Quick primer on SCNA callling

Fong Chun Chan

Quick primer on SCNA calling using high-throughput sequencing data

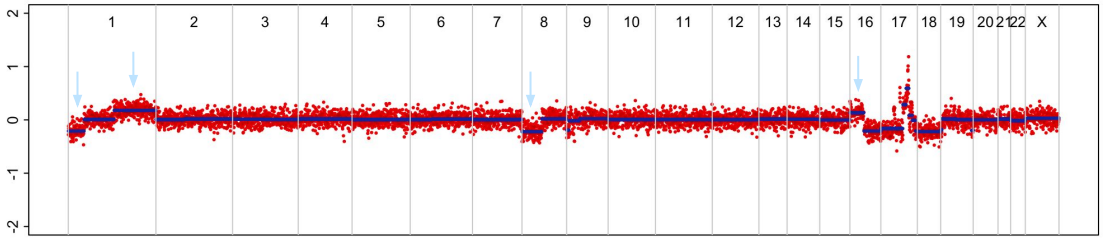
- Most SCNA callers use a read-depth based approach:
  - Contrasted to SV callers (e.g. LUMPY) that use split and spanning reads
- Two main input channels:
  - Log2 ratio (logR): Relative depth

From matched normal or panel of normals

$$RDR_i = \frac{\text{Tumour Coverage}}{\text{Normal Coverage}}$$

$$logR_i = log_2\left(RDR_i\right)$$

S28 from ASCAT example data



Quick primer on SCNA calling using high-throughput sequencing data

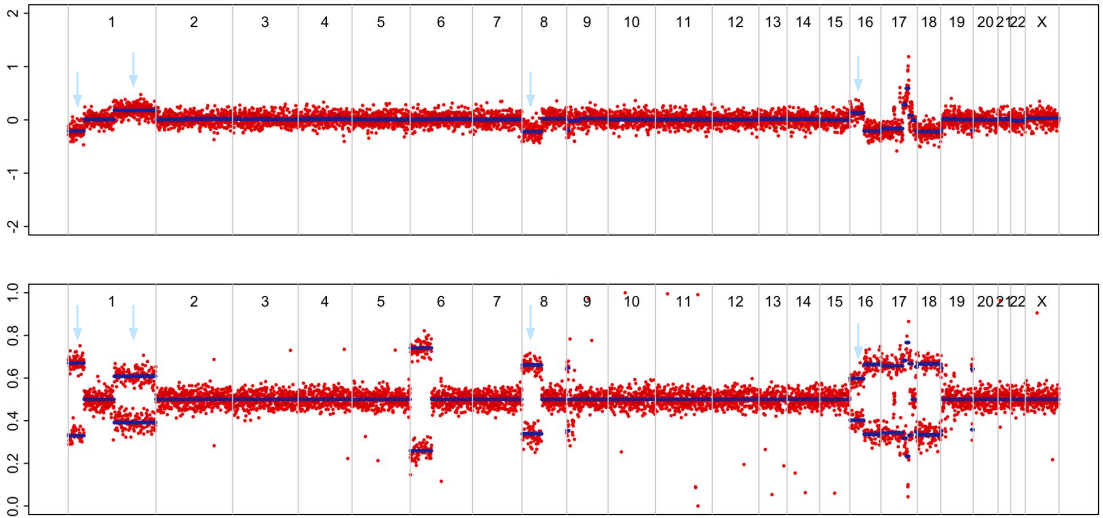
- Most SCNA callers use a read-depth based approach:
  - Contrasted to SV callers (e.g. LUMPY) that use split and spanning reads
- Two main input channels:
  - Log2 ratio (logR): Relative depth
  - B-allele fraction (BAF): Allelic imbalance
    - Focus on SNPs
    - **Gives the ability to call allele-specific copy number**
- Data are segmented into regions of constant copy number (i.e. blue lines)

