Bastien JOB
2021.01.18

Genomic Copy-Number Anomalies in Cancer

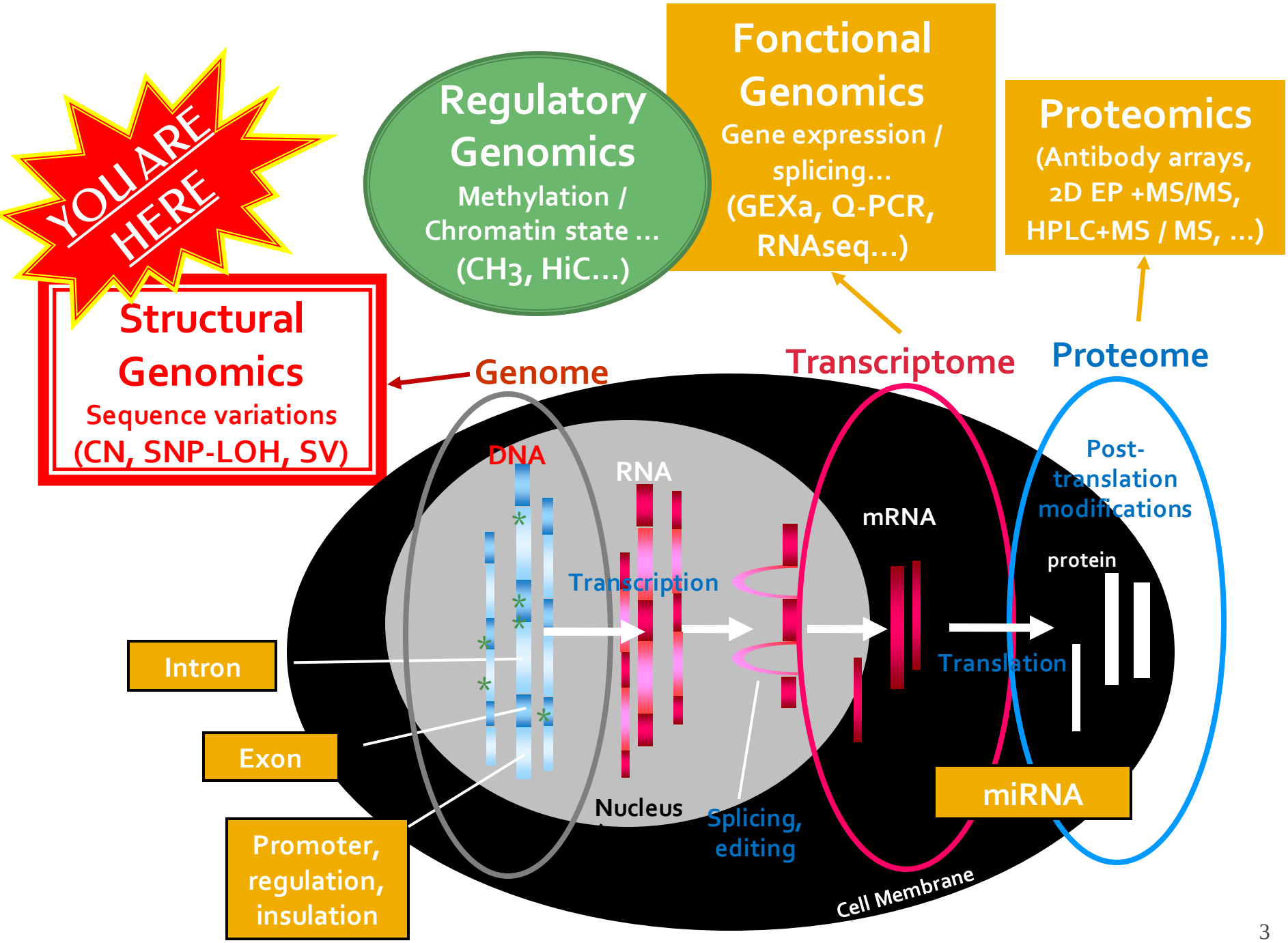
Their study through microarray-based and next-generation sequencing technologies

Inserm

GUSTAVE ROUSSY
CANCER CAMPUS
GRAND PARIS

Introduction

First things first


YOU ARE HERE
Regulatory Genomics
Methylation / Chromatin state … (CH_3, HiC…)
Fonctional Genomics
Gene expression / splicing…
(GEXa, Q-PCR, RNAseq…)
Proteomics
(Antibody arrays, 2D EP +MS/MS, HPLC+MS / MS, …)
Structural Genomics
Sequence variