Leeds Institute of Molecular Medicine

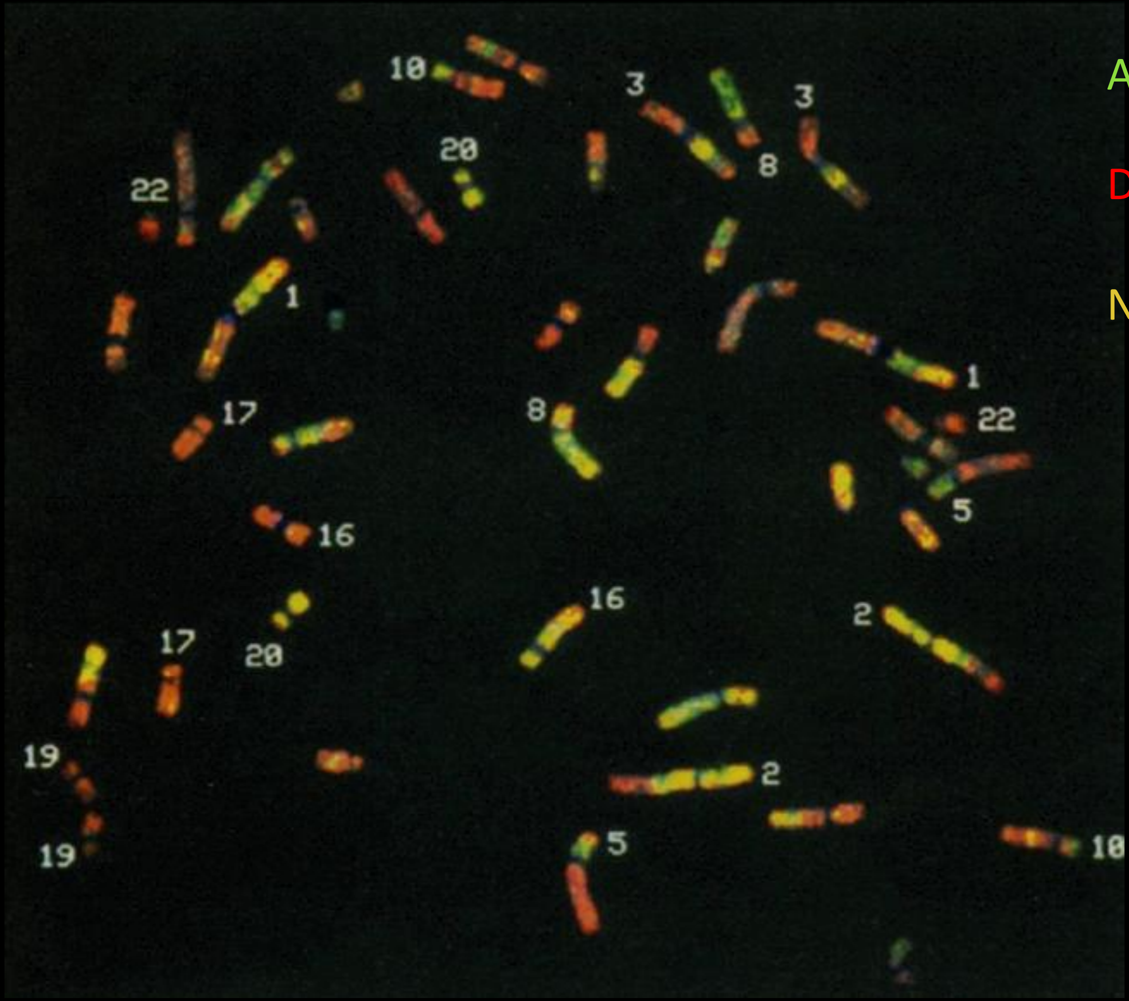
UNIVERSITY OF LEEDS

An algorithm to detect Copy Number Aberrations in cancer genomes of tumour specimens.

Arief Gusnanto, ***Stefano Berri***,
Henry M. Wood and Pamela Rabbitts

The cancer genome is often aneuploid

UNIVERSITY OF LEEDS



Amplifications

Deletions

Normal

Hartwell and Kastan. Science, 1994

Detecting abnormalities

UNIVERSITY OF LEEDS

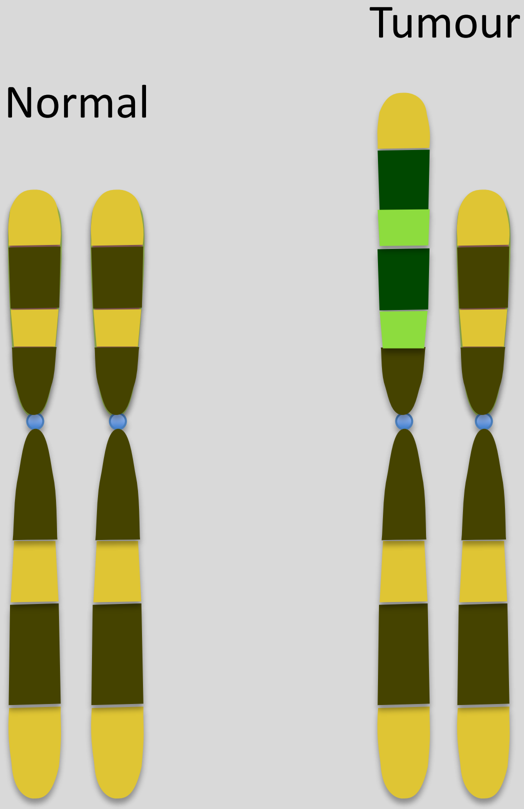
Why?

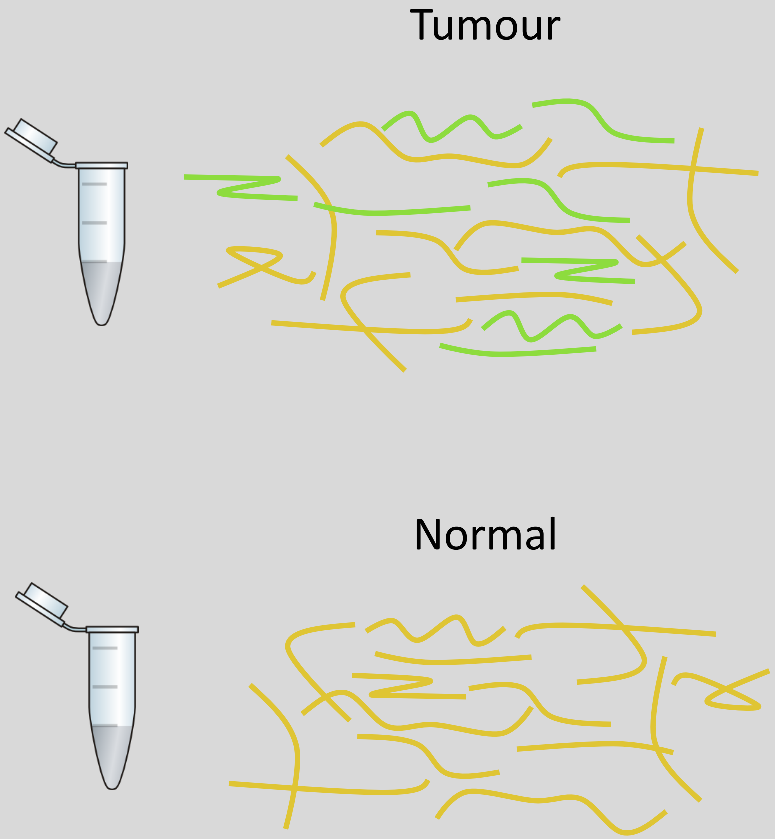
- Molecular characterisation and classification of tumours
- Diagnostic, prognostic and predictive tool
- Understand the biology of cancer

How?

- CGH
- aCGH (BAC or oligo)
- SNP microarray
- “NextGen” Sequencing
  - ✓ Tuneable resolution/cost
  - ✓ Re-use of data
  - ✓ Flexible platform
  - ✓ Technical independence Test – Control
  - ✓ Might become very cheap

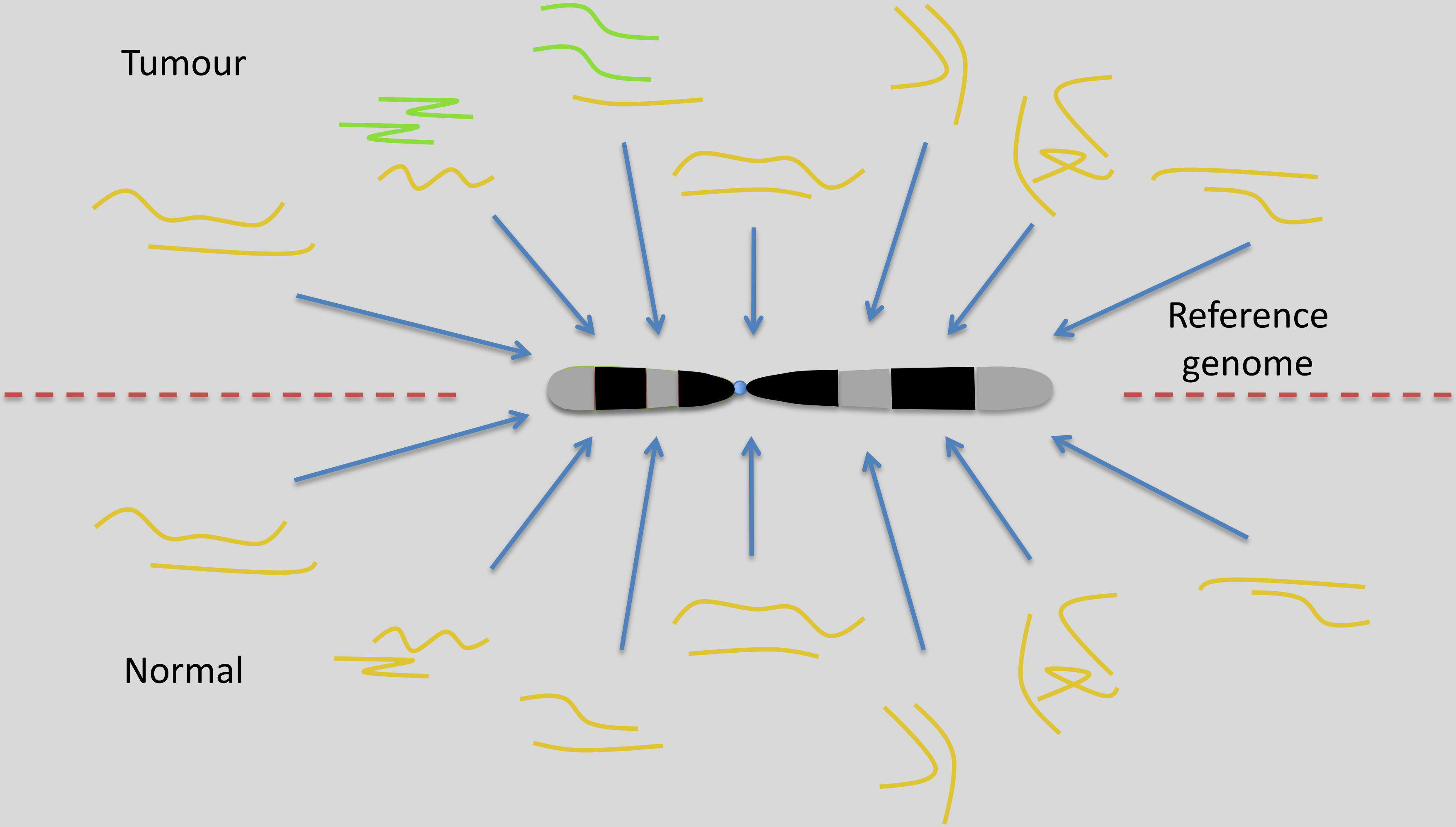
Copy number by “NextGen” Sequencing





“NextGen” sequencing and reads mapping.