From matched normal or panel of normals

$$RDR_i = \frac{\text{Tumour Coverage}}{\text{Normal Coverage}}$$

$$logR_i = log_2\left(RDR_i\right)$$

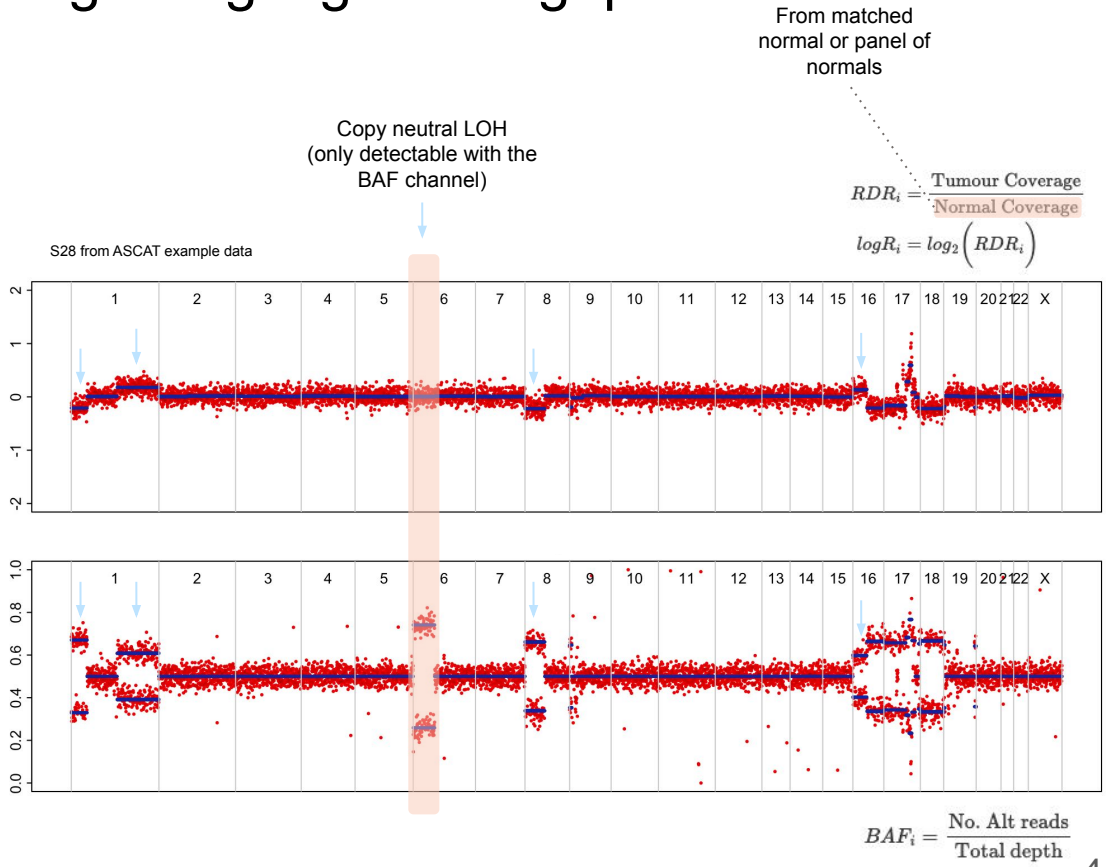
S28 from ASCAT example data



$$BAF_i = \frac{\text{No. Alt reads}}{\text{Total depth}}$$

Quick primer on SCNA calling using high-throughput sequencing data

- Most SCNA callers use a read-depth based approach:
  - Contrasted to SV callers (e.g. LUMPY) that use split and spanning reads
- Two main input channels:
  - Log2 ratio (logR): Relative depth
  - B-allele fraction (BAF): Allelic imbalance
    - Focus on SNPs
    - **Gives the ability to call allele-specific copy number**
- Data are segmented into regions of constant copy number (i.e. blue lines)

