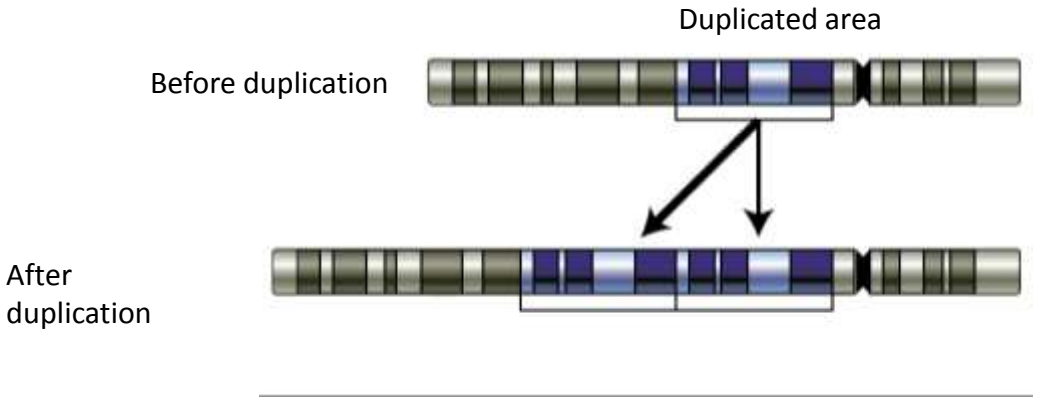
Slides presented:
December 7 & 8, 2015 @ European Bioconductor Developers Meeting: Cambridge, UK

What QDNAseq can do

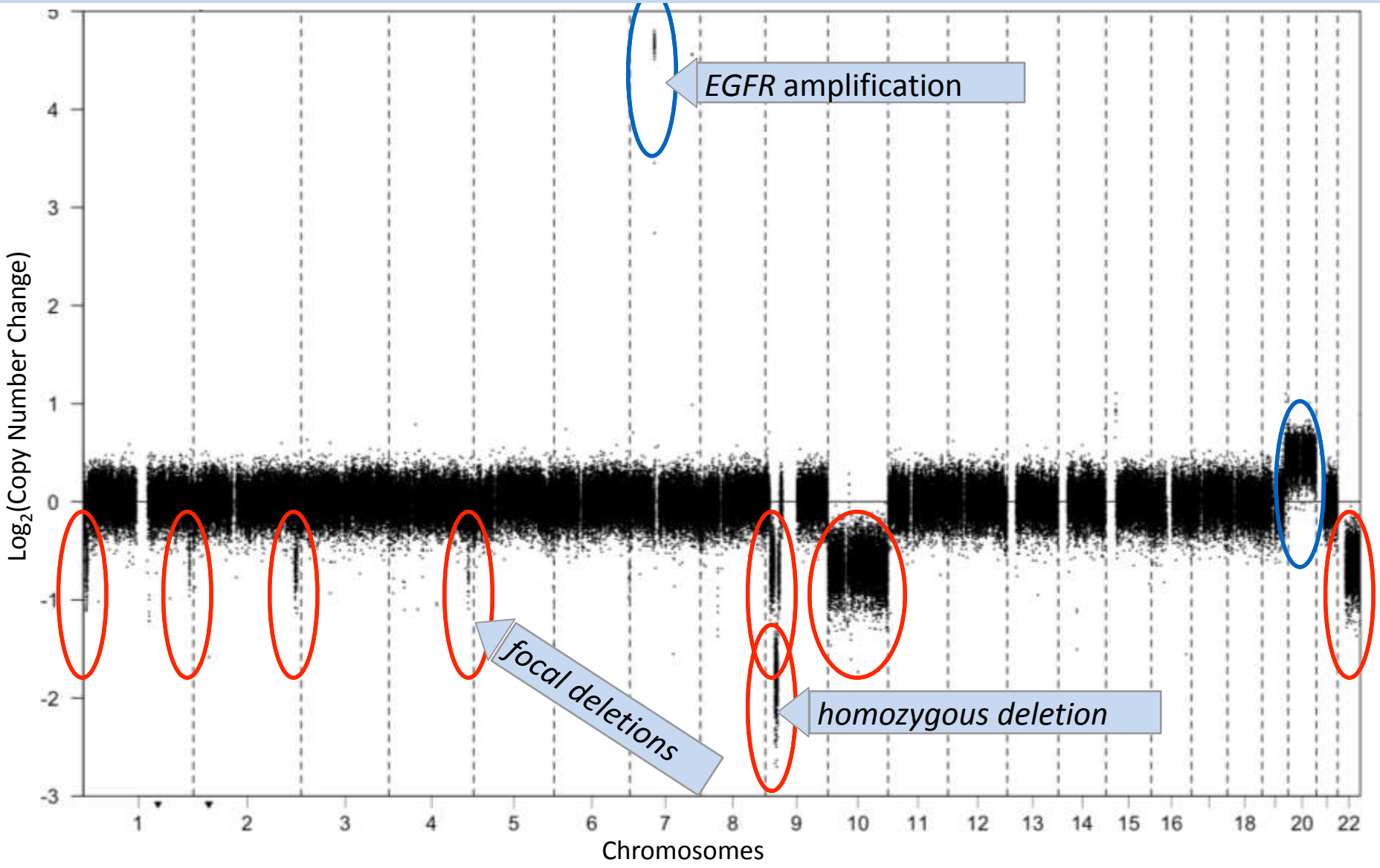
Quantify and visualize Chromosomal Copy Number Aberrations

Background:

In cancer cells, chromosomal regions can be amplified or deleted



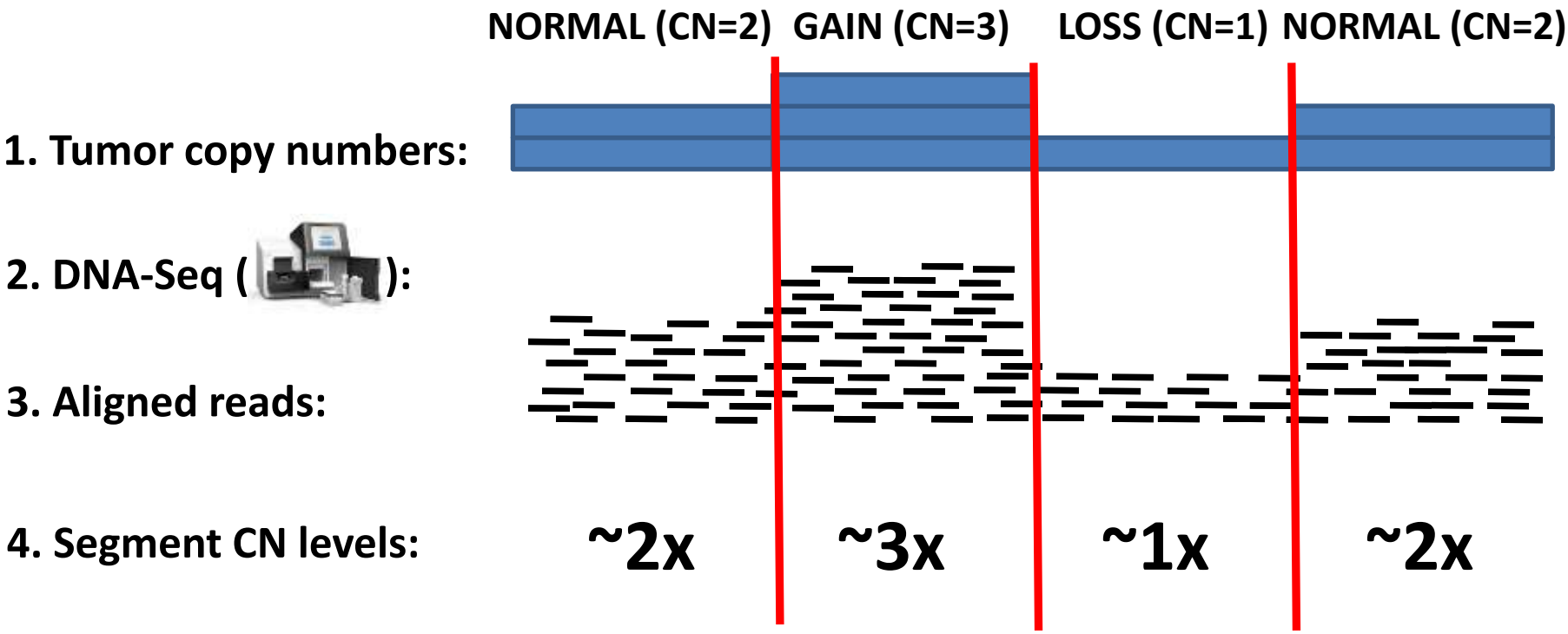
Example Copy Number Plot by QDNAseq



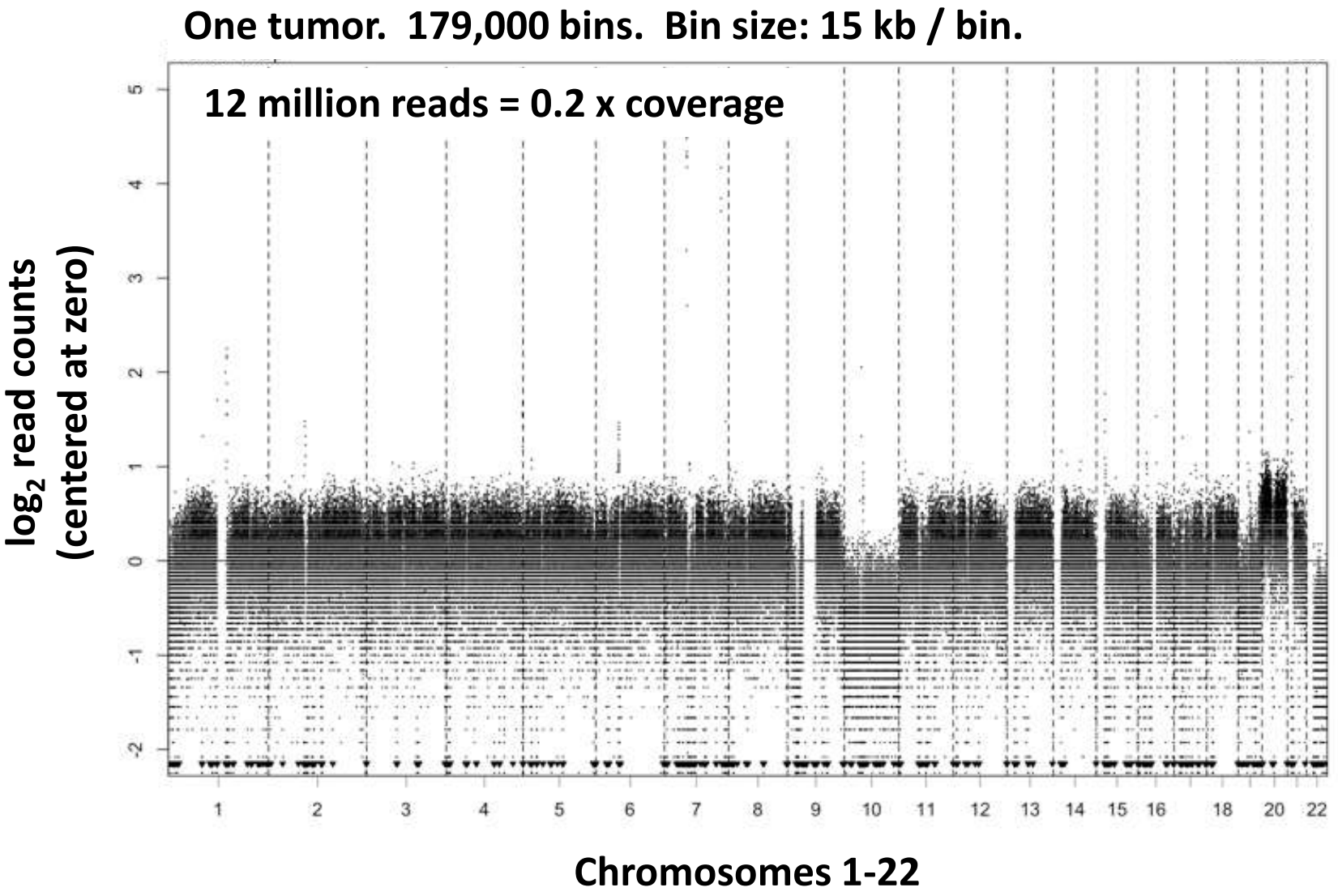
Input to QDNAseq

- Shallow whole genome seq data (0.1x / 6 Million reads)
- 50 bp reads
- Human or mouse
- Select bin size 15, 30, 100, or 1000 kB

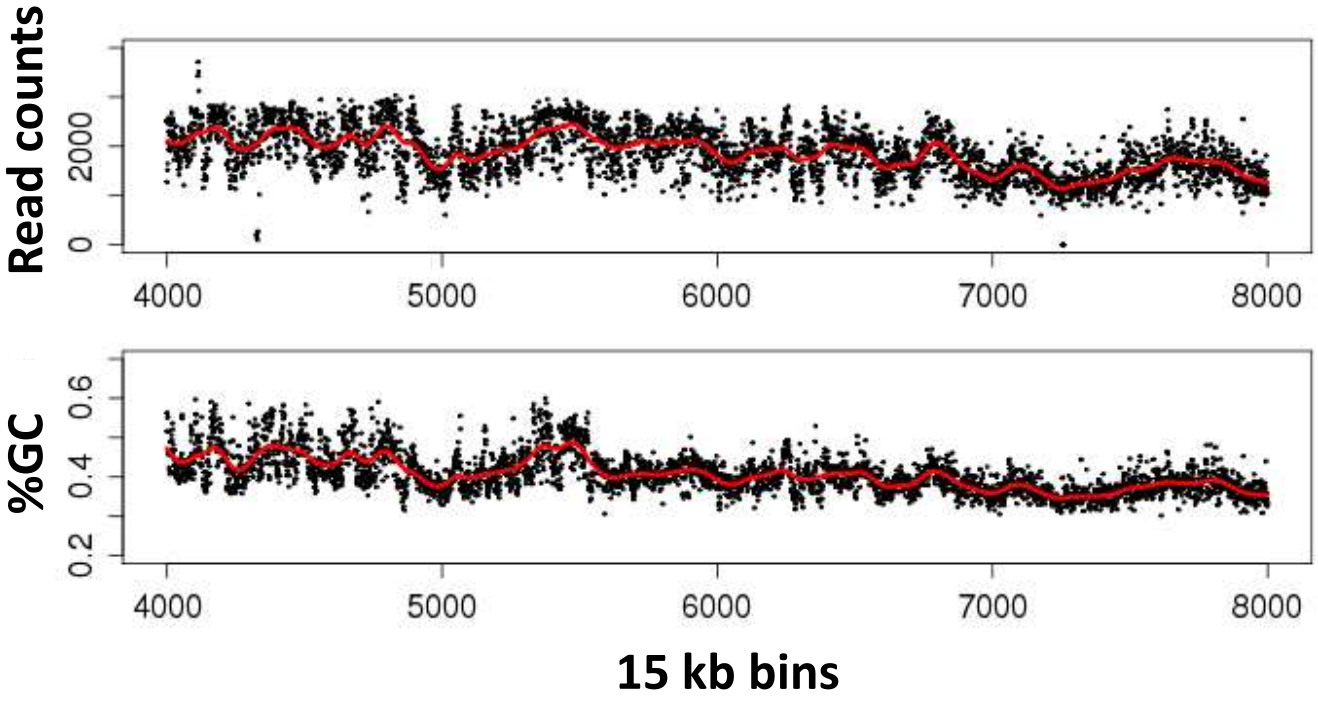
More DNA copies in a region gives more sequencing reads

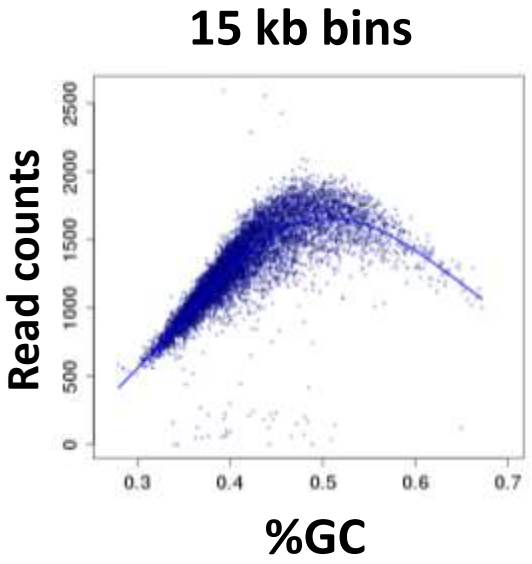


Read counts reveal copy numbers



Read count is a function GC content (%)