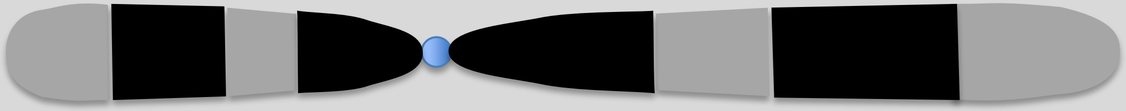
UNIVERSITY OF LEEDS



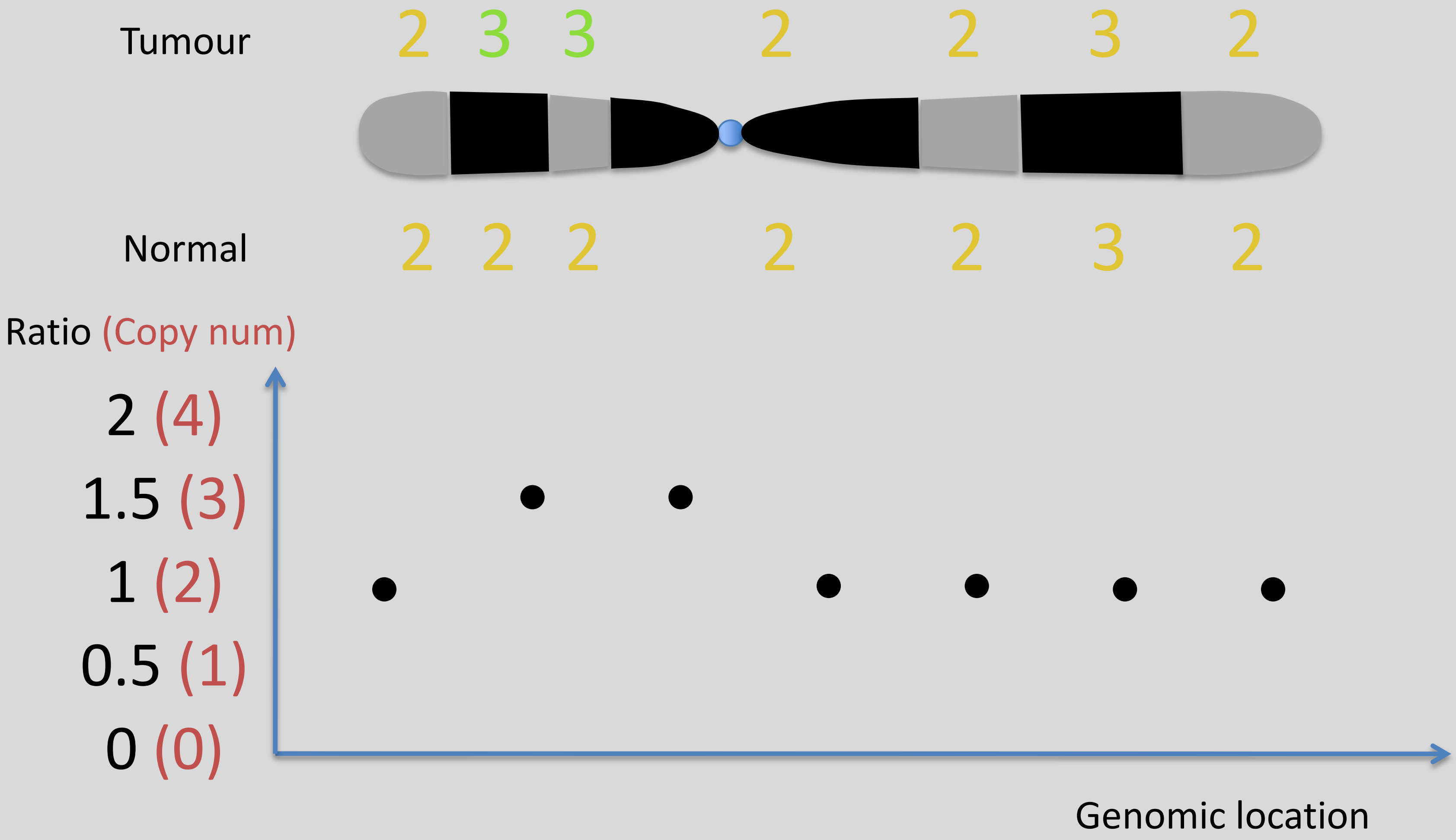
Counting number of sequences for each window