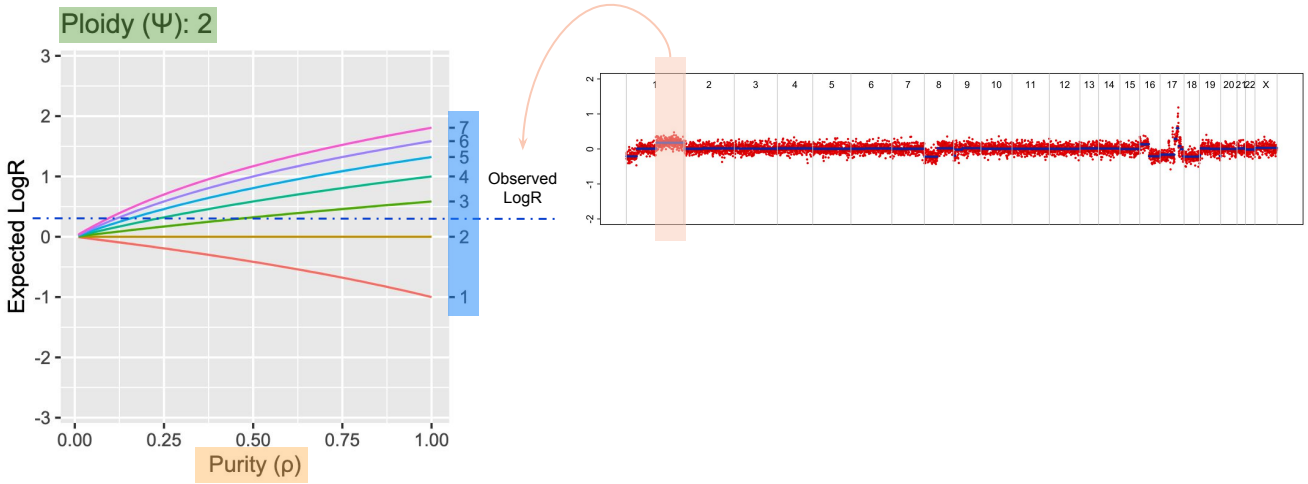


Statistical framework around SCNA calling

Determine the likelihood of observed LogR

Total copy number

$$\text{Expected } LogR = \frac{2(1-\rho) + \rho C}{2(1-\rho) + \rho\psi}$$



Statistical framework around SCNA calling

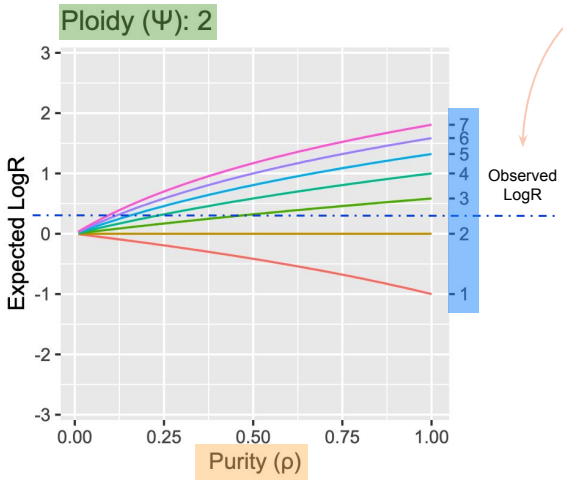
Determine the likelihood of observed BAF