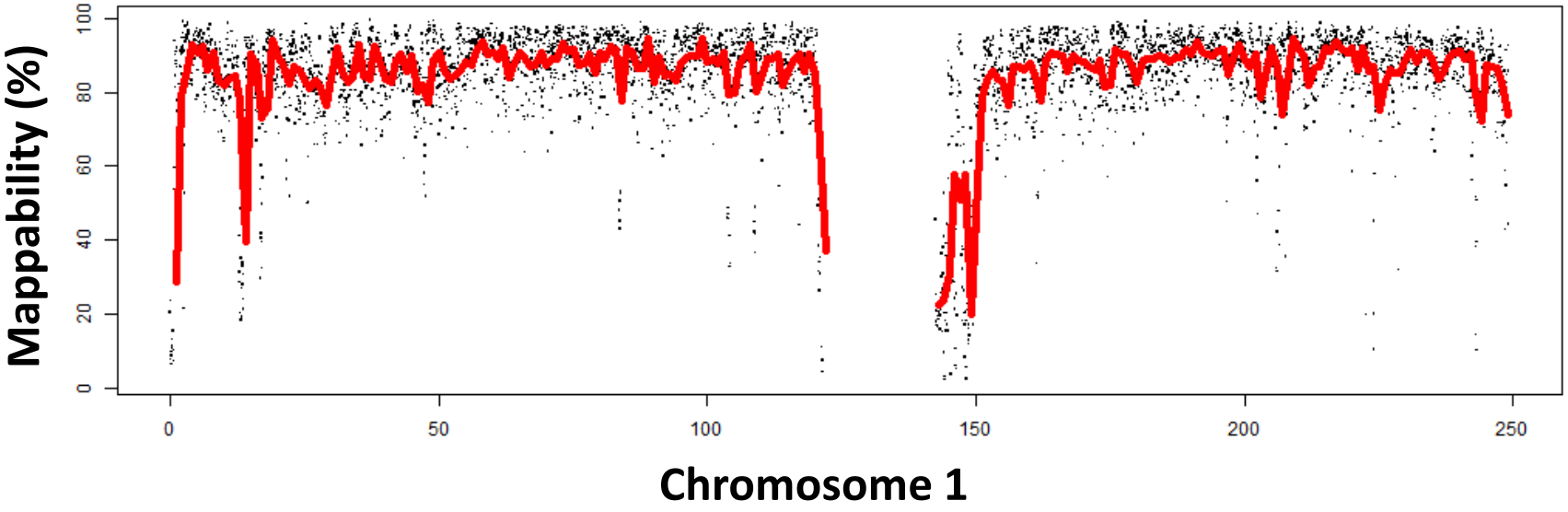




Y. Benjamini (2010)

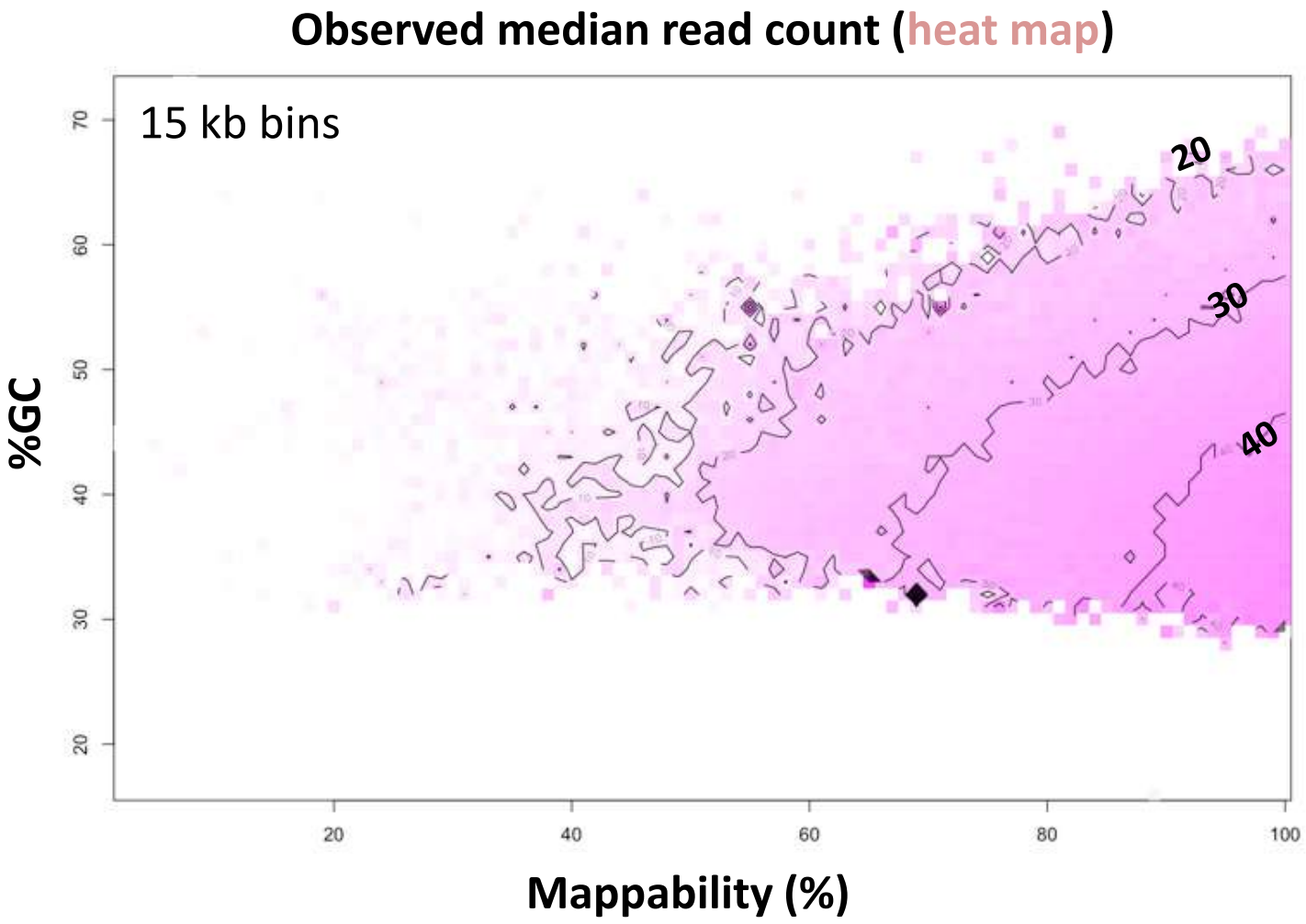
Read count is also a function mappability

Mappability ≈ How uniquely a read maps to a certain location
We calculate the average mappability per bin



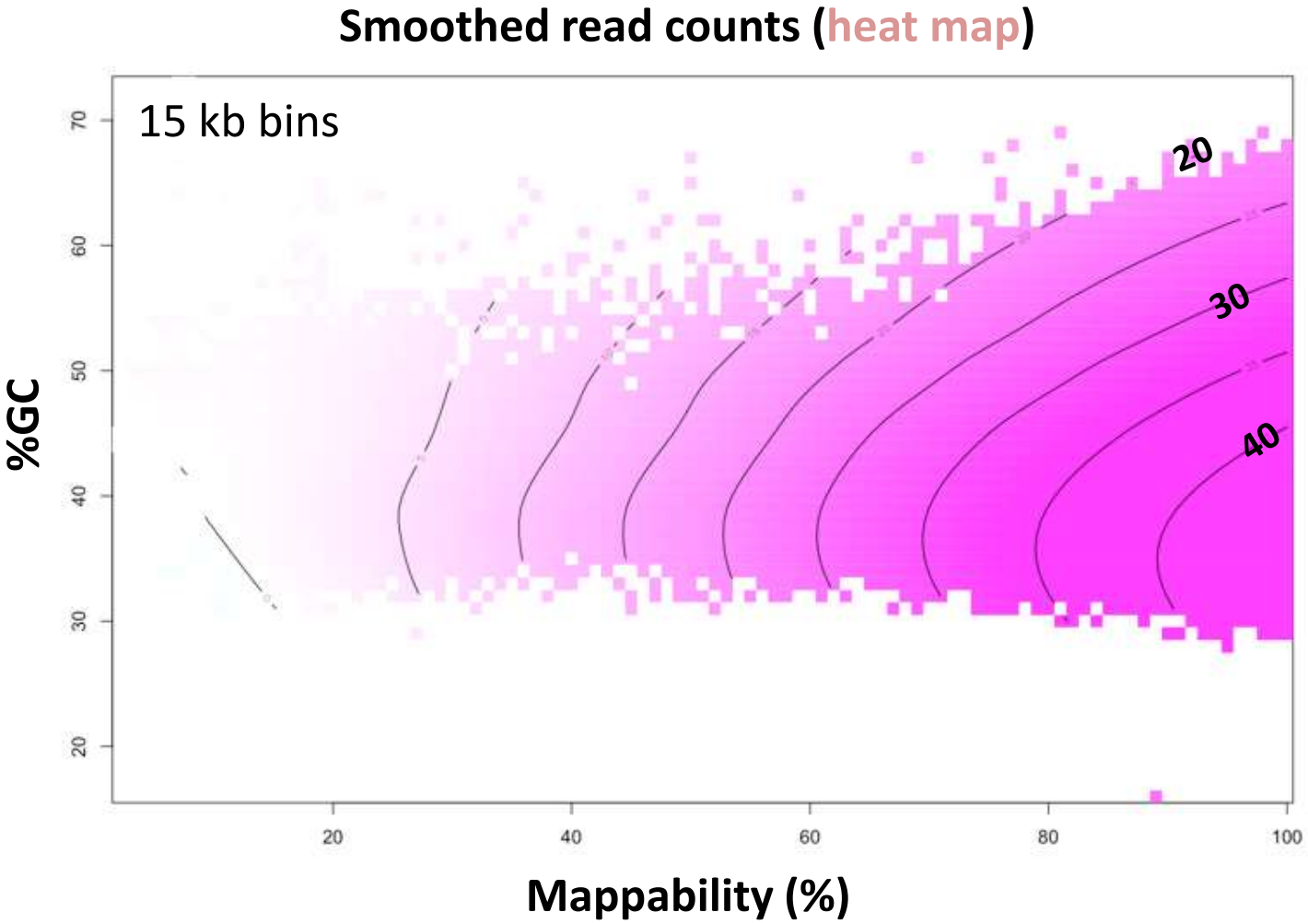
Derrien et al. PLOS one (2012)