UNIVERSITY OF LEEDS

Tumour	2	3	3	2	2	3	2
Normal	2	2	2	2	2	3	2

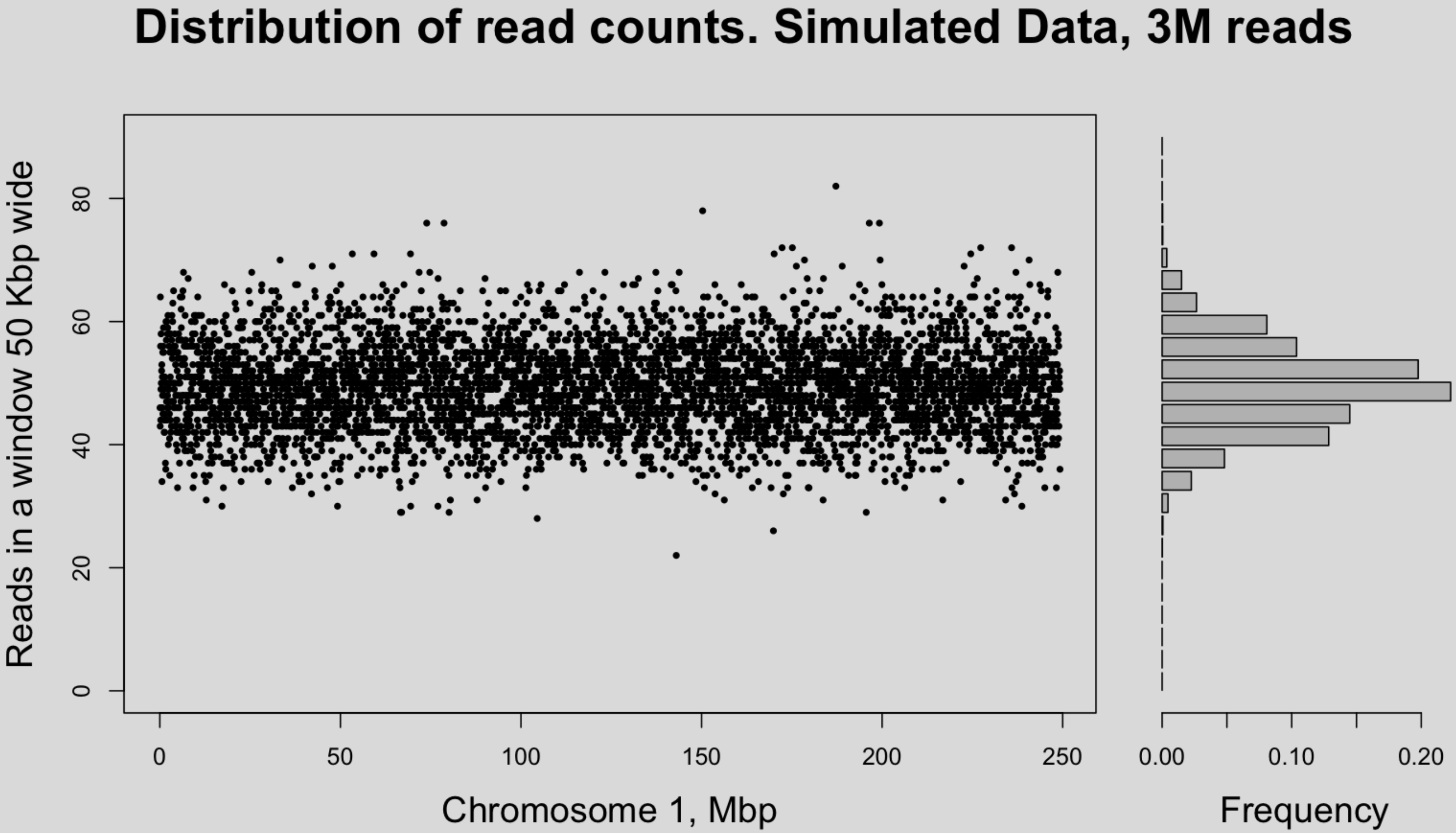


Counting number of sequences for each window

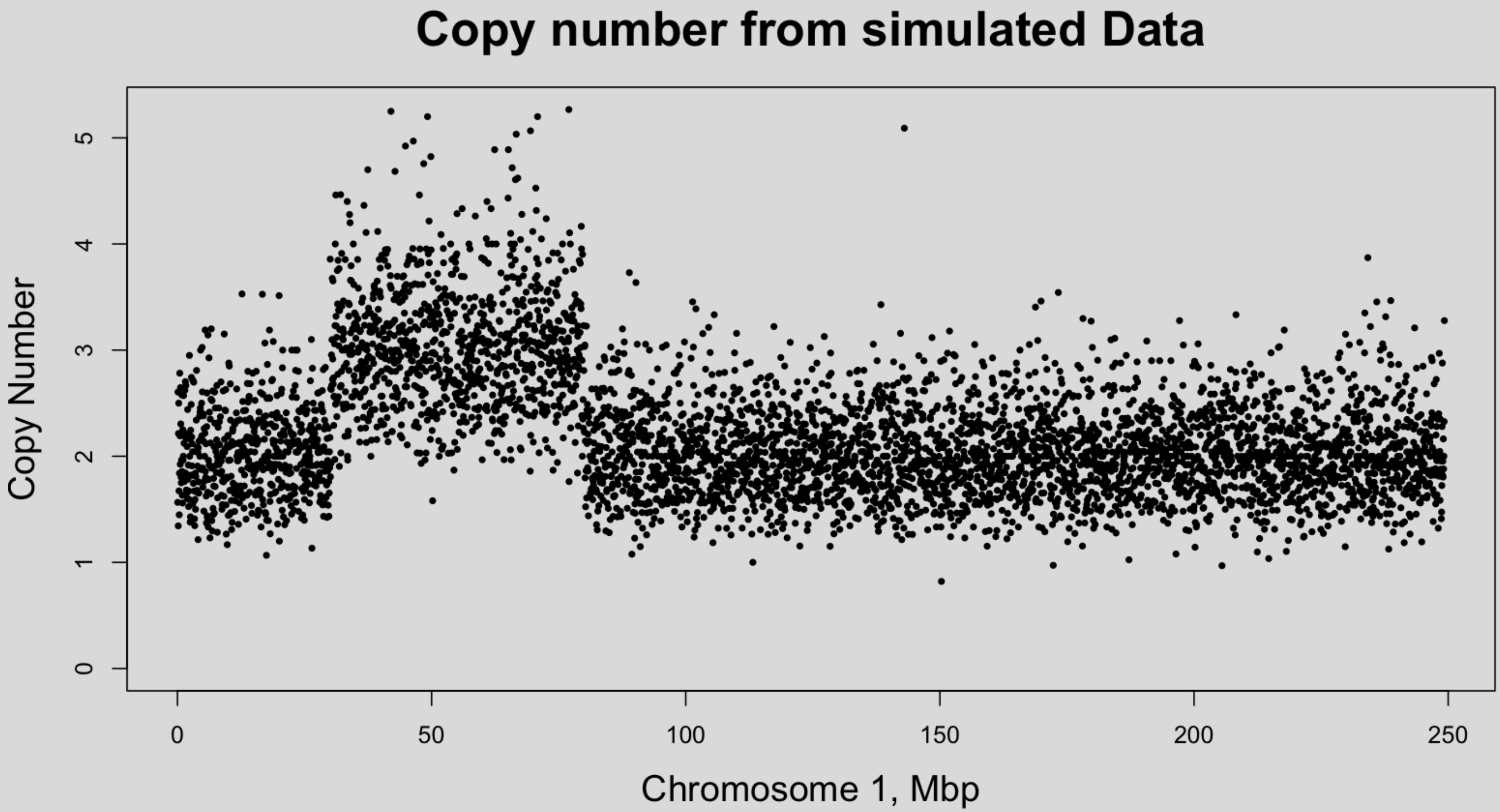


Toward the real data

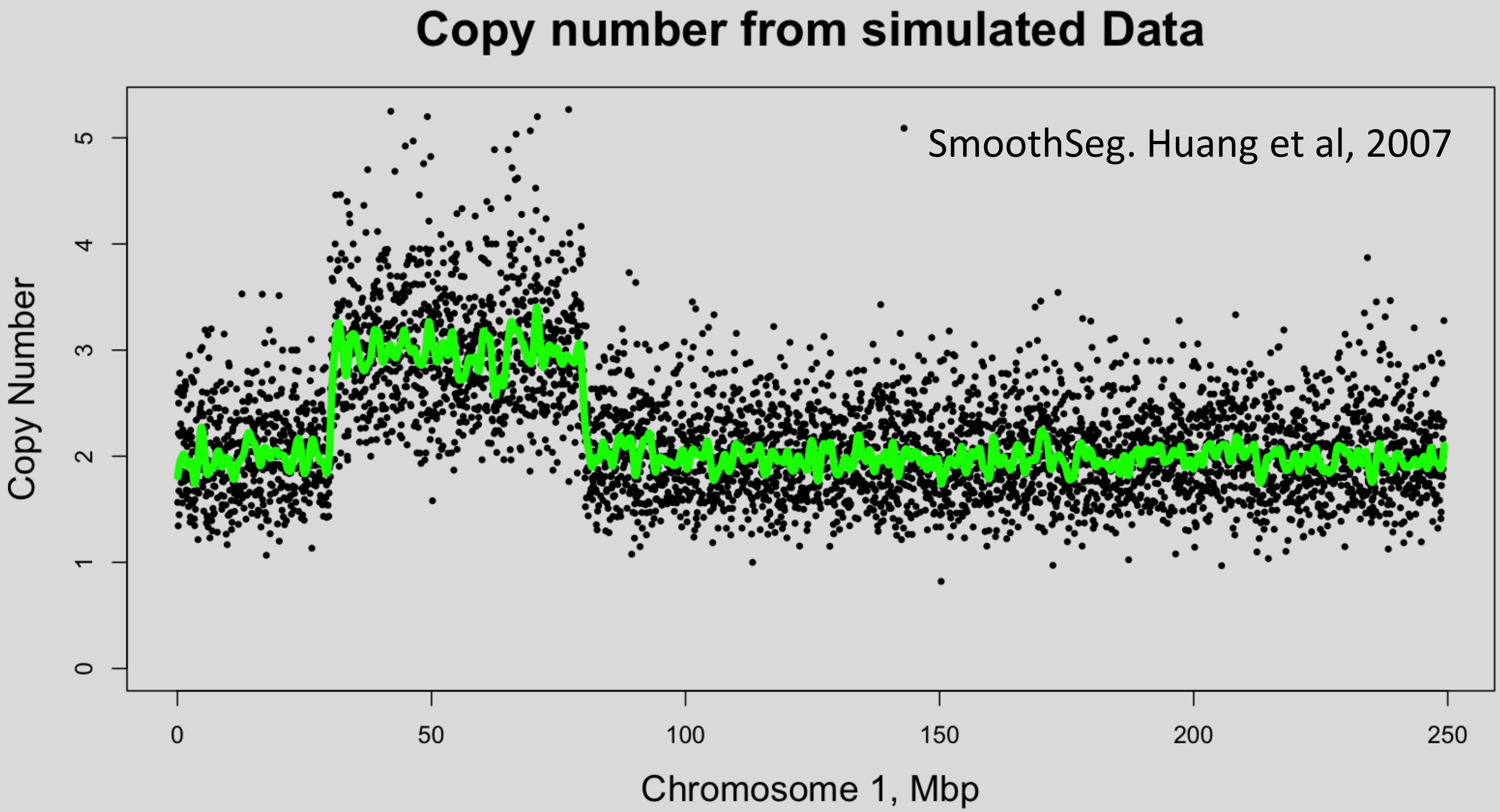
UNIVERSITY OF LEEDS



Ratio Test/Control



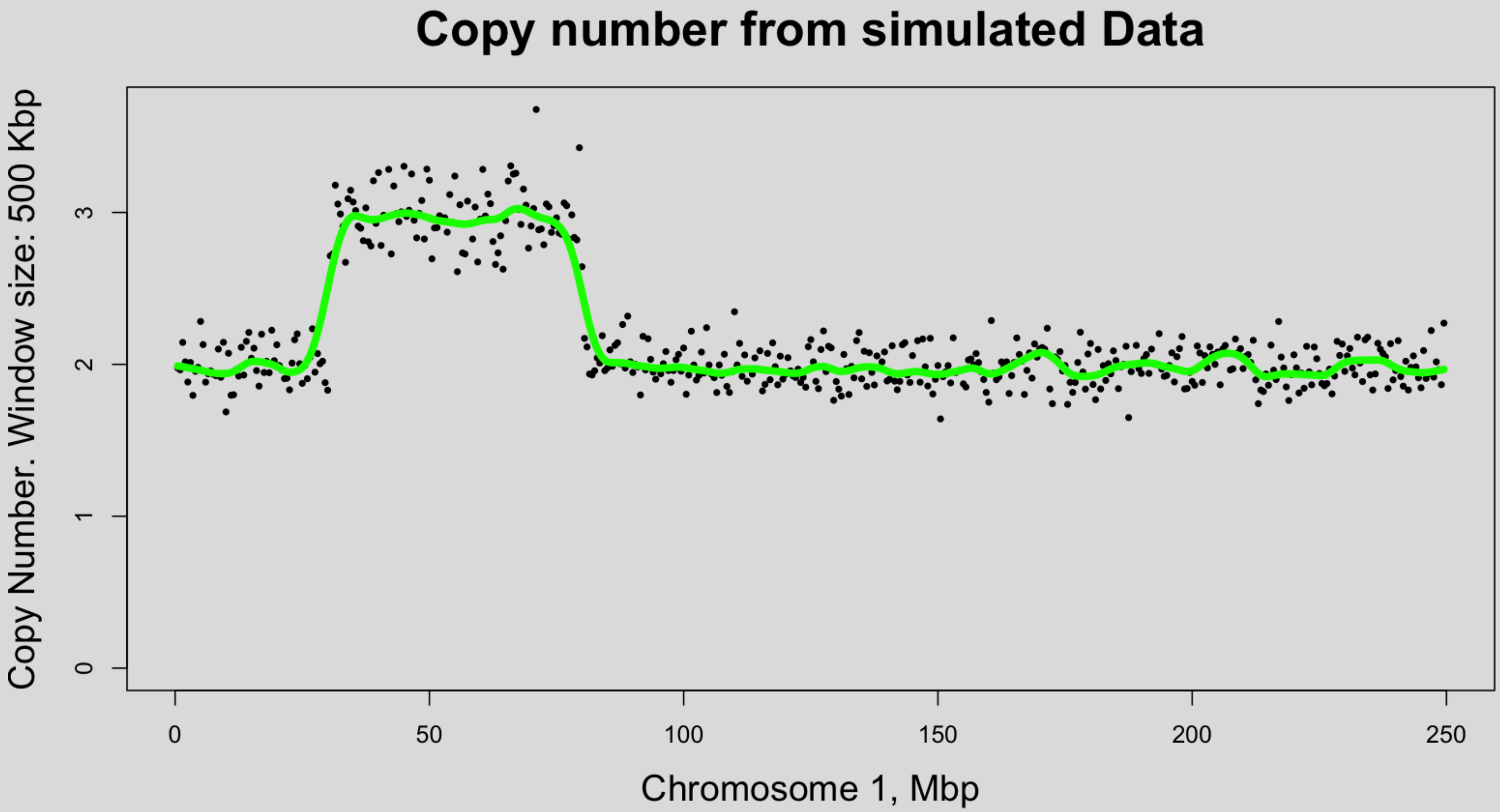
Ratio Test/Control



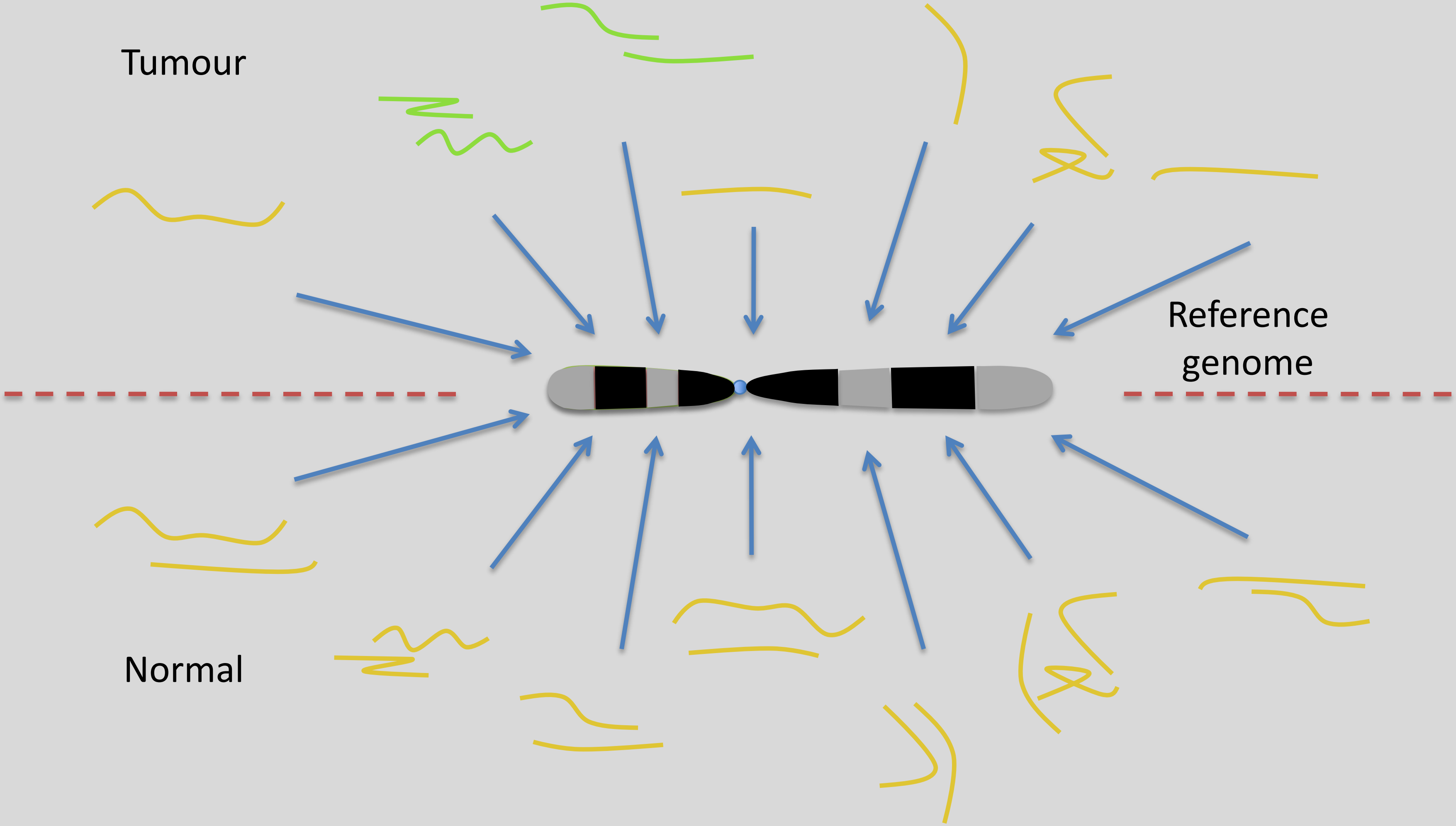
SmoothSeg. Huang et al, 2007

Resolution and noise go together!