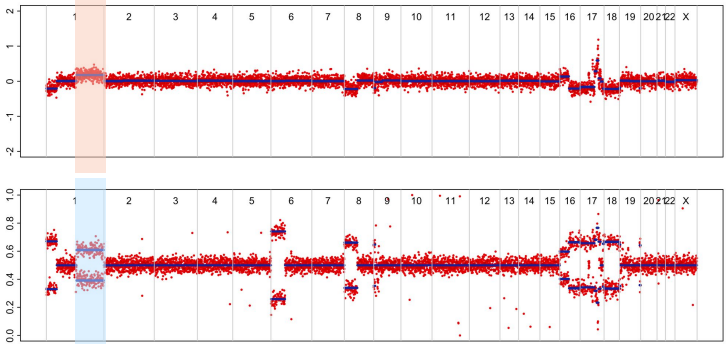
Total copy number

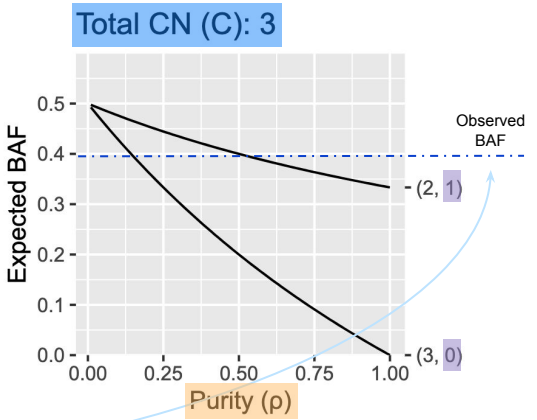
$$\text{Expected } LogR = \frac{2(1-\rho)+\rho C}{2(1-\rho)+\rho\psi}$$