Read count is a function of both GC content and mappability



(concurrently also reported by Yu et al. (2014); CLImAT)

Read count is a function of both GC content and mappability

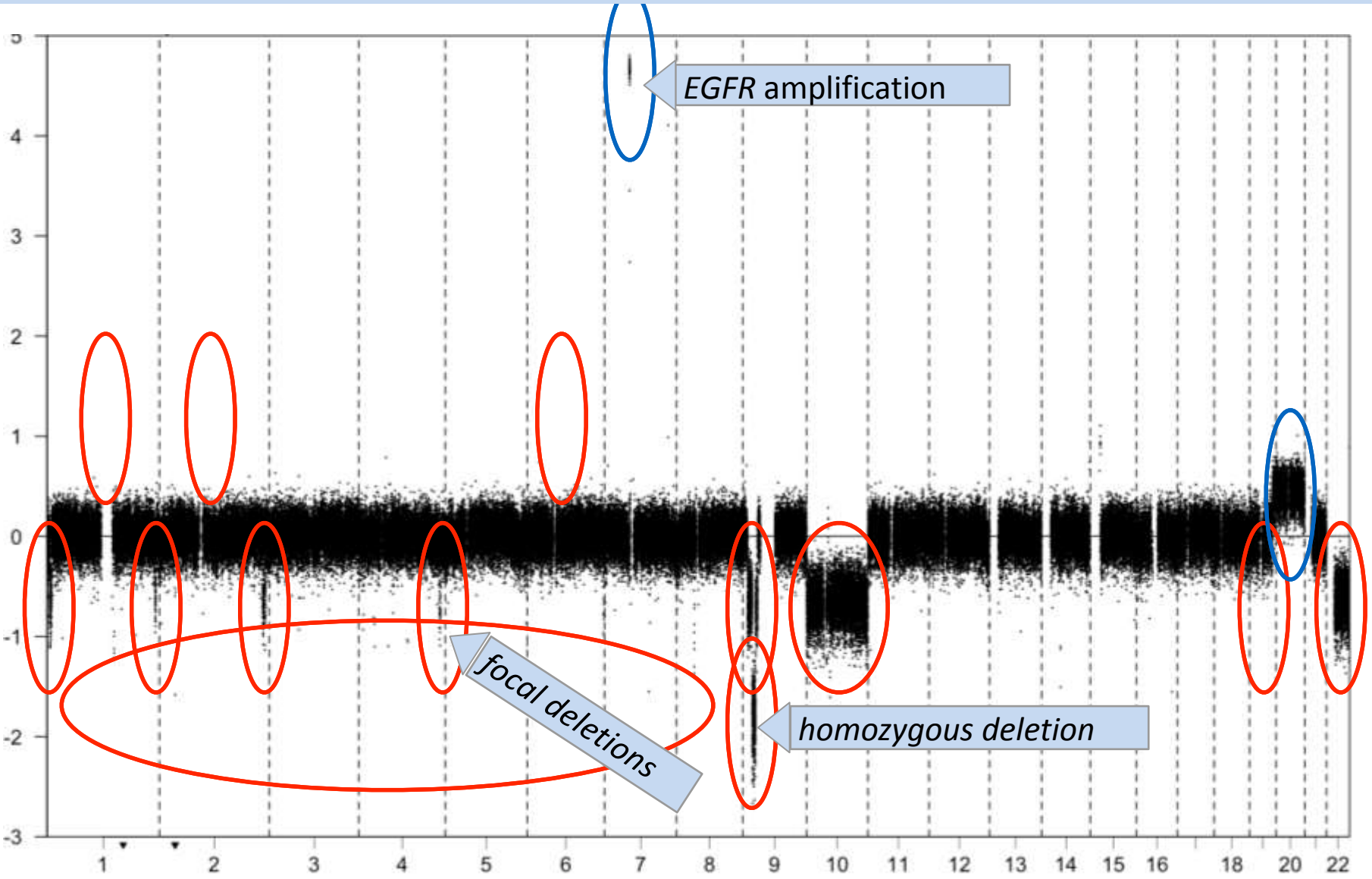


Corrections and Filters used by QDNAseq

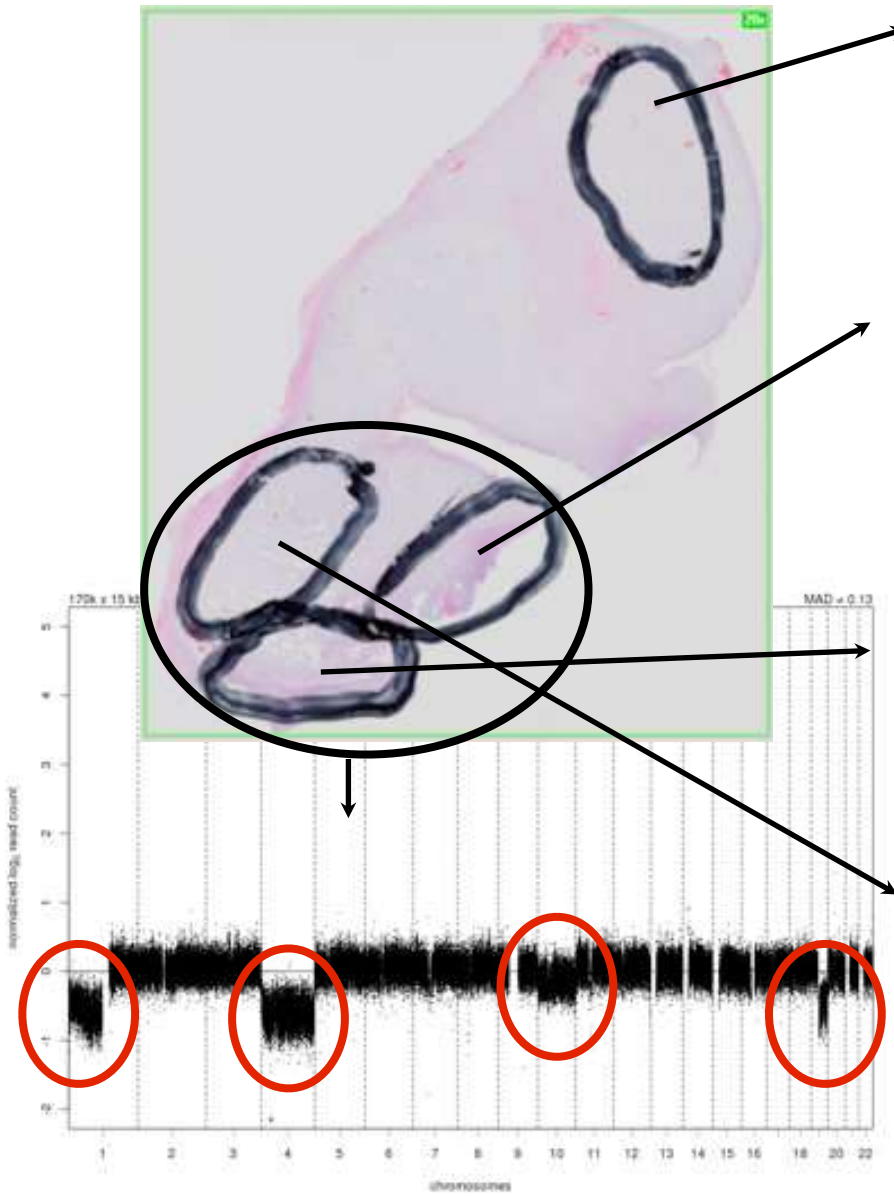
- Correct **read count** for **GC content** and **mappability**