UNIVERSITY OF LEEDS

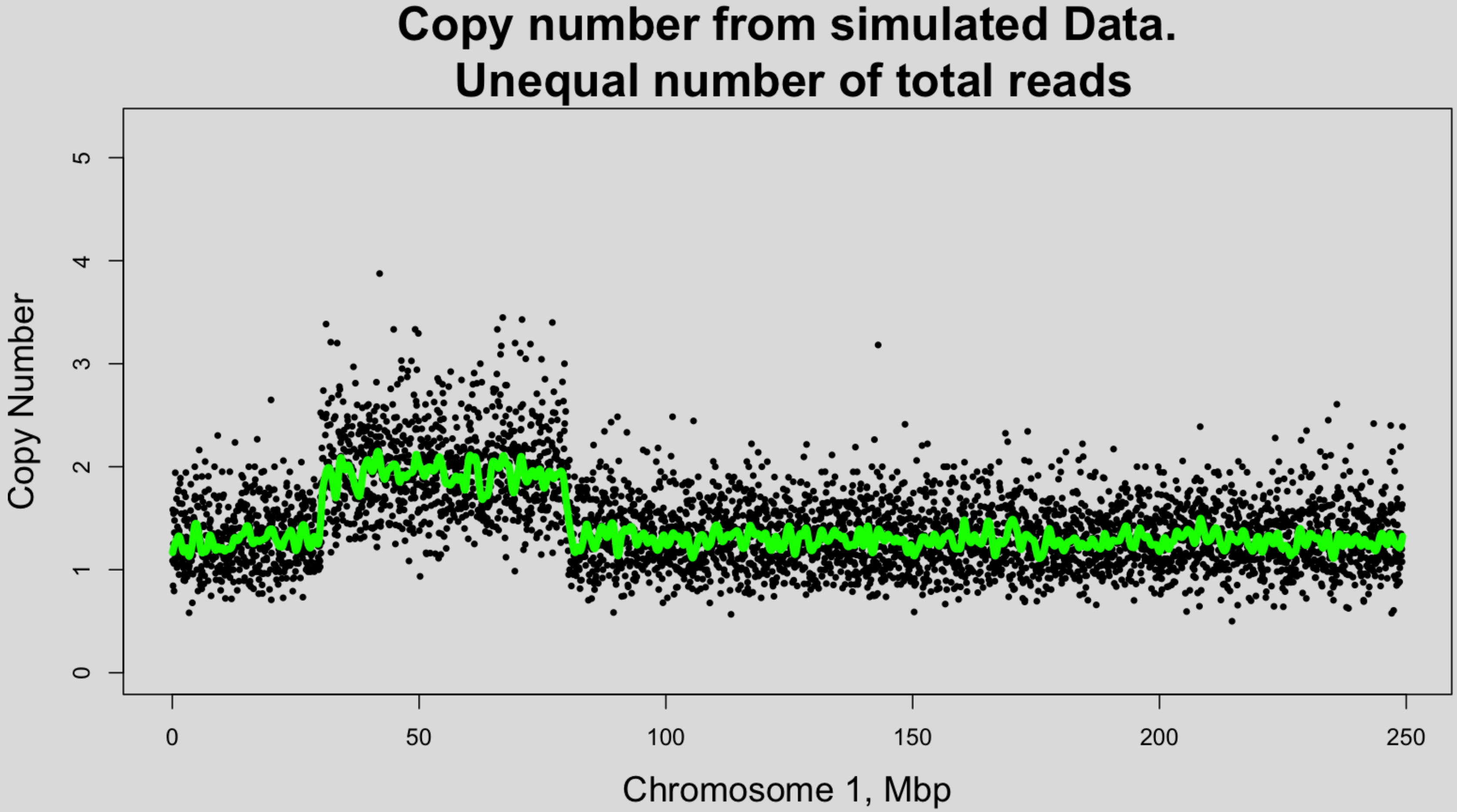


Different number of total reads



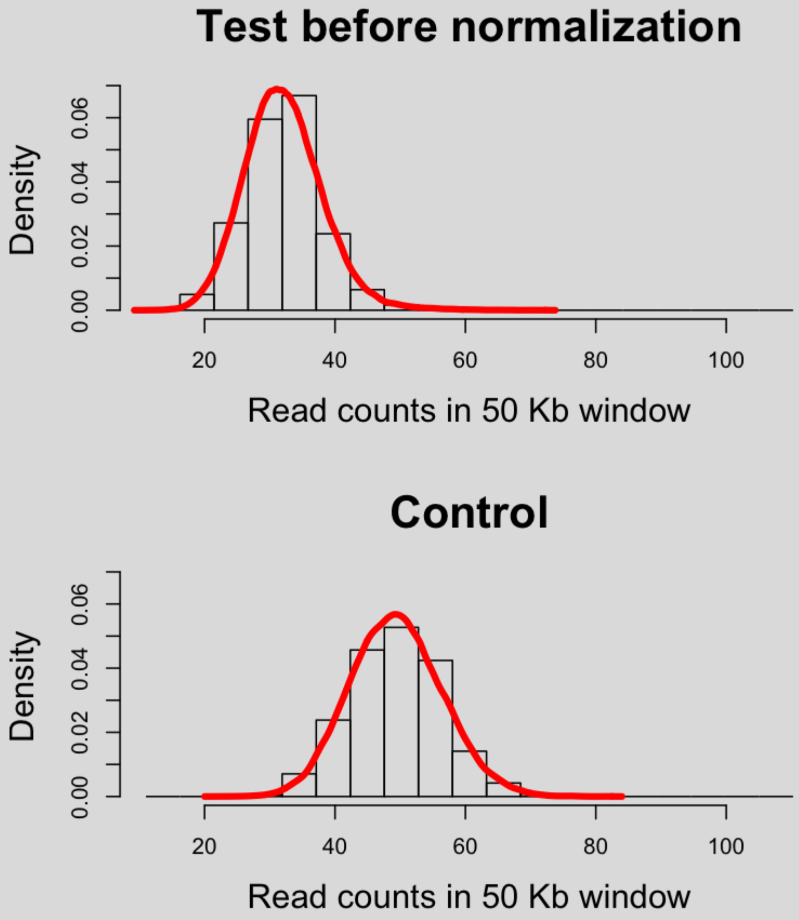
Total number of reads varies.