Minor copy number

$$\text{Expected BAF} = \frac{(1-\rho)+\rho n_b}{C\rho+2(1-\rho)}$$

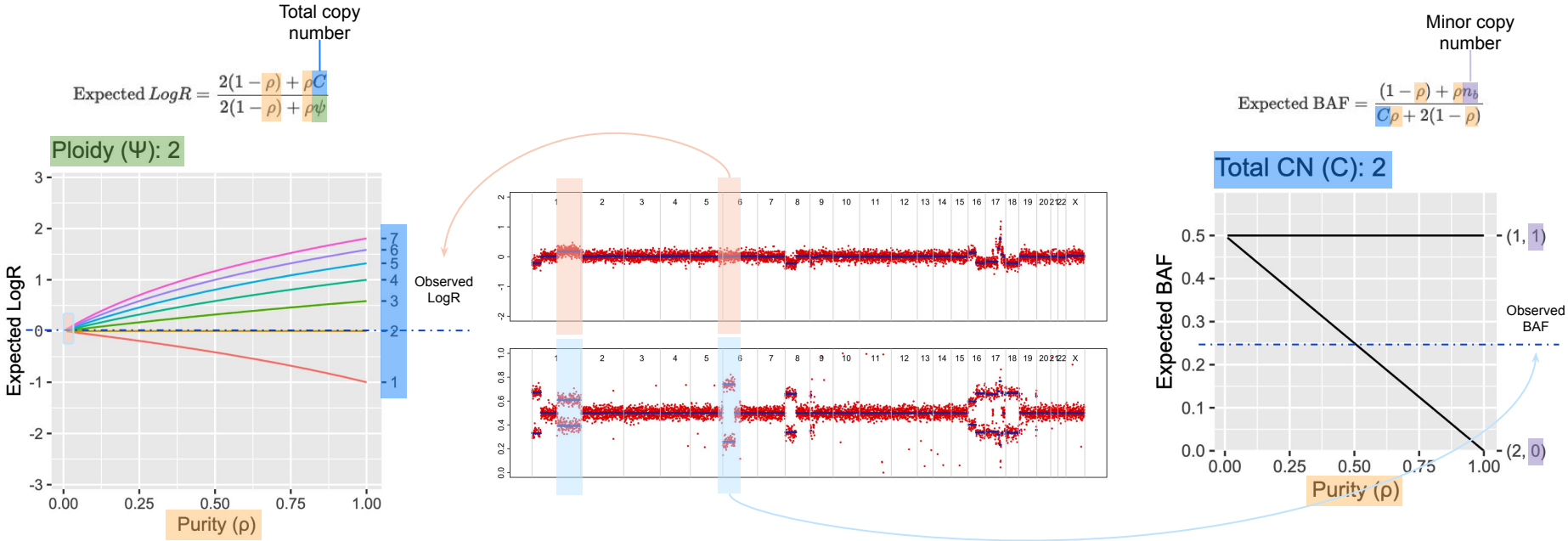






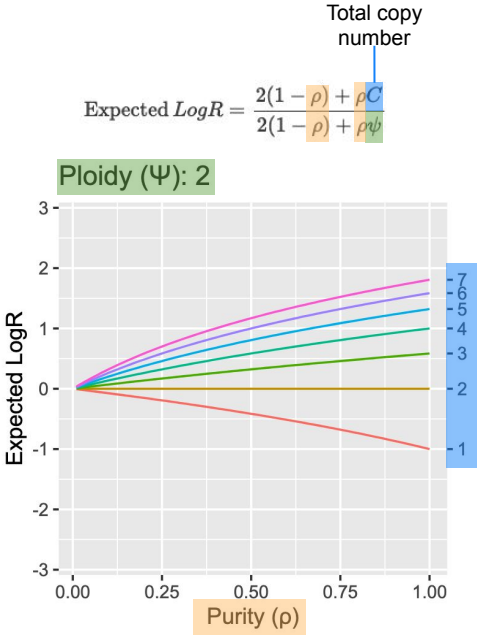
Statistical framework around SCNA calling

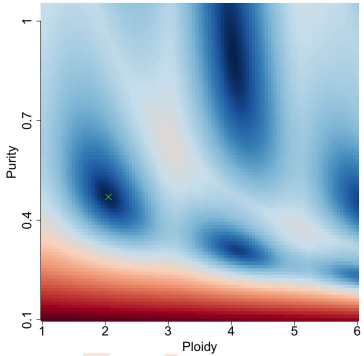
Determine the joint likelihood of observed LogR and BAF for each genomic loci *i*

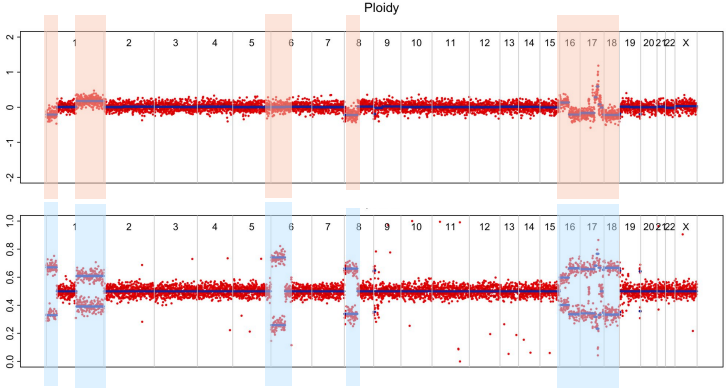


Statistical framework around SCNA calling

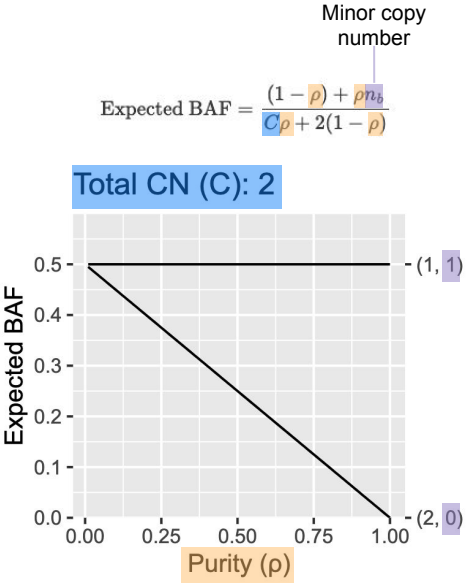
Determine the purity/ploidy total likelihood grid