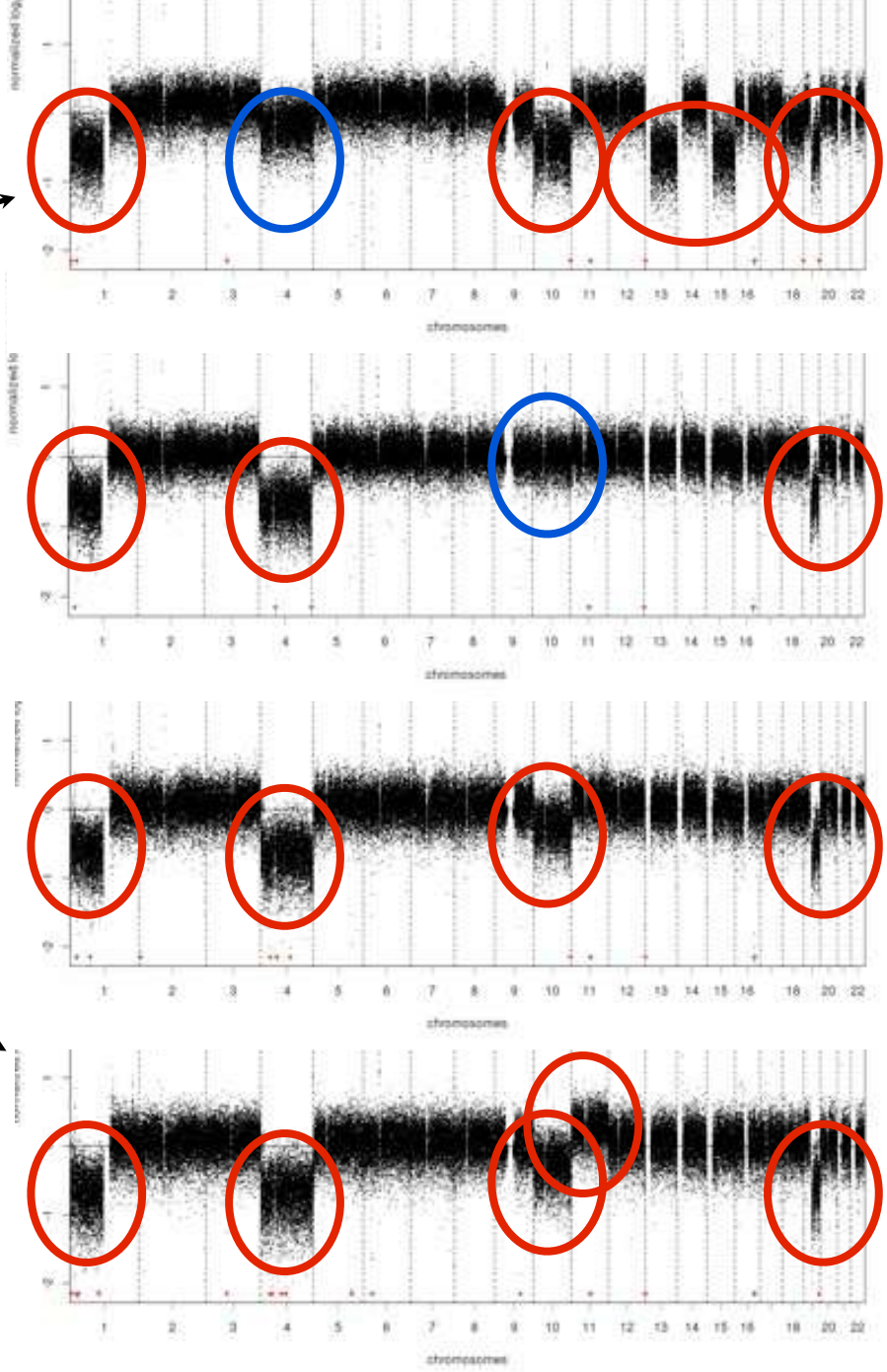
- **Black listing:** Drop bins (7%) that are known to be ‘outliers’, i.e. have large variation of read counts in 38 normal genomes of the 1000 genomes project

QDNAseq: effect of all Corrections and Filters



Example: Tumor heterogeneity in low-grade glyomas detected via CNA





van Thuijl et al., Genome Biology 2014

Acknowledgements

VUmc, Amsterdam:

Bauke Ylstra

Ilari Scheinin

Daoud Sie

Paul Eijk

Hendrik van Essen

Gerrit Meijer

Jaap Reijneveld

François Rustenburg

Hinke van Thuijl

Pieter Wesseling

Mark van de Wiel

NYU, New York:

Donna Albertson

Dan Pinkel

UCSF, San Francisco:

Henrik Bengtsson

Adam Olshen

Availability of QDNAseq:

Bioconductor R package (all platforms)

