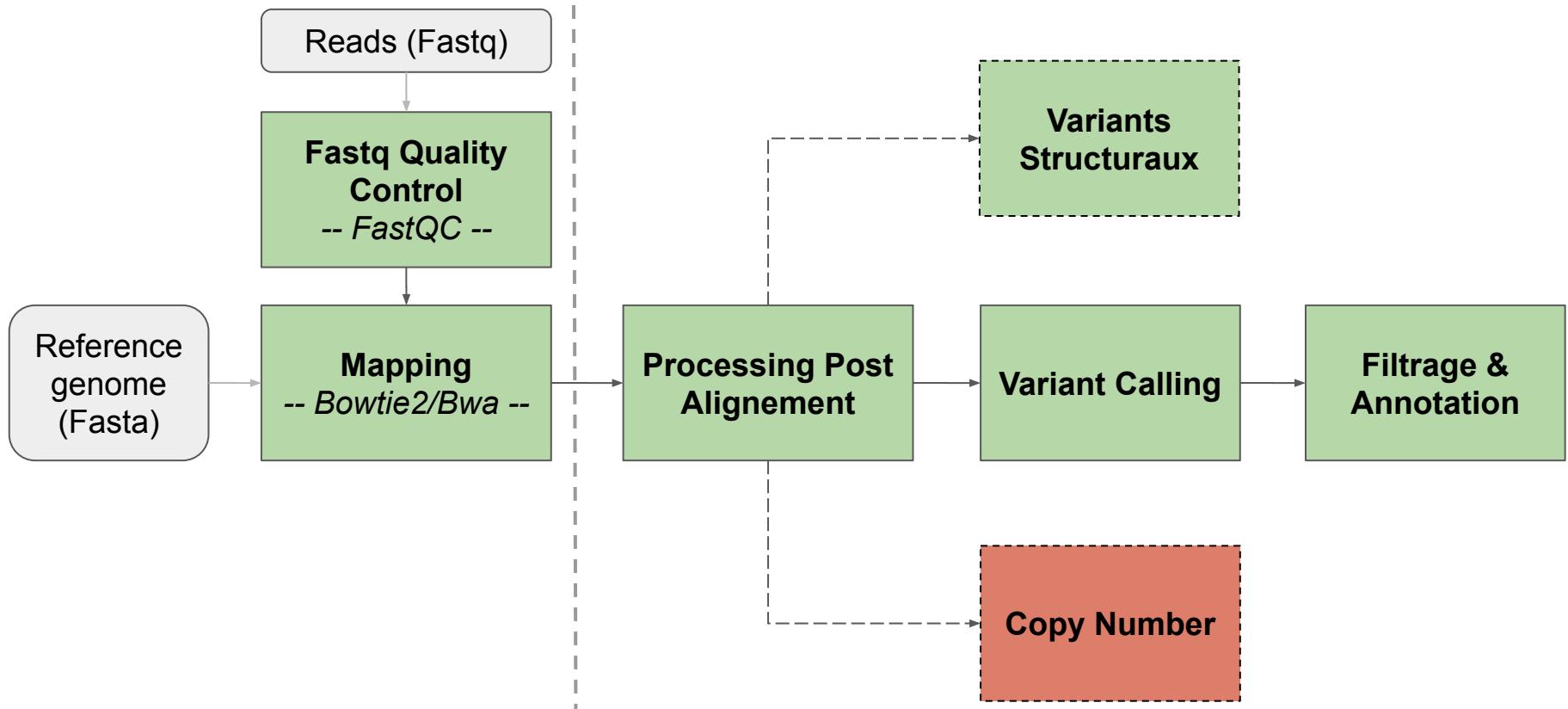


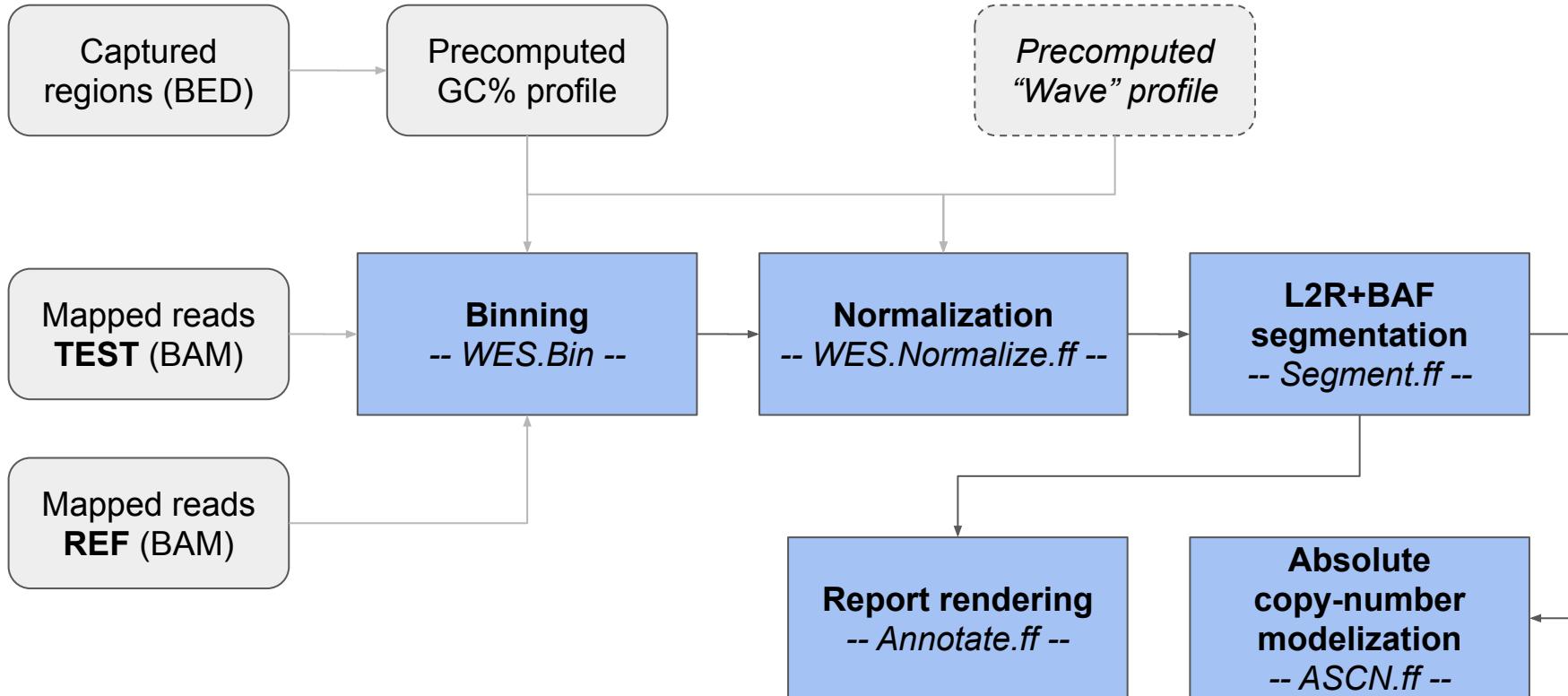
Analysis of Genomic Copy Number Alterations

Bastien Job - INSERM / Gustave Roussy

DNAseq Workflow



Copy Number Alteration Workflow

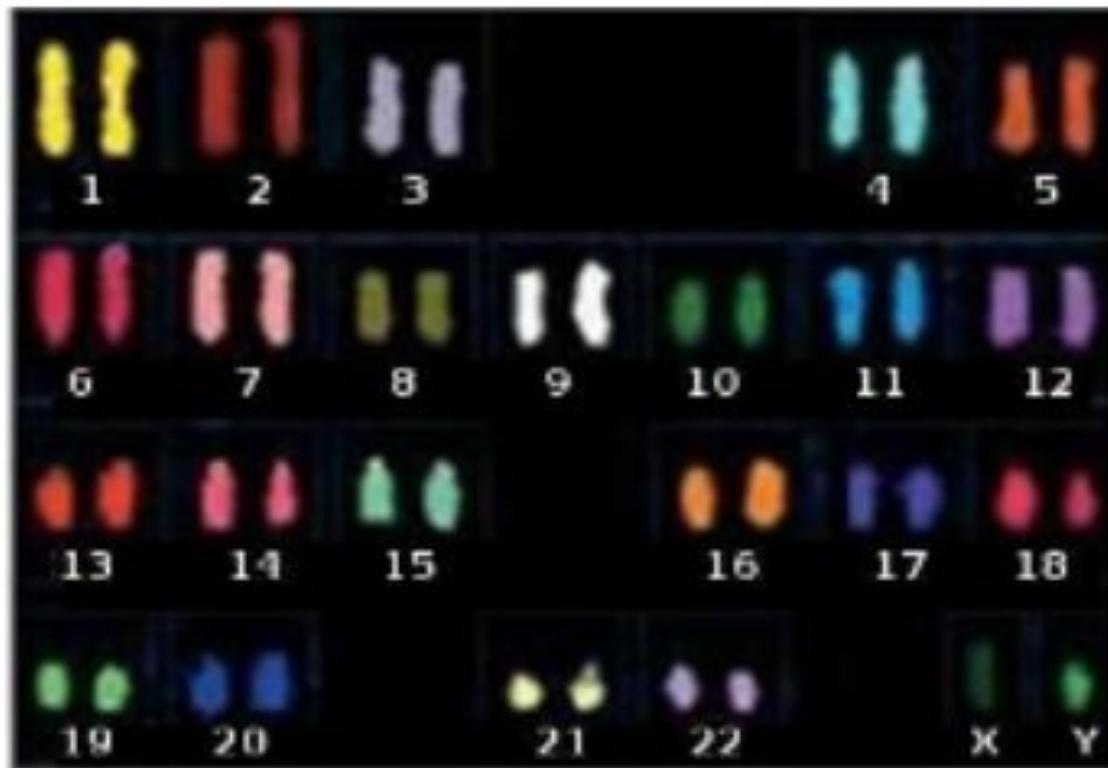


A bit of vocabulary

- "CNV" (Copy Number Variation) includes :
 - "CNA" :
 - Copy Number *Anomaly*
 - Copy Number *Alteration*
 - Copy Number *Aberration*
 - Copy Number *Abnormality*
 - Large-scale (> 1Kb) polymorphisms (often called ... "CNV")

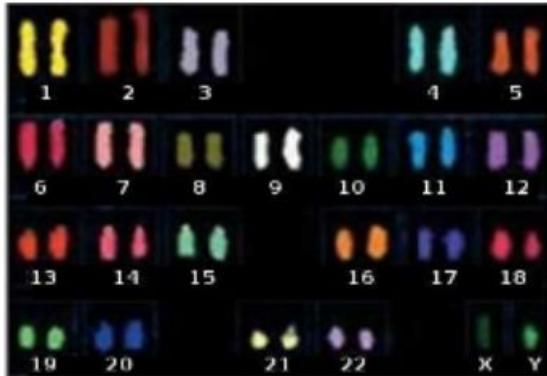
Copy Number Alterations (and Cancer)

Normal cell

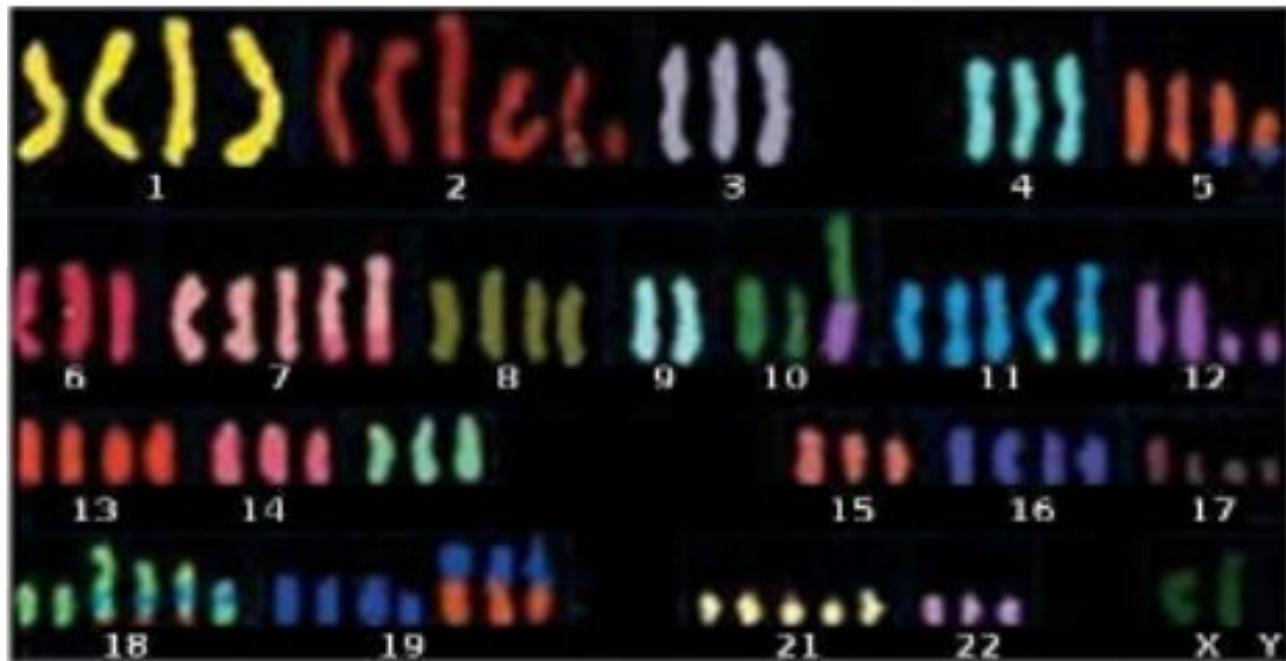


Copy Number Alterations (and Cancer)

Normal cell

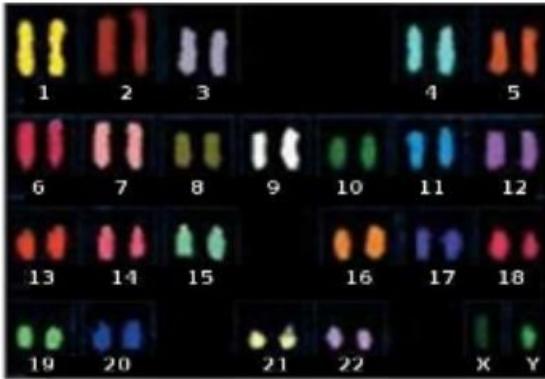


Tumor cell

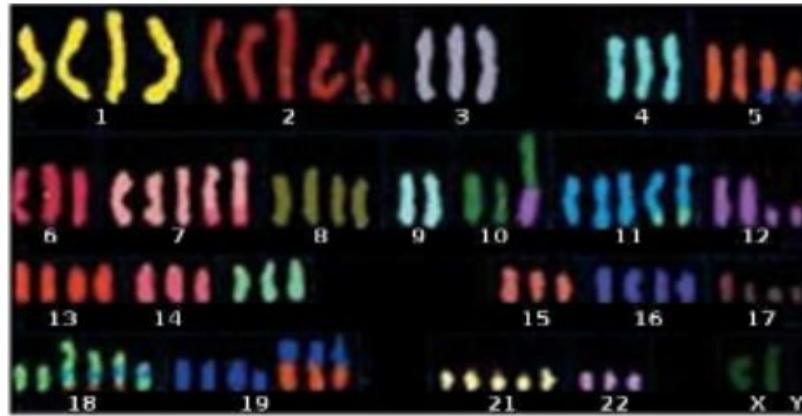


Copy Number Alterations (and Cancer)

Normal cell



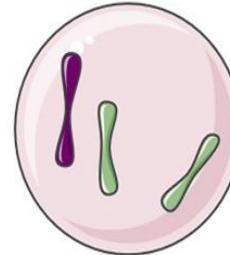
Tumor cell



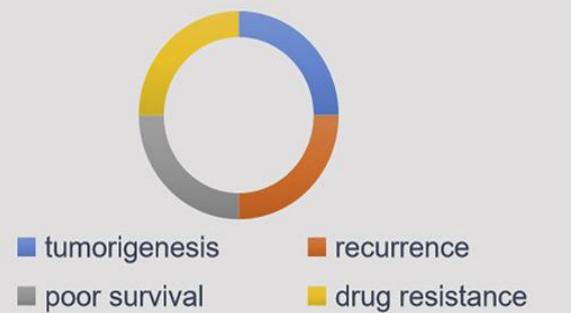
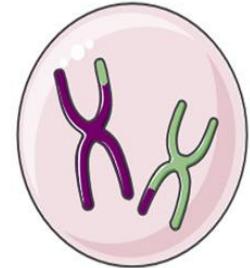
Chromosome
Abnormalities

Genomic
Instability

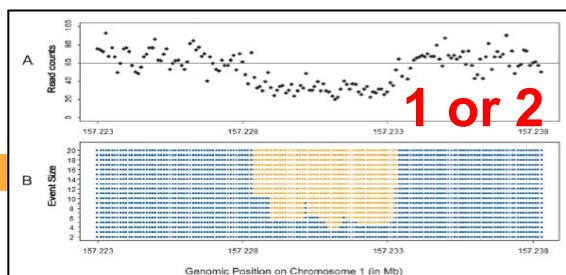
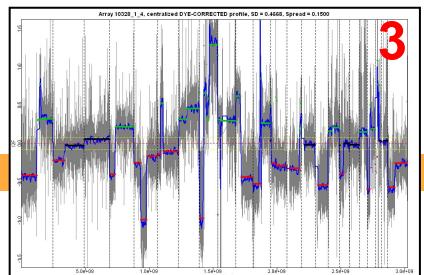
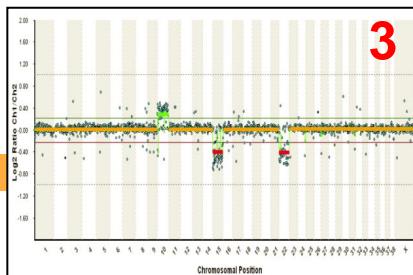
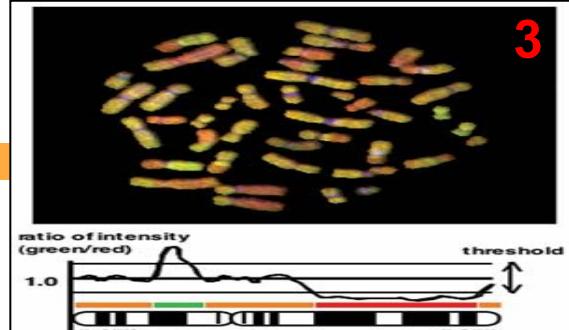
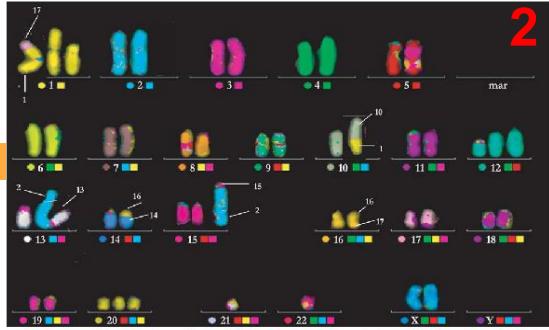
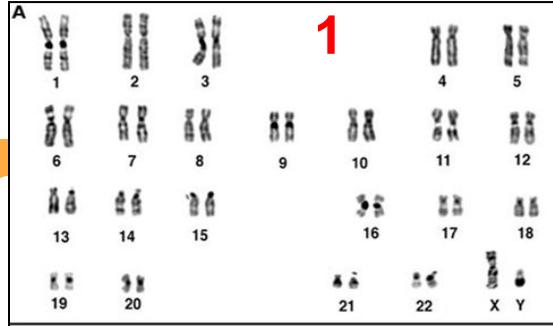
Numerical Chromosome
Abnormalities



Structural Chromosome
Abnormalities

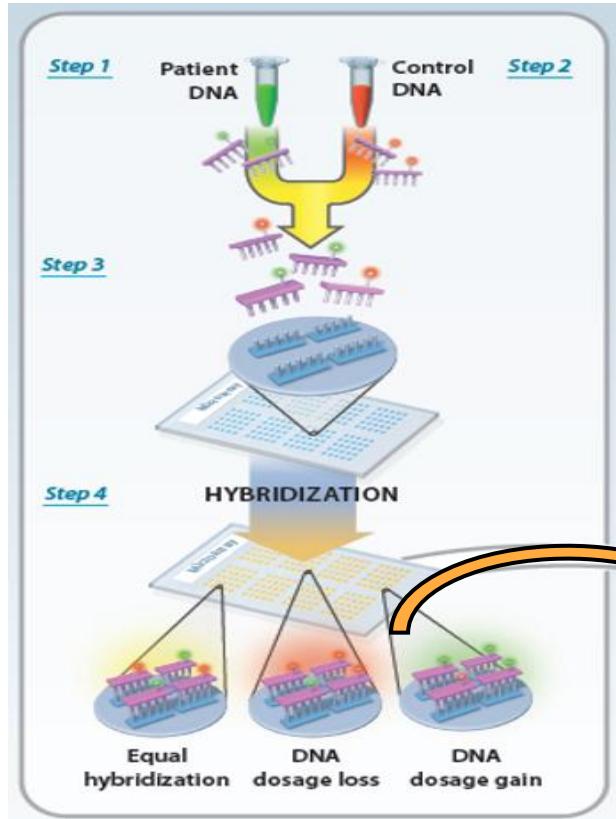


A Bit of History

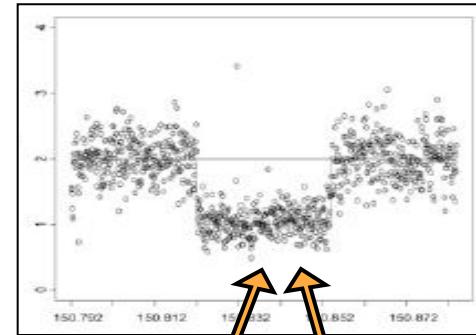


Technical Principle

A Microarray



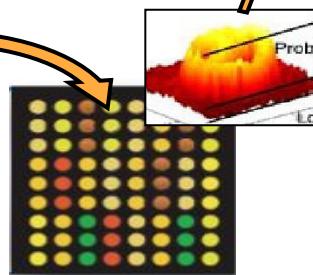
CNA Detection



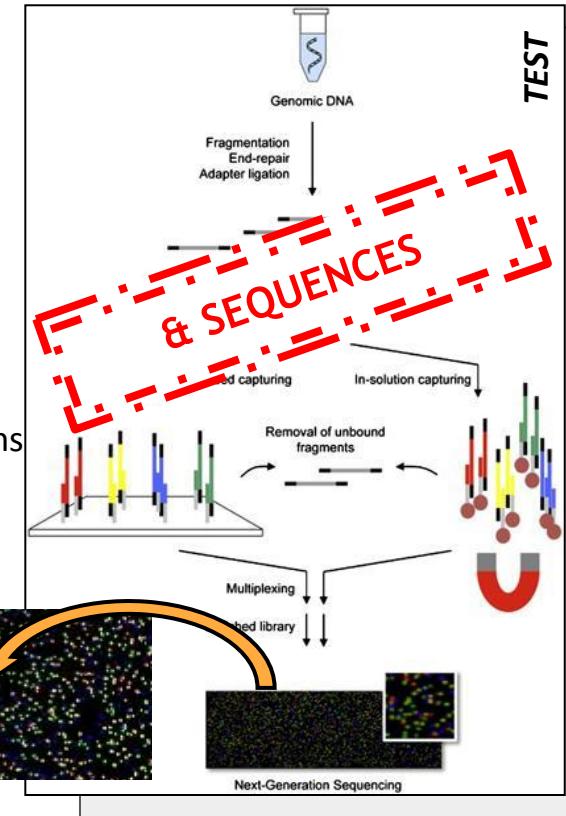
Probe
intensities

Read depths

Maps



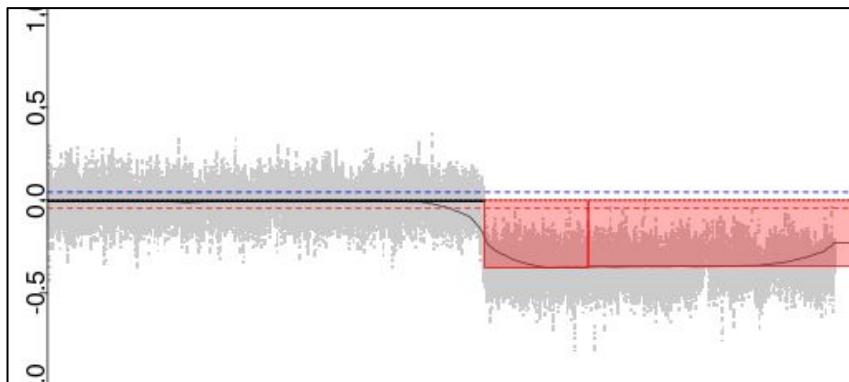
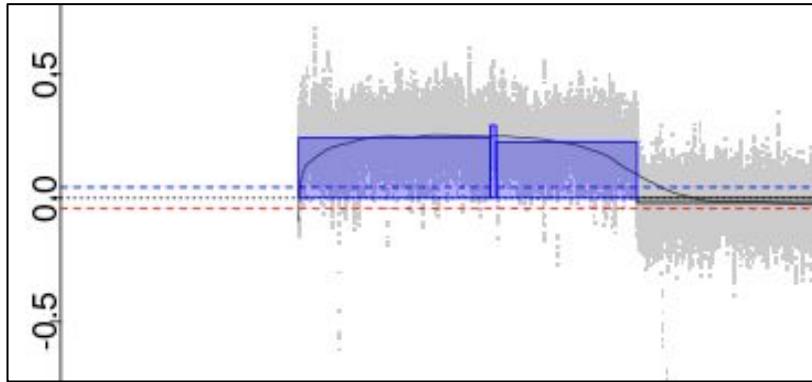
S HTS



REFERENCE

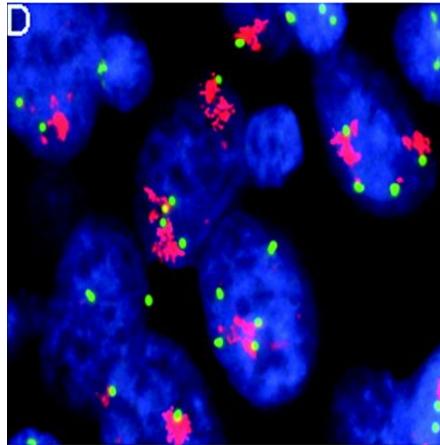
A Family of Events : Gain and Loss

- First cases of abnormality

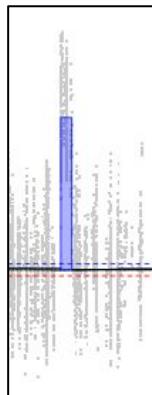


A Family of Events : Amplification

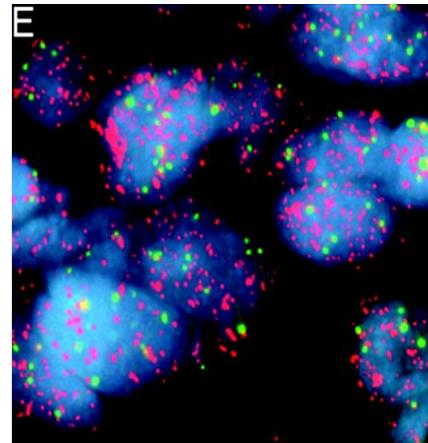
- Extreme gain case
- Theoretical level : $L2R \geq 1.5$:
 - 3 additional copies from diploidy
 - $3 + 2 = 5$ copies
 - $\log_2(5) \approx 1.58$
 - rounded to 1.5
- Classically focal event
- Two sorts
 - Homogeneously stained regions (HSR) : multiple tandem
 - Double-minute chromosomes (up to thousands of copies)



EGFR amplification in lung cancer as **HSR** (*homogeneously stained region*)



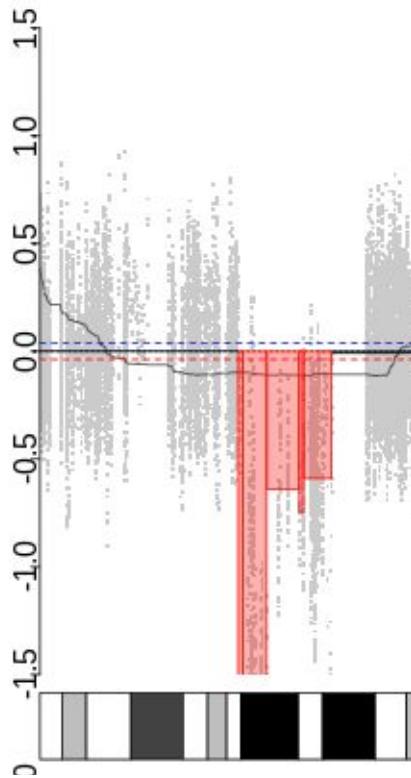
EGFR amplification in lung cancer as **double-minutes**



Varella-Garcia et al, J Clin Pathol 2009

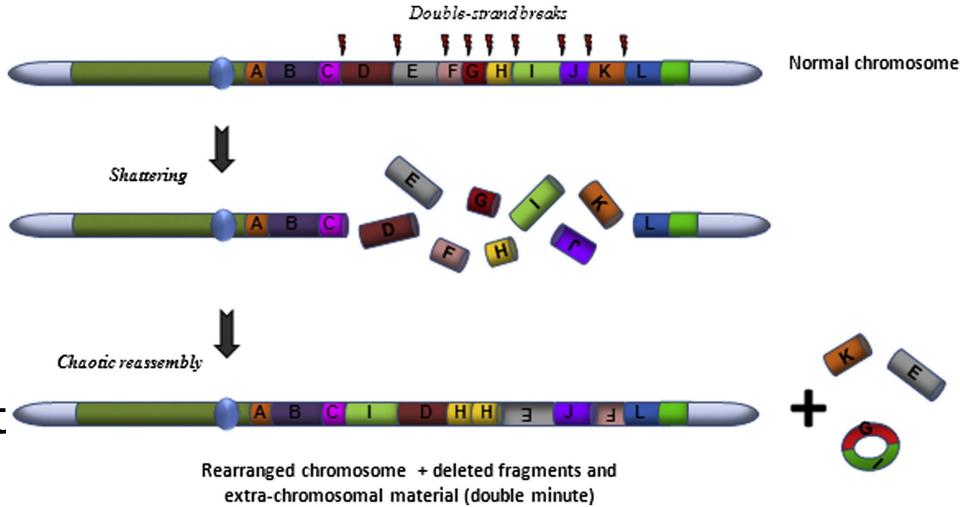
A Family of Events : Deletion

- Extreme loss case (no remaining copy)



A Family of Events : Chromothripsis

- Extreme, catastrophic event
- Up to thousands of fragments involved
- Single temporal event
- Classically arm- or whole chromosome-level
- Can result in external double-minut chromosomes
- Alternance of 2, sometimes 3 copy levels
- Few locations (up to two chromosomes)
- If more locations : *chromoplexy*

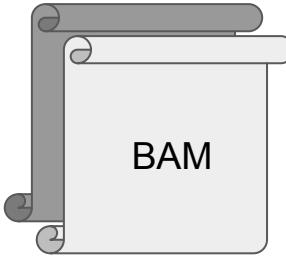


A Family of Events : Loss of Heterozygosity

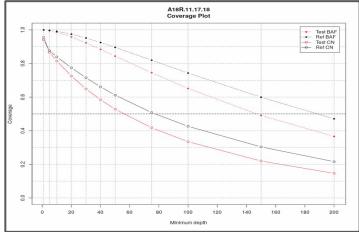
- Measured using L2R and the local frequency of parental alleles (**BAF**, AD, beta-score), thanks to SNVs.
- Quantitative loss of one of the two parental alleles : **LOH**
 - Obvious case : single copy loss or gain from an even ploidy
 - Extreme loss case : single copy loss from diploidy : **haplody**
- Not mandatorily a copy number event :
 - loss or gain of a parental allele, with completion from the other one : **copy-neutral LOH**
 - Extreme case : total loss of a parental allele, recovering using the remaining one : **unisomy**