UNIVERSITY OF LEEDS



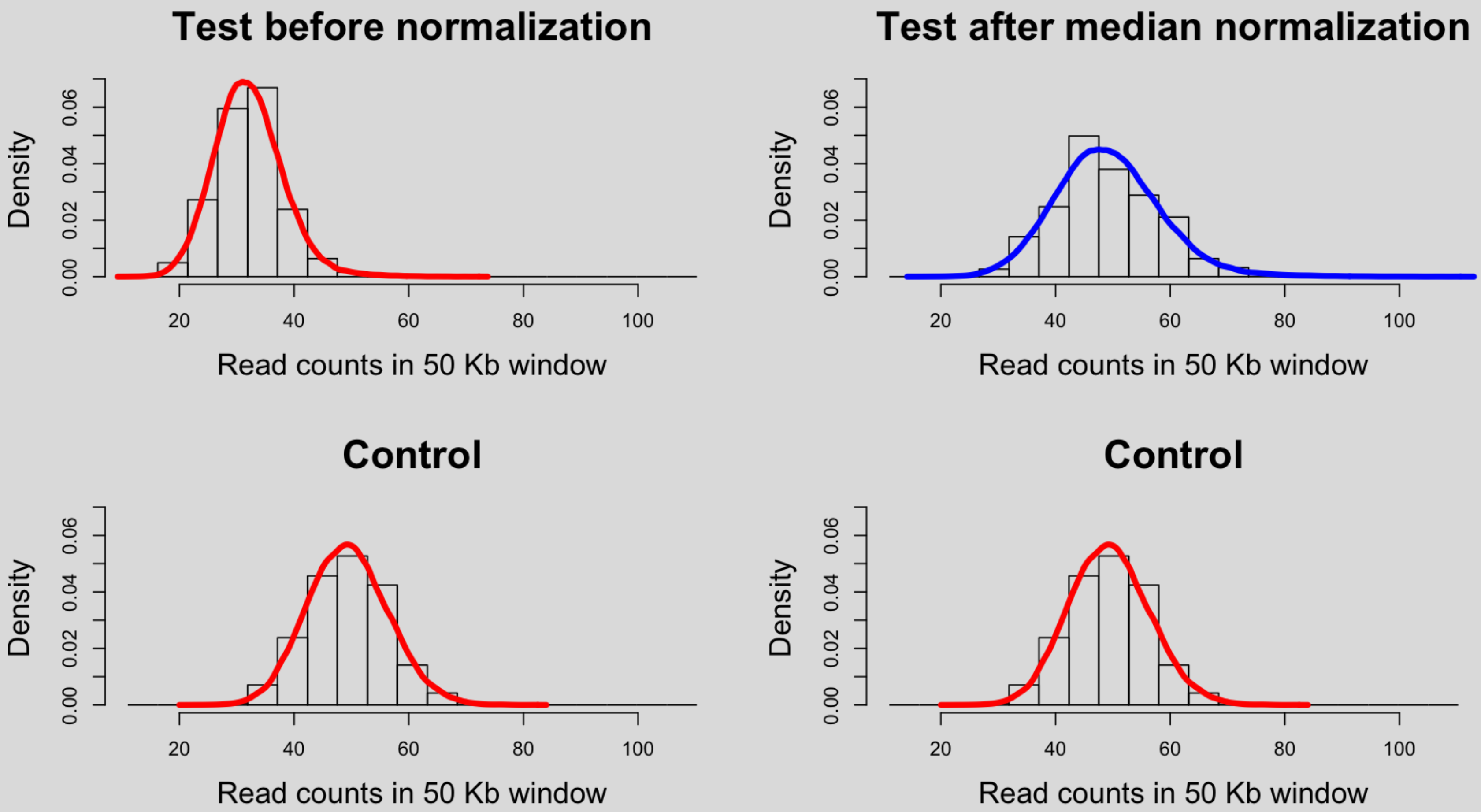
Normalization. A crucial step

UNIVERSITY OF LEEDS



Normalization. A crucial step

UNIVERSITY OF LEEDS



Test before normalization

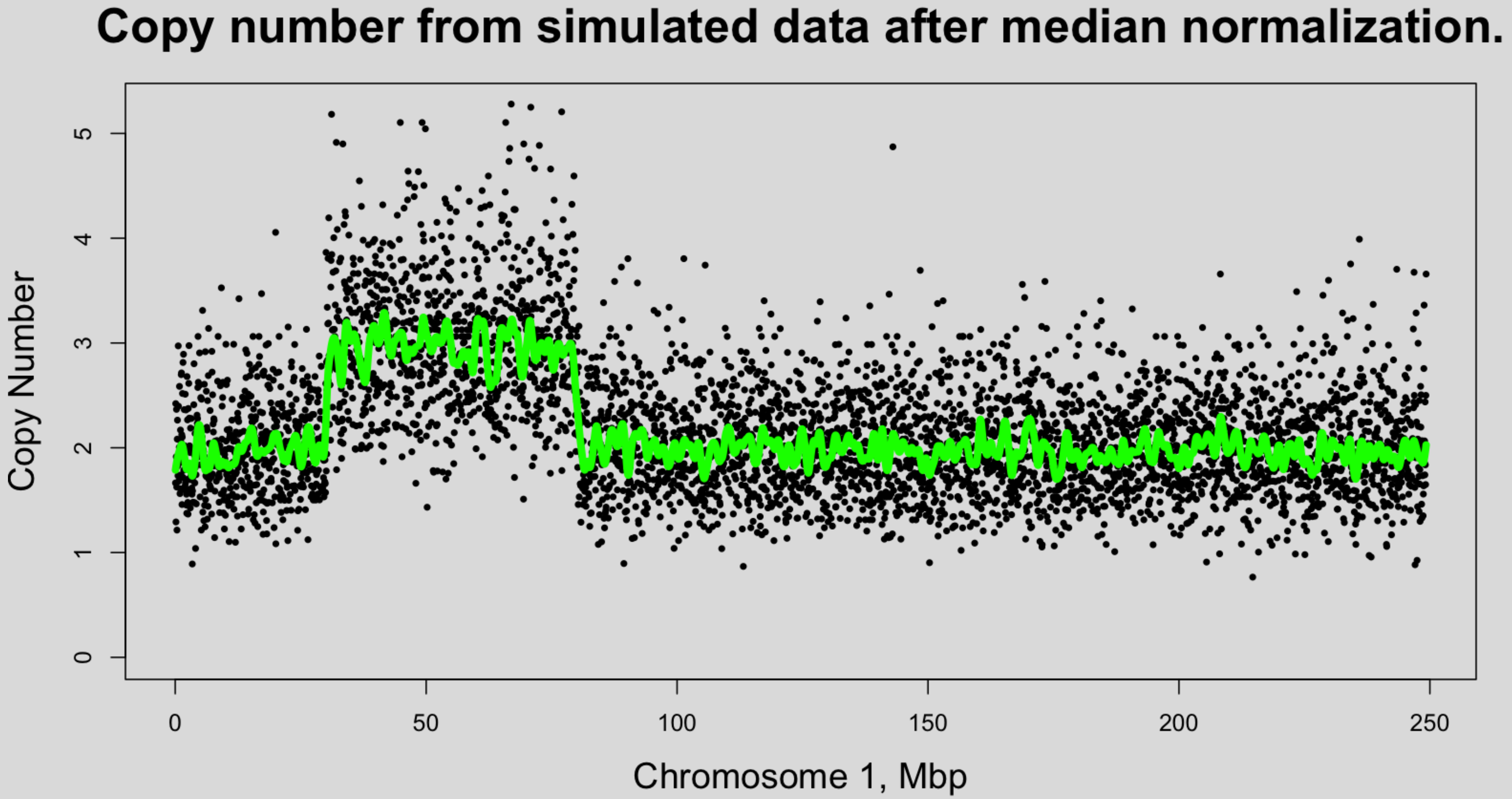
Test after median normalization

Control

Control

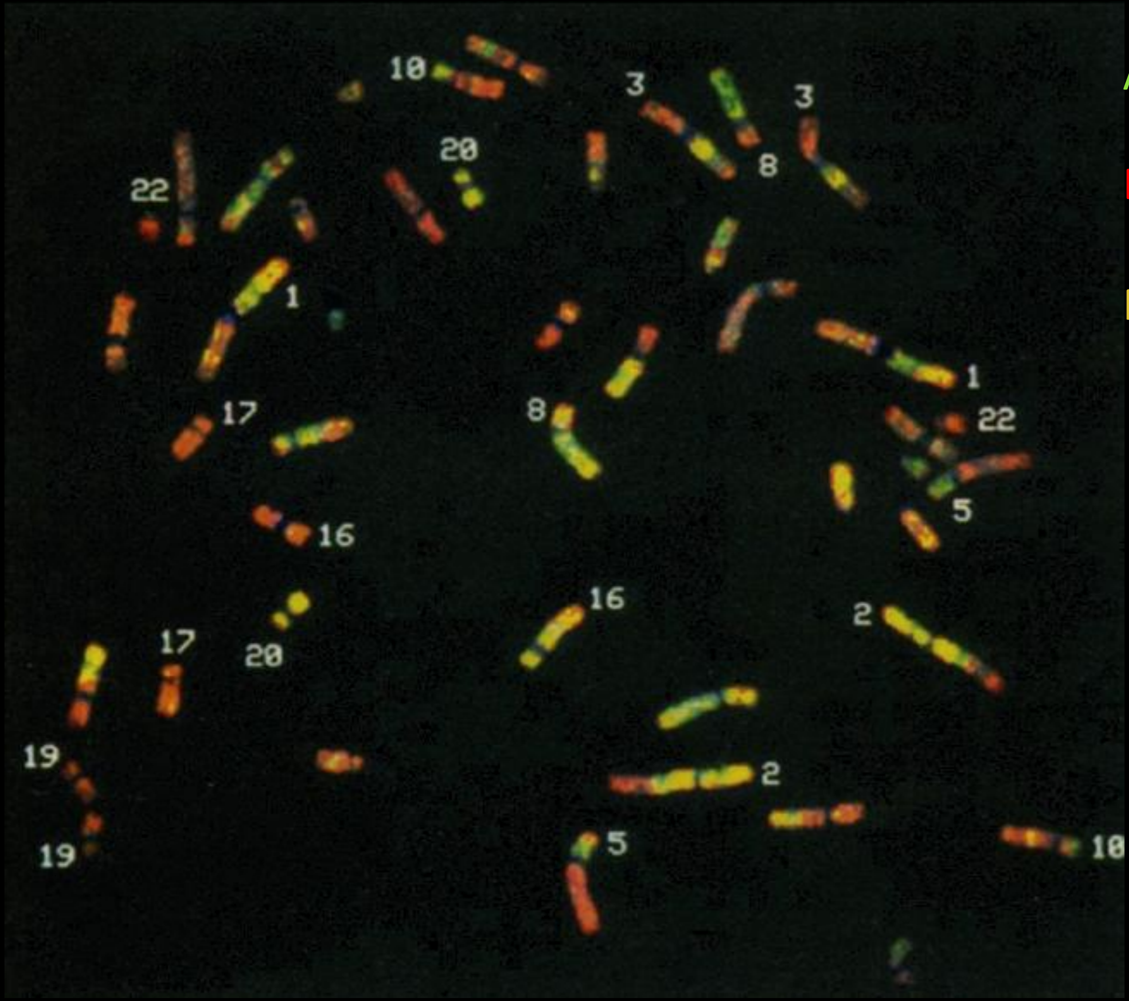
Normalization. A crucial step

UNIVERSITY OF LEEDS



The cancer genome is often aneuploid

UNIVERSITY OF LEEDS



Amplifications

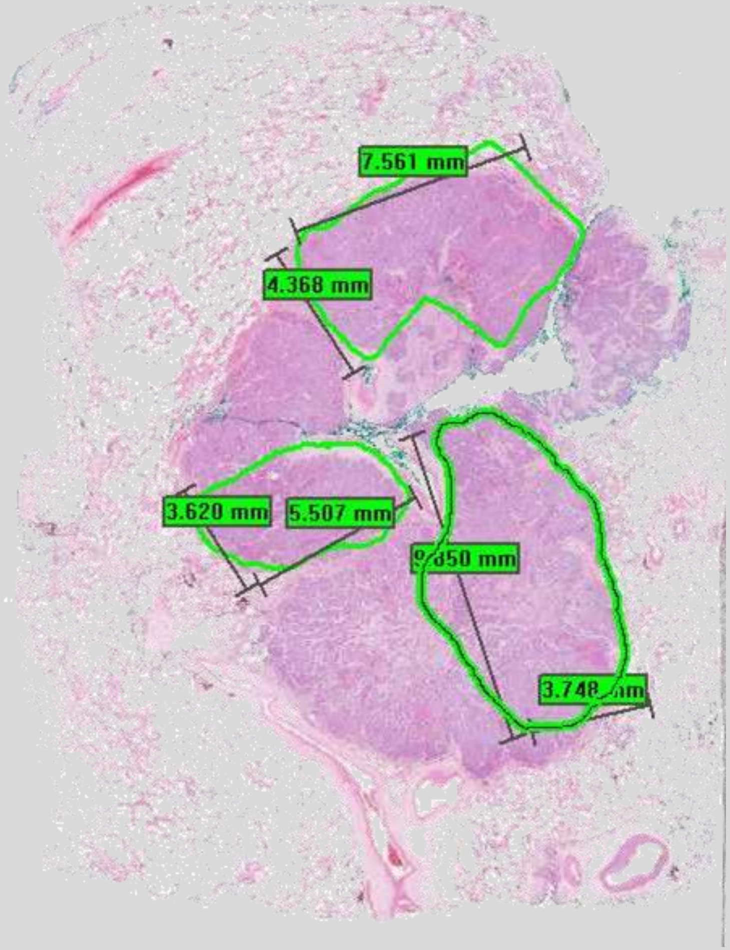
Deletions

Normal

Many amplifications and deletions!