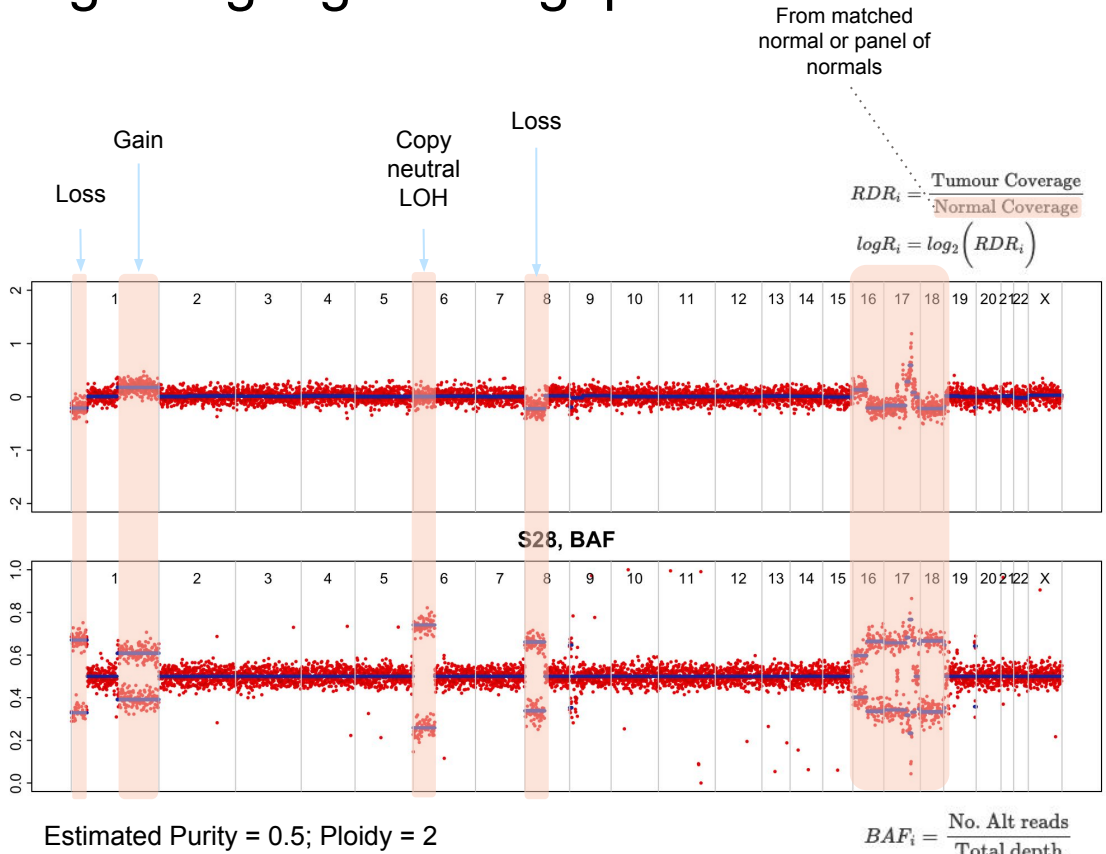


SCNA calling has a model identifiability issue. Multiple parameterization (purity, ploidy) can explain the same observed data



Quick primer on SCNA calling using high-throughput sequencing data

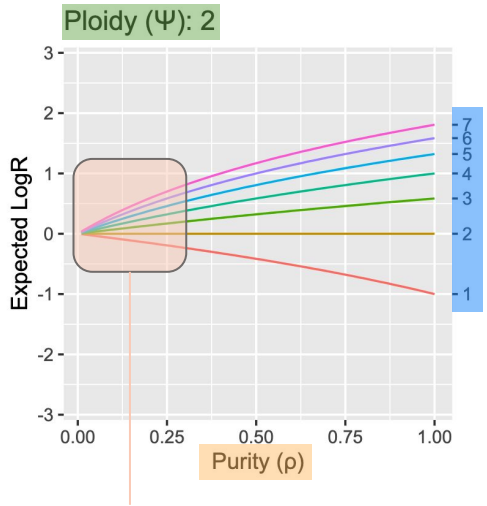
- Most SCNA callers use a read-depth based approach:
  - Contrasted to SV callers (e.g. LUMPY) that use split and spanning reads
- Two main input channels:
  - Log2 ratio (logR): Relative depth
  - B-allele fraction (BAF): Allelic imbalance
    - Focus on SNPs
    - **Gives the ability to call allele-specific copy number**
- Data are segmented into regions of constant copy number (i.e. blue lines)
- Segments are classified into copy number events



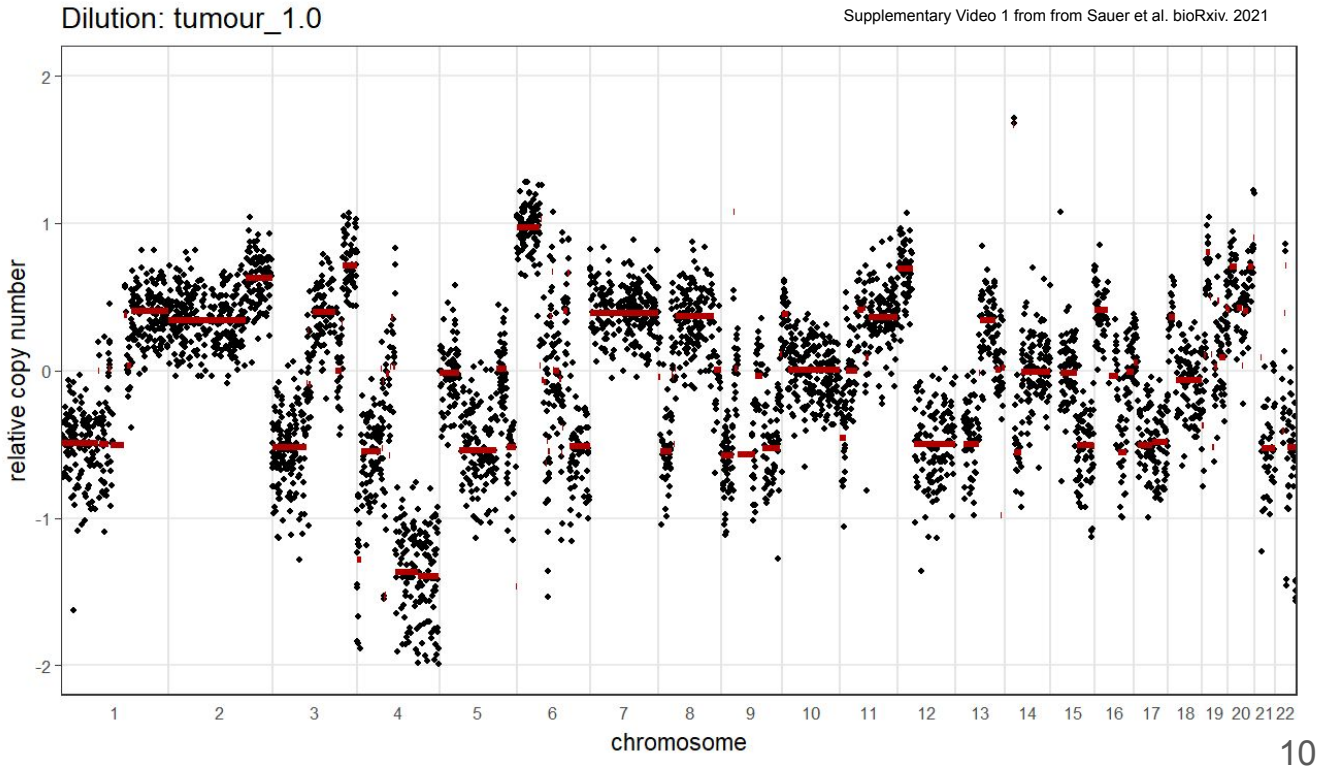
Tumour purity plays a large role in SCNA signal

Lower purity means fewer cells harbour the SCNA events -> weaker signal (i.e. signal to noise ratio is decreased)

$$\text{Expected } LogR = \frac{2(1-\rho) + \rho C}{2(1-\rho) + \rho\psi}$$



At < 30% purity, distinguishing between different copy number states becomes very challenging



Supplementary Video 1 from from Sauer et al. bioRxiv. 2021