Analysis Workflow

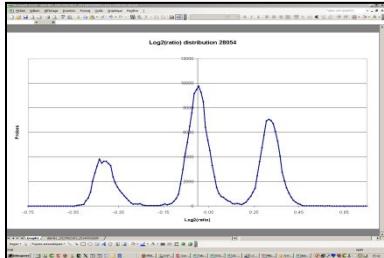
Acquisition



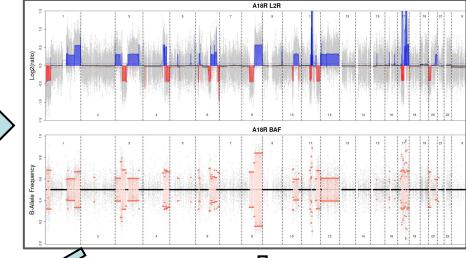
Binning + QC



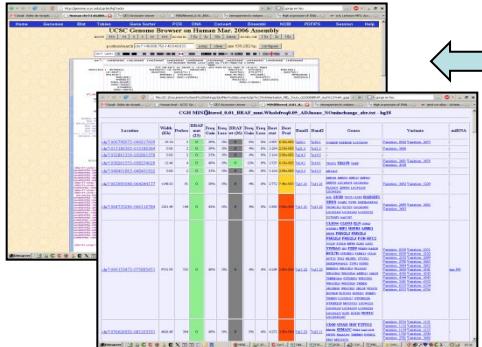
Normalization, centralization



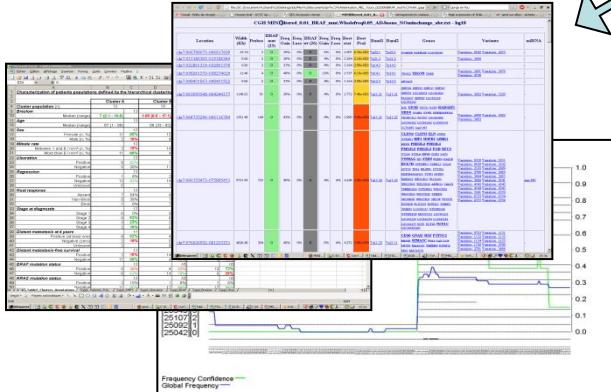
Segmentation, calling



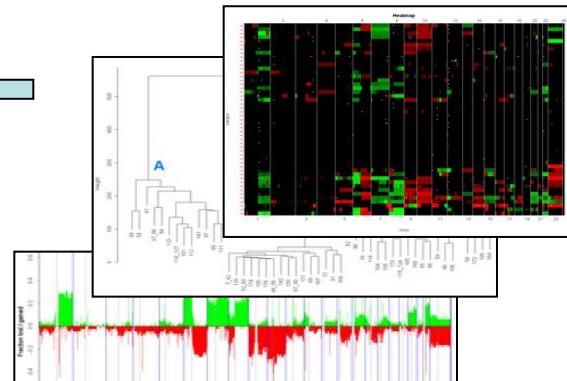
Annotation



Genomic regions of interest



Cohort analysis





- Bivariate segmentation (L2R + BAF), choice of 3 different algorithms : ASCAT, FACETS, SEQUENZA
- Total (TCN) and allele-specific (ASCN) copy number modeling, ploidy and tumor cellularity estimation
- Compatibility with WES and Affymetrix microarrays (SNP6, OncoScan family, CytoScan family)
- From “raw” (BAMs / CELs) data to annotated segments
- Rendering of a QC / results HTML report
- Compatible with any genome that can be handled by BSgenome
- Full R, quite recent (open source since 2018-05)
- Built for bioinformaticians and researchers with starting R knowledge
- <https://github.com/gustaveroussy/EaCoN>

Our Training Dataset

- Data source : The Cancer Genome Atlas <https://cancergenome.nih.gov/>
- Pathology : breast cancer (BRCA)
- Sample name : BH-A18R
- Sequencer : Illumina HiSeq 2000 (2011/05)
- Sequencing kit : Illumina Paired-End 2 x 101 pb
- Capture kit : “NIMBLEGEN exome version 2”
- Mapper : BWA
- Reference genome build : hs37d5 (\approx hg19 without “chr” in chr names)
- Restrictions (for execution time) :
 - Reads from chr11, chr17 and chr18 only
 - Depth dropped to 20% of original (Tumor : 4,790,417 ; Normal : 5,643,177)

PRACTICE : Warm-up (IFB cluster)

```
# Requesting and interactive shell with needed resources
$ srun --cpus=4 --mem=16G -J session_<user_name> --pty bash

# Loading the already prepared EaCoN v0.3.5 execution environment
$ module load r-eacon/0.3.5

# Building a local directory for this training session and copying data
$ mkdir ~/tp_cna
$ cp -r /shared/projects/ebai2021_n2/data/dna_seq/cna/* ~/tp_cna
$ cd ~/tp_cna

# Building our output directory
$ mkdir -p ~/tp_cna/RESULTS/REDUX

# Opening and interactive R session
$ R
```

PRACTICE : Warm-up (R)

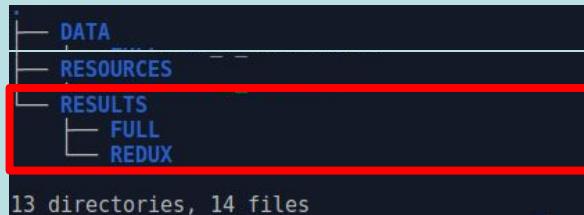
Once in R :

- a. Move to your output directory

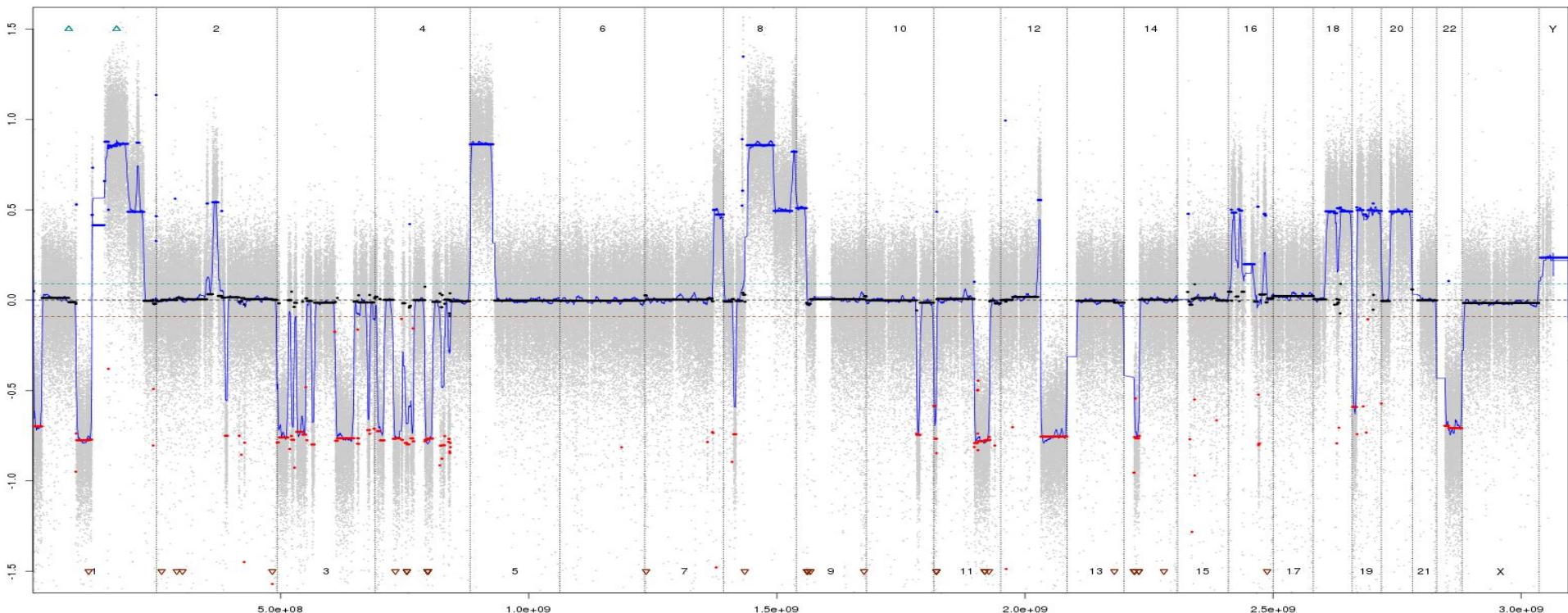
```
setwd("~/tp_cna/RESULTS/REDUX")
```

- b. Charger le package EaCoN :

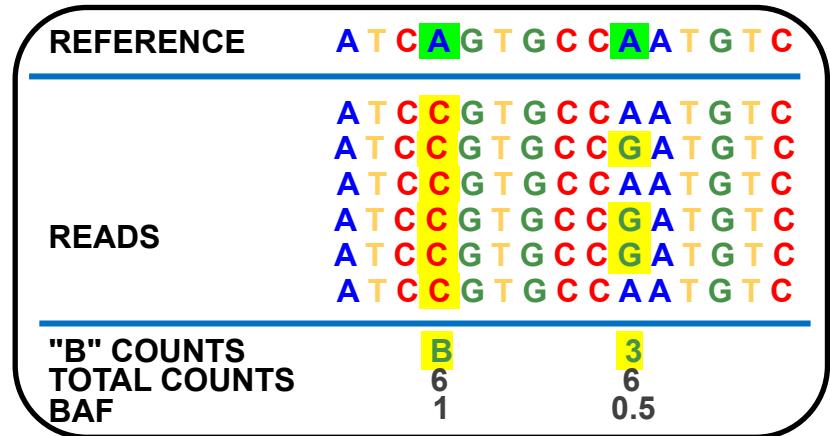
```
library(EaCoN)
```



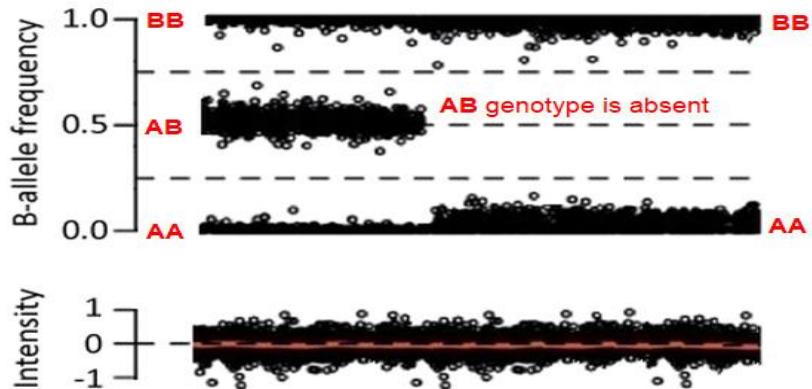
Our aim : a CNA profile (L2R)



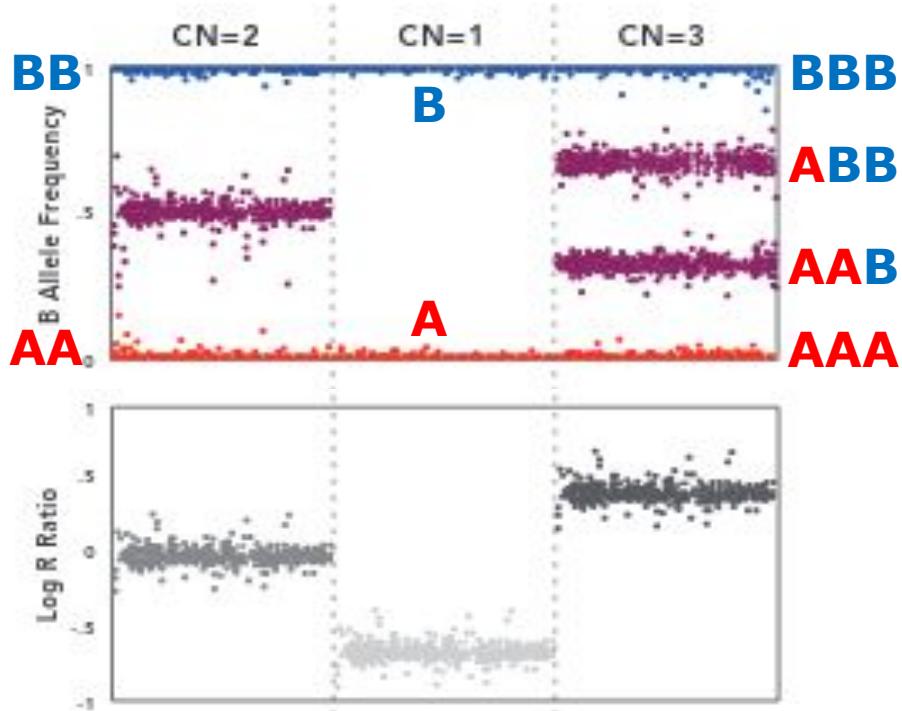
Our aim : a CNA profile (BAF)



loss of heterozygosity

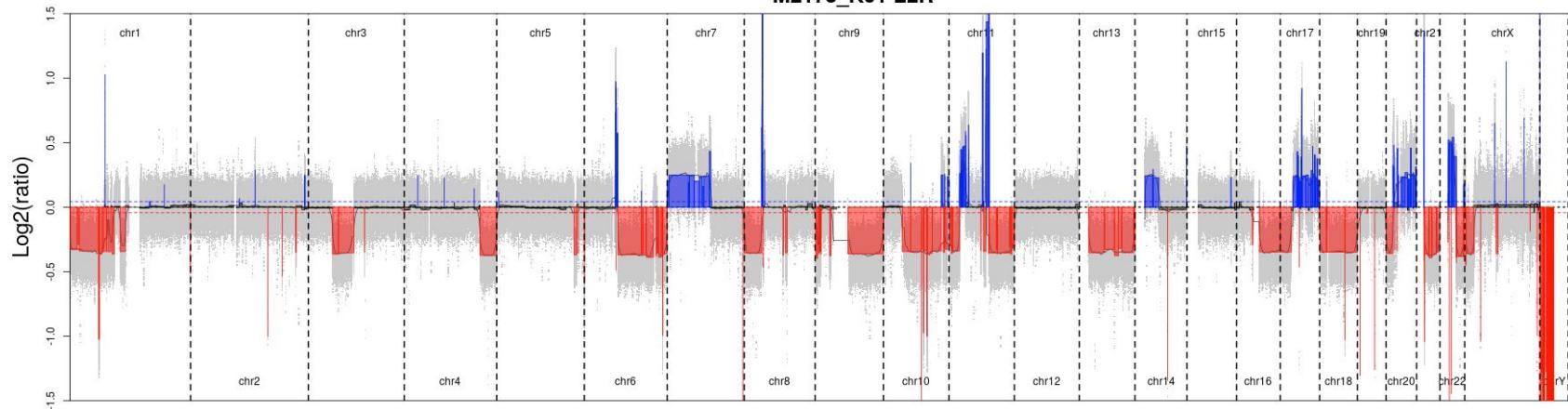


Copy Number Analysis

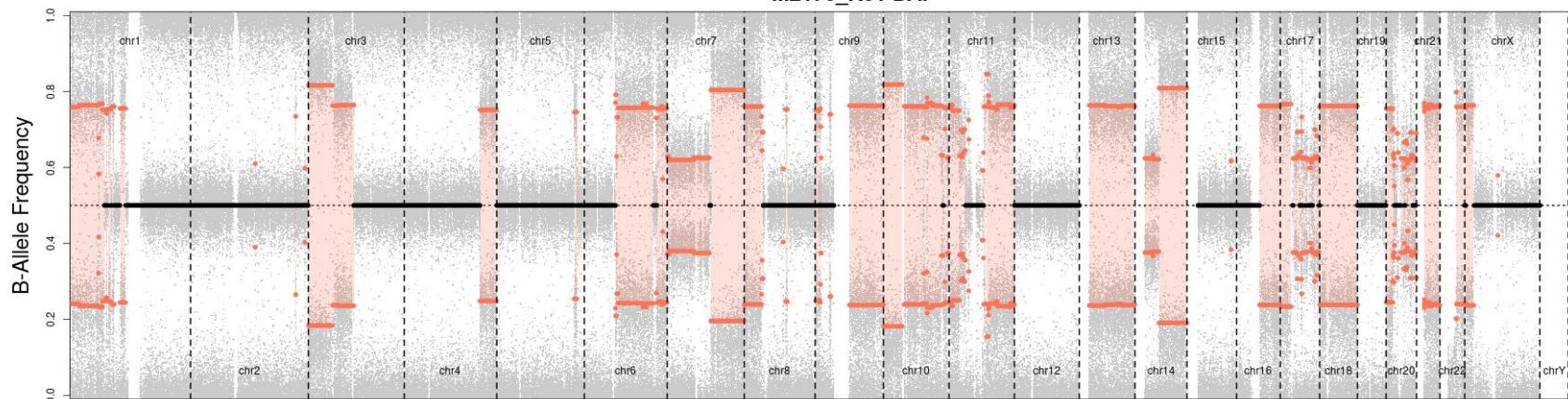


Our aim : a CNA profile

M2173_K01 L2R



M2173_K01 BAF



Data Reduction

-

Binning

Data reduction (binning) : Why ?

- WES = fragmented data by nature (capture)
- Reduction is necessary :
 - Computational time. An example for hg19 WES :
 - Capture BED width \approx 65 Mb
 - Reads \approx 100 M, length \approx 75 b
 - CPU times :
 - 50 b bins (1 CPU) \approx 5h
 - Without binning \approx $50 \times 5h = 250 h = 10j\ 10h !$
 - Some incompatible methodologies. Ex : GC% normalisation :
 - Computed on windows (score 0 <> 100), impossible at the single nucleotide level (A or T = 0%, C or G = 100%)
- Drawbacks :
 - Lowering the breakpoint precision

1	65409	65725
1	65731	66073
1	69381	69700
1	721281	722042
1	752816	753135
1	761995	762665
1	777159	777742
1	782961	783251
1	792170	792546
1	861166	861596
1	865482	865887
1	866231	866607
1	870964	871362
1	874267	874916
1	876385	876819

Data reduction (binning) : How ?

- Source data : BAM files
- Required data :
 - base-level depths
 - variants position and frequency
- Tool : *Rsamtools* package
 - Allows to efficiently read from BAM files
 - Performs mpileup generation :
 - Converts into an easy to use R object
- Additional steps :
 - Aggregating base-level depths to bins
 - Computing BAF values

Generating a “BINpack”

- BED entries (exons ?) are :
 - split if > bin size
 - kept as is if < bin size
- GC% computed using the genome sequence for different windows around each bin coordinates
 - by default : 0, 50, 100, 200, 400, 800, 1600, 3200, 6400 b
- Only done once per (bed + bin size + genome build) combo
- Optionally include other bin-level tracks (ie : wave effect)
- Results stored on disk into a RDS (R Data Storage) package
- *For our course, already prepared to spare few minutes of our time for something more interesting*

PRACTICE : Data binning

Input data files :

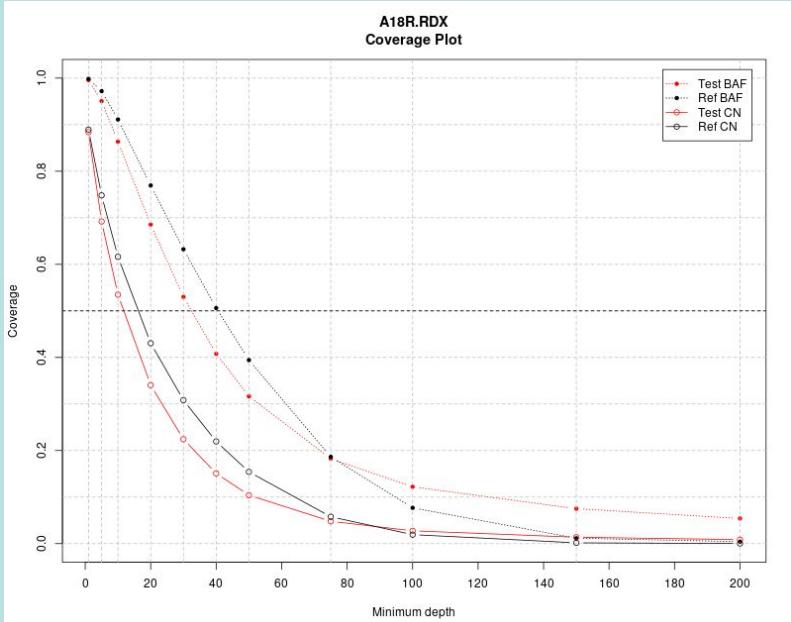
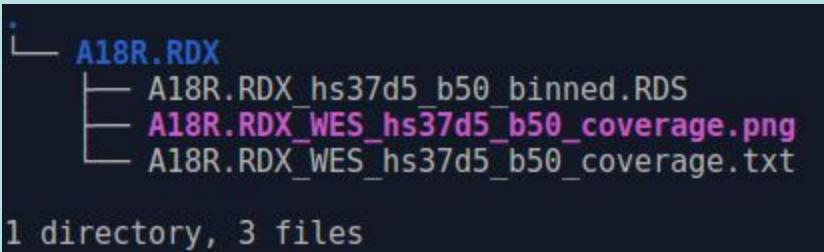
- Test BAM (tumor) :
 - ~/tp_cna/DATA/REDUX/A18R_T_RDX.bam
- Reference BAM (normal, same patient) :
 - ~/tp_cna/DATA/REDUX/A18R_N_RDX.bam
- Precomputed BINpack :
 - ~/tp_cna/RESOURCES/REDUX/SSCREp_RDX_b50.GC.rda

```
WES.Bin(testBAM = ".../..../DATA/REDUX/A18R_T_RDX.bam",
refBAM = ".../..../DATA/REDUX/A18R_N_RDX.bam",
BINpack = ".../..../RESOURCES/REDUX/SSCREp_RDX_b50.GC.rda",
samplename = "A18R.RDX", nsubthread = 3)
```

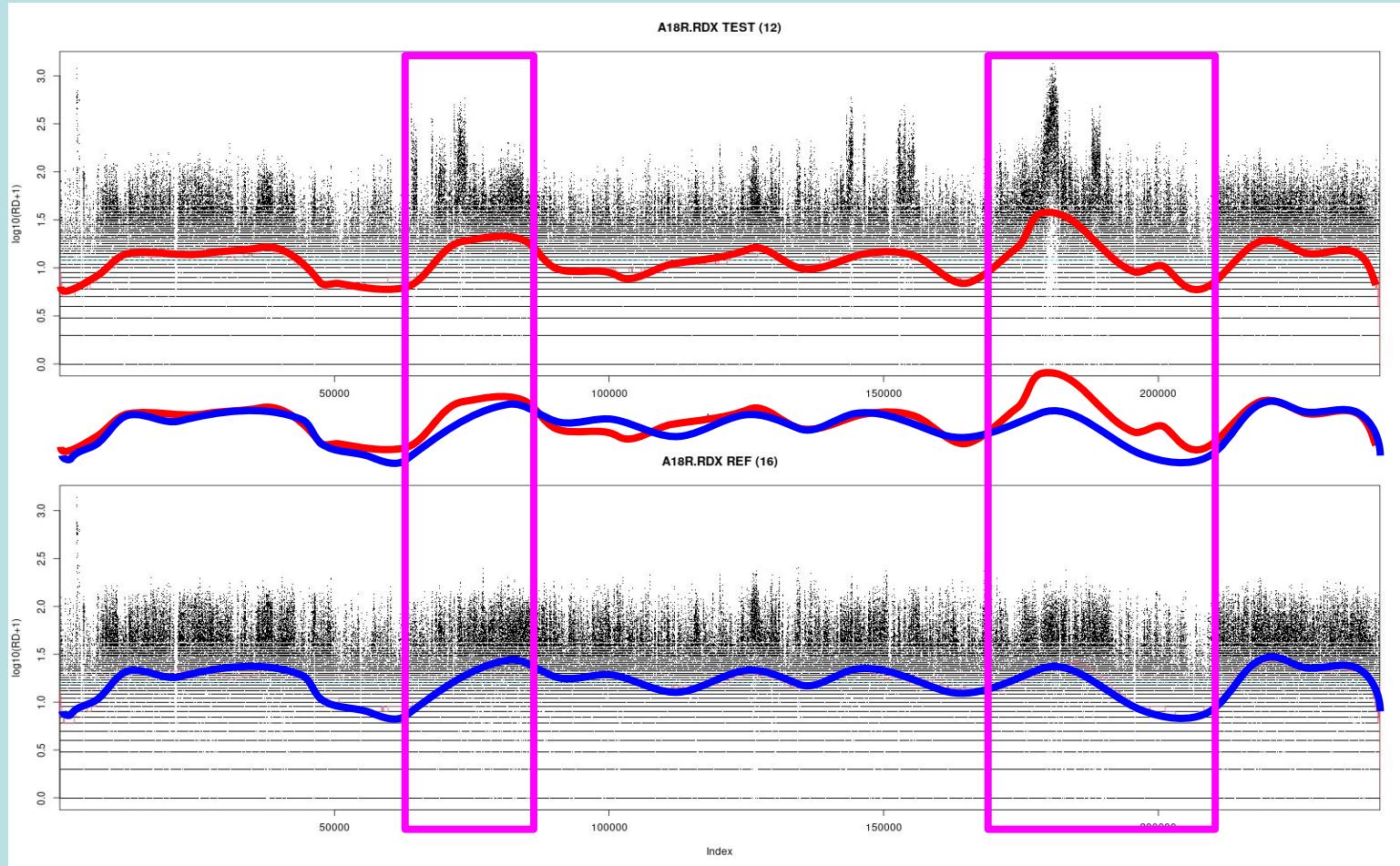
```
system("tree -sf")
```

Data Binning : Outputs

```
List of 3
$ RD :Classes 'tbl_df', 'tbl' and 'data.frame': 240351 obs. of 6 variables:
..$ chr      : Factor w/ 3 levels "11","17","18": 1 1 1 1 1 1 1 1 ...
..$ start    : int [1:240351] 192951 193001 193051 193101 193151 193201 193251
..$ end      : int [1:240351] 193000 193050 193100 193150 193200 193250 193300
..$ bin      : int [1:240351] 1000125 1000126 1000127 1000128 1000129 1000130
..$ tot_count.test: int [1:240351] 7 8 10 14 15 10 7 5 4 7 ...
..$ tot_count.ref: int [1:240351] 11 9 18 20 14 6 6 1 0 7 ...
$ SNP :Classes 'tbl_df', 'tbl' and 'data.frame': 436147 obs. of 7 variables:
..$ chr      : Factor w/ 3 levels "11","17","18": 1 1 1 1 1 1 1 1 ...
..$ pos      : int [1:436147] 192981 192995 192997 193023 193034 193078 193087
..$ bin      : int [1:436147] 1000125 1000125 1000125 1000126 1000126 1000127
..$ tot_count.test: int [1:436147] 9 8 6 7 7 12 12 11 12 19 ...
..$ alt_count.test: int [1:436147] 1 0 2 0 0 0 1 1 ...
..$ tot_count.ref: int [1:436147] 13 11 13 6 9 20 18 20 21 19 ...
..$ alt_count.ref: int [1:436147] 0 1 4 1 1 1 0 16 9 0 ...
$ meta:List of 2
..$ basic:List of 9
...$ samplename   : chr "A18R.RDX"
...$ source       : chr "WES"
...$ source.file  : List of 3
...$ refBAM       : chr ".../DATA/REDUX/A18R_N_RDX.bam"
...$ testBAM      : chr ".../DATA/REDUX/A18R_T_RDX.bam"
...$ BINpack      : chr ".../RESOURCES/REDUX/SSCREp_RDX_b50.GC.rda"
...$ type         : chr "WES"
...$ manufacturer : chr "illumina"
...$ species      : chr "Homo sapiens"
...$ genome       : chr "hs37d5"
...$ genome.pkg   : chr "BSgenome.Hsapiens.1000genomes.hs37d5"
...$ predicted.gender: chr "NA"
$ WES :List of 8
...$ testBAM.header      : chr "list(targets = c(249250621, 243199373,
364022, 141213")| _truncated_
...$ refBAM.header       : chr "list(targets = c(249250621, 243199373,
364022, 141213")| _truncated_
...$ samtools.Q          : num 20
...$ bin.size            : num 50
...$ BIN.tot.count.test.mean.summary: Named num [1:6] 0 3 11 22.2 27 ...
...$ attr(*, "names")= chr [1:6] "min" "q25" "median" "mean" ...
...$ BIN.tot.count.ref.mean.summary: Named num [1:6] 0 4 15 24 35 ...
...$ attr(*, "names")= chr [1:6] "min" "q25" "median" "mean" ...
...$ SNP.tot.count.test.summary: Named num [1:6] 0 16 32 60.5 59 ...
...$ attr(*, "names")= chr [1:6] "min" "q25" "median" "mean" ...
...$ SNP.tot.count.ref.summary: Named num [1:6] 0 21 40 48.2 65 ...
...$ attr(*, "names")= chr [1:6] "min" "q25" "median" "mean" ...
```



Data Binning : Outputs

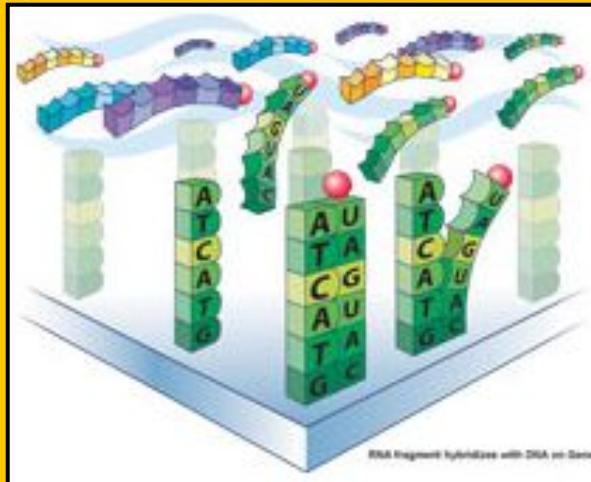


Normalization

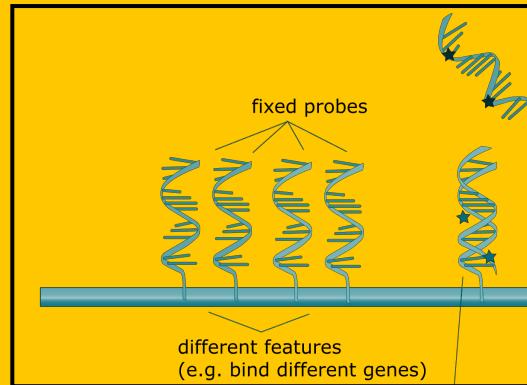
Reducing sources of bias

QUIZ TIME!

GC Bias ? Resolve an old case!