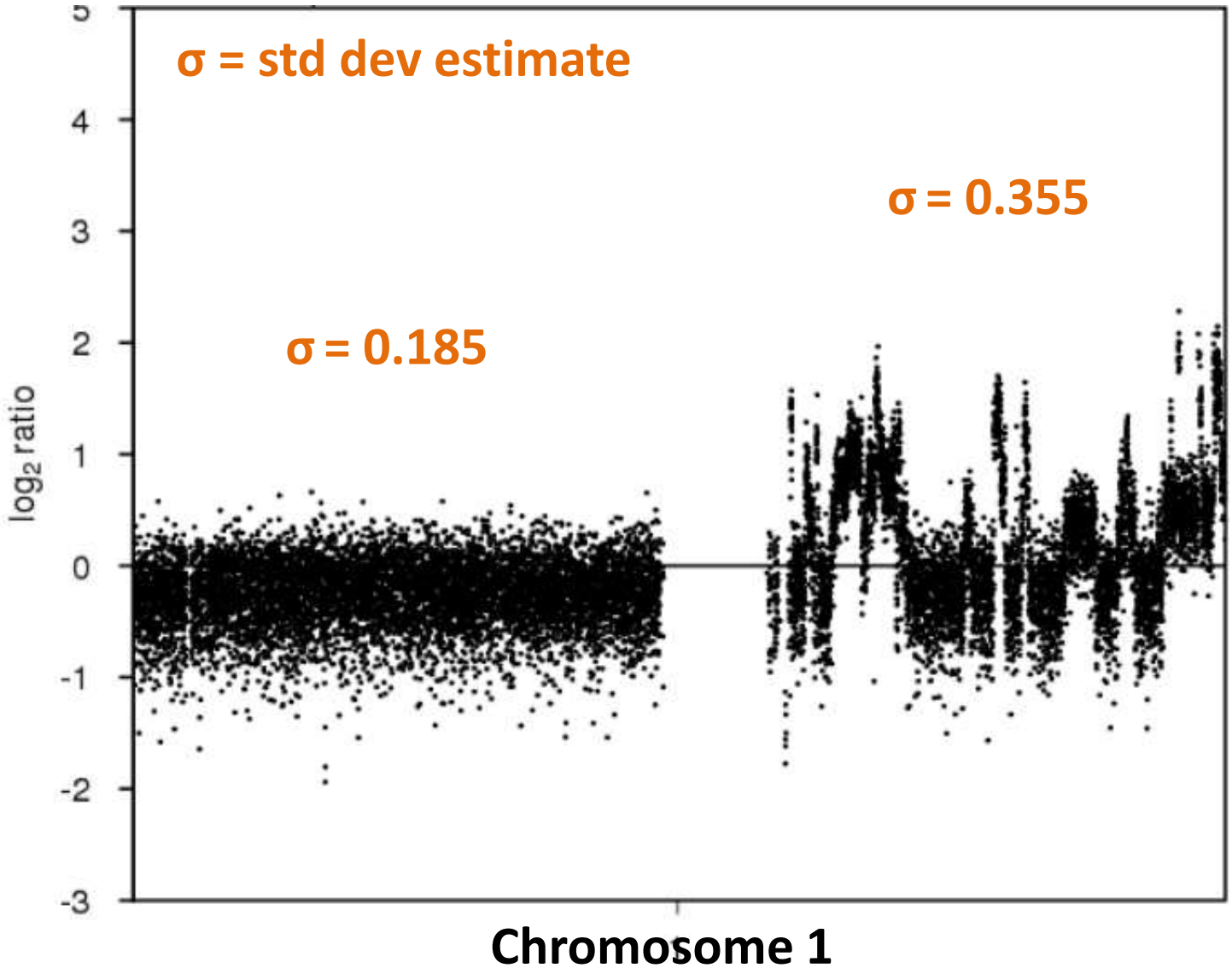
GitHub source code and issue tracker

We embrace bug reports!

Thank you!

Backup Slides

Sample standard deviation, MAD, … are poor estimators when there are aberrations



First-order successive difference noise estimator is robust against genomic aberrations

1894: E. Vallier uses successive differences to estimate dispersion: E. VALLIER, *Balistique Experimentale*, Paris, 1894, p. 166.

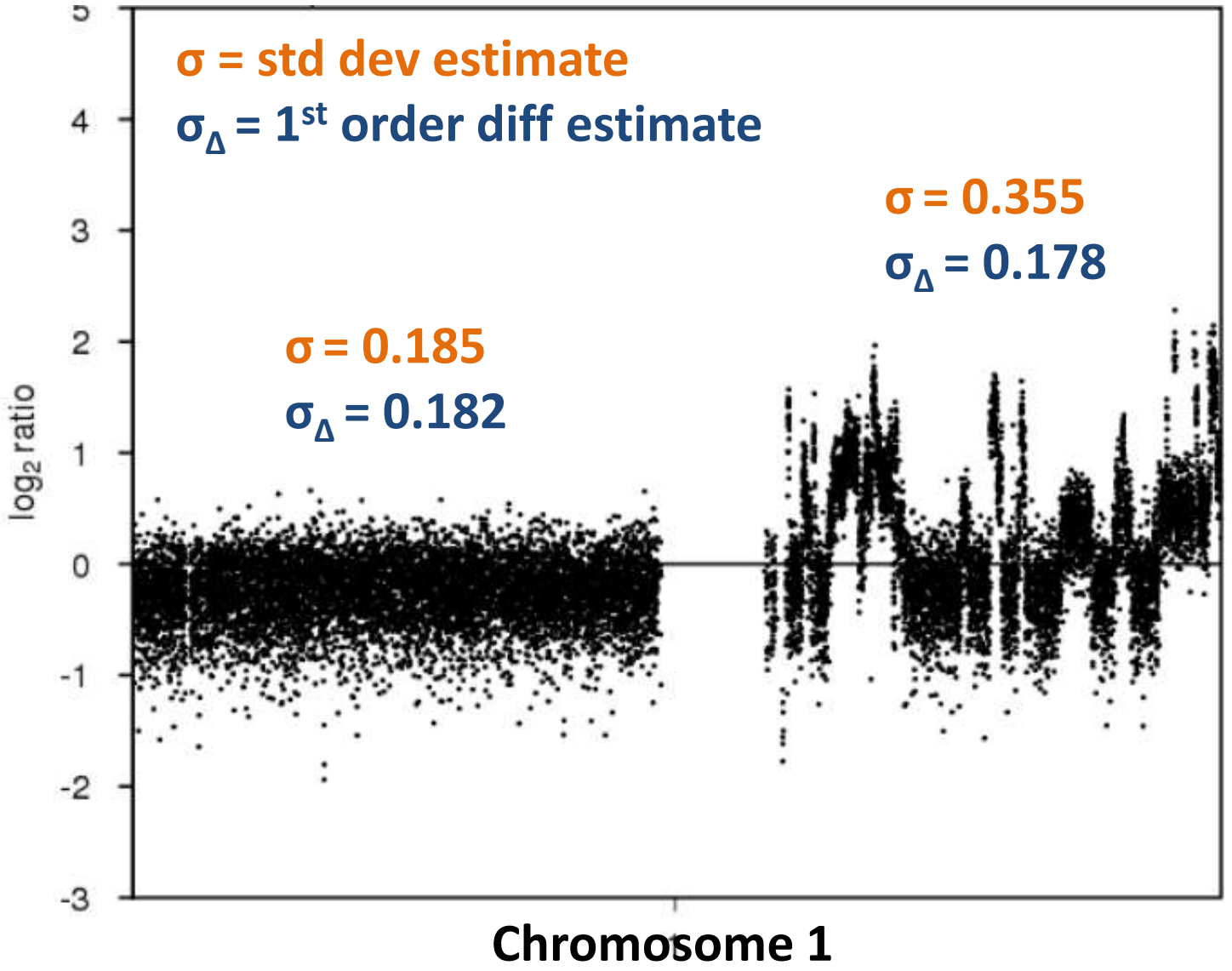
1941: J. von Neumann proposes first-order variance estimator (non-robust):

$$\frac{\delta^2}{2} = \frac{\sum_{i=1}^{n-1} (x_{i+1} - x_i)^2}{2(n-1)}$$

Robust version for standard deviation:

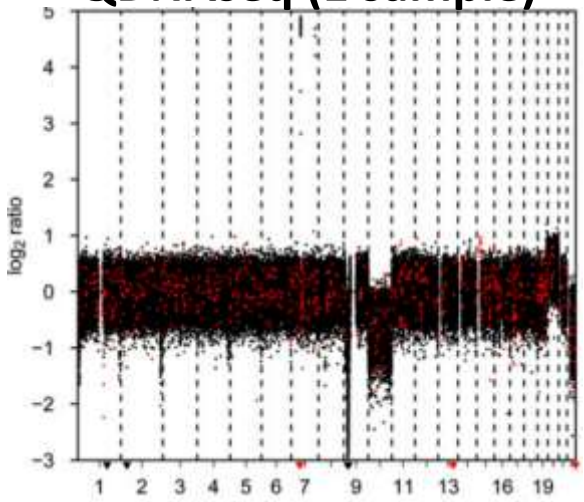
$$\sigma_\Delta = 1.486/\mathrm{sqrt}(2) \cdot \mathrm{median}_i \, |x_{i+1} - x_i|$$

First-order successive difference noise estimator is robust against genomic aberrations

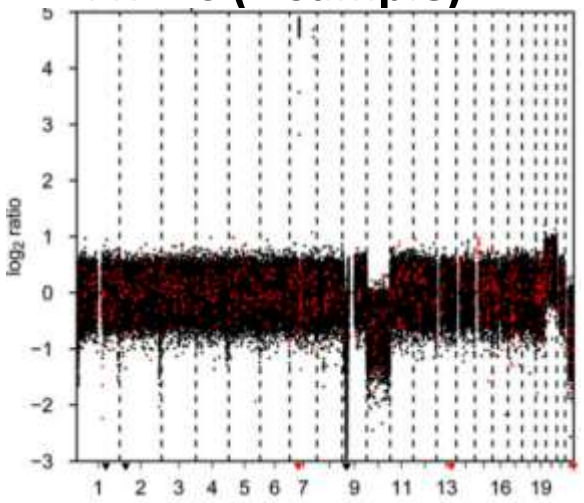


QDNAseq gives stronger signal than FREEC

QDNAseq (1 sample)



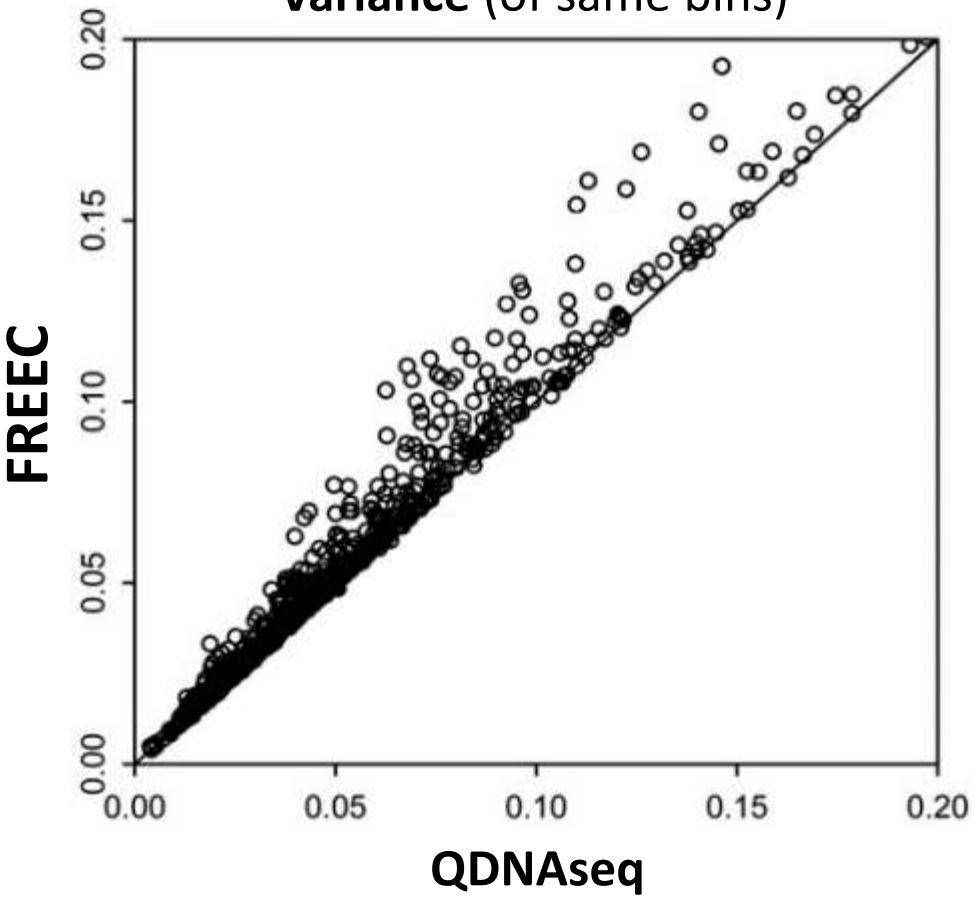
FREEC (1 sample)



Boeva et al. (2011) – Control-FREEC

> **1,000 FFPE samples**
(~ 25 institutions)

Variance (of same bins)



QDNAseq is cheap & works

1. Tumor DNA sample (archival DNA from FFPE or fresh)
   (Standard DNA library preparation)
2. Shallow DNA-Seq (0.1-0.5x coverage)
   (multiplex - samples per sequence run,
   50bp single-end reads)
3. Read alignment (we use BWA)
4. Bin counting
5. Correcting for systematic effects
6. Excluding poor bins
7. Copy-number segmentation

-- Scheinin et al. DNA copy number analysis of fresh and formalin-fixed specimens by shallow whole-genome sequencing with identification and exclusion of problematic regions in the genome assembly. *Genome Research,* 2014.