Hartwell and Kastan. Science, 1994

Patient’s tumour samples

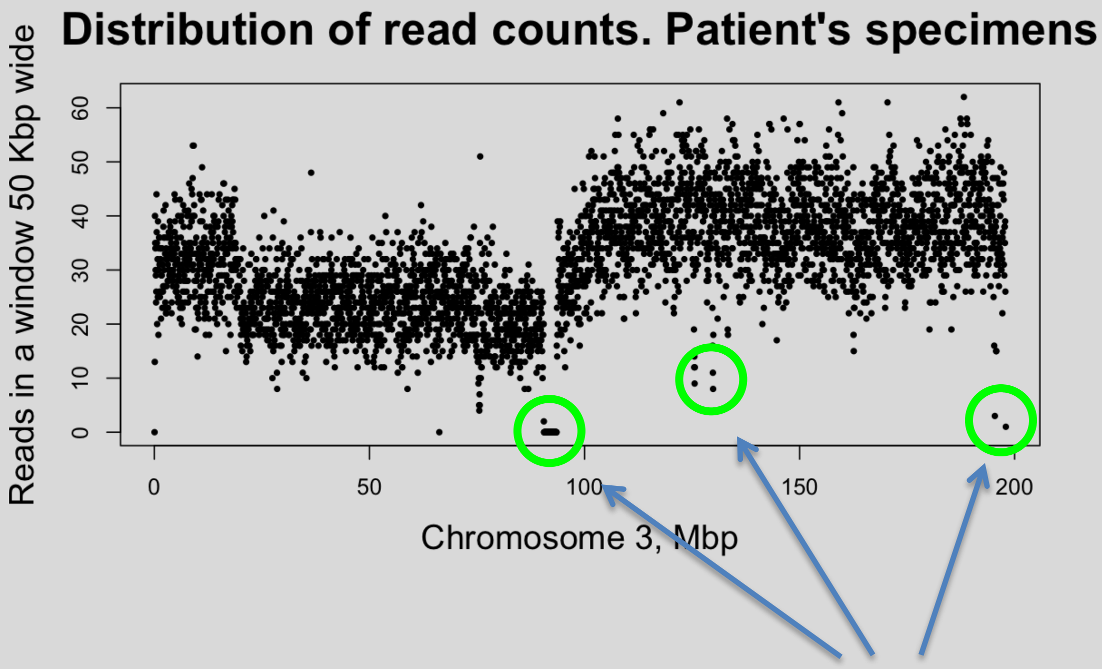


- Contamination with stroma, inflammatory cells...

Lung tumour

The real samples. A lot noisier

UNIVERSITY OF LEEDS

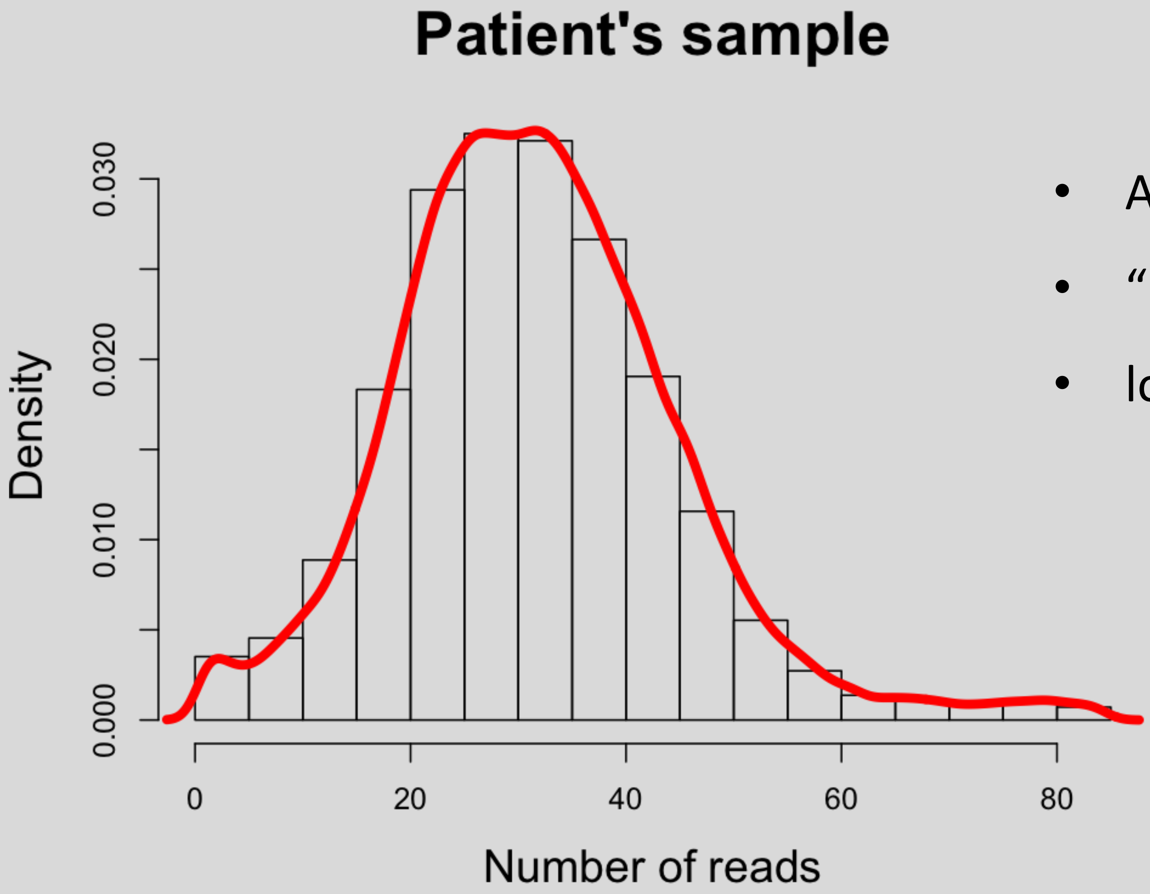


Aligning artefacts

- Some sequences cannot be aligned (repeated regions)
- GC content bias
- Unequal number of total reads.
- Extra noise of unknown origin