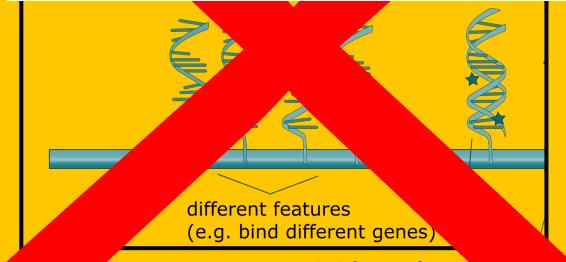
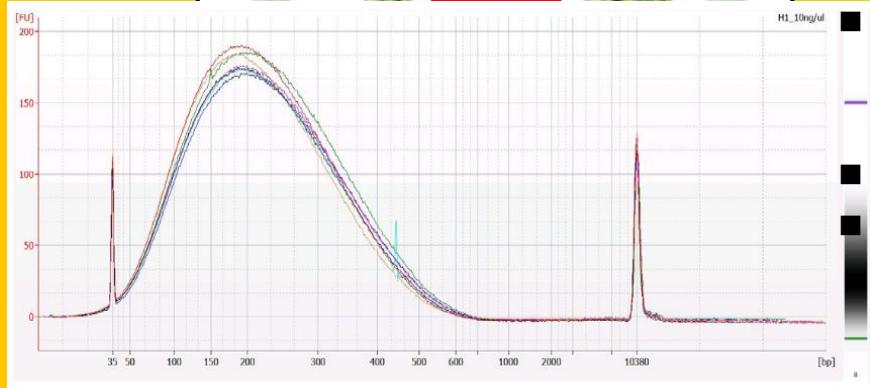
1. Genomic DNA is fragmented (restriction enzymes cocktail, sonication, fine needle shattering, FFPE, ...)
2. ... then labelled by random priming
3. ... then melt and put to hybridization on microarray (*see pictures on the left*)
4. ... then intensities are read (scan)
5. $\log_2(\text{test}/\text{ref})$ is computed
6. Regression of precomputed probes GC% (25 ~ 55 nt) versus log2ratio profile is performed
7. **FAILURE ! (WHY ?)**



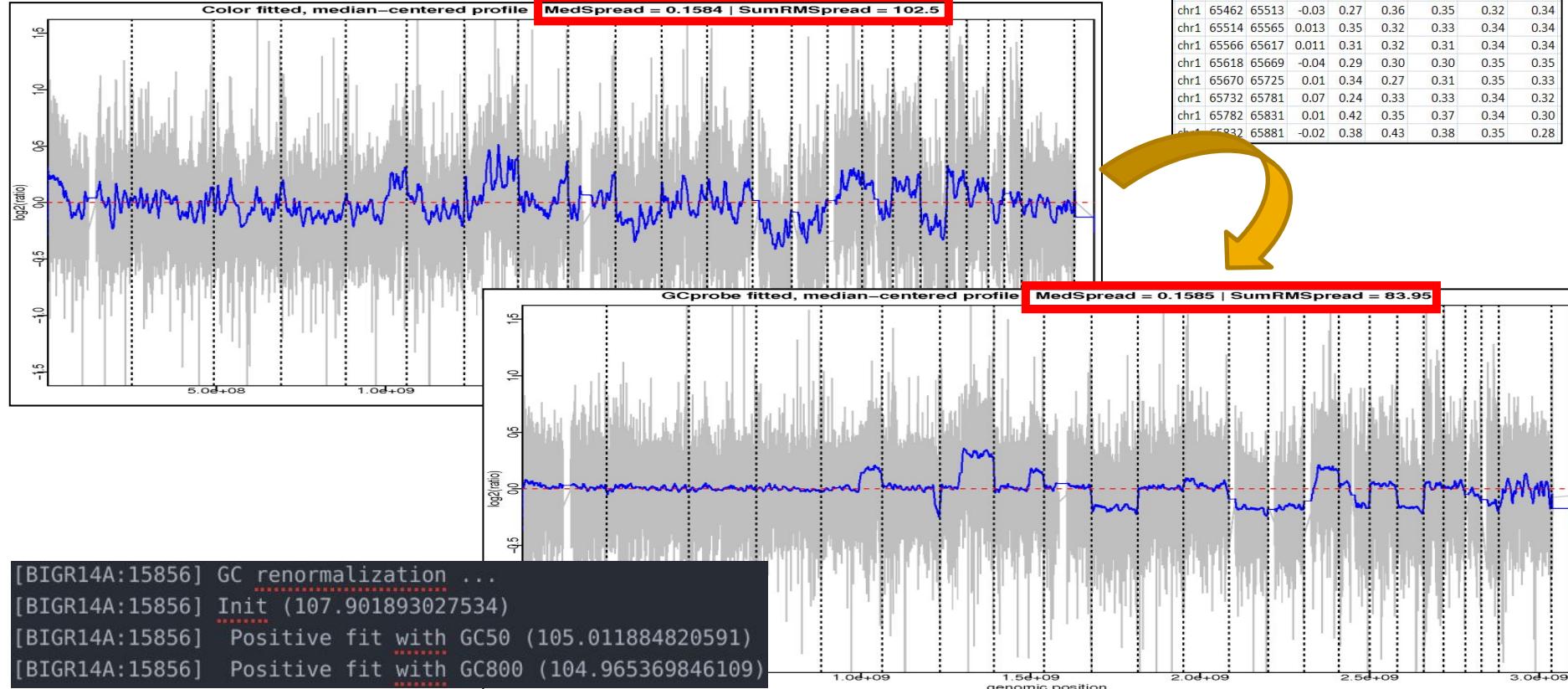
GC Bias ? Resolve an old case!



- Labelled fragments are way longer than probes !
- GC% to use is not the one from the probe, but from the longer generated DNA fragments
- How ? Higher GC% : higher chance of auto-hybridization (shadowed signal : lower intensity / depth)
- Took years (lost...) to figure it out !!
- An epic example of the requirement of precise, detailed communication between lab researchers / techs and bioinformaticians
- *The culprit was colonel [figures], in the [poorly reviewed publications] using weapon [bad communication]*
- *He's still running... (cf Wikipedia plot)*

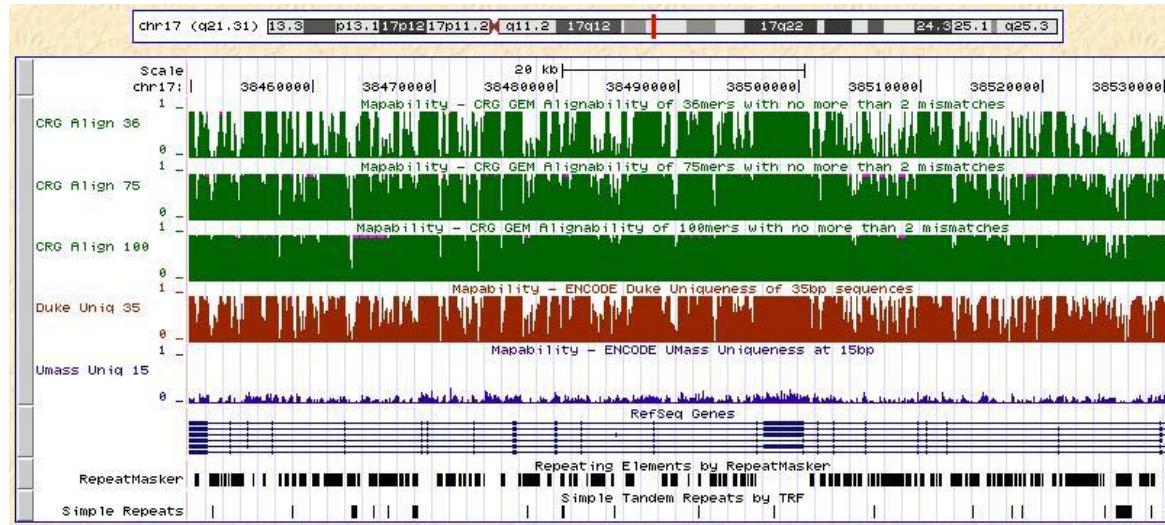


GC% Normalization by Recursive Lowess Regression



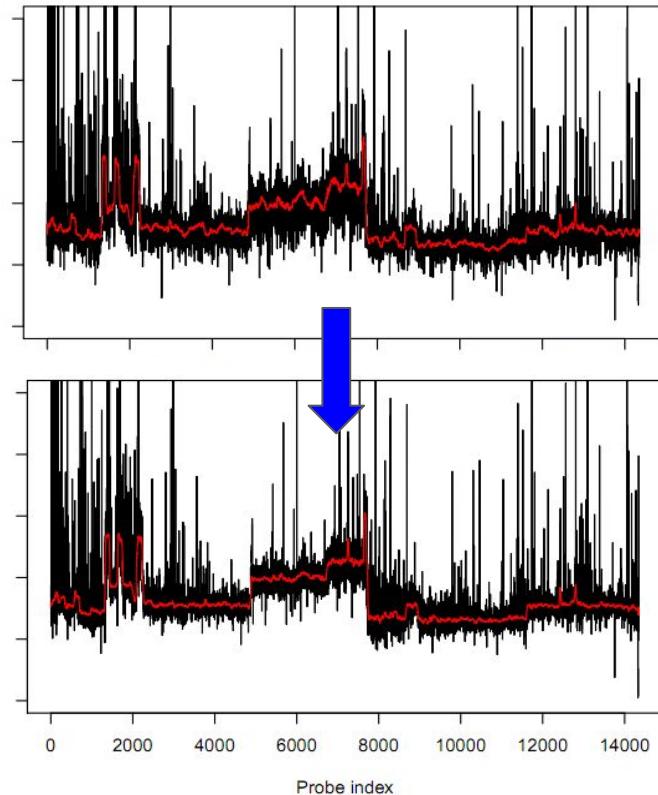
Normalization : “Mappability” Probability

- “Mappability” :
 - Genomic regions with notably **few reads sequenced**
 - First identified by occurrence, now modelled from genome composition (GEM)
 - Has effects on very low coverage BAMs only (<50x)



Normalization : “Wave effect”

- Residual wavelets after all knowledge-base (GC%, mappability scores) normalizations
- Corresponds to *unknown* sources of bias
- Can occur at different levels in multiple samples : can be inferred !
 - R Package *cghseg*
 - 5~10 samples minimum required
 - *huge* CPU/RAM/time resources required

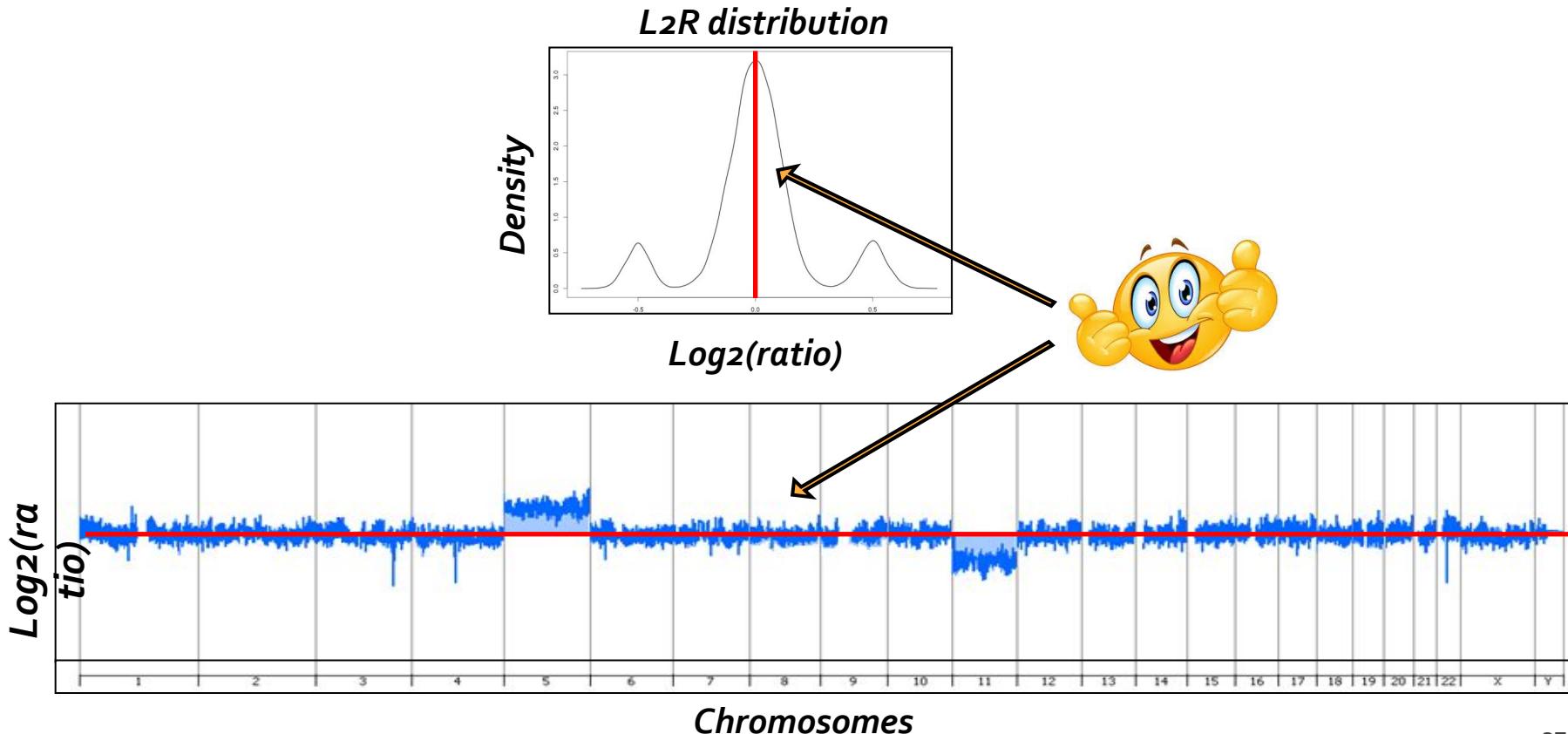


Centralization

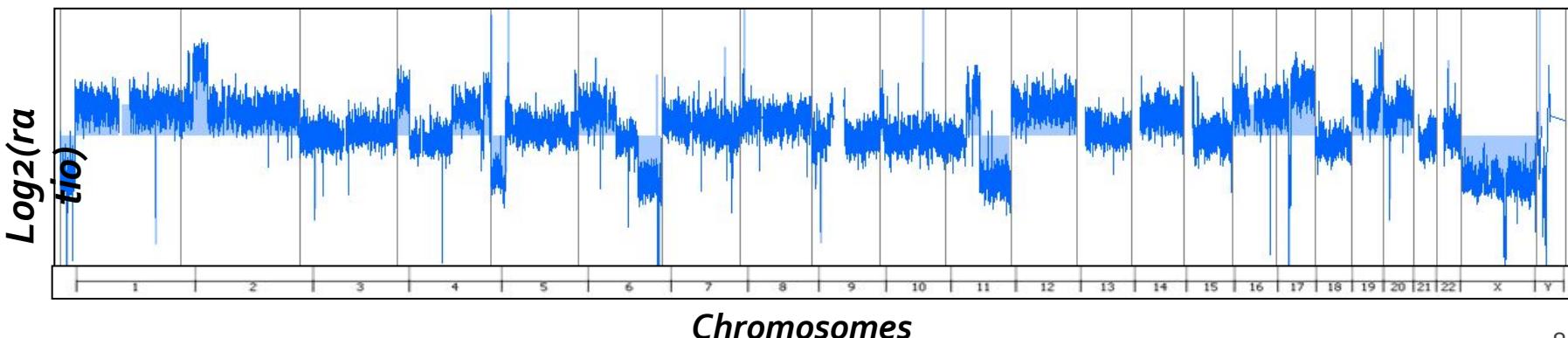
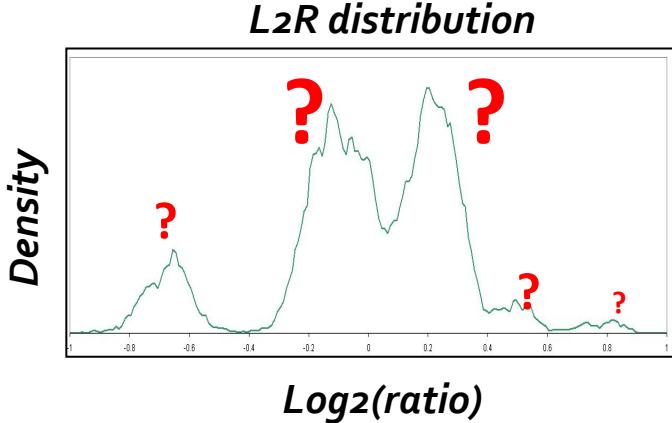
Normality should always be the reference



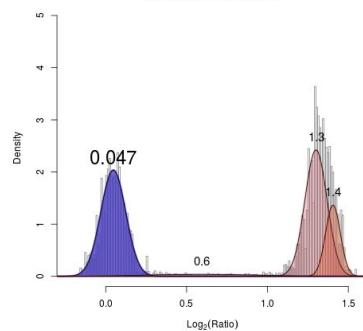
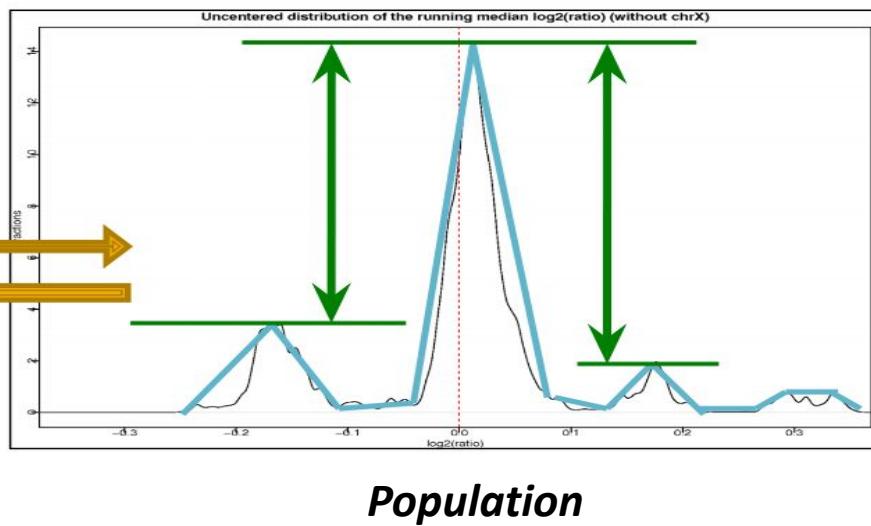
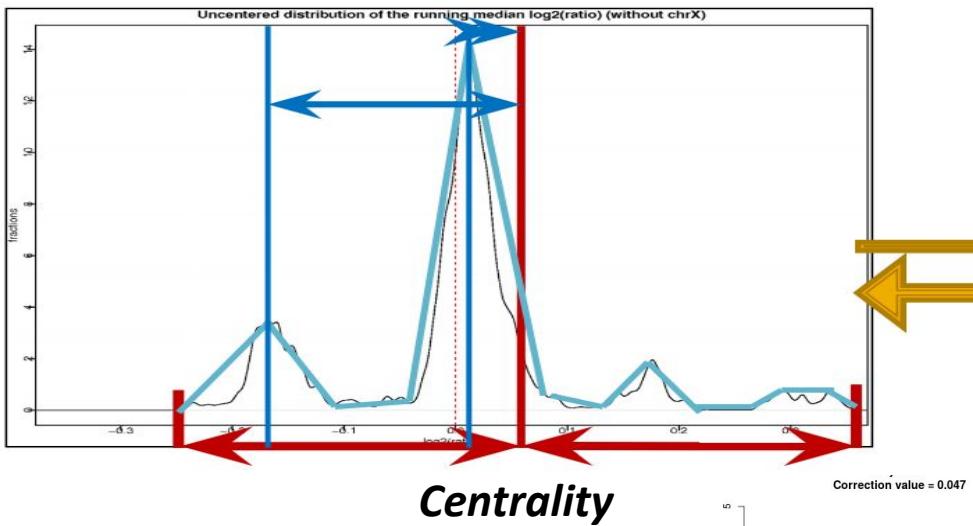
Centralization : An Easy Synthetic Example



Centralization : A Real-life Cancer Example



Centralization : Centrality / Density Trade-off



PRACTICE : Normalization & Centralization

Input data :

- The binned data you generated :
 - o ~ /tp_cna/RESULTS/REDUX/A18R.RDX/A18R.RDX_hs37d5_b50_binned.RDS
- Precomputed **GC%** tracks from the BINpack :
 - o ~ /tp_cna/RESOURCES/REDUX/SSCREp_RDX_b50.GC.rda
- Precomputed “**Wave**” tracks from public datasets :
 - o ~ /tp_cna/RESOURCES/REDUX/SSCREp_RDX_b50.Wave.rda

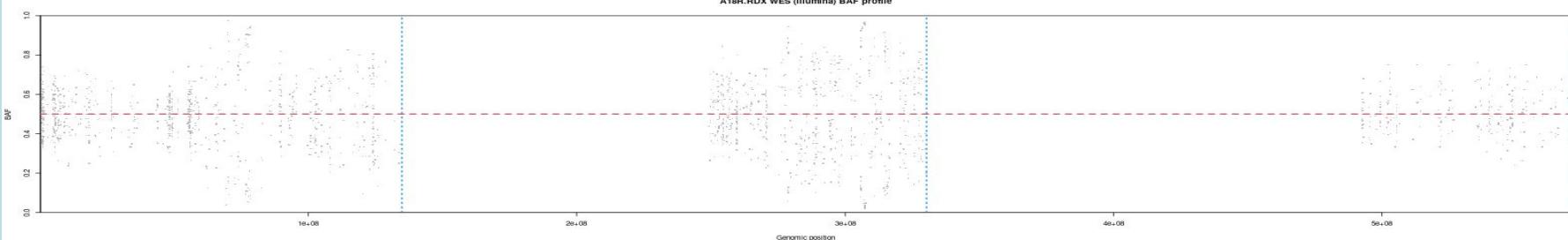
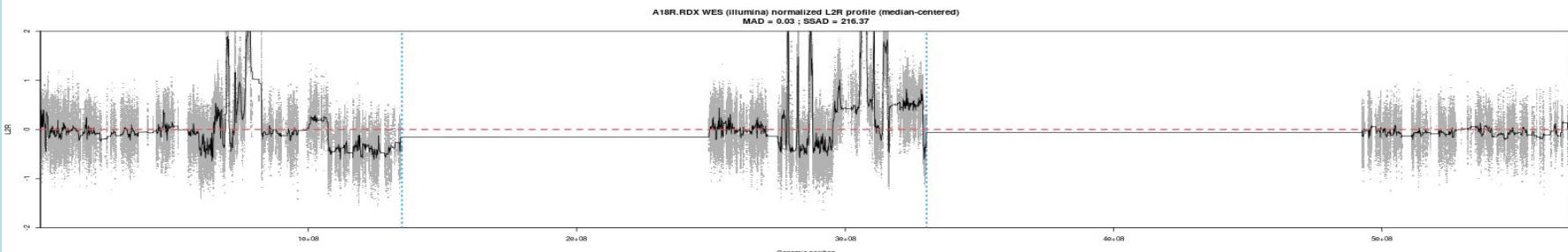
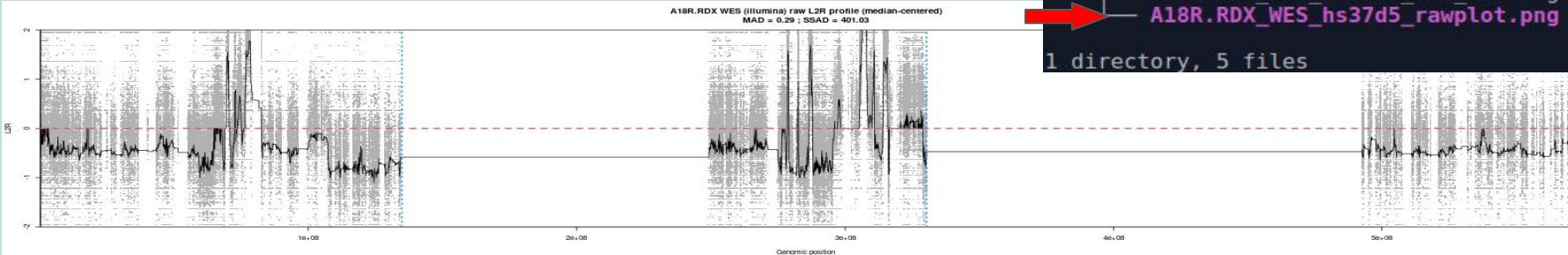
```
WES.Normalize.ff(BIN.RDS.file = "A18R.RDX/A18R.RDX_hs37d5_b50_binned.RDS",
BINpack = ".../.../RESOURCES/REDUX/SSCREp_RDX_b50.GC.rda",
wave.rda = ".../.../RESOURCES/REDUX/SSCREp_RDX_b50.Wave.rda",
wave.renorm = TRUE)

system("tree -sh")
```

PRACTICE : Outputs

```
A18R.RDX
├── A18R.RDX_hs37d5_b50_binned.RDS
├── A18R.RDX_hs37d5_b50_processed.RDS
├── A18R.RDX_WES_hs37d5_b50_coverage.png
└── A18R.RDX_WES_hs37d5_b50_coverage.txt
    └── A18R.RDX_WES_hs37d5_rawplot.png
```

1 directory, 5 files

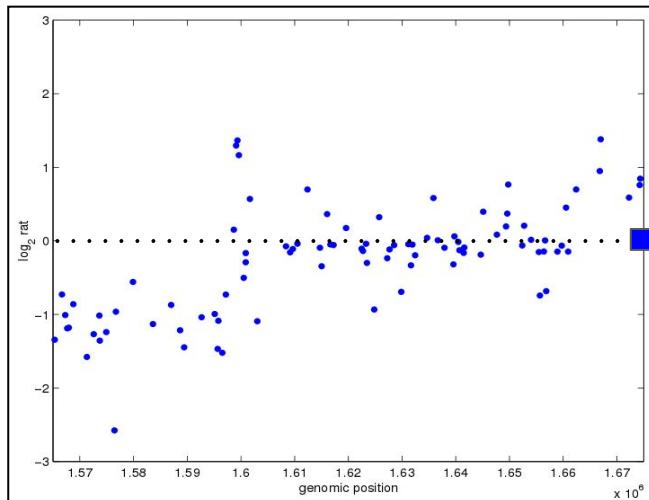


Segmentation

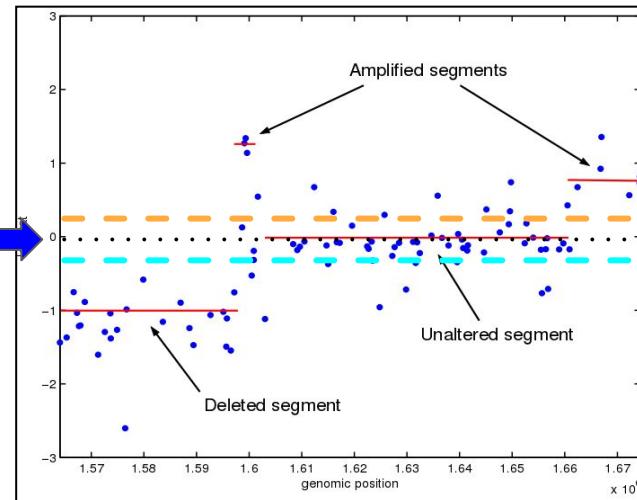
From numerous, noisy, and punctual local measures to limited,
denoised and larger genomic intervals

Segmentation (and calling)

- **Segmentation** = longitudinal data reduction
 - From numerous, punctual, noisy, technical **measures** (bins) ...
 - ... to limited, continuous **intervals** with a single value and closer to the biological reality
- **Calling** = defining the limits of normality



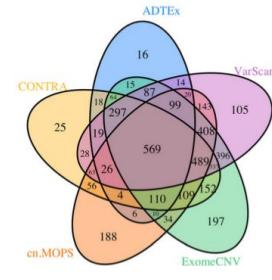
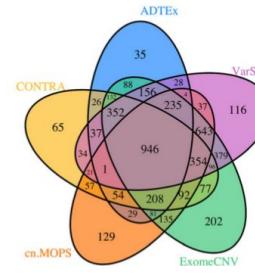
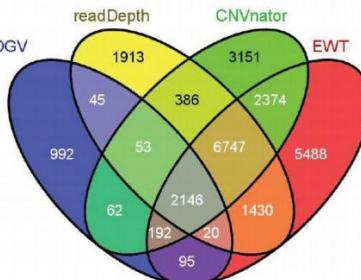
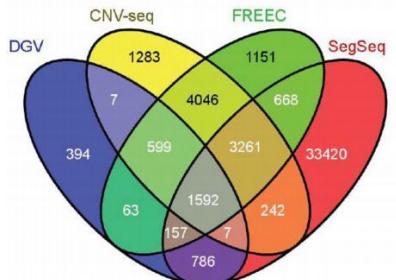
$M = 100$



$S = 4$

Segmentation : Numerous Tools Available ...

Method	Reference	Language	Control required?	Input format	GC correction	single-end/pair-end	Methodology characteristics
CNV-seq	[15]	R, perl	Yes	hits	No	single-end	statistical testing
FREEC	[21]	C	Optional	SAM,BAM,bed,etc.	Optional	both	LASSO regression
readDepth	[22]	R	No	bed	Yes	both	CBS, LOESS regression
CNVnator	[23]	C	No	BAM	Yes	both	mean shift algorithm
SegSeq	[14]	Matlab	Yes	bed	No	single-end	statistical testing,CBS
EWT (RDXplorer)	[11]	R, python	No	BAM	Yes	single-end	statistical testing
cnD	[16]	D	No	SAM,BAM	No	both	HMM, Viterbi algorithm
CNVer	[17]	C	No	BAM	Yes	pair-end	maximum-likelihood, graphic flow
CopySeq	[18]	Java	No	BAM	Yes	pair-end	MAP estimator
rSW-seq	[19]	NA	Yes	NA	Yes	single-end	Smith-Waterman algorithm
CNAseg	[20]	R	Yes	BAM	No	pair-end	wavelet transform and HMM
CNAnorm	[24]	R	Yes	SAM,BAM	Yes	both	linear regression or CBS
cn.MOPS	[26]	R, C++	multiple samples	BAM or data matrix	No	both	mixture of Poissons, MAP, EM, CBS
JointSLM	[27]	R, Fortran	multiple samples	data matrix	Yes	both	HMM, ML estimator, Viterbi algorithm



Duan et al.
Plos One
2013

... and a very poor consensus

Zare et al.
BMC Bioinformatics
2017

PRACTICE : Segmentation and calling [ASCAT]

Input data :

- The normalized data RDS you generated :
 - ~ /tp_cna/RESULTS/REDUX/A18R.RDX/A18R.RDX_hs37d5_b50_processed.RDS

```
Segment.ff(RDS.file = "A18R.RDX/A18R.RDX_hs37d5_b50_processed.RDS",
segmenter = "ASCAT", smooth.k = 5, nrf = 10, SER.pen = 5)
```

PRACTICE : Outputs [ASCAT]

```
└── A18R.RDX
    ├── A18R.RDX_hs37d5_b50_binned.RDS
    ├── A18R.RDX_hs37d5_b50_processed.RDS
    ├── A18R.RDX_WES_hs37d5_b50_coverage.png
    ├── A18R.RDX_WES_hs37d5_b50_coverage.txt
    ├── A18R.RDX_WES_hs37d5_rawplot.png
    └── ASCAT
        └── L2R
            ├── A18R.RDX.Cut.cbs
            ├── A18R.RDX.NoCut.cbs
            ├── A18R.RDX.Rorschach.png
            ├── A18R.RDX.SEG.ASCAT.png
            ├── A18R.RDX.SEG.ASCAT.RDS
            └── A18R.RDX.SegmentedBAF.txt

3 directories, 11 files
```

QUIZ
TIME!

Name these event types !

1

2

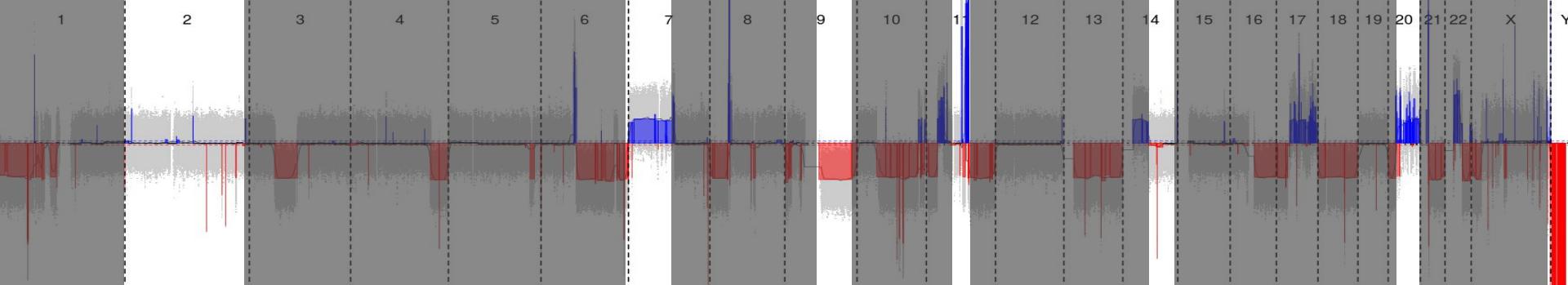
4

3

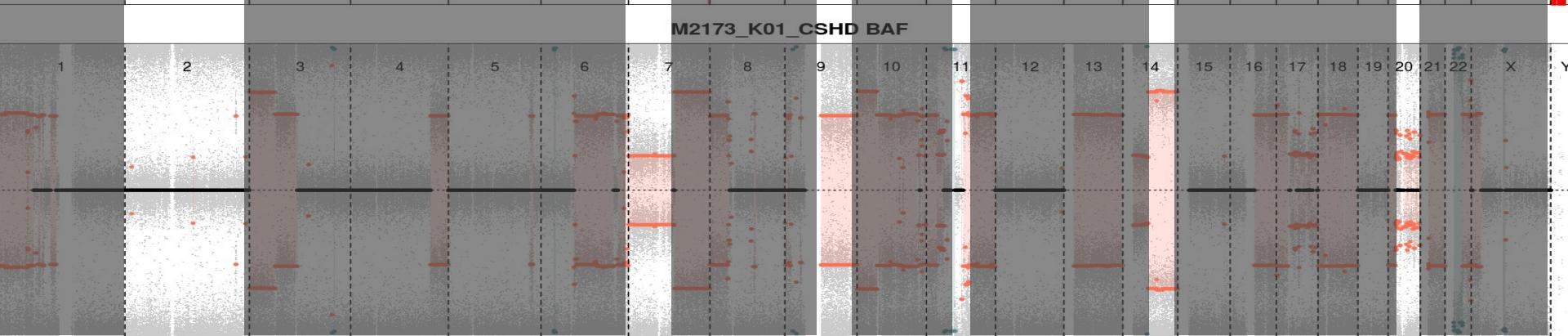
7

6

5

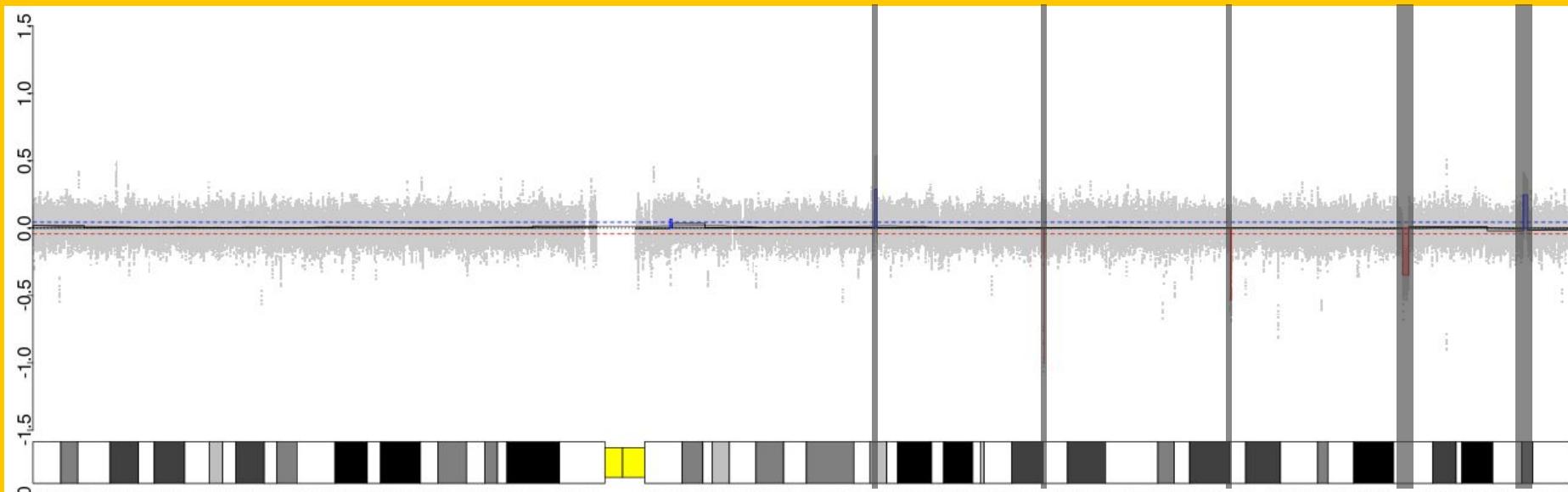


M2173_K01_CSHD BAF



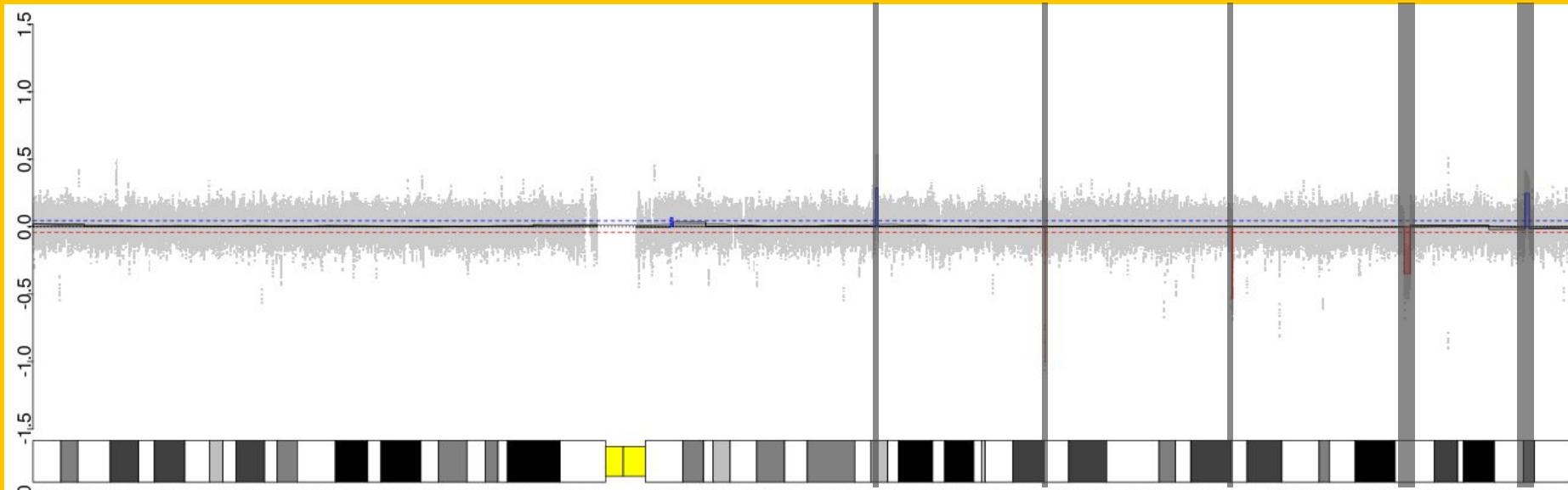
Name these putative event cases !

1. chr2



Name these putative event cases !

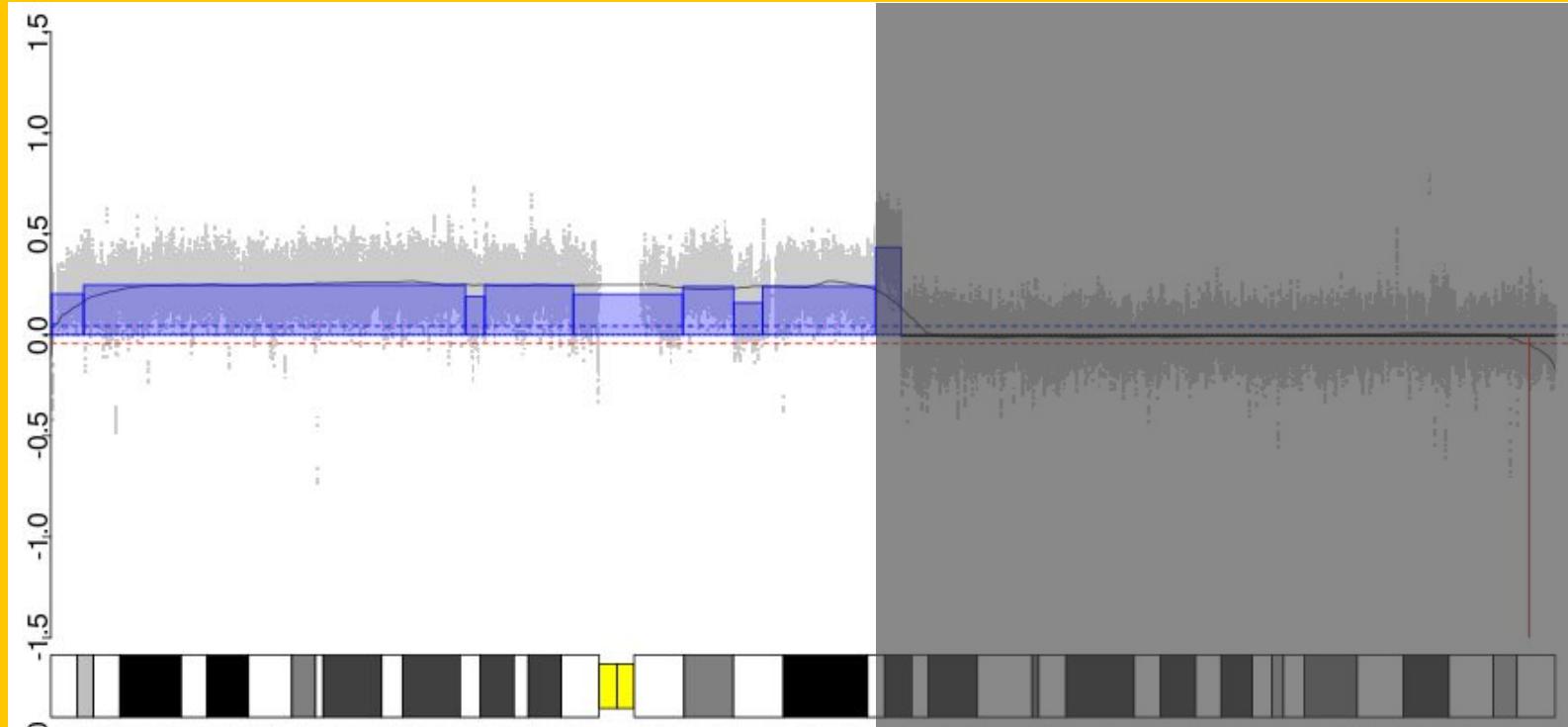
1. chr2 : NORMALITY



QUIZ
TIME!

Name these putative event cases !

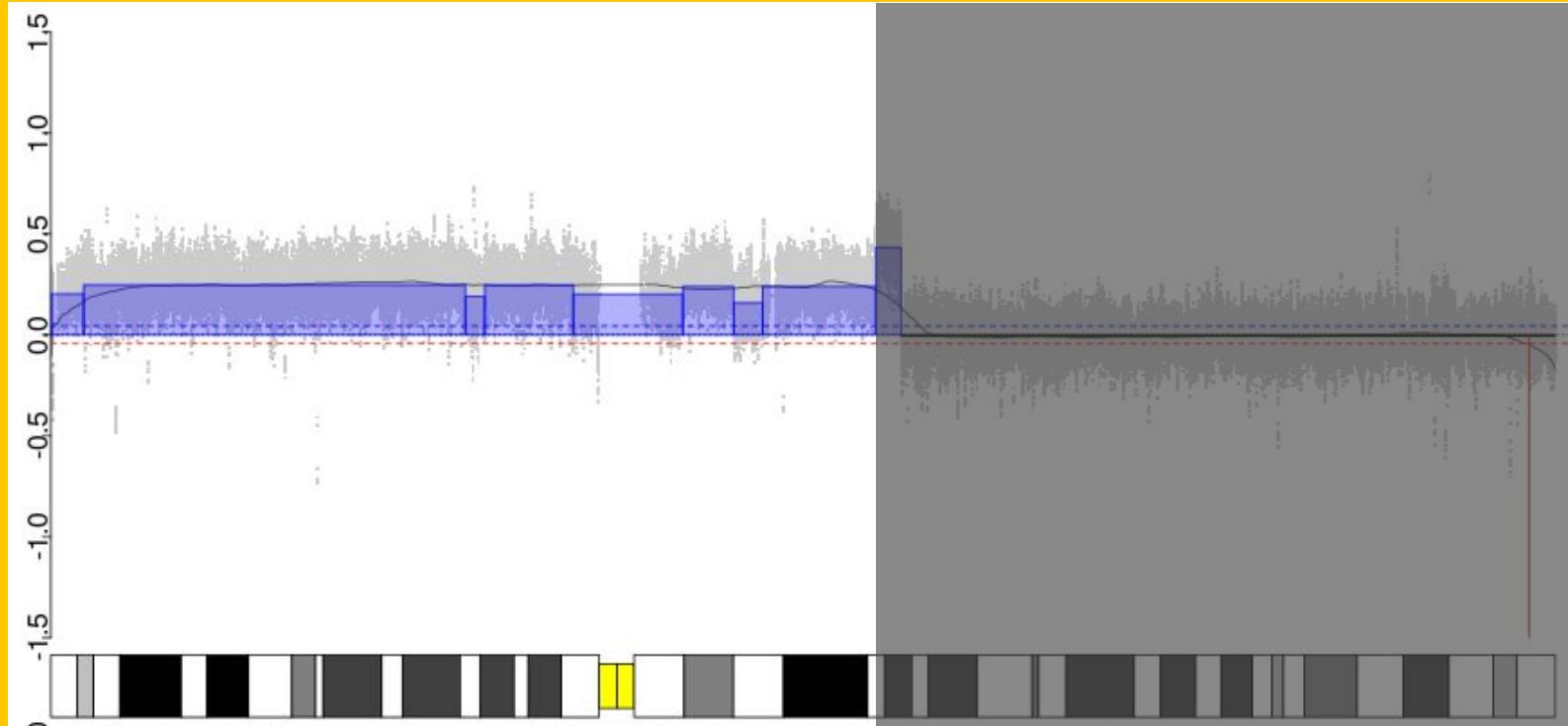
2. chr7



QUIZ
TIME!

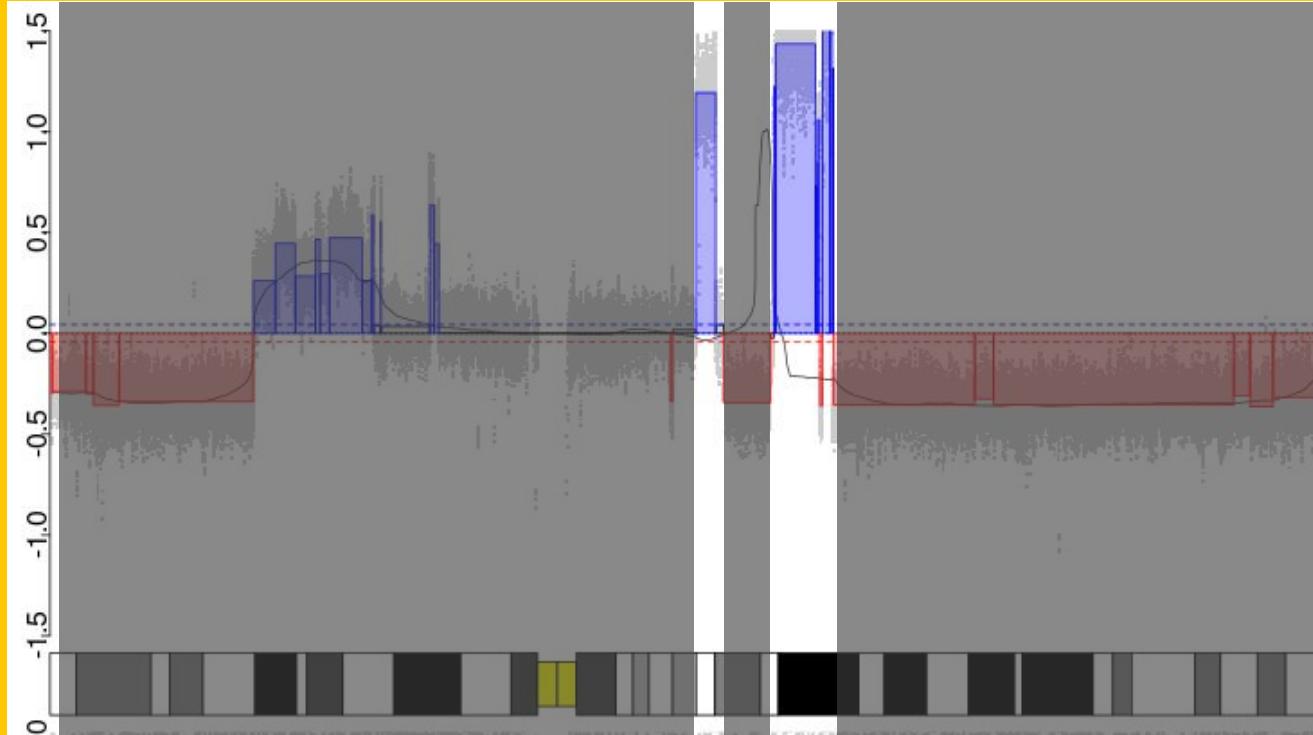
Name these putative event cases !

2. chr7 : GAIN



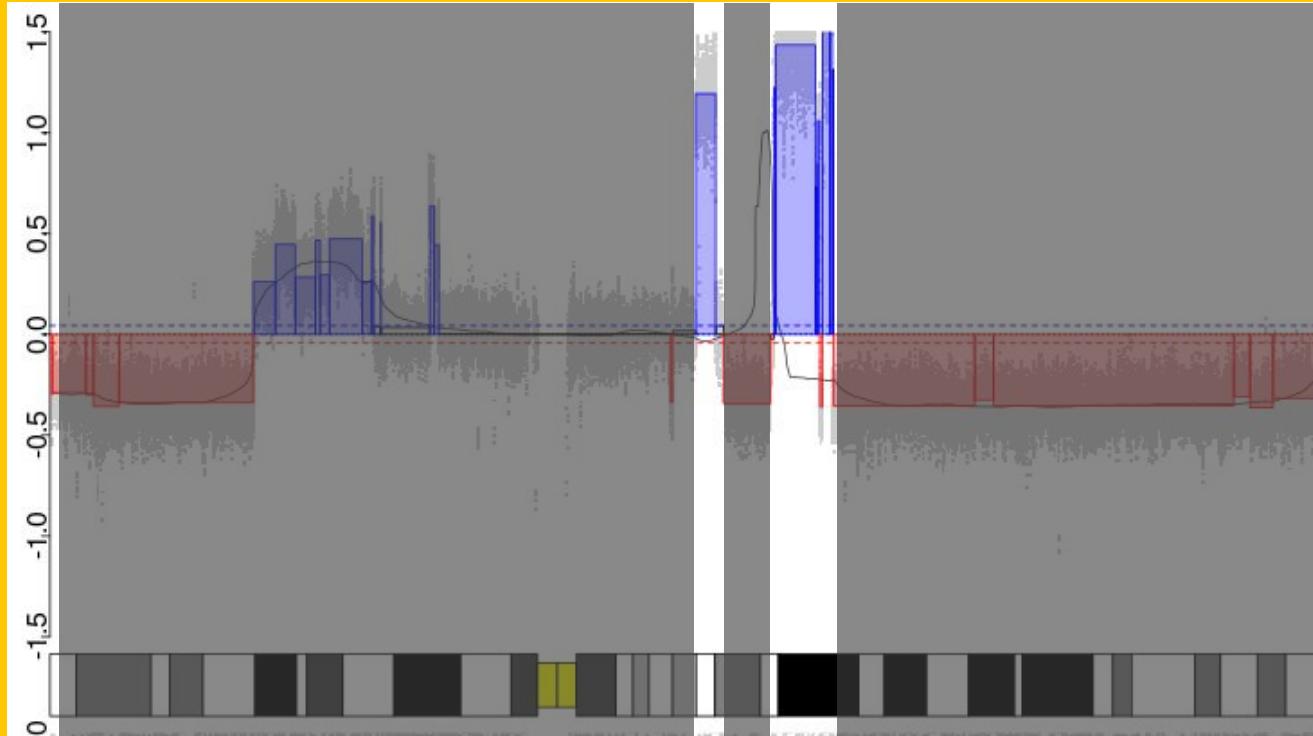
Name these putative event cases !

3. chr11



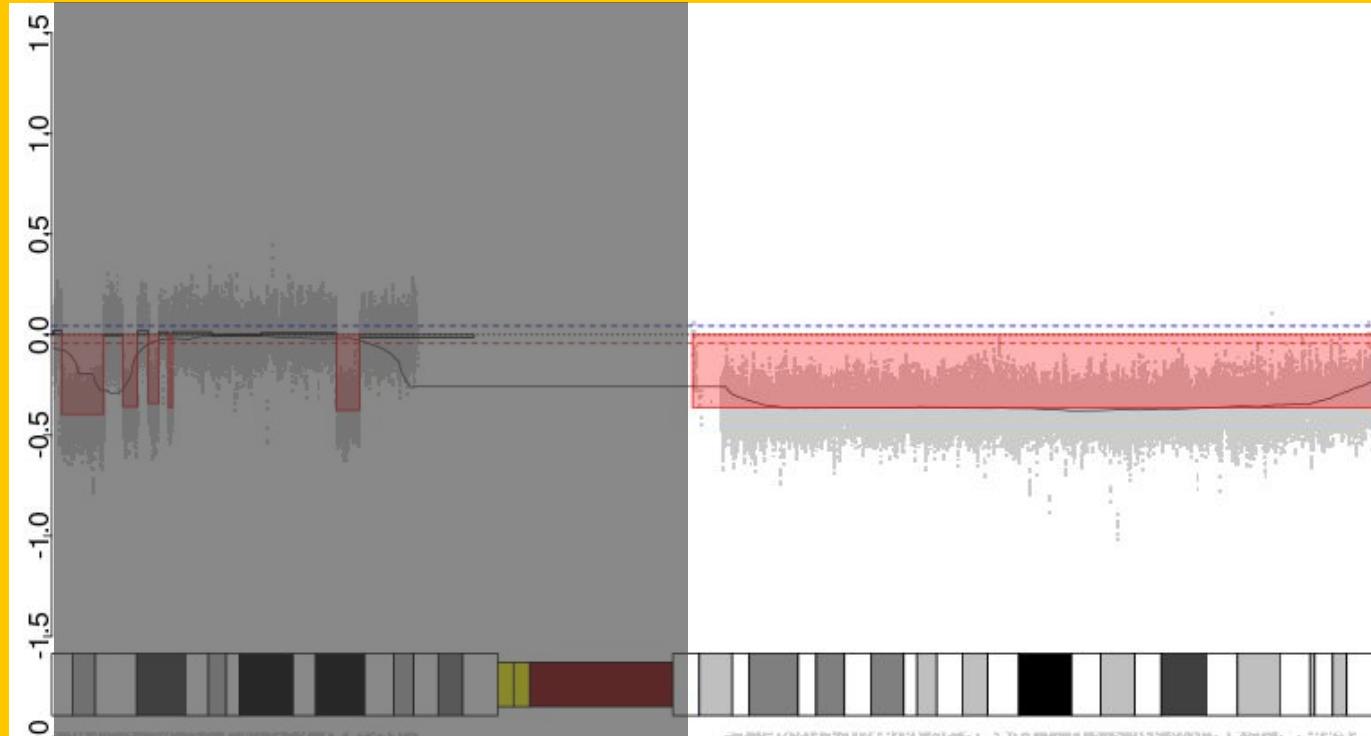
Name these putative event cases !

3. chr11 : AMPLIFICATION



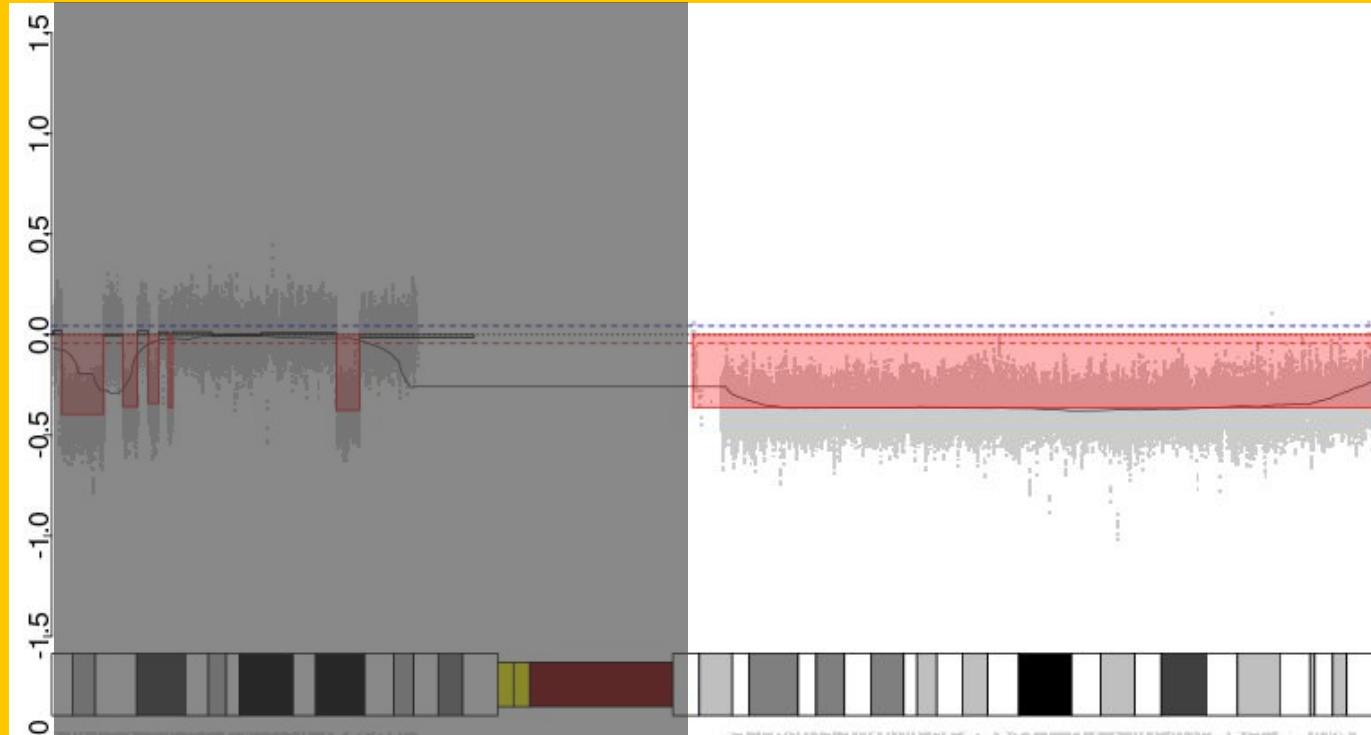
Name these putative event cases !

4. chr9



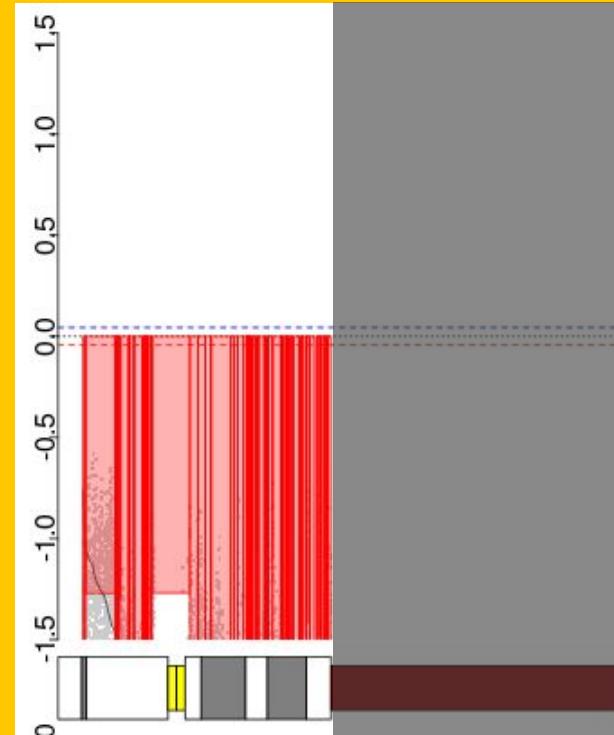
Name these putative event cases !

4. chr9 : LOSS



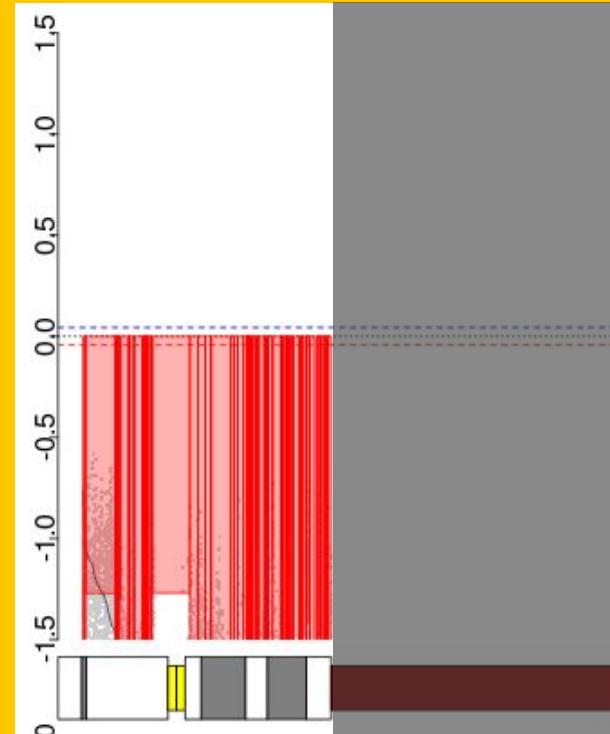
Name these putative event cases !

5. chrY



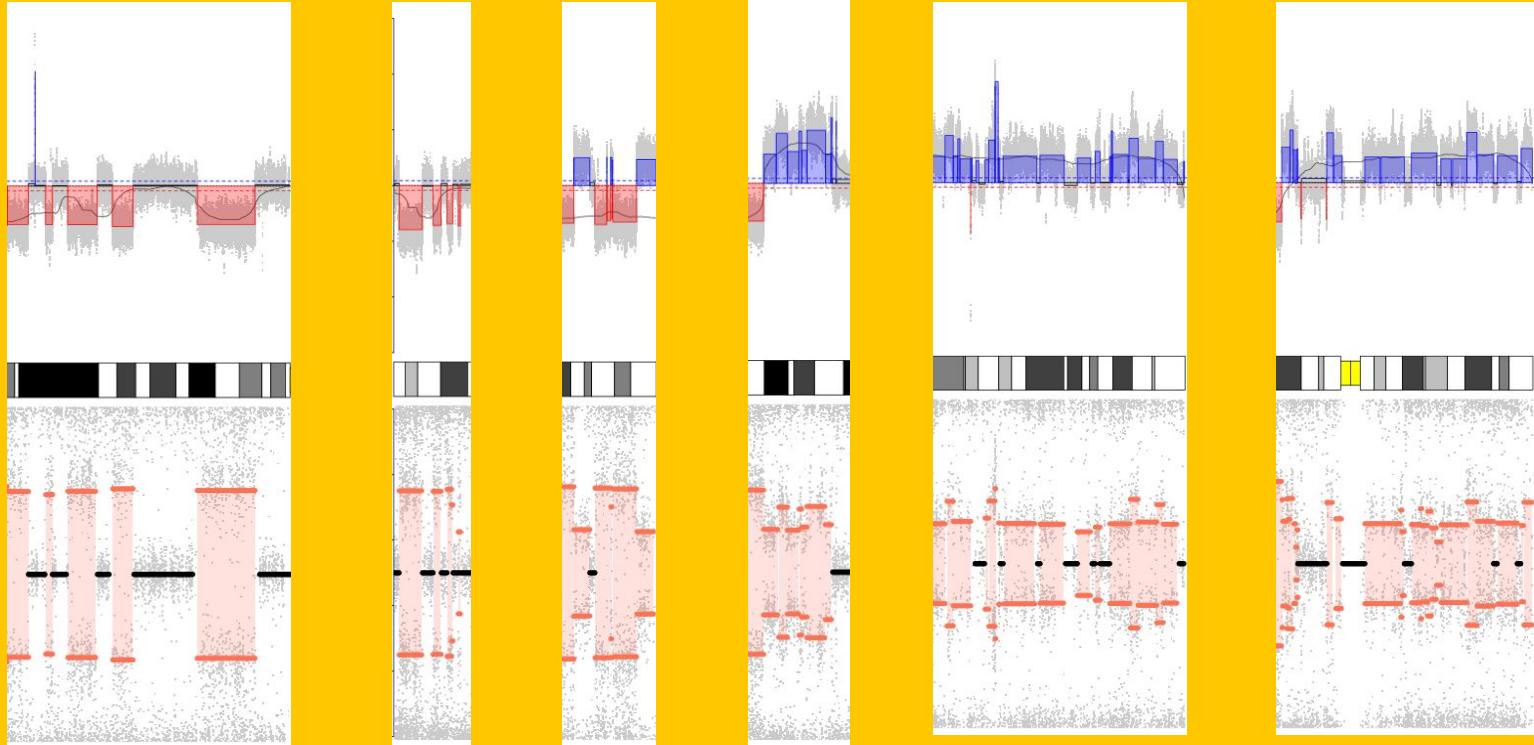
Name these putative event cases !

5. chrY : DELETION



Name these putative event cases !