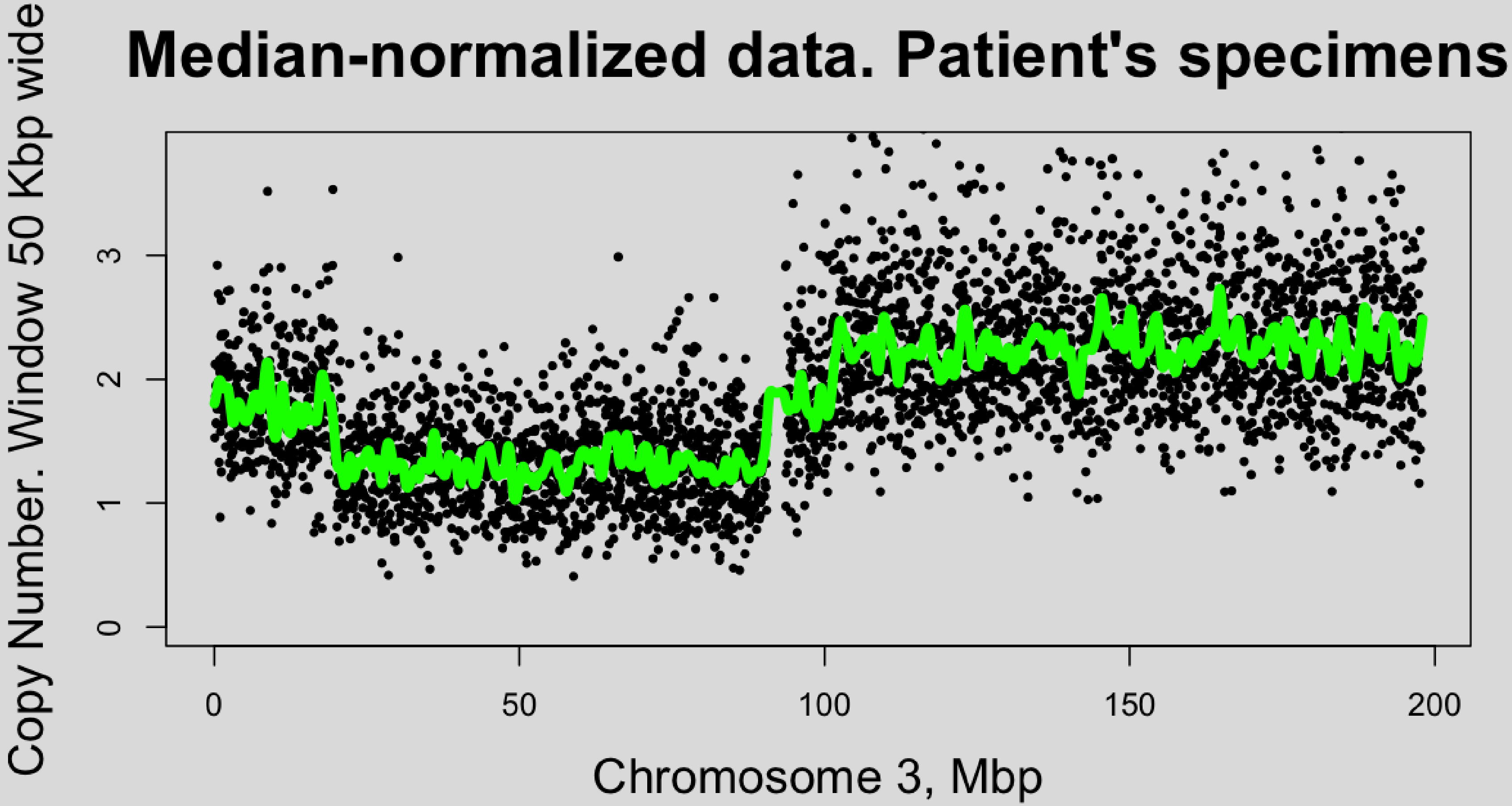
The median might be meaningless

UNIVERSITY OF LEEDS

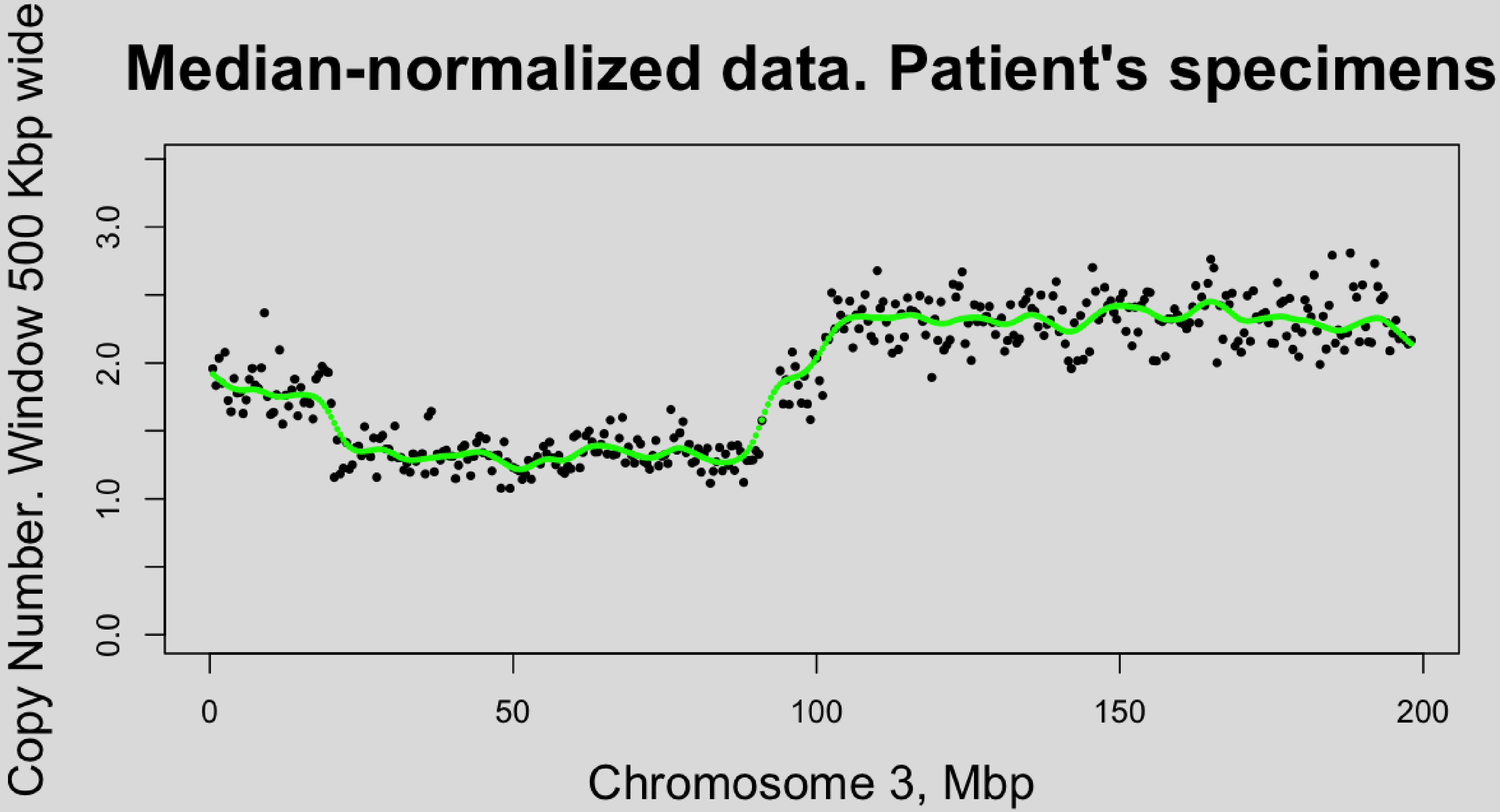


- Asymmetric distribution
- “Flat top”
- long tail

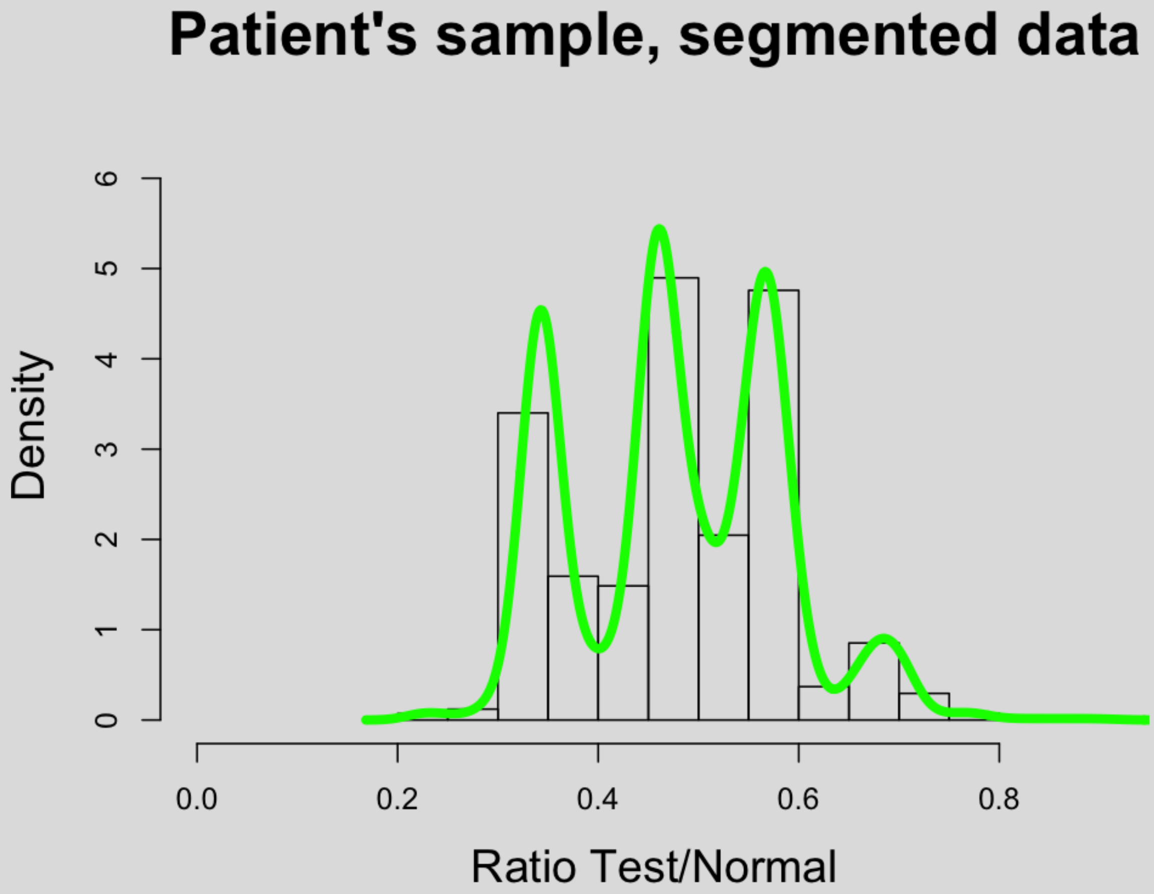
Median normalisation



Trade resolution for noise

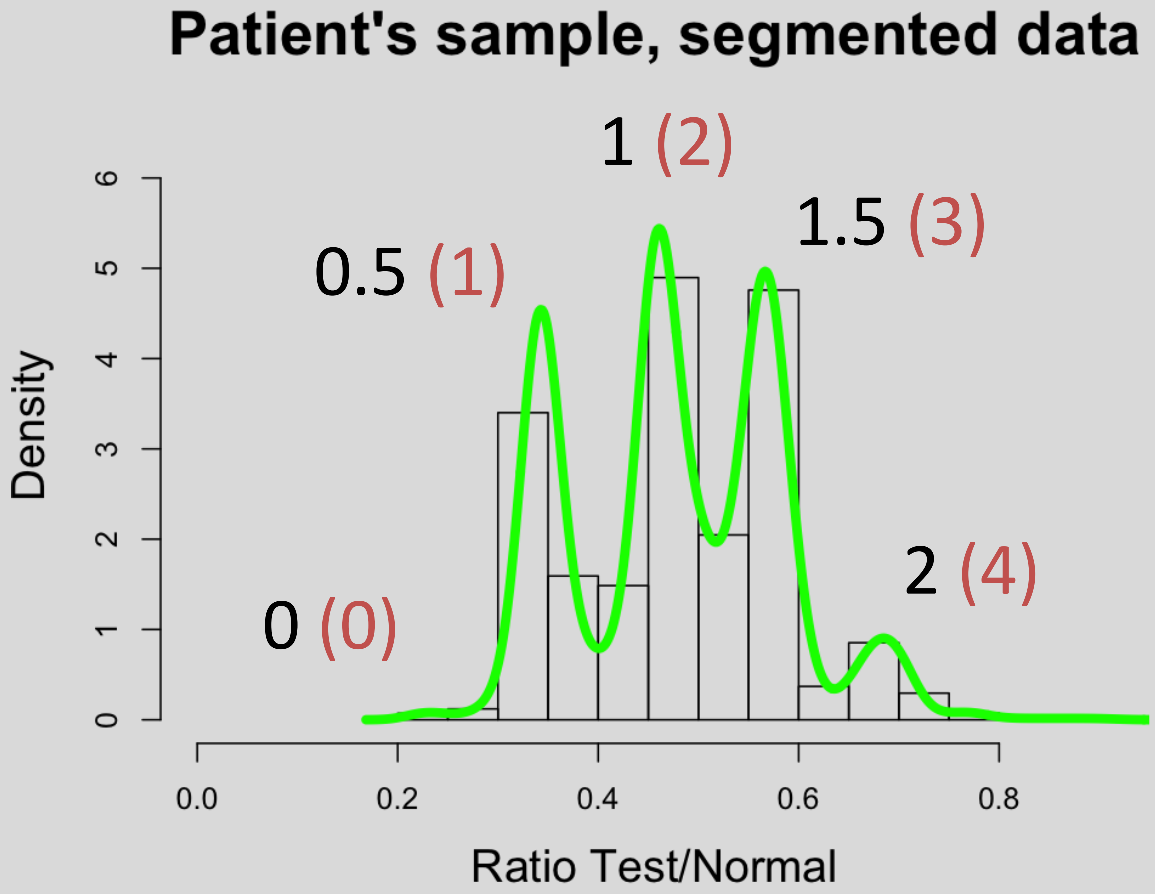


Discrete data normalisation



Discrete data normalisation

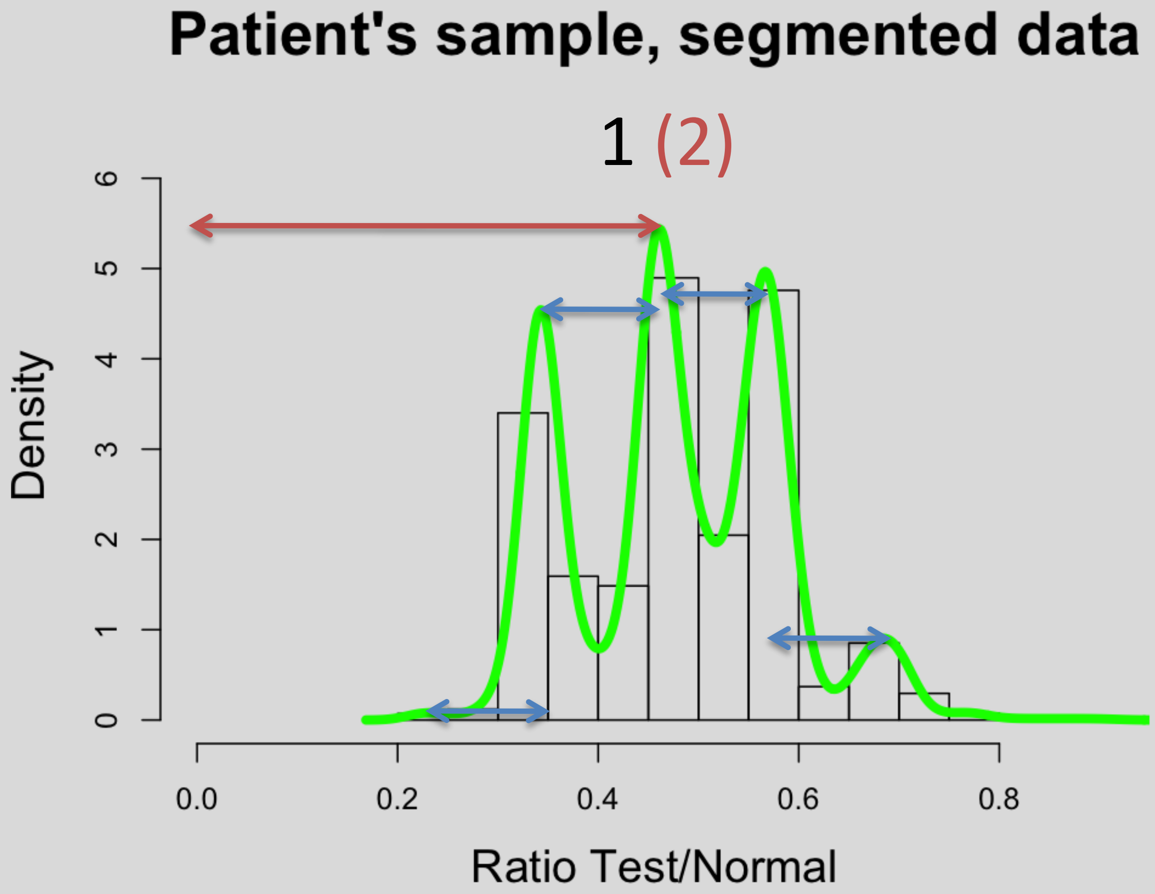
UNIVERSITY OF LEEDS



Ratio (Copy num)