5. chr1/9/10/11/17/20

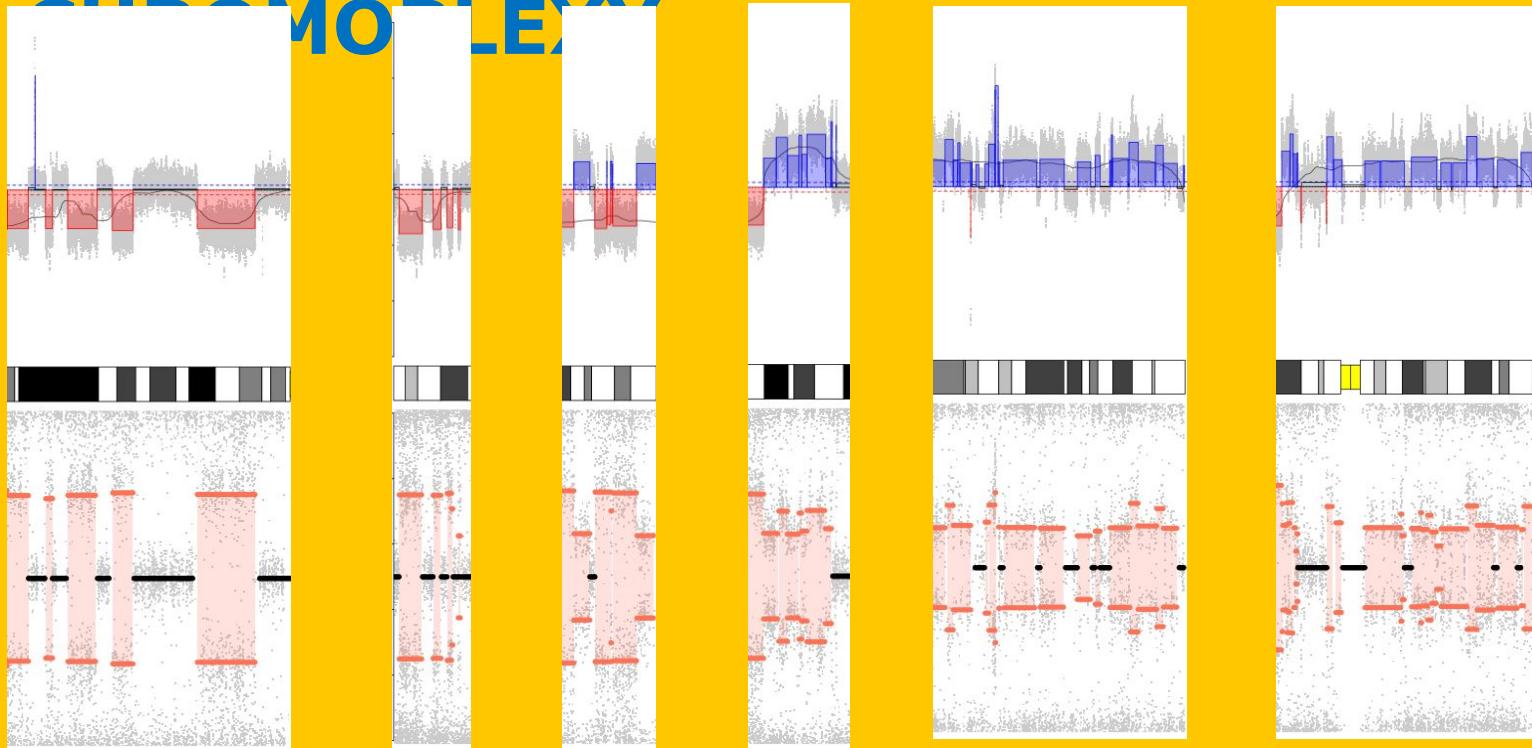


QUIZ
TIME!

Name these putative event cases !

5. chr1/9/10/11/17/20 :

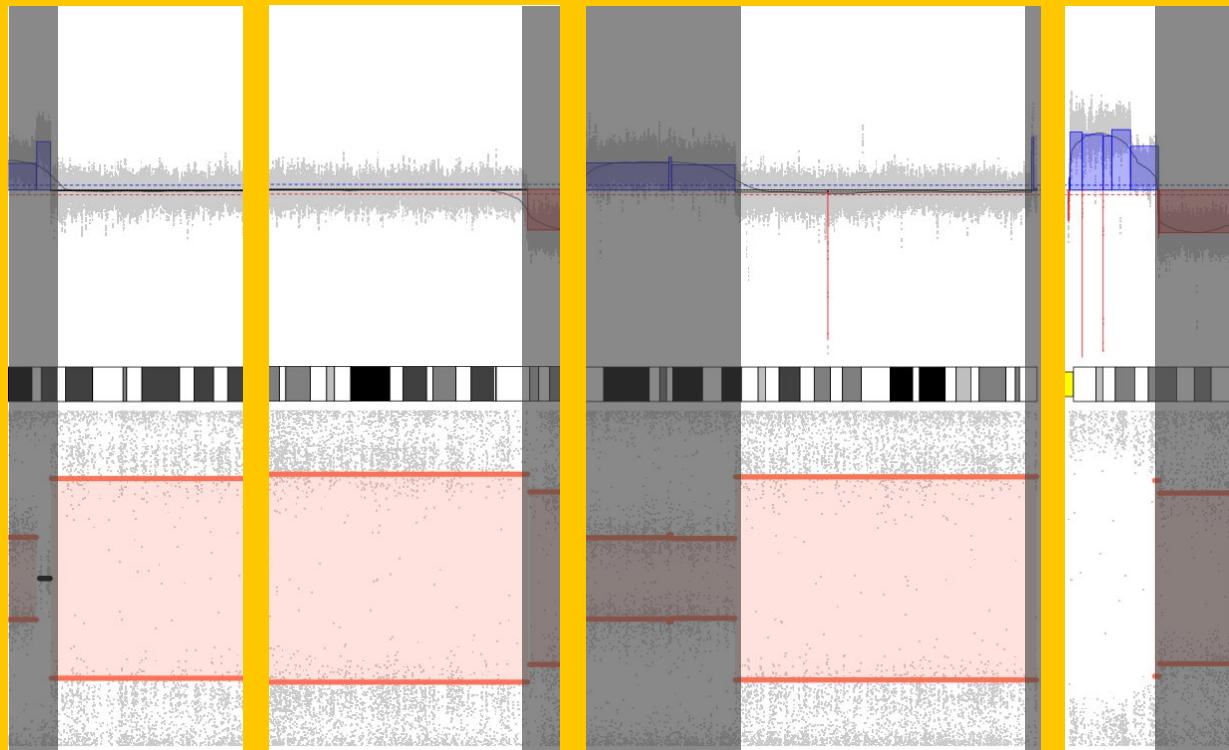
CHROMOPLEX



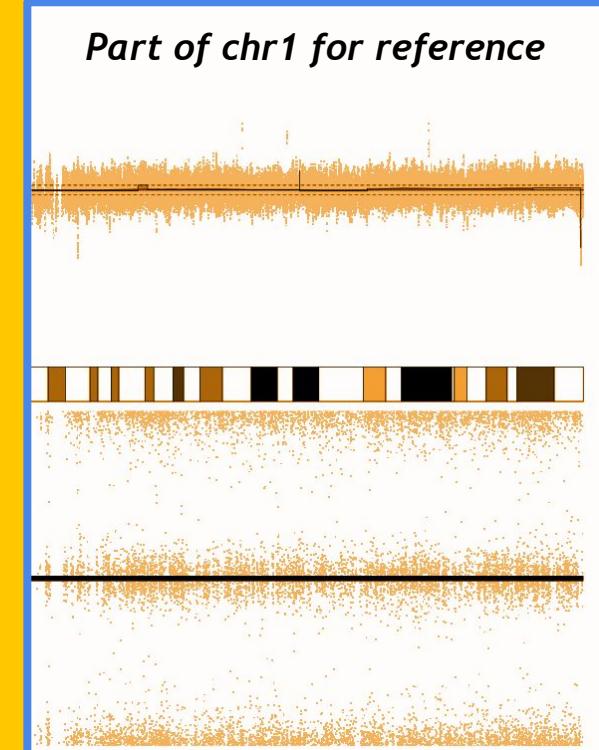
QUIZ
TIME!

Name these putative event cases !

7. chr3/7/14/22

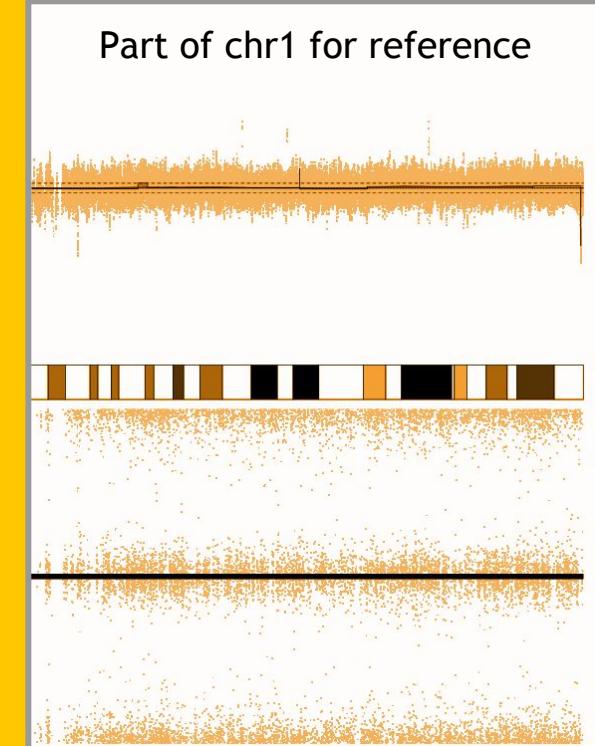
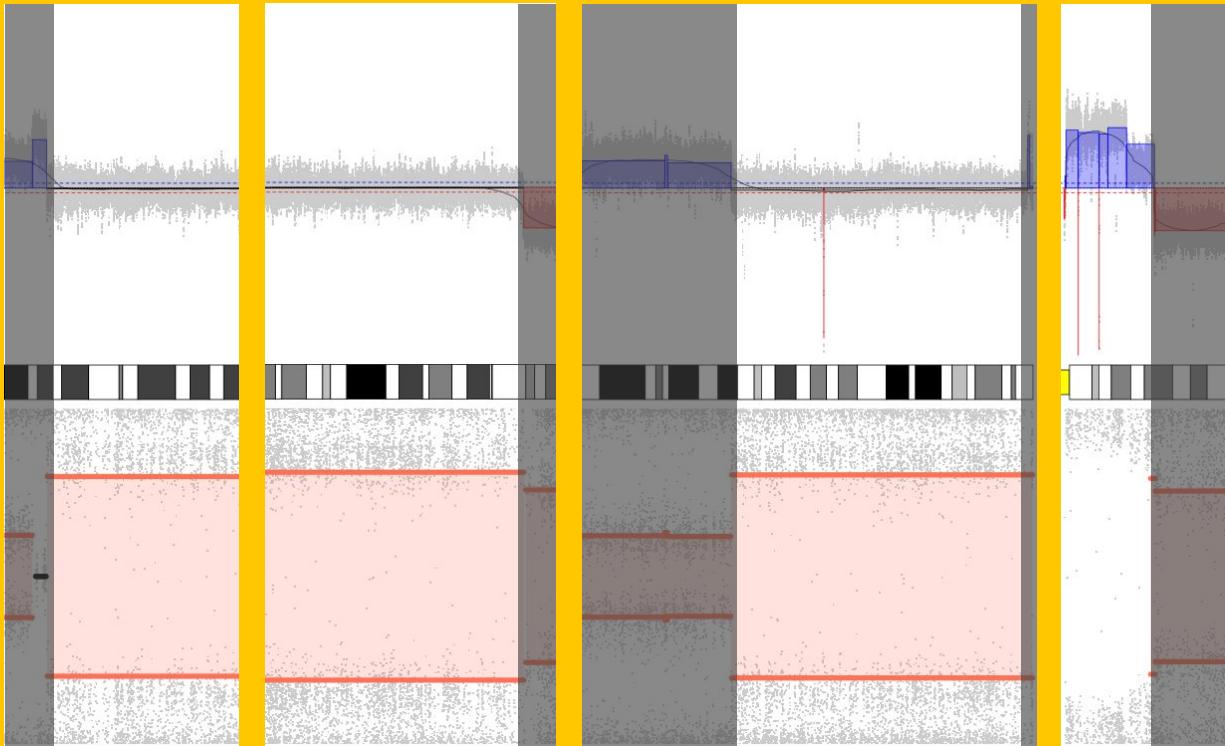


Part of chr1 for reference



Name these putative event cases !

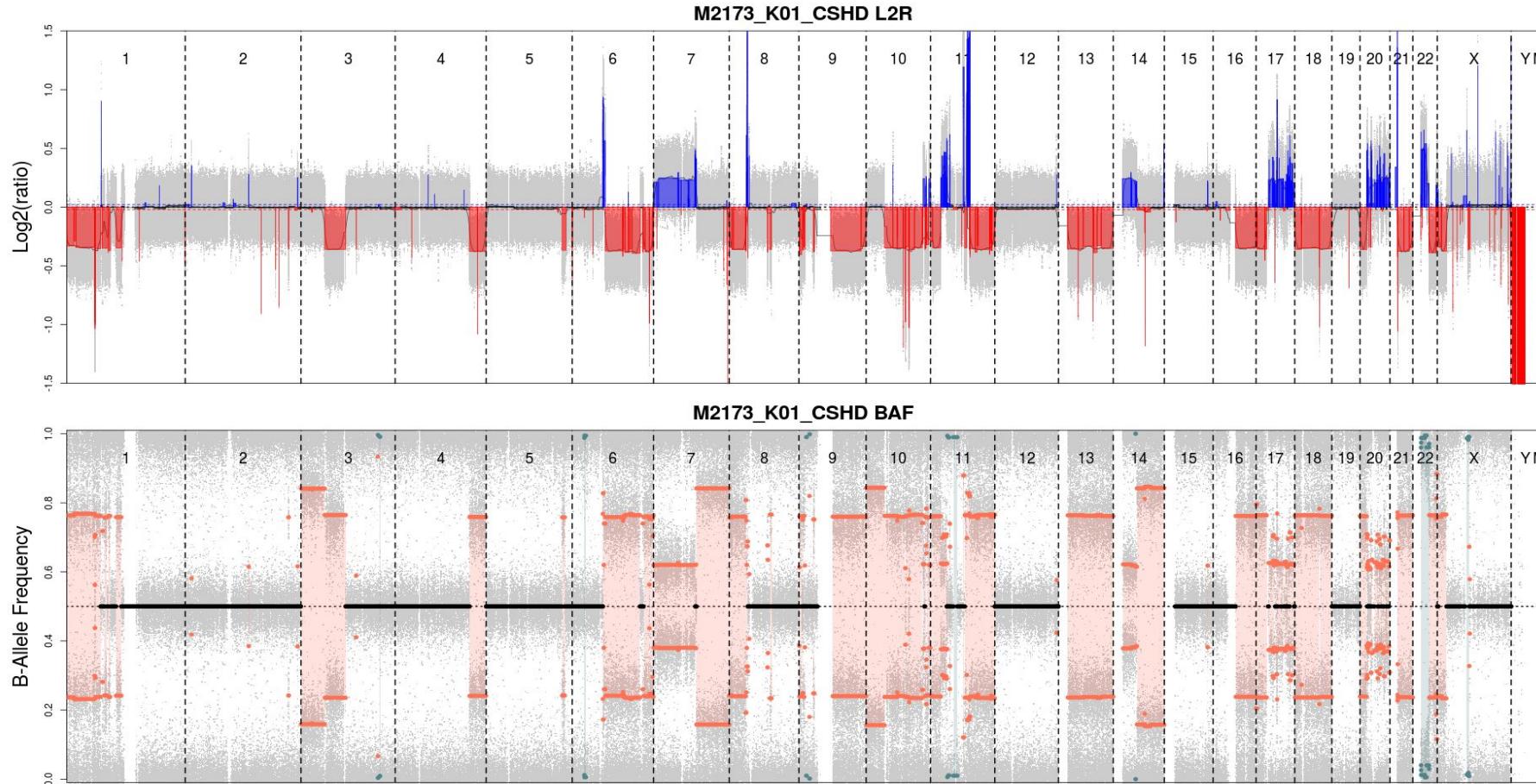
7. chr3/7/14/22 : UNISOMY



Total Copy Number

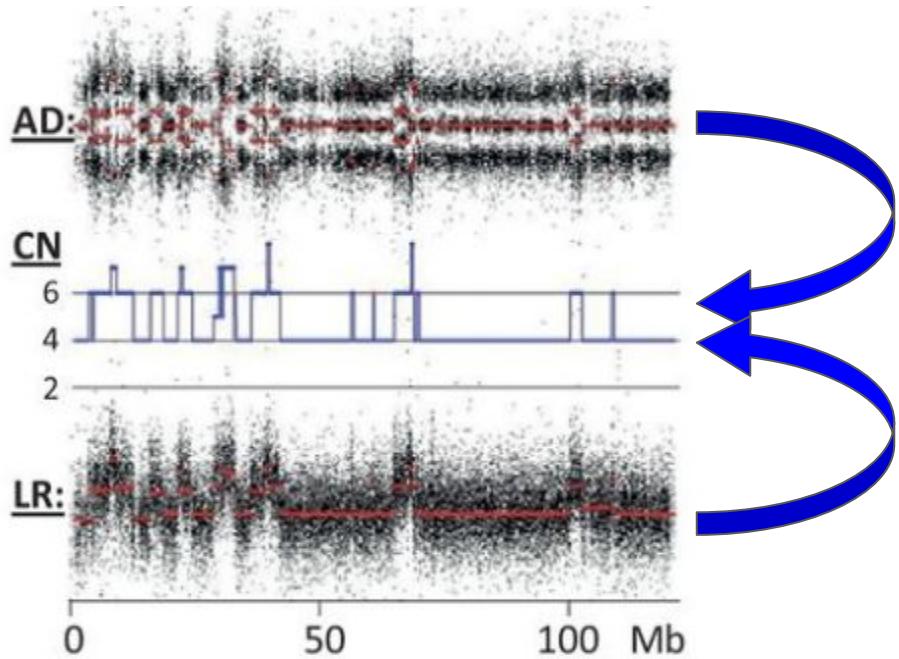
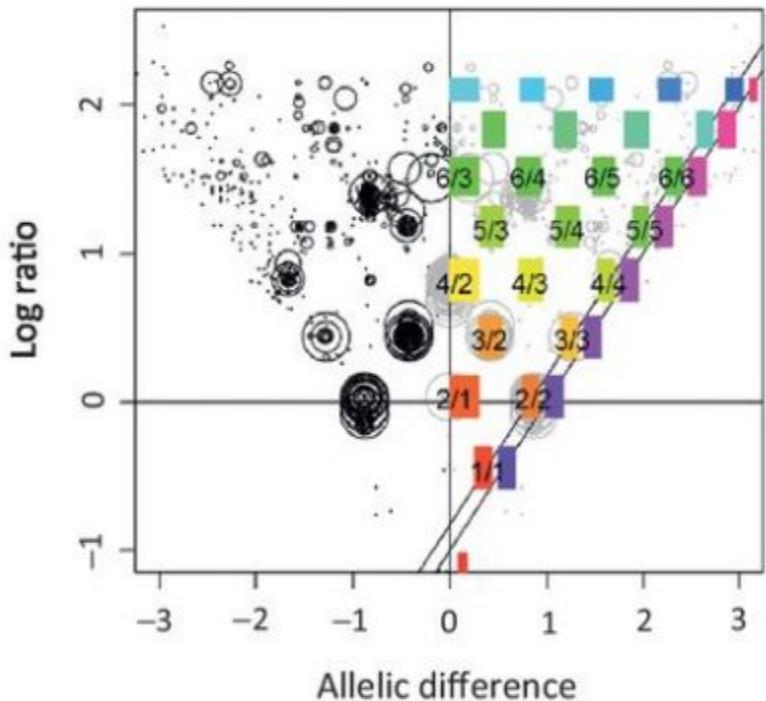
Beyond the L2R Profile

Using Both L2R and BAF



Modelization of Absolute Copy Number

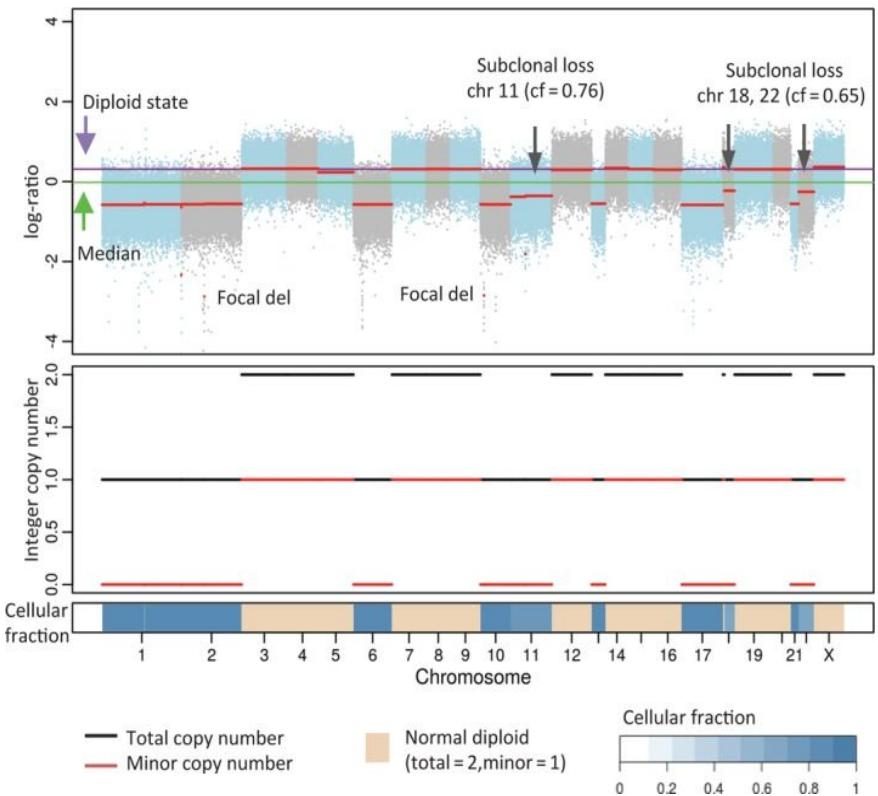
A mathematical combination of L2R and BAF/AD signals



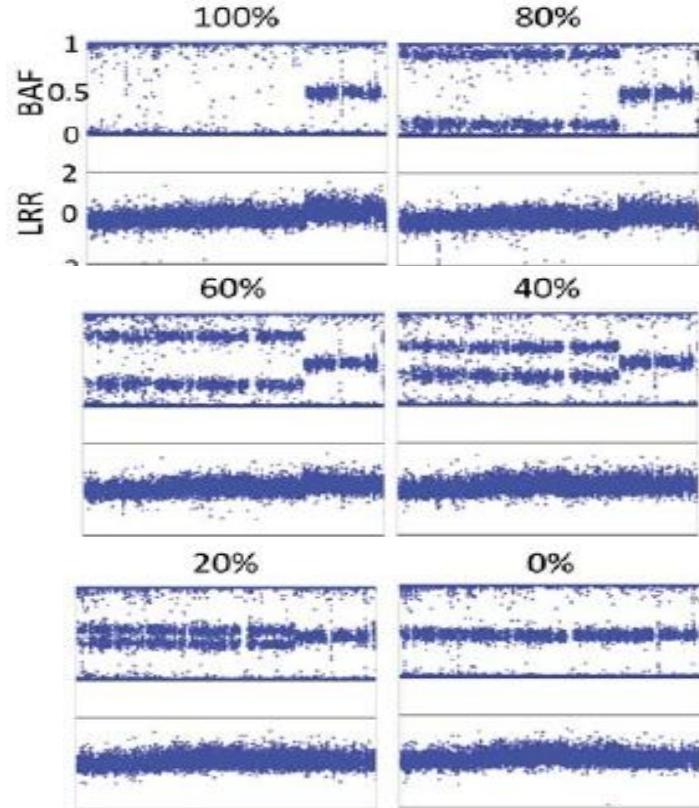
Popova et al., Cancer Res, 2016

Estimation of Global Ploidy and Cellularity

Ploidy : Width-ponderated TCN



Cellularity : dilution of L2R/BAF signals



PRACTICE : Absolute Copy Number [ASCAT]

Input data :

- The **segmented** data you generated :
 - ~/tp_cna/RESULTS/REDUX/A18R.RDX/**ASCAT/L2R/A18R.RDX.SEG.ASCAT.RDS**

```
ASCN.ff(RDS.file = "A18R.RDX/ASCAT/L2R/A18R.RDX.SEG.ASCAT.RDS", nsubthread  
= 3)  
  
system("tree -sh")
```

PRACTICE : Outputs [ASCAT]

```
└── A18R.RDX
    ├── A18R.RDX_hs37d5_b50_binned.RDS
    ├── A18R.RDX_hs37d5_b50_processed.RDS
    ├── A18R.RDX_WES_hs37d5_b50_coverage.png
    ├── A18R.RDX_WES_hs37d5_b50_coverage.txt
    ├── A18R.RDX_WES_hs37d5_rawplot.png
    └── ASCAT
        └── ASCN
            ├── A18R.RDX.gammaEval.png
            ├── A18R.RDX.gammaEval.txt
            ├── gamma0.55
            ├── gamma0.60
            ├── gamma0.65
            ├── gamma0.70
            ├── gamma0.75
            ├── gamma0.80
            ├── gamma0.85
            ├── gamma0.90
            └── gamma0.95
        └── L2R
            ├── A18R.RDX.Cut.cbs
            ├── A18R.RDX.NoCut.cbs
            ├── A18R.RDX.Rorschach.png
            ├── A18R.RDX.SEG.ASCAT.png
            ├── A18R.RDX.SEG.ASCAT.RDS
            └── A18R.RDX.SegmentedBAF.txt

```

13 directories, 13 files

PRACTICE : Annotation and HTML Report [ASCAT]

Input data :

- The **segmented** data you generated :
 - ~/tp_cna/RESULTS/REDUX/A18R.RDX/ASCAT/L2R/A18R.RDX.SEG.ASCAT.RDS

```
Annotate.ff(RDS.file = "A18R.RDX/ASCAT/L2R/A18R.RDX.SEG.ASCAT.RDS")
```

```
system("tree -sh")
```

PRACTICE : Outputs [ASCAT]

```
A18R.RDX
└── A18R.RDX_hs37d5_b50_binned.RDS
└── A18R.RDX_hs37d5_b50_processed.RDS
└── A18R.RDX_WES_hs37d5_b50_coverage.png
└── A18R.RDX_WES_hs37d5_b50_coverage.txt
└── A18R.RDX_WES_hs37d5_rawplot.png
└── ASCAT
    └── L2R
        ├── A18R.RDX.BAF.png
        ├── A18R.RDX.Cut.acbs
        ├── A18R.RDX.Cut.cbs
        └── A18R.RDX.Instab.txt
        ├── A18R.RDX.L2R.G.png
        ├── A18R.RDX.L2R.K.png
        ├── A18R.RDX.NoCut.cbs
        └── A18R.RDX.REPORT.html
        ├── A18R.RDX.Rorschach.png
        ├── A18R.RDX.SEG.ASCAT.png
        ├── A18R.RDX.SEG.ASCAT.RDS
        ├── A18R.RDX.SegmentedBAF.txt
        └── A18R.RDX.TargetGenes.txt
        └── A18R.RDX.TruncatedGenes.txt
        └── chromosomes
            ├── 11.png
            ├── 17.png
            └── 18.png
```



BONUS 1

Segmentation of Complete WES Data

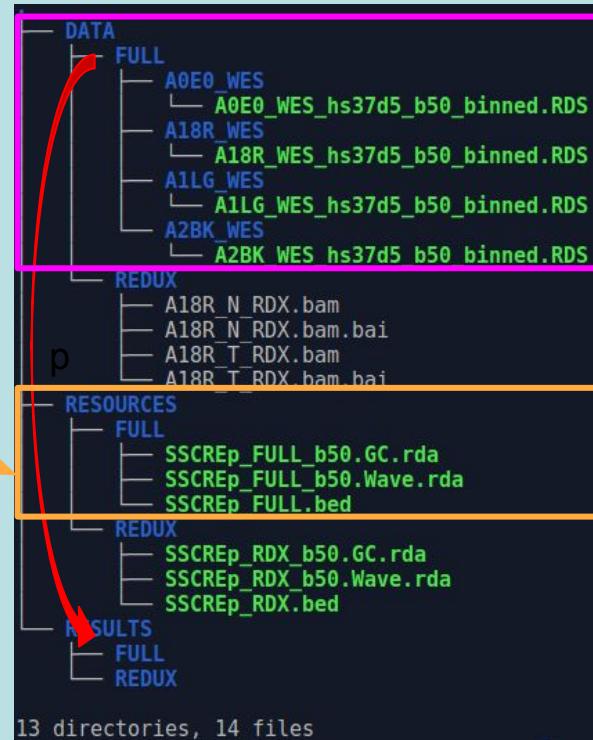
PRACTICE : Complete Pre-normalized WES Data [ASCAT]

Input data :

- Binned datasets
- “GC%” & “Wave” packs corresponding to all WES regions

Exercise :

- In a shell, copy the content of `~/tp_cna/DATA/FULL` into `~/tp_cna/RESULTS/FULL`
- In R/Rstudio, set the working directory to `~/tp_cna/RESULTS/FULL/`
- Reproduced all the analysis steps performed earlierly (since normalization) on these samples, still using ASCAT.



PRACTICE : Complete Pre-normalized WES Data [ASCAT]

To make it faster, we will perform it using multiple threads (CPUs)

```
WES.Normalize.ff.Batch(BINpack = ".../..../RESOURCES/FULL/SSCREp_FULL_b50.GC.rda",
wave.rda = ".../..../RESOURCES/FULL/SSCREp_FULL_b50.Wave.rda", wave.renorm = TRUE,
nthread = 4)
```

```
Segment.ff.Batch(segmenter = "ASCAT", smooth.k = 5, nrf = 1, SER.pen = 5,
nthread = 4)
```

```
ASCN.ff.Batch(RDS.files = list.files(pattern = "SEG.ASCAT.RDS", recursive =
TRUE), nthread = 4)
```

WARNING : DO NOT use the **nsubthread** parameter seen in your preceding use of the **ASCN.ff()** function : this would run the analysis using $4 \times 4 = 16$ CPUs, but you did not requested it to the file cluster scheduler !

```
Annotate.ff.Batch(RDS.files = list.files(pattern = "SEG.ASCAT.RDS", recursive =
TRUE), nthread = 4)
```

BONUS 2

Comparison with microarray
profiles

PRACTICE : Affymetrix SNP6.0 Microarrays Results



Result files :

- /shared/projects/ebai2021_n2/correction/dna_seq/tp_cna/SNP6/A0E0.snp6/
- /shared/projects/ebai2021_n2/correction/dna_seq/tp_cna/SNP6/A18R.snp6/
- /shared/projects/ebai2021_n2/correction/dna_seq/tp_cna/SNP6/A1LG.snp6/
- /shared/projects/ebai2021_n2/correction/dna_seq/tp_cna/SNP6/A2BK.snp6/

Exercise :

- Compare, thanks to plots and the HTML report, the segmentation results obtained from WES and SNP6 for the same samples.
- ***NOTE : Expected WES results are also available here :***
 - /shared/projects/ebai2021_n2/correction/dna_seq/tp_cna/WES/REDUX/
 - /shared/projects/ebai2021_n2/correction/dna_seq/tp_cna/WES/FULL/

BONUS 2

-
WES with another segmenter
: FACETS

PRACTICE : Segmentation & calling [FACETS]

BONUS

Input data :

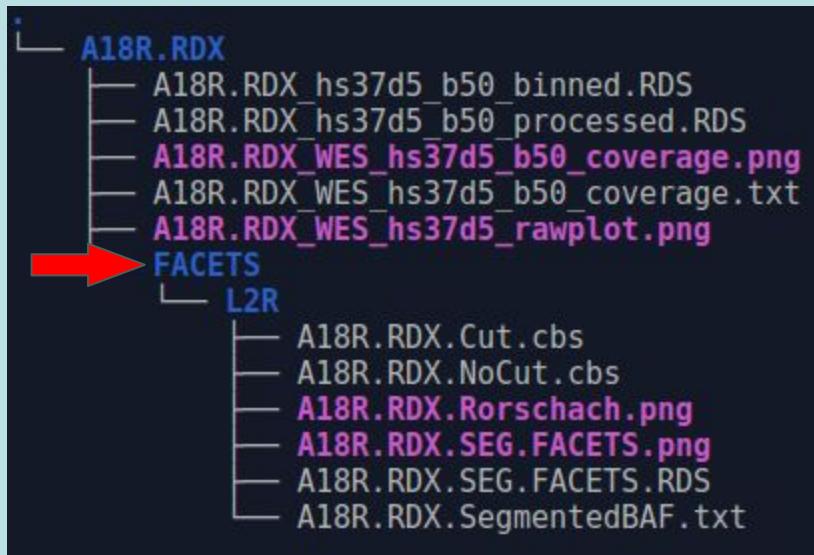
- The **normalized RDS dataset** :
 - o ~/tp_cna/RESULTS/REDUX/A18R.RDX/A18R.RDX_hs37d5_b50_processed.RDS

Run this command in R (*same as with ASCAT, changing the **segmenter** parameter*) :

```
Segment.ff(RDS.file = "A18R.RDX/A18R.RDX_hs37d5_b50_processed.RDS",
segmenter = "FACETS", smooth.k = 5, nrf = 1, SER.pen = 5)
```

PRACTICE : Outputs [FACETS]

BONUS



PRACTICE : Absolute Copy Number [FACETS]



Input data :

- The **segmented** dataset you generated :
 - o ~/tp_cna/RESULTS/REDUX/A18R.RDX/FACETS/L2R/A18R.RDX.SEG.FACETS.RDS

Run in R (*same as with ASCAT*) :

```
ASCN.ff(RDS.file = "A18R.RDX/FACETS/L2R/A18R.RDX.SEG.FACETS.RDS")
```

PRACTICE : Outputs [FACETS]

BONUS

```
. └── A18R.RDX
    ├── A18R.RDX_hs37d5_b50_binned.RDS
    ├── A18R.RDX_hs37d5_b50_processed.RDS
    ├── A18R.RDX_WES_hs37d5_b50_coverage.png
    ├── A18R.RDX_WES_hs37d5_b50_coverage.txt
    ├── A18R.RDX_WES_hs37d5_rawplot.png
    └── FACETS
        └── ASCN
            ├── A18R.RDX.ASCN.FACETS.png
            ├── A18R.RDX.ASCN.FACETS.RDS
            ├── A18R.RDX.cn
            ├── A18R.RDX_model.txt
            └── A18R.RDX.TCNvsL2R.png
        └── L2R
            ├── A18R.RDX.Cut.cbs
            ├── A18R.RDX.NoCut.cbs
            ├── A18R.RDX.Rorschach.png
            ├── A18R.RDX.SEG.FACETS.png
            ├── A18R.RDX.SEG.FACETS.RDS
            └── A18R.RDX.SegmentedBAF.txt
```

PRACTICE: Annotation & HTML Report

BONUS

Input data :

- The **segmented** dataset you generated :
 - `~/tp_cna/RESULTS/REDUX/A18R.RDX/FACETS/L2R/A18R.RDX.SEG.FACETS.RDS`

Run in R (same as with ASCAT) :

```
Annotate.ff(RDS.file = "A18R.RDX/FACETS/L2R/A18R.RDX.SEG.FACETS.RDS")
```

(you can also try the 3 other samples !)

BONUS 3

-
WES with another segmenter
: SEQUENZA

PRACTICE : Segmentation & calling [SEQUENZA]



Input data :

- The normalized RDS dataset :
 - ~/tp_cna/RESULTS/REDUX/A18R.RDX/A18R.RDX_hs37d5_b50_processed.RDS

```
Segment.ff(RDS.file = "A18R.RDX/A18R.RDX_hs37d5_b50_processed.RDS",
segmenter = "SEQUENZA", smooth.k = 5, nrf = 1, SER.pen = 5)
```

PRACTICE : Absolute Copy Number [SEQUENZA]

BONUS

Input data :

- The **segmented** dataset you generated :
 - ~/tp_cna/RESULTS/REDUX/A18R.RDX/**SEQUENZA/L2R/A18R.RDX.SEG.SEQUENZA.RDS**

```
ASCN.ff(RDS.file = "A18R.RDX/SEQUENZA/L2R/A18R.RDX.SEG.SEQUENZA.RDS")
```

PRACTICE : Annotation & HTML Report

BONUS

Input data :

- The **segmented** dataset you generated :
 - ~/tp_cna/RESULTS/REDUX/A18R.RDX/**SEQUENZA/L2R/A18R.RDX.SEG.SEQUENZA.RDS**

```
Annotate.ff(RDS.file = "A18R.RDX/SEQUENZA/L2R/A18R.RDX.SEG.SEQUENZA.RDS")
```

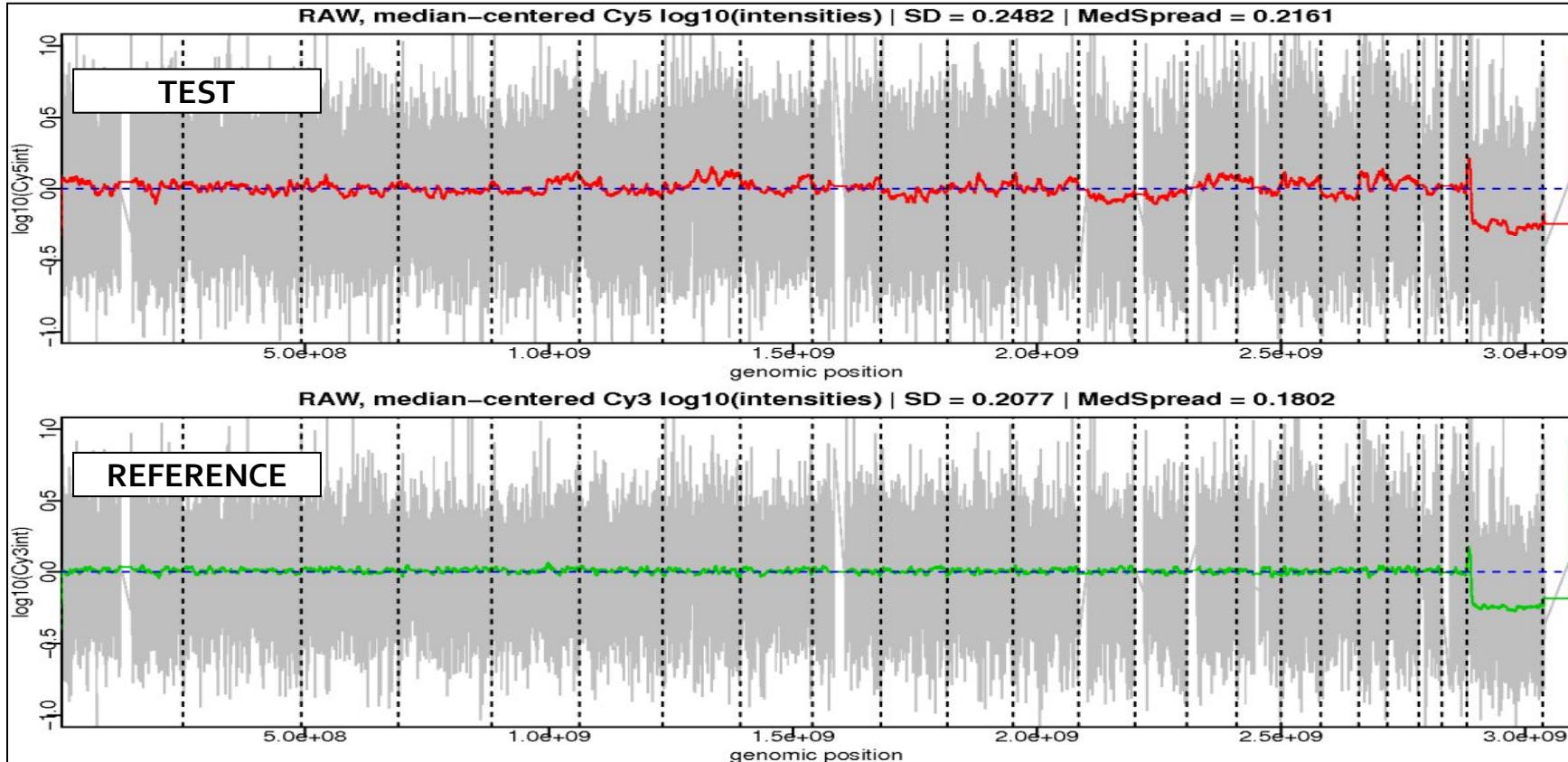
(and you can still play with the 3 other samples !)

APPENDIX

Normalization : Source (Dye / Library / Run ...) Bias

A

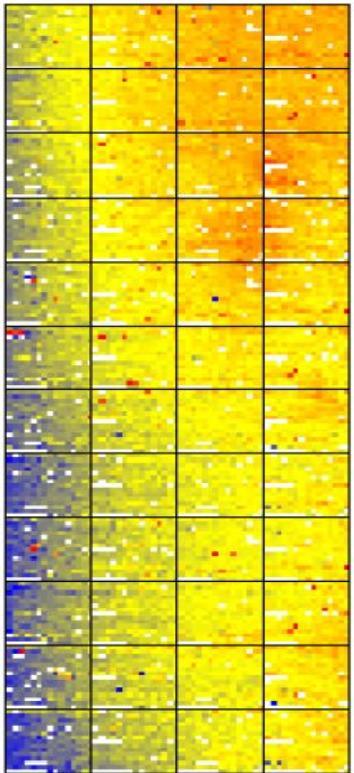
S



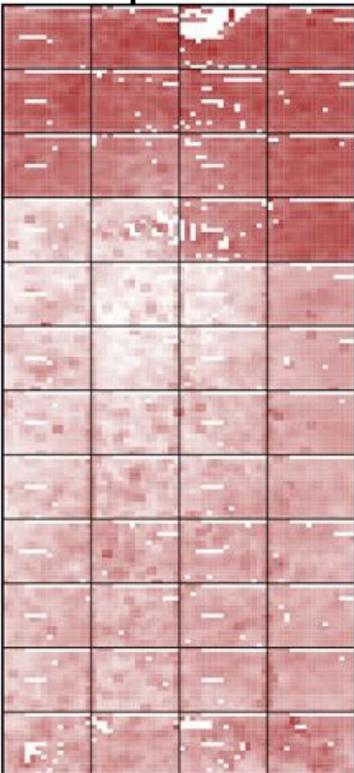
Normalization : Spatial Bias Sources

A

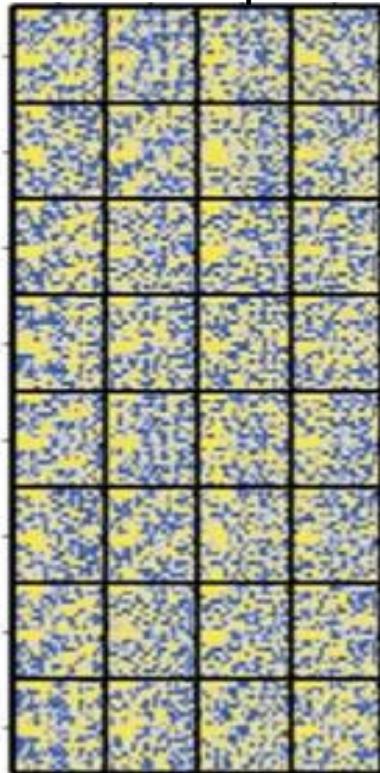
Gradient



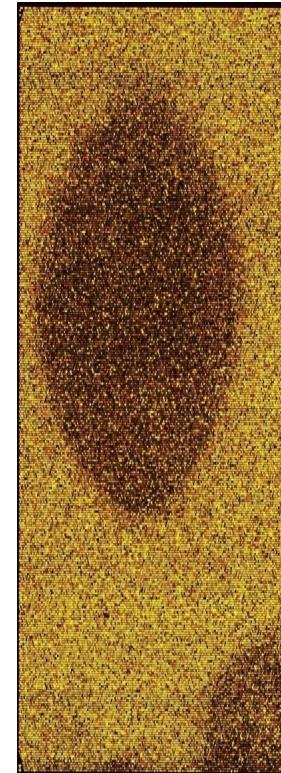
Spotter



Print-tip

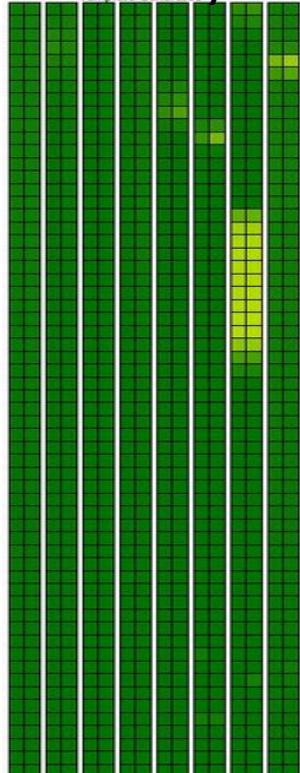


Leak



S

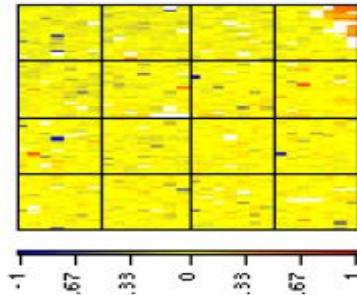
Density



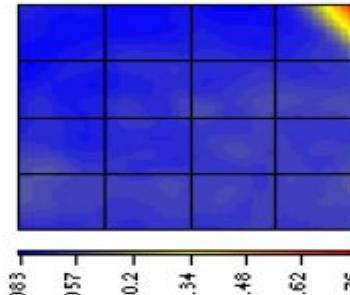
Normalization : Spatial Bias Correction

A

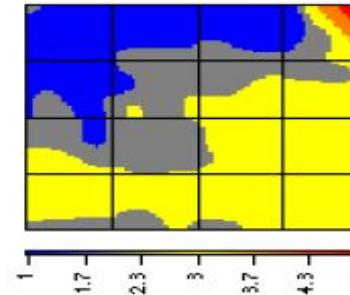
(a) Centered log-ratios



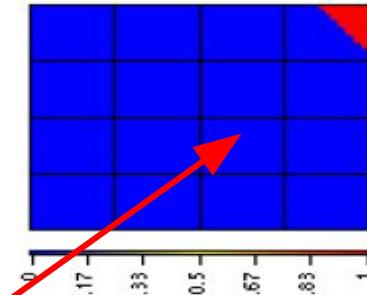
(b) Array trend



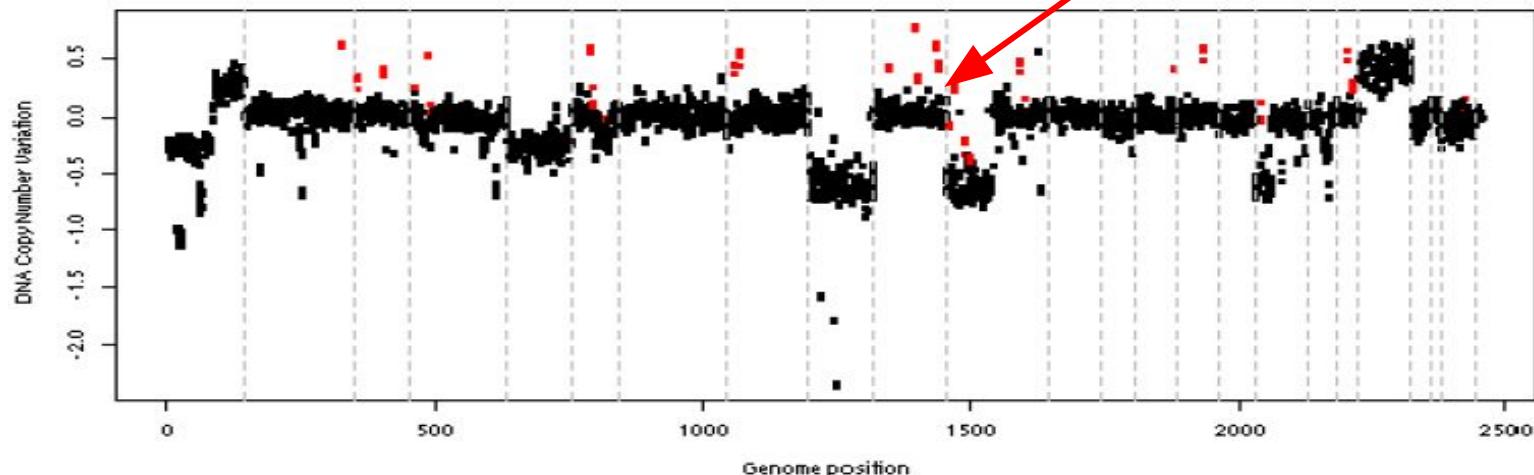
(c) Spatial clustering



(d) Detected areas



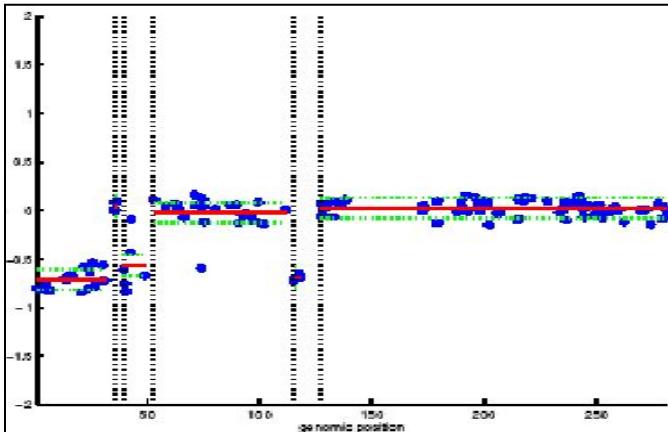
(e) Pan-genomic profile



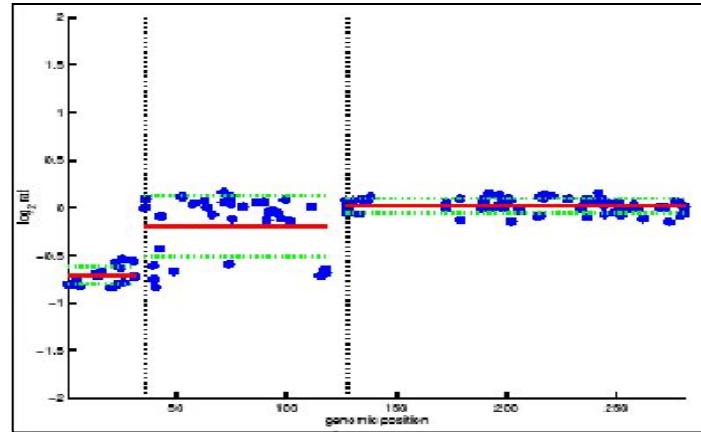
Segmentation : A Computational Challenge

- Two unknowns for breakpoints :
 - Localization
 - Quantity
- Three families of algorithms :
 - Smoothers (wavelet)
 - Change-point
 - Binary segmentation (CBS)
 - Optimal partitionning (PELT)
 - HMM modeling (bioHMM)

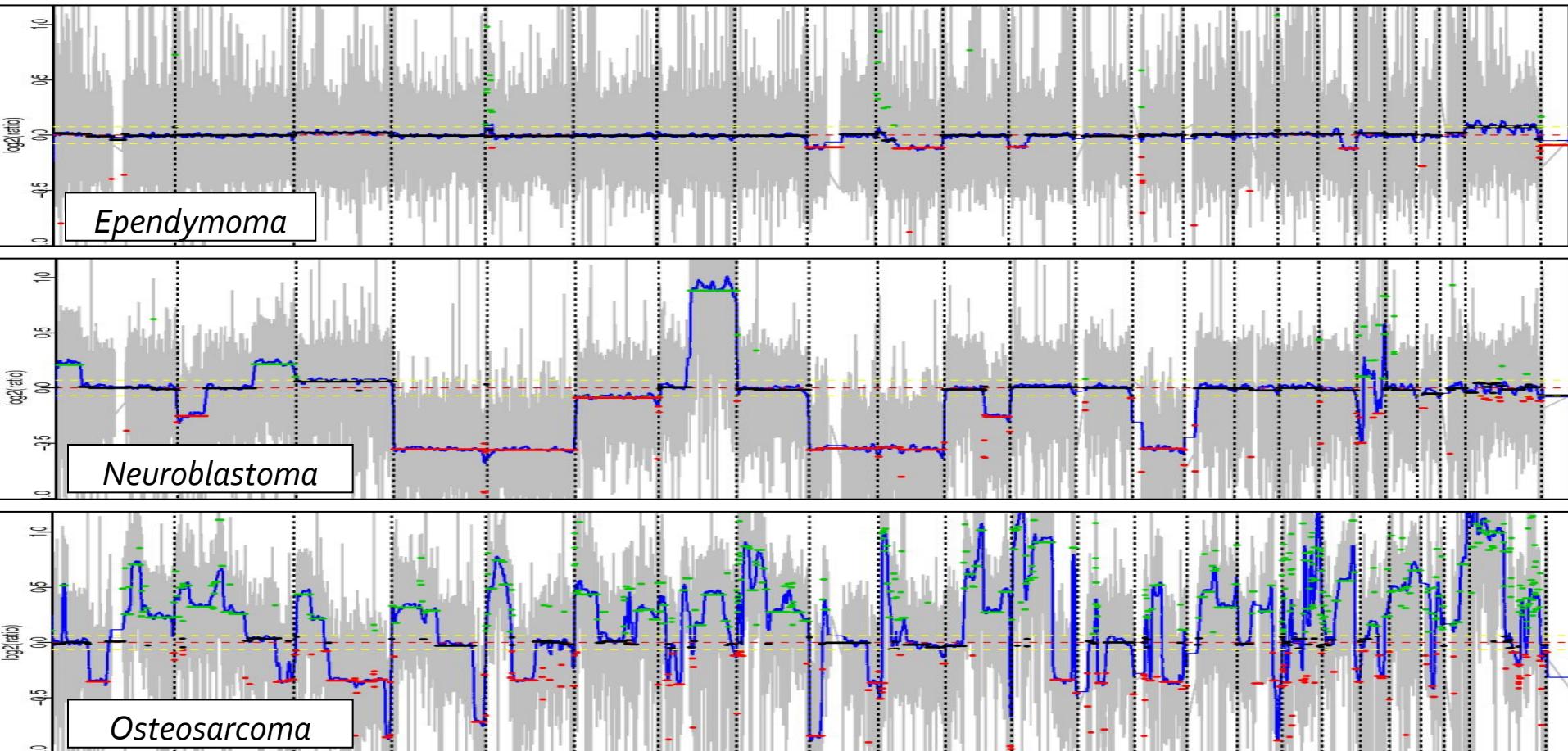
Homoscedastic (m)



Heteroscedastic (m, V)

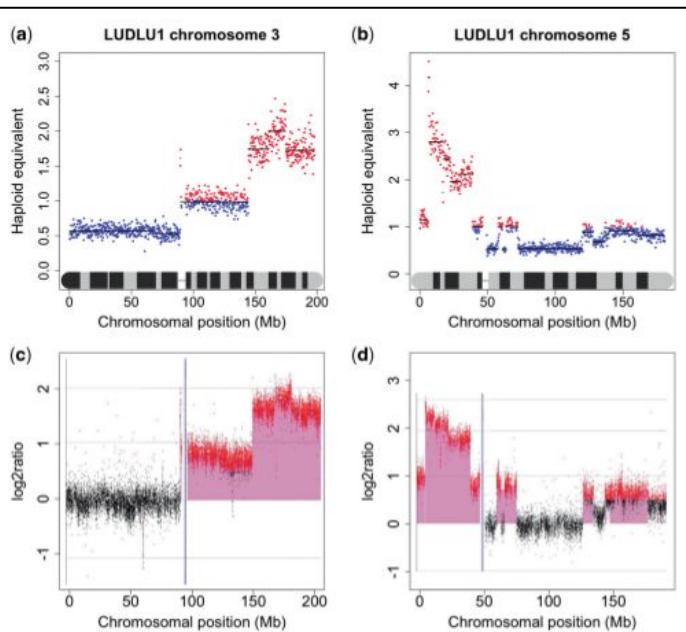


Segmentation : Variations in Complexity



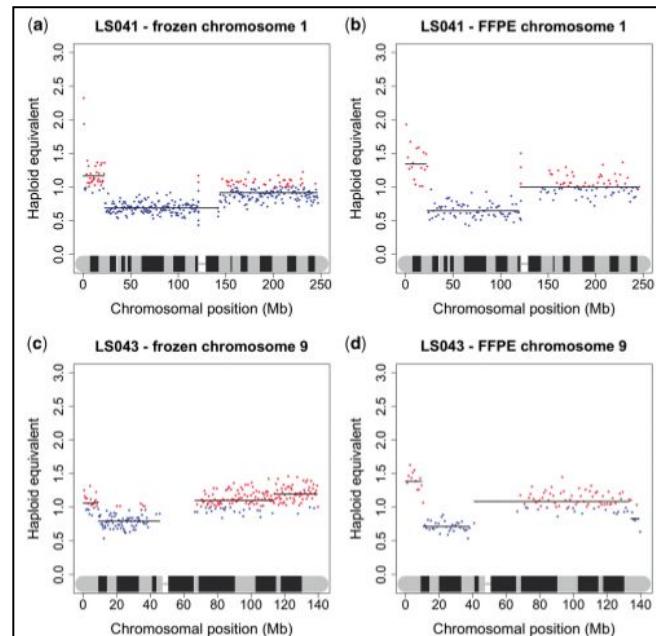
NGS Beyond Microarrays

NGS : Low Input, FFPE



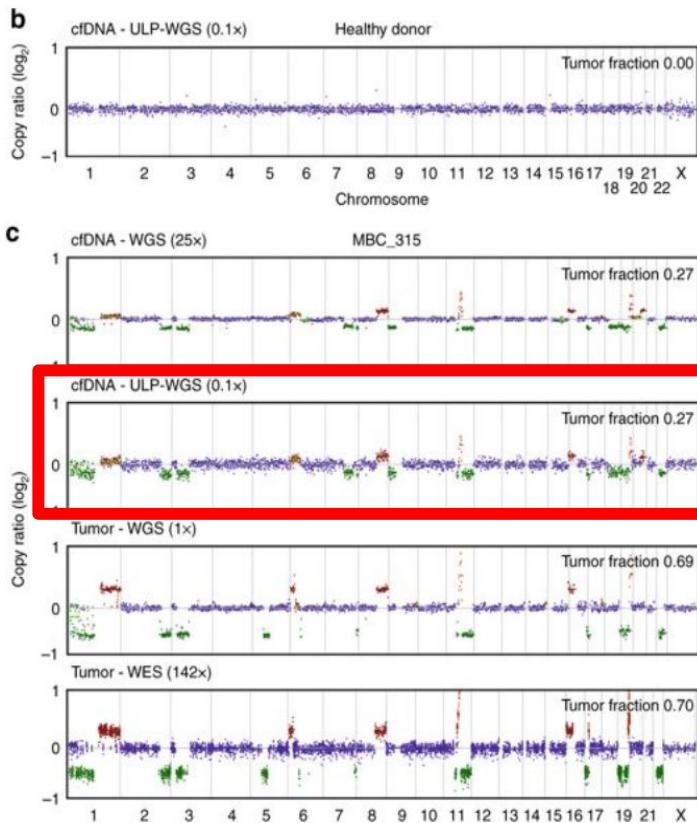
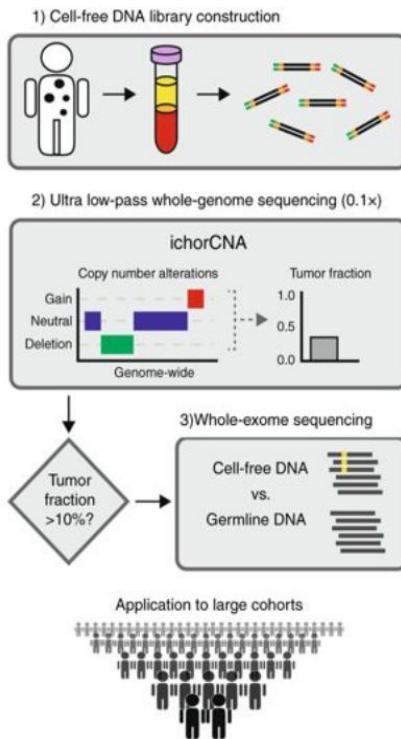
- 2 to 5 ng of DNA

- **FFPE** (Formalin-fixed paraffin-embedded) samples



NGS : Cell-free DNA Shallow WGS

IchorCNA (Adalsteisson, Nature Com, 2017)



- Cell-free DNA
- ULP-WGS
- 0.1X coverage

NGS : Cell-free DNA shallow WGS

WisecondorX (Raman et al, NAR, 2018)

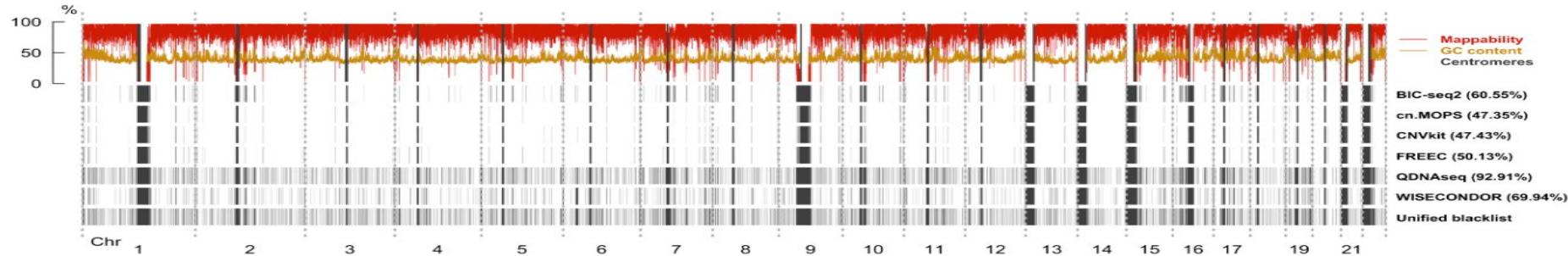


Figure S3. Representation of the blacklists across the considered tools at 30 kb.

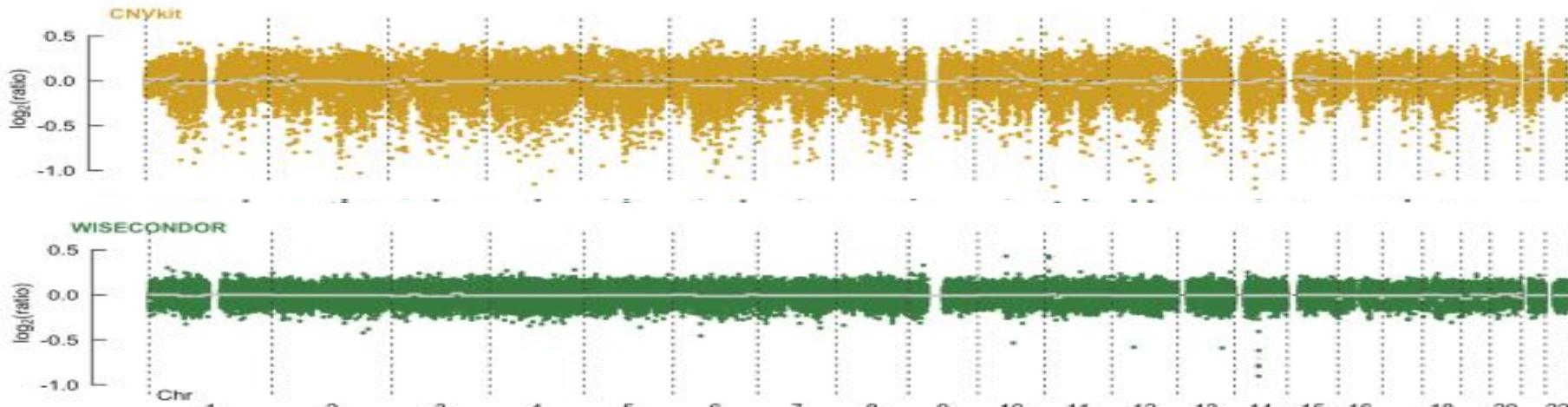
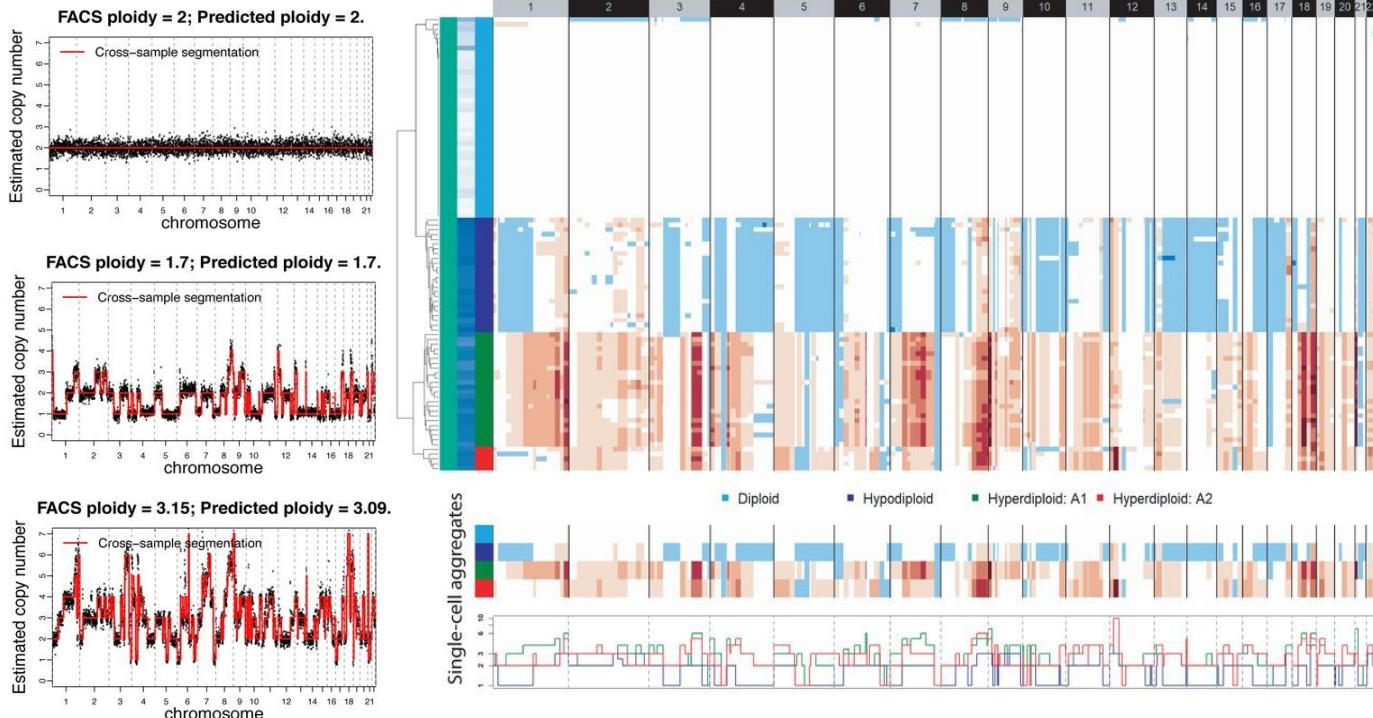
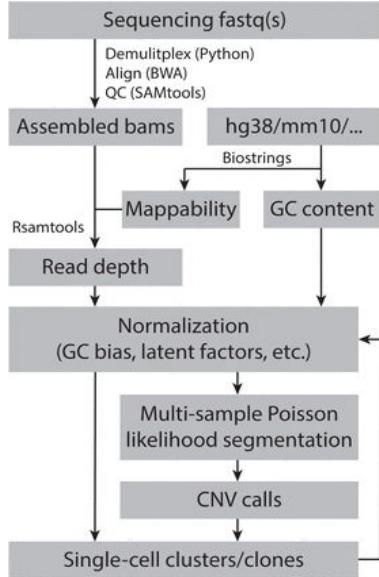


Figure S8. Autosome-wide profile comparison of problematic sample gDNA-3.

NGS : Single Cell CNA (SCOPE)

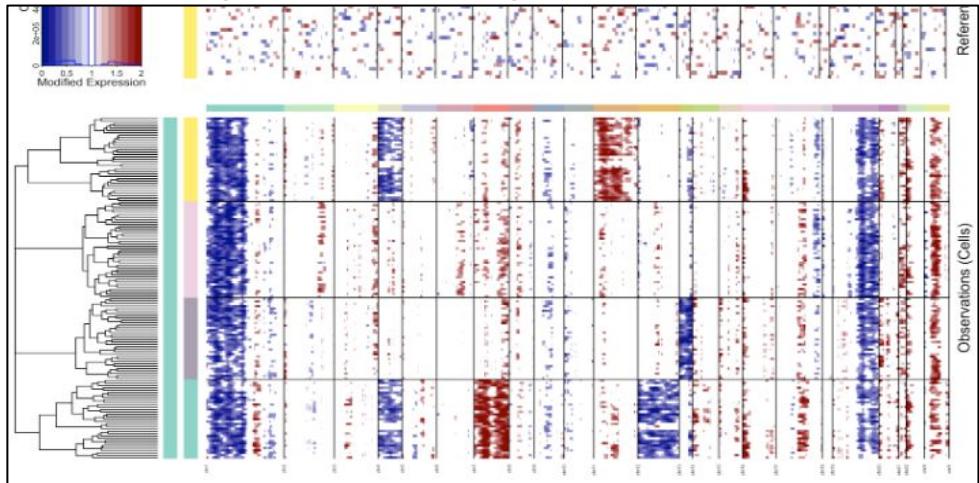


WARNING :

- Limited resolution : > 2 Mb (binning)
- Requires > 750,000 reads / cell

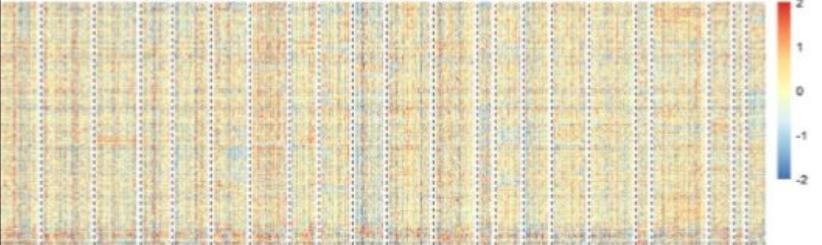
NGS : Single Cell CNA from scRNAseq (InferCNV / CaSpER)

InferCNV (Broad Institute)

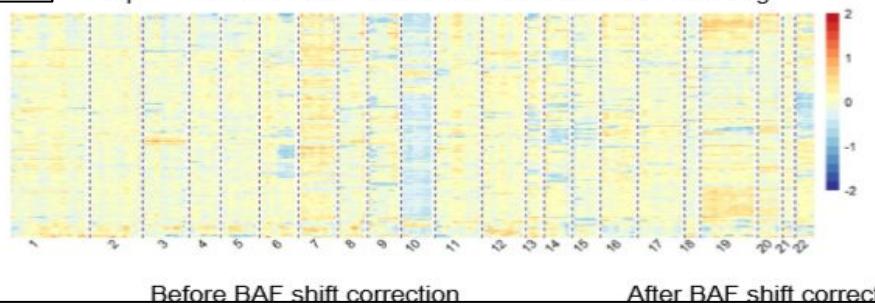


CaSpER (Armanci et al, BioRxiv 2019)

Control normalized original expression matrix



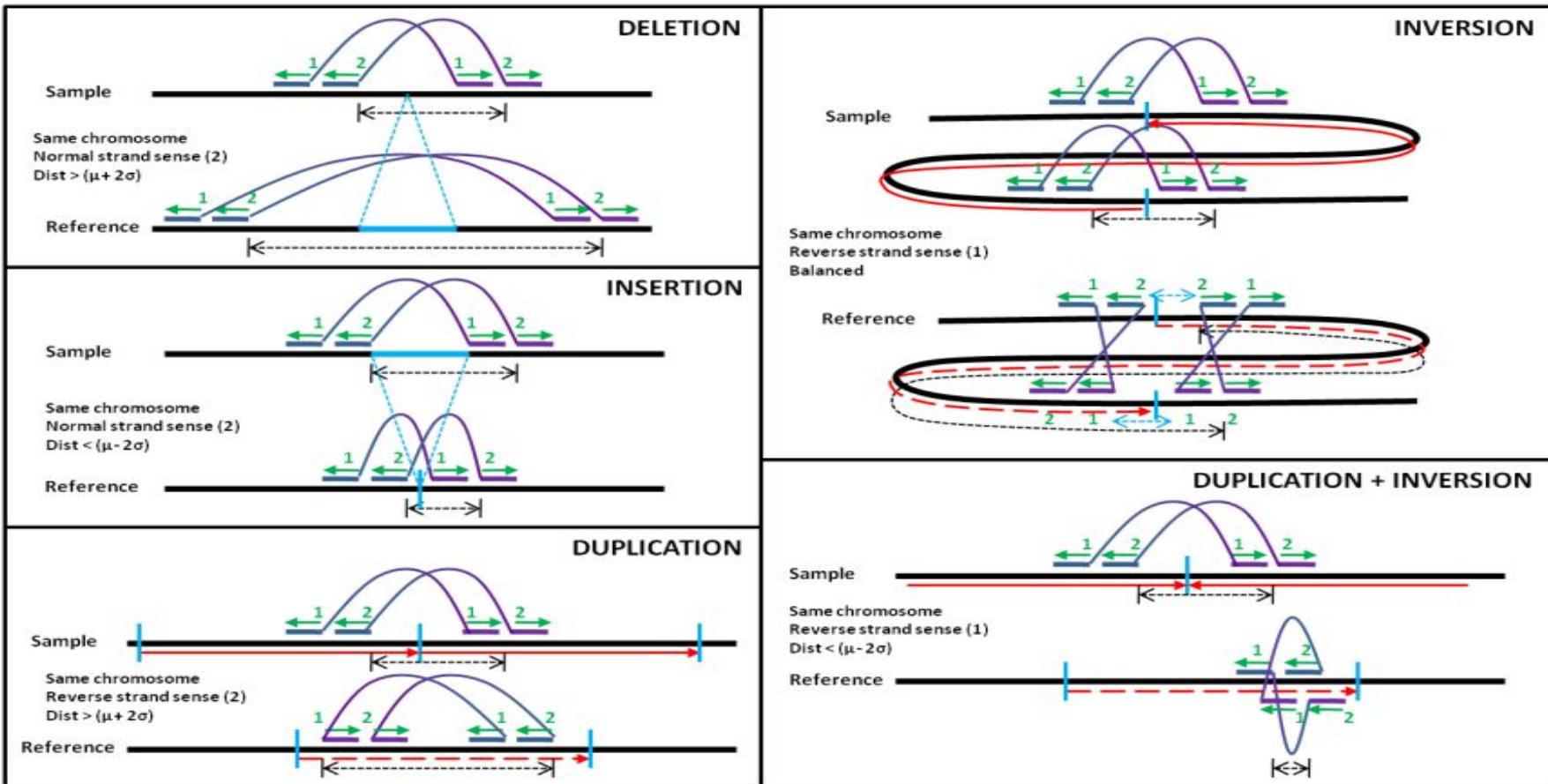
Expression Matrix after Recursive Multiscale Median Filtering



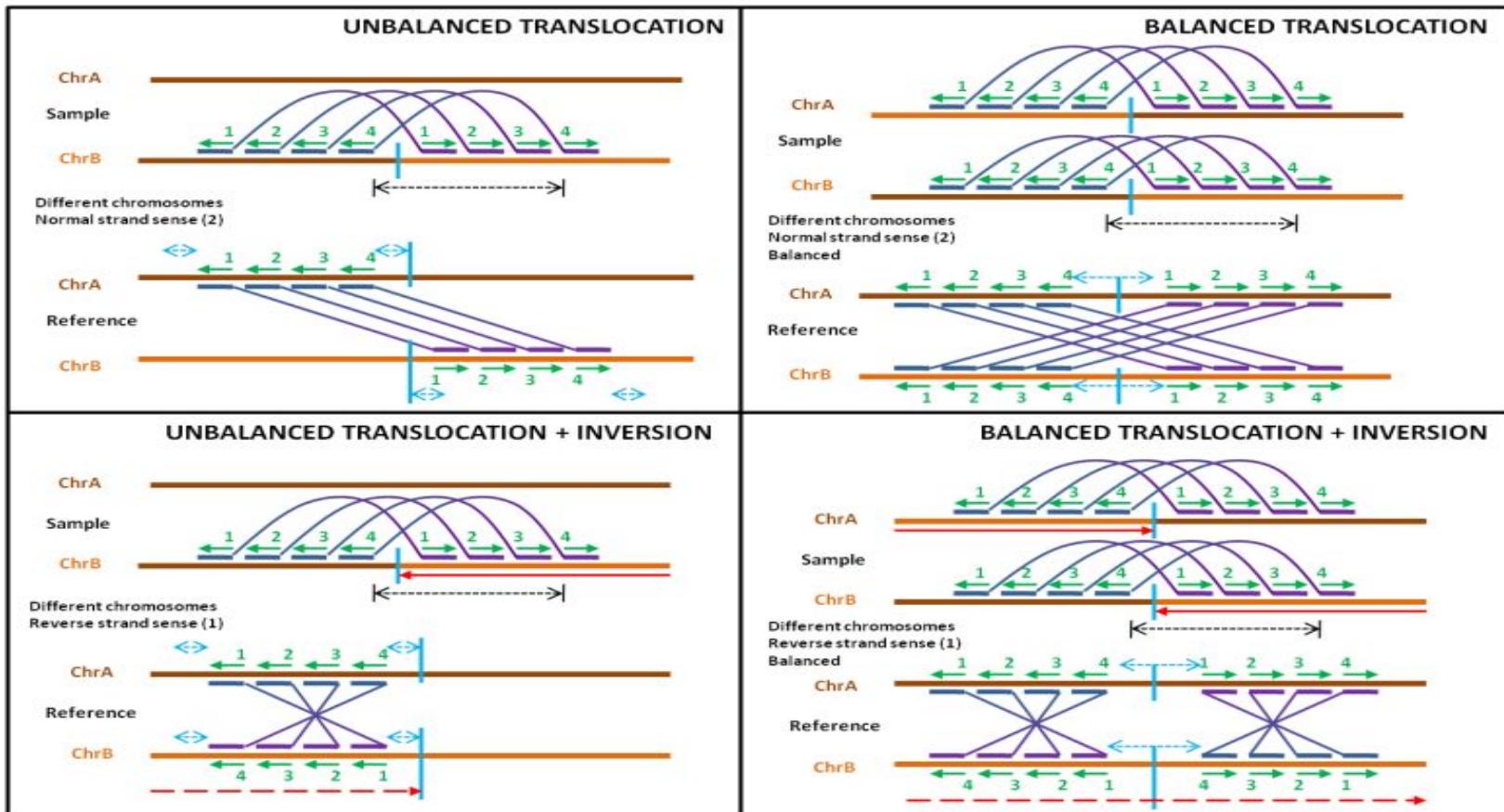
WARNING :

- Coarse grain (> 10 Mb)
- Requires > 75,000 reads / cell

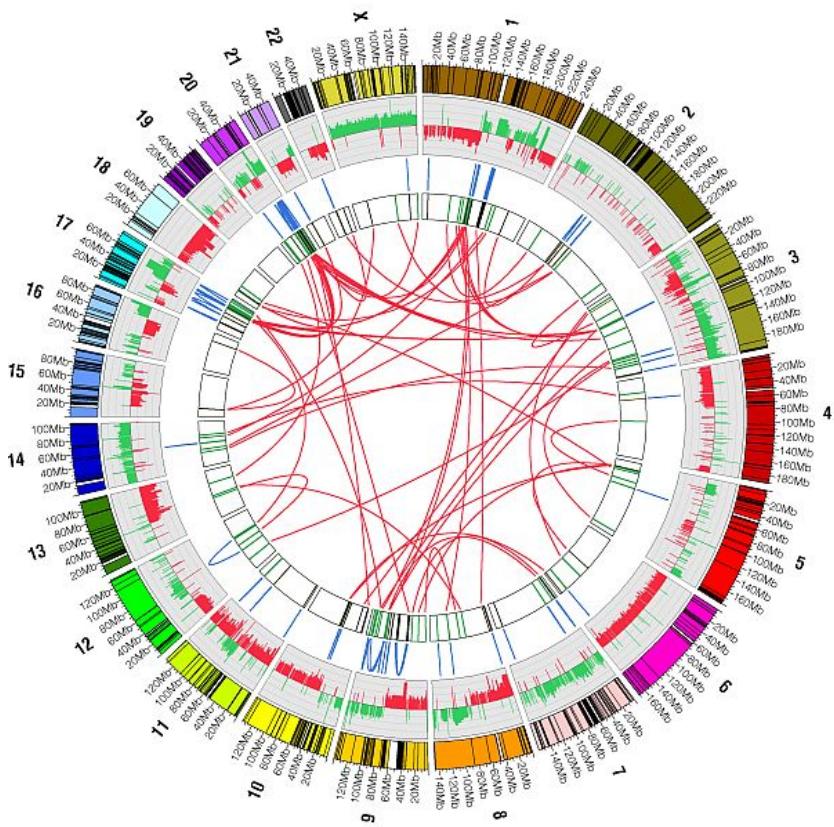
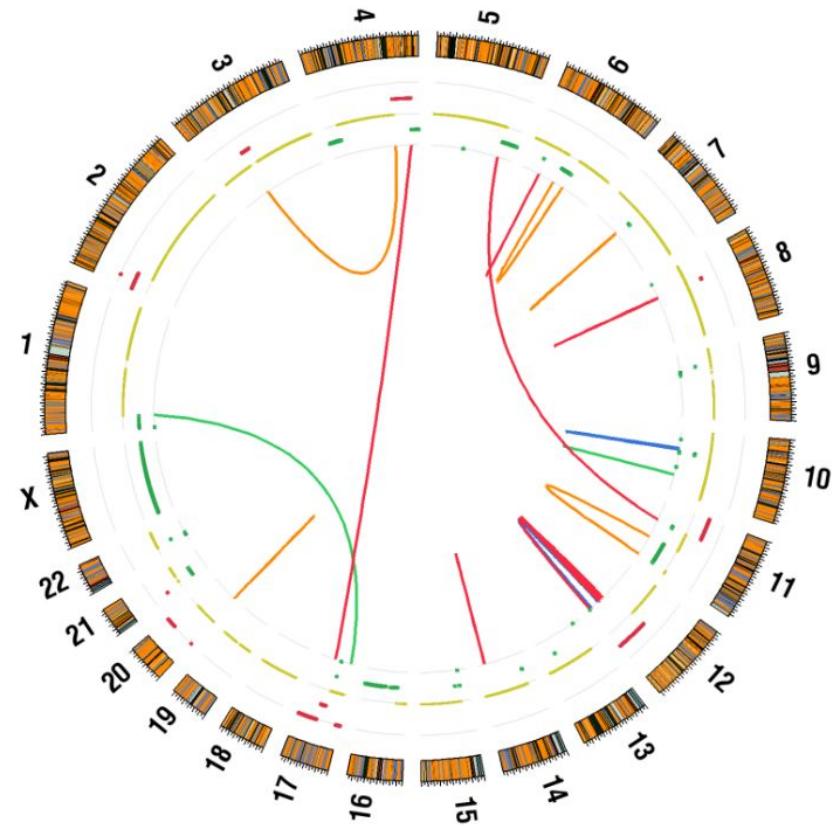
WGS : Intra-chromosomal Structural Variations



WGS : Inter-chromosomal Structural Variations



Visualization : Circos-plots



NGS vs Microarrays

	Microarray	NGS (WES / WGS)
Physical entity	Array on a glass slide	Lane in a flowcell
Measurement entity	Spot of probes	Cluster of fragments
Measurement unit	Luminous intensity per genomic position	Read depth per genomic bin
Data distribution	Log-normal	Negative binomial
Data transformation	Log ratio of intensities Test / Ref	Log ratio of depths Test / Ref
Bias main sources	Spatial effects, dye, GC-content	Library effects, spatial effects, coverage, GC-content, mappability
CNV information	Normality, gains and losses relative to the reference	Normality, gains and losses relative to the reference, absolute and allele-specific copy number levels
CNV event precision	Up to ~3 Kb	~50 b
Structural information	Large-scale deletions	Insertions, deletions, inversions, balanced translocations
SNP information	Known SNPs (if specific probes)	All kinds of SNPs, position and allele frequency
SNV (mutation) information	No / some*	All SNVs
Sequence information	No	Full covered sequence

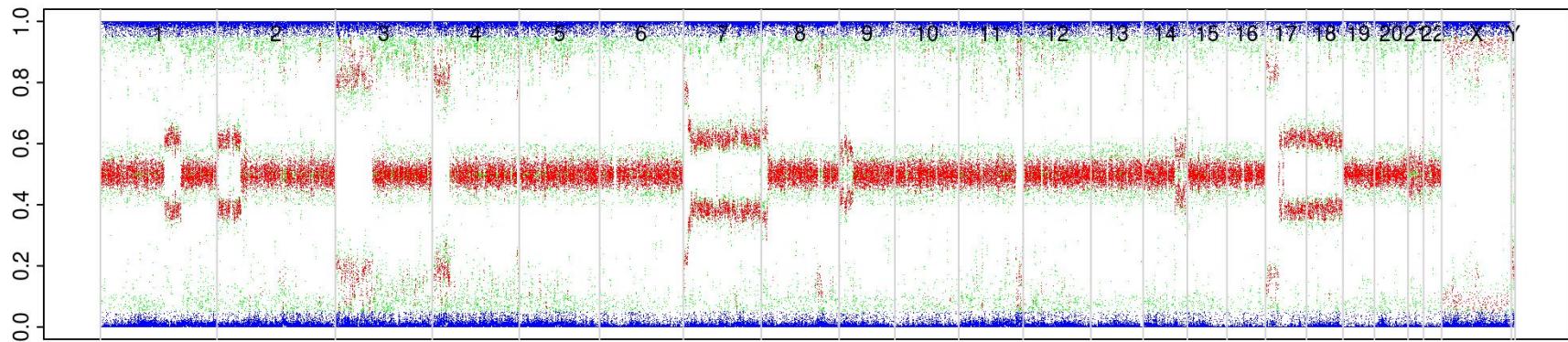
Microarrays are still alive !



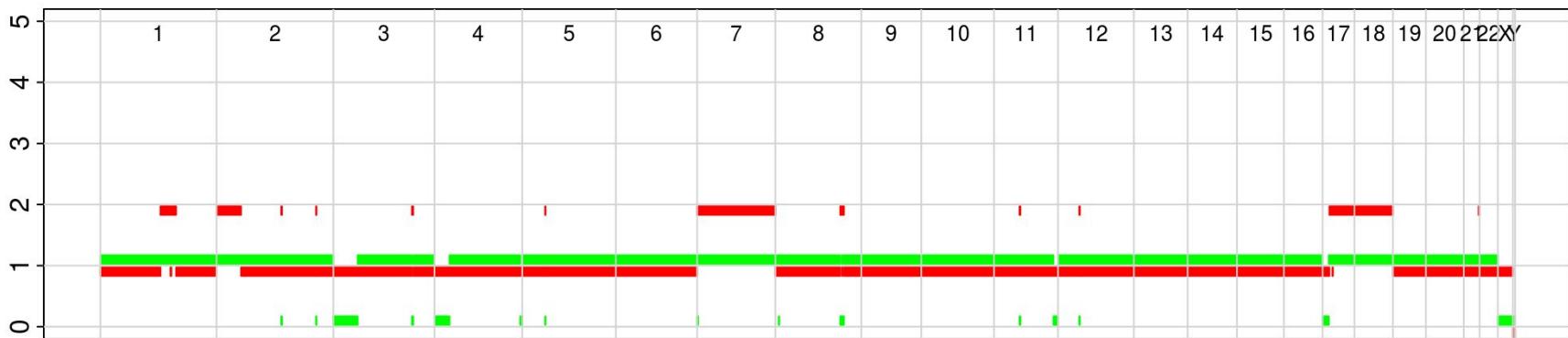
A

FFPE Samples

M1084_PED 58370 129246



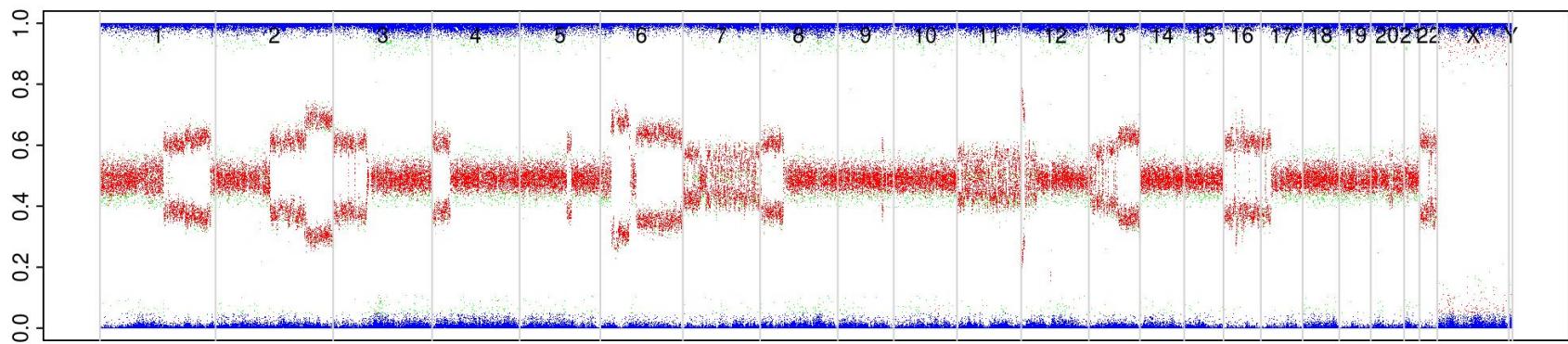
Ploidy: 2.17, aberrant cell fraction: 82%, goodness of fit: 94.7%



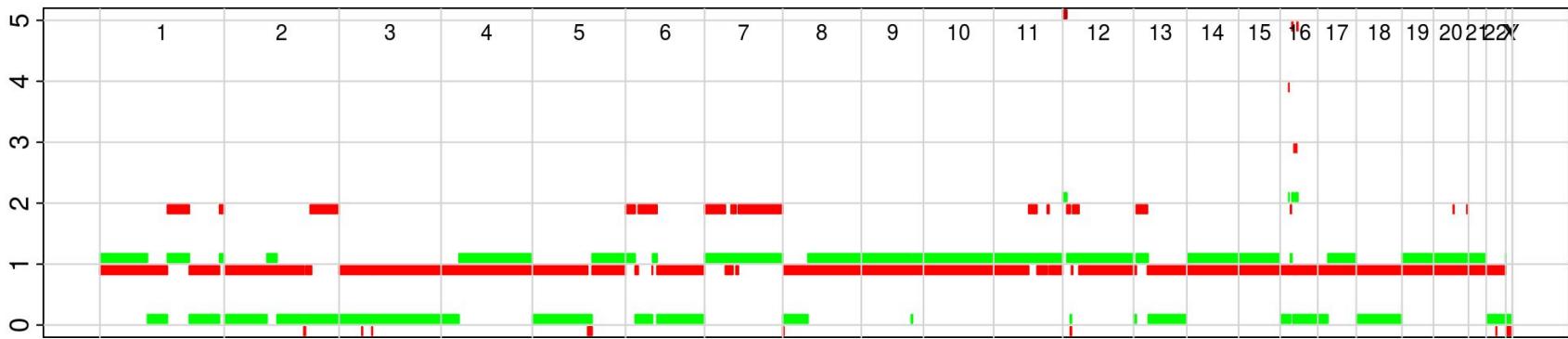
A

ctDNA

M782_circ 54608 156269



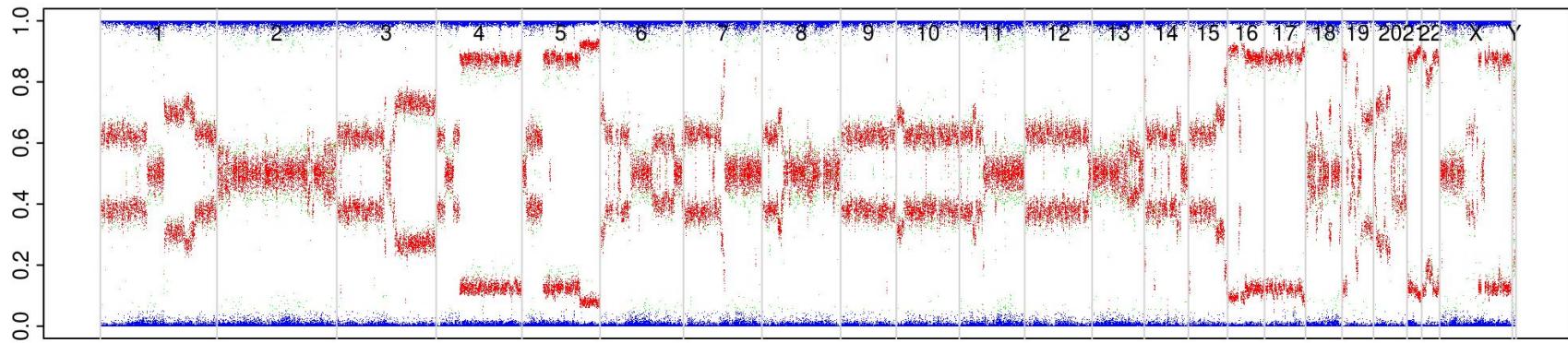
Ploidy: 1.61, aberrant cell fraction: 47%, goodness of fit: 89.8%



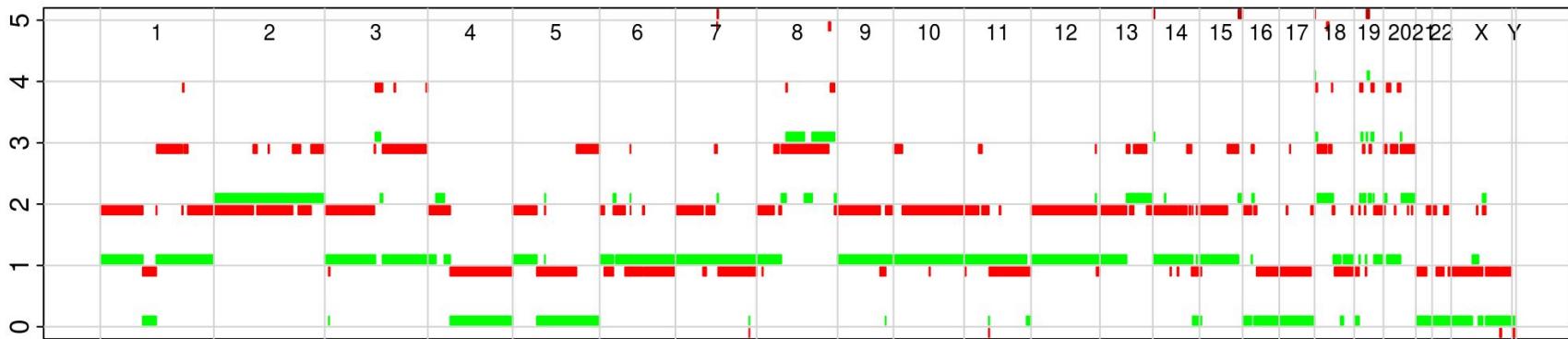
A

Ascite DNA

A26 67028 140247

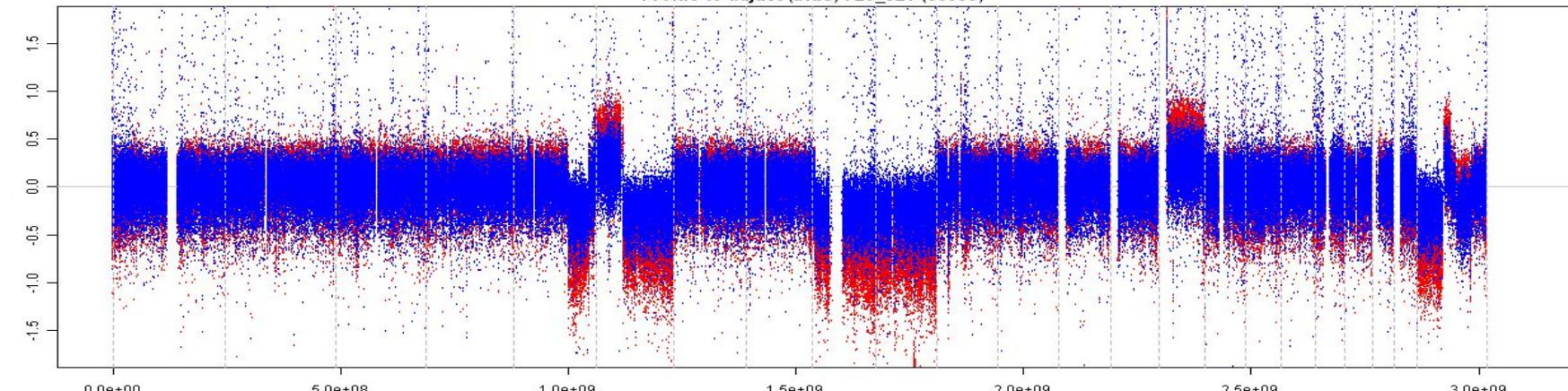


Ploidy: 2.98, aberrant cell fraction: 86%, goodness of fit: 93.5%

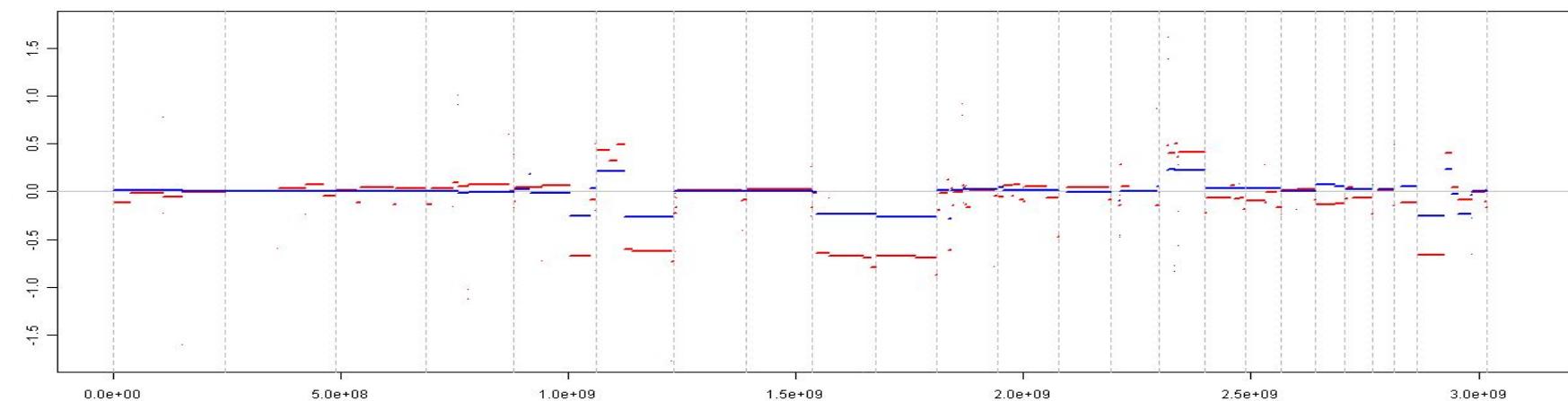


Further Analyses

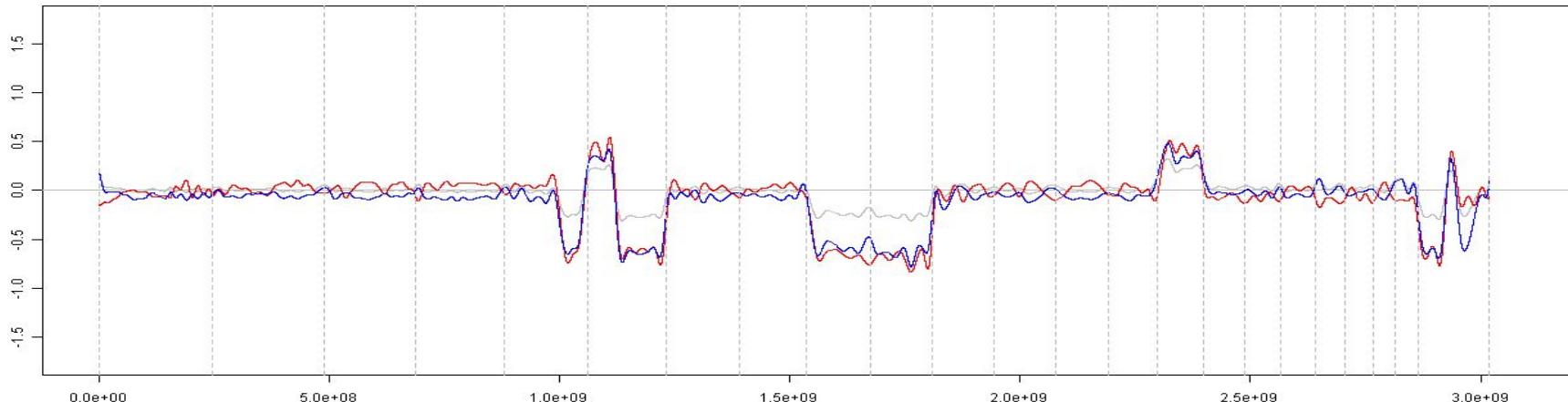
Comparison of a Pair of Profiles (scaling)



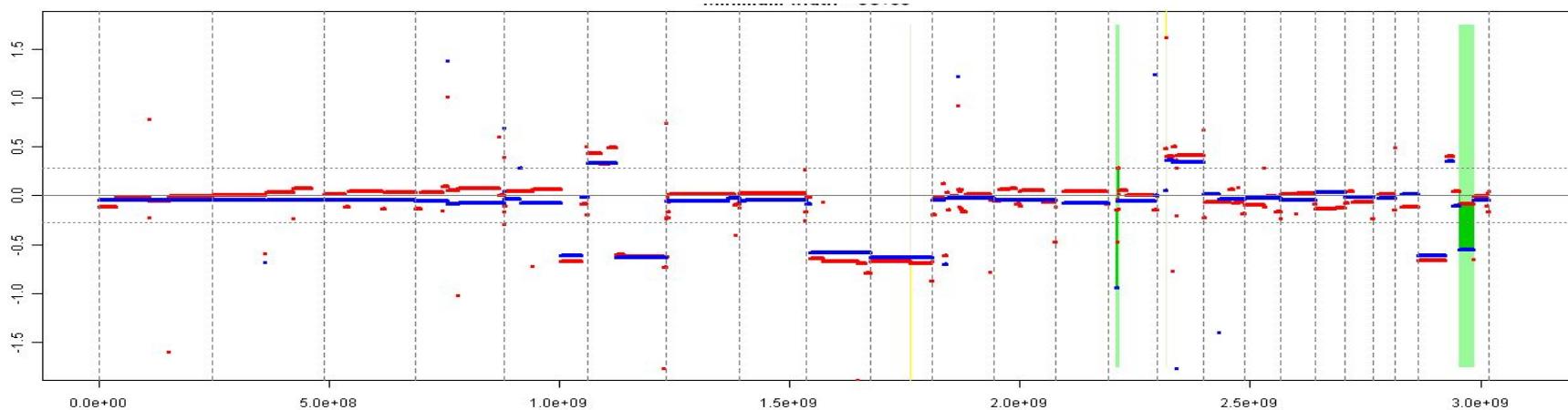
d_0
 d_{21}



Comparison of a Pair of Profiles (scaled)



$d0$
 $d21$



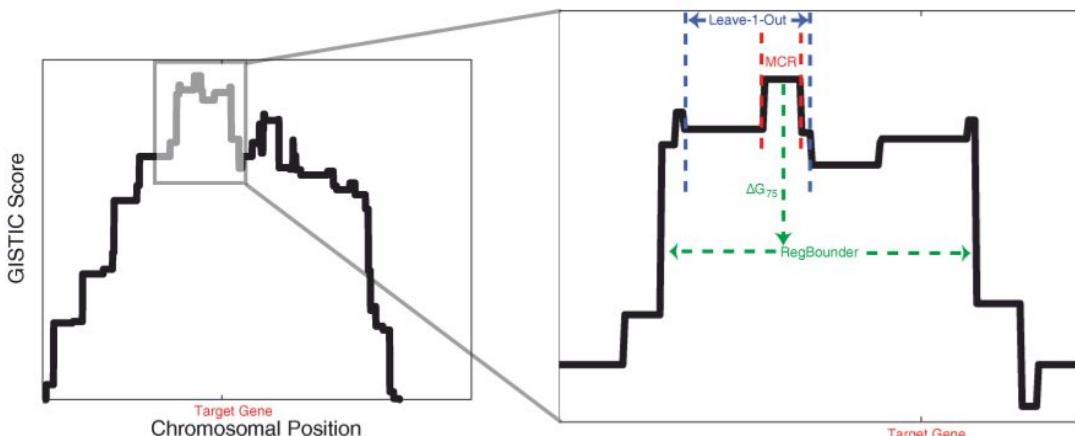
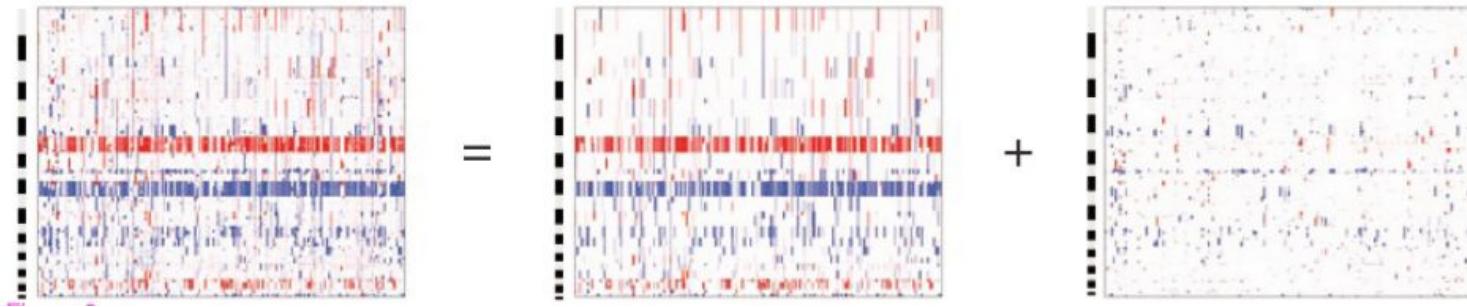
Minimal Common Regions (MCR)

(c)

All Data

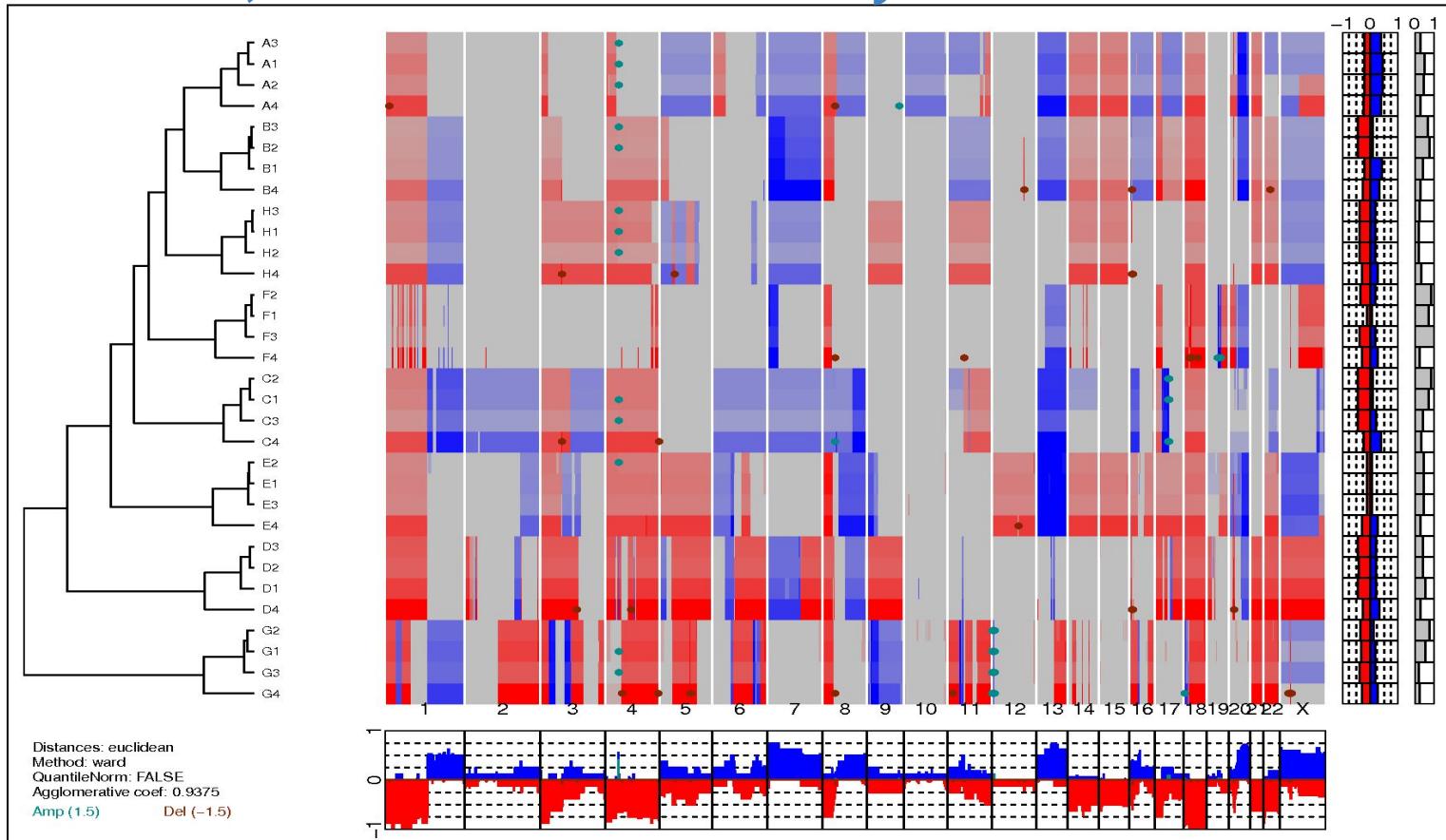
Arm-level SCNAs

Focal SCNAs



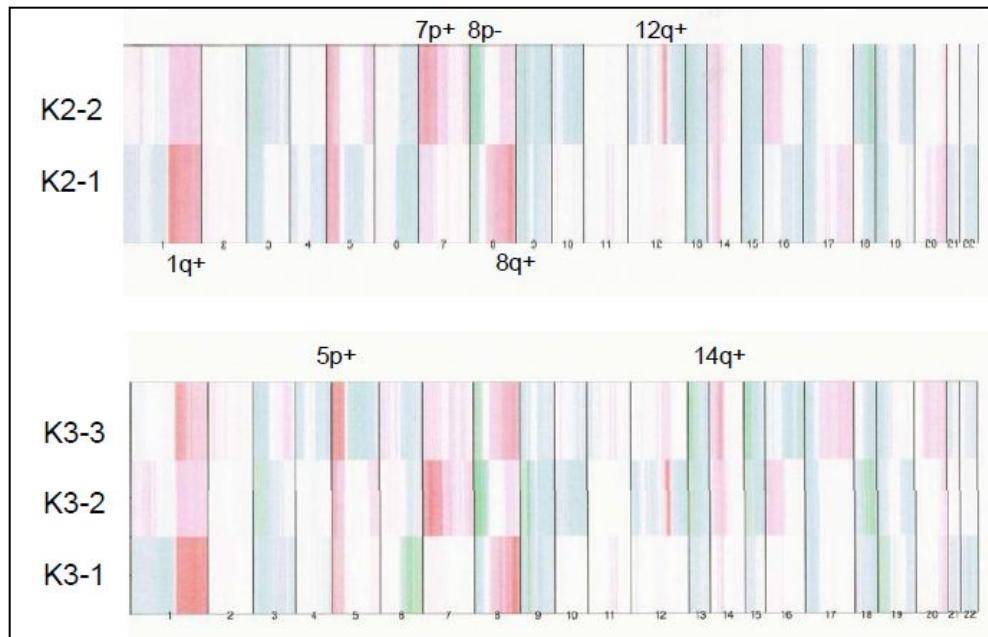
Hierarchical Clustering, Heatmap, Frequency of Aberrations, Genomic Instability Scores

A S

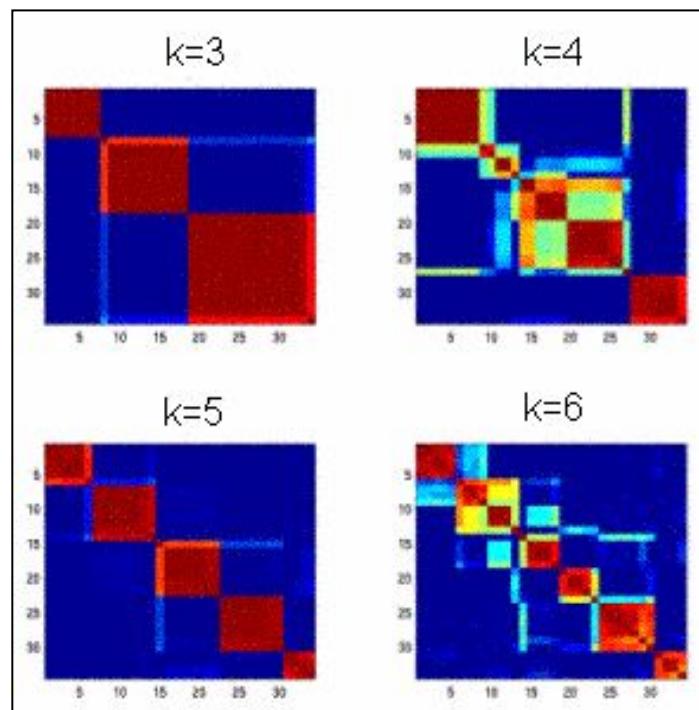


Other Clustering Methods

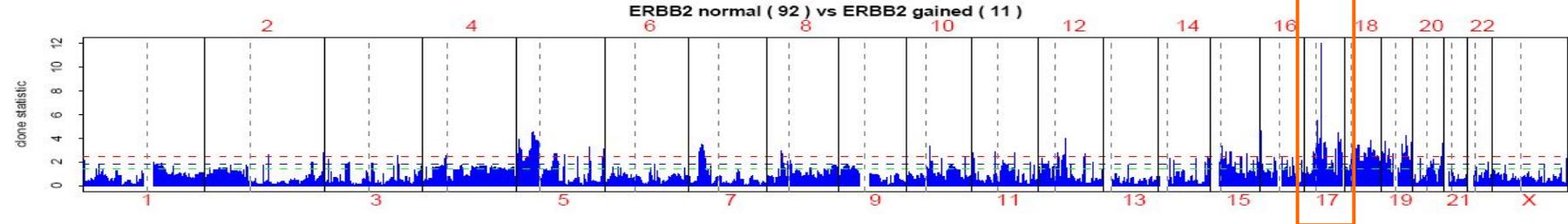
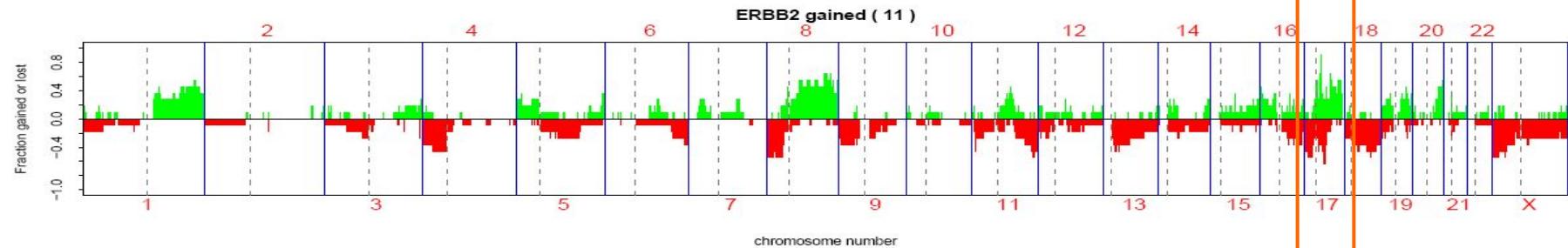
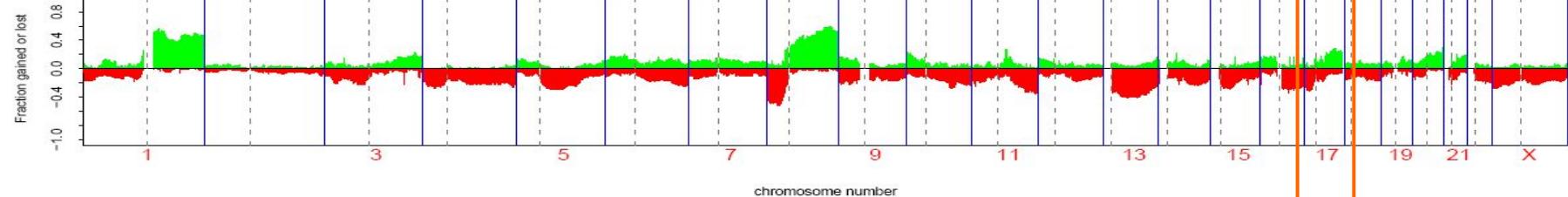
K-means



Non-negative Matrix Factorization



Differential Analysis (in Subpopulations)



Annotation of Genomic Regions

Loc	Width	Band1	Band2	Num.probes	Status	log2(ratio)	Ratio	Genes	CNV	miRNA	CpGisI
8:161471-6914076	6.75Mb	8p23.3	8p23.1	409	L	-1.13	0.46	2 12 32	1949	15	120
8:6939250-7786708	847.46Kb	8p23.1	8p23.1	10	L	-0.34	0.79	3 23	135	-	11
8:8100383-22878739	14.78Mb	8p23.1	8p21.3	879	L	-1.09	0.47	1 15 39 109	1560	28	120
8:22888308-24757290	1.87Mb	8p21.3	8p21.2	102	L	-0.45	0.73	3 7 15	105	-	16
8:24773594-26994749	2.22Mb	8p21.2	8p21.2	160	L	-1.09	0.47	1 9 12	190	9	17
8:2701529-27667961	652.43Kb	8p21.2	8p21.1	53	L	-0.41	0.75	2 5 12	34	11	10
8:27678088-33627376	5.95Mb	8p21.1	8p12	369	L	-1.07	0.48	2 1 16 37	284	6	34
8:33665709-34086359	420.65Kb	8p12	8p12	16	G	0.58	1.49	-	11	-	-
8:34129287-34595586	466.30Kb	8p12	8p12	17	G	1.38	2.61	-	35	-	-
8:34615562-35126922	511.36Mb	8p12	8p12	22	G	2.23	4.70	1 1	27	-	2
8:35137186-37228379	2.09Mb	8p12	8p11.23	94	L	-1.07	0.47	2 2	75	-	-
8:37281736-38008581	726.85Kb	8p11.23	8p11.23	49	G	2.13	4.38	9 11	35	-	11
8:38021058-39195522	1.17Mb	8p11.23	8p11.22	91	G	2.52	5.72	1 1 7 16	56	-	14

Gene	Chr	Start	End	Width	Description	Pathways	CTD	CNV
LSM1	8	38,020,838	38,034,248	13.41Kb	LSM1, U6 small nuclear RNA associated	Gene Expression Metabolism of RNA RNA degradation	6	1
BAG4	8	38,034,105	38,070,819	36.72Kb	BCL2-associated athanogene 4	-	11	1
DDHD2	8	38,089,008	38,120,287	31.28Kb	DDHD domain containing 2	-	4	2
PPAPDC1B	8	38,120,649	38,126,738	6.09Kb	phosphatidic acid phosphatase type 2 domain containing 1B	Immune System	14	-
WHSC1L1	8	38,132,560	38,239,790	107.23Kb	Wolf-Hirschhorn syndrome candidate 1-like 1	Lysine degradation	9	5
LETM2	8	38,243,958	38,266,062	22.11Kb	leucine zipper-EF-hand containing transmembrane protein 2	-	7	-
FGFR1	8	38,268,655	38,326,352	57.70Kb	fibroblast growth factor receptor 1	Adherens junction Developmental Biology Disease Immune System MAPK signaling pathway Melanoma Pathways in cancer Prostate cancer Regulation of actin cytoskeleton Signal Transduction	51	-

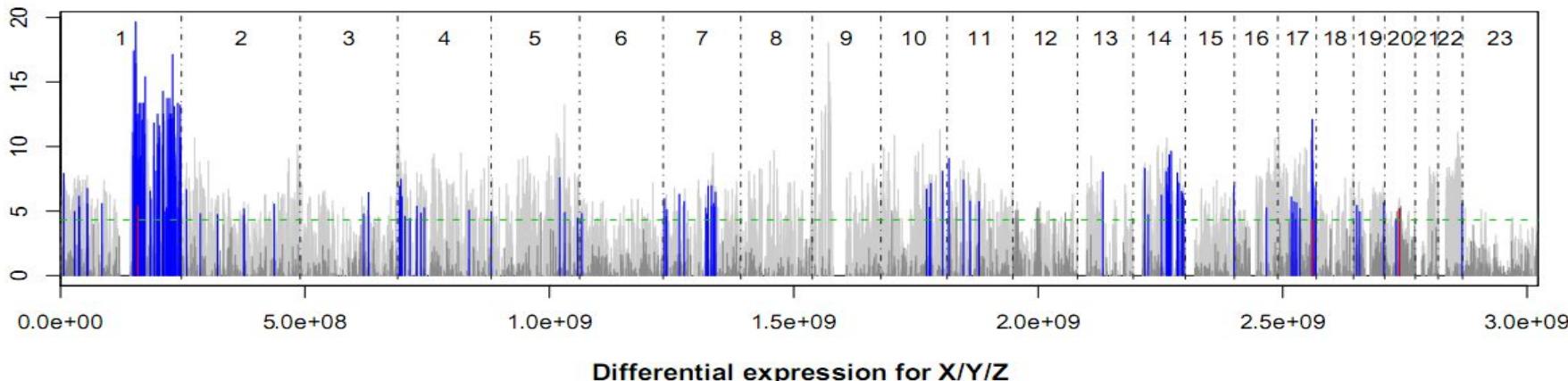
Sample Classes and Clinical Variables

	Hclust A		Hclust B		Hclust C	
Stage at diagnostic		13		16		20
Stage 1	0	0%	9	56%	4	20%
Stage 2	8	62%	4	25%	13	65%
Stage 3	3	23%	3	19%	3	15%
Stage 4	2	15%	0	0%	0	0%
Distant metastasis at 4 years		11		8		11
Positive (at least one)	9	82%	4	50%	3	27%
Negative (zero)	2	18%	4	50%	8	73%
Distant metastasis-free survival		13		16		20
Positive	2	15%	12	75%	15	75%
Negative	11	85%	4	25%	5	25%
Death		13		15		19
Positive	10	77%	6	40%	9	47%
Negative	3	23%	9	60%	10	53%

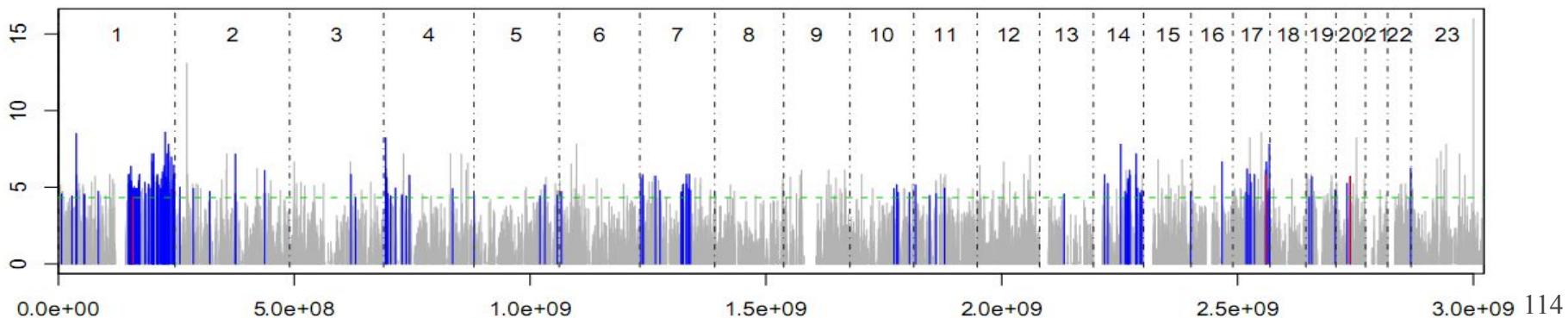
CNA + GEX : Associative Analysis (correlation)

A S

CGH-GE Correlation (spearman)
200 probes identified (197 pro, 3 anti)

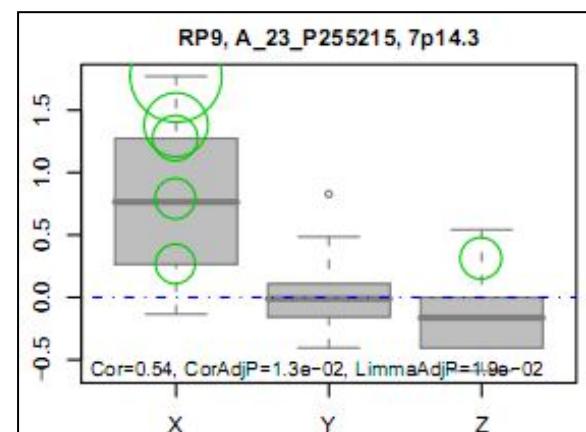
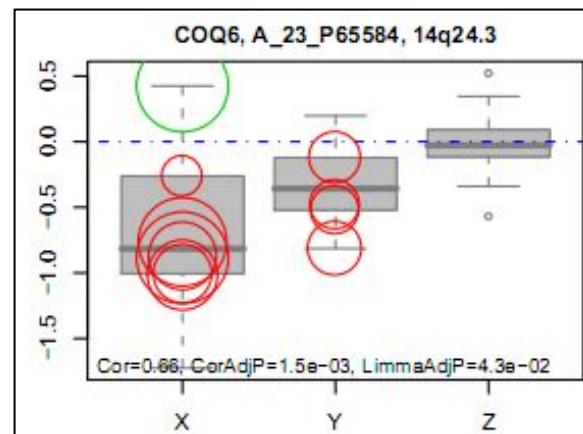
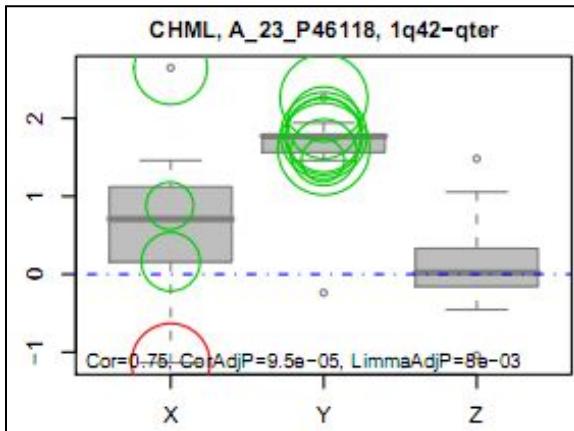


Differential expression for X/Y/Z



CNA + GEX : Associative Analysis (correlation)

A S



CNA + GEX : Integrative Analysis (multiblocs)

A S

RGCCA, SGCCA

(a)

X_1	Gene ₁	...	Gene ₁₅₂₀₁
Patient ₁	0.18		-0.73
Patient ₂	1.15		0.27
:			
Patient ₅₃	1.39		-0.17

X_2	CGH ₁	...	CGH ₁₂₂₉
Patient ₁	0.00		-0.55
Patient ₂	-0.30		0.00
:			
Patient ₅₃	0.00		0.43

Design 1

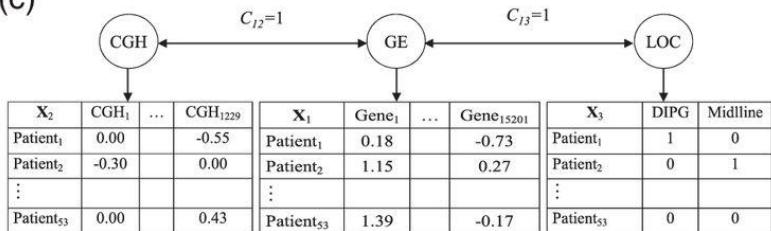
(b)

X_1	Gene ₁	...	Gene ₁₅₂₀₁
Patient ₁	0.18		-0.73
Patient ₂	1.15		0.27
:			
Patient ₅₃	1.39		-0.17

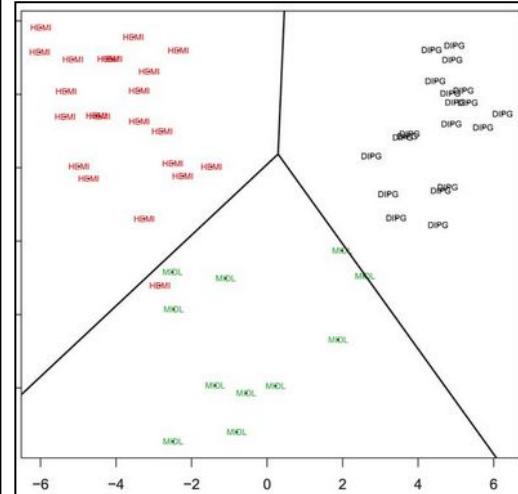
X_2	CGH ₁	...	CGH ₁₂₂₉
Patient ₁	0.00		-0.55
Patient ₂	-0.30		0.00
:			
Patient ₅₃	0.00		0.43

Design 2

(c)



Design 3



CNA + GEX : Integrative Analysis (factorial)