Discrete data normalisation

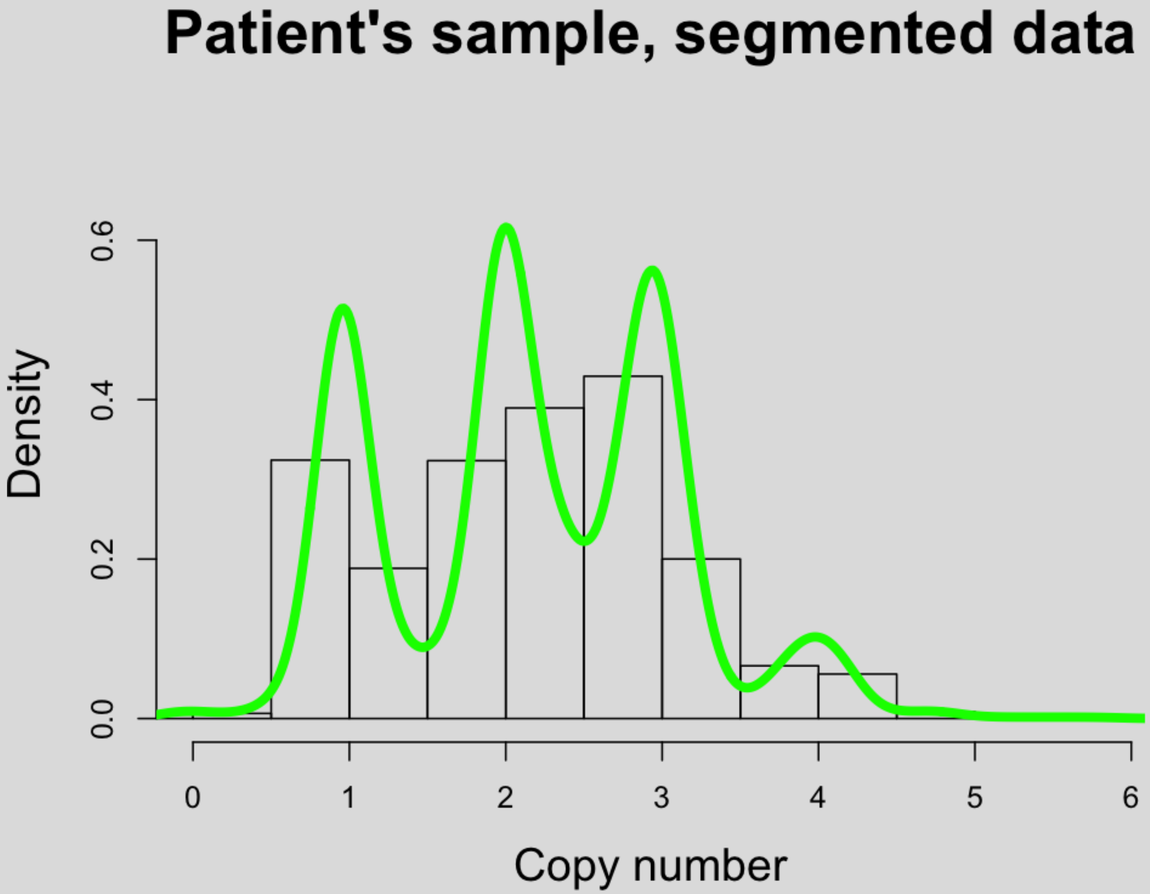
UNIVERSITY OF LEEDS



Ratio (Copy num)

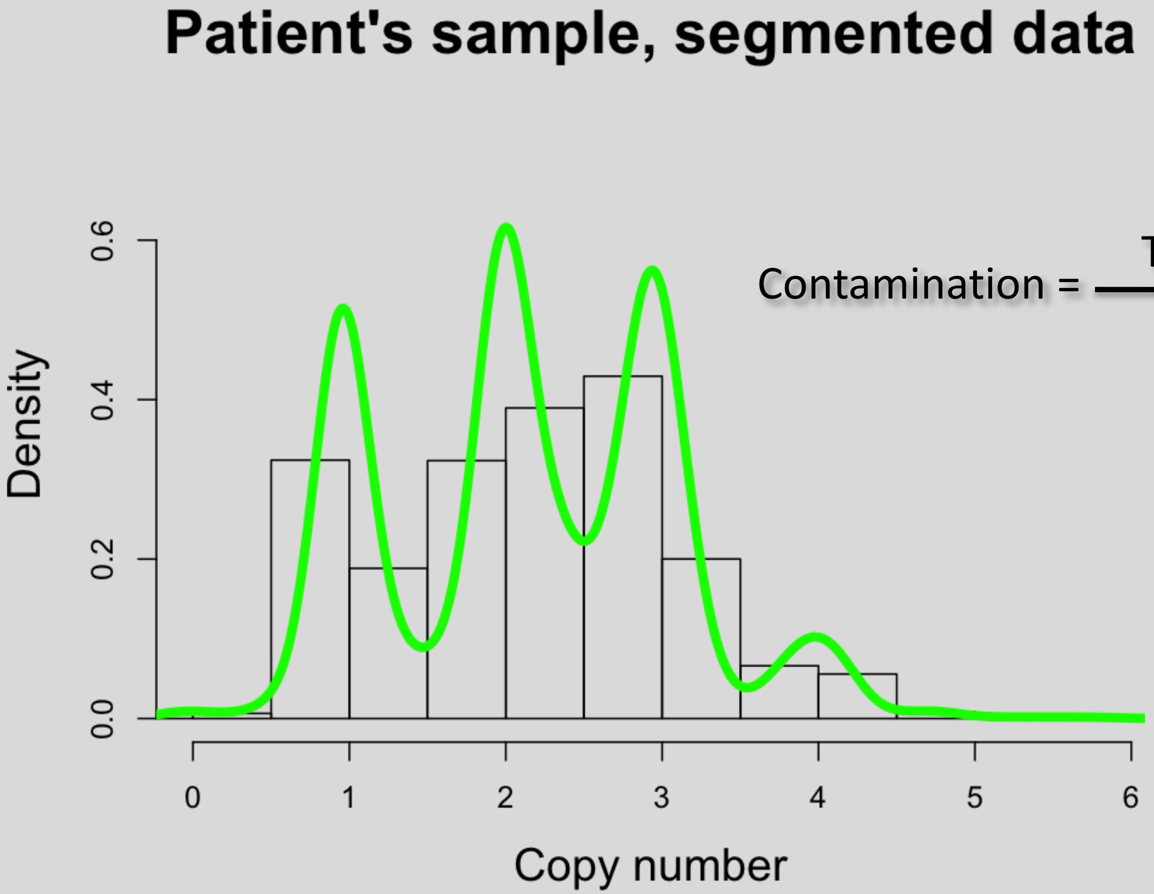
Discrete data normalisation

UNIVERSITY OF LEEDS



Discrete data normalisation

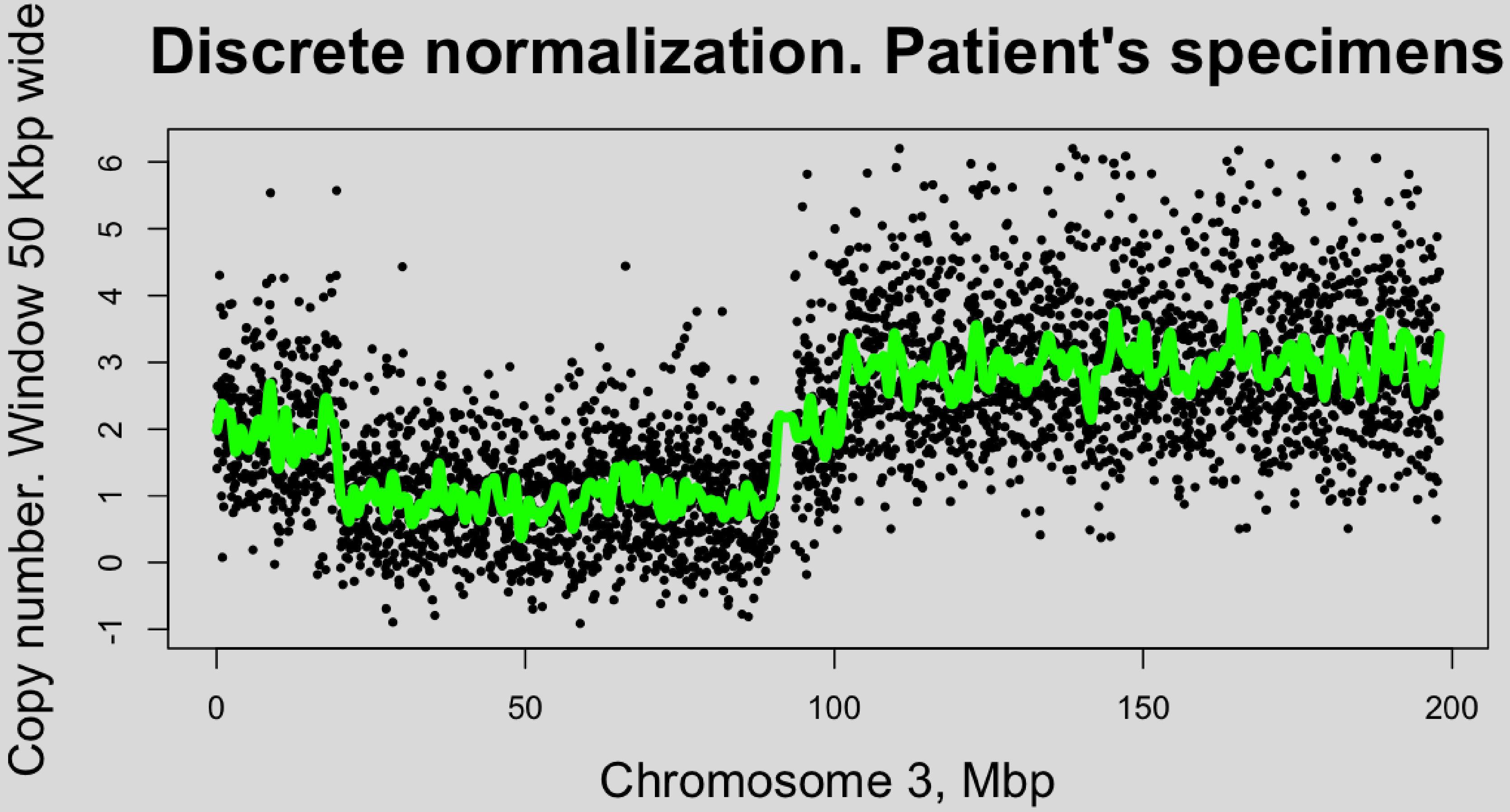
UNIVERSITY OF LEEDS



$$\text{Contamination} = \frac{\text{Tumour tissue}}{\text{Total tissue}} = 51\%$$

Keep high resolution and normalise

UNIVERSITY OF LEEDS



Conclusions

- Develop a novel normalisation method for “NextGen” data that can cope with
  - ✓ Highly abnormal genomes
  - ✓ Tumour samples contaminated by normal cells
- We can estimate contamination percentage.

Limits

- Contamination is allowed, but otherwise the tumour should be homogeneous.
- Process might require human supervision when calling discrete states.

Acknowledgements

Department of Statistics, School of Mathematics

Arief Gusnanto

Leeds Institute of Molecular Medicine, St. James's University Hospital

Pamela Rabbitts, Henry M. Wood

Catherine Daly, Ornella Belvedere

Joanne Morgan, Graham Taylor

Yorkshire Cancer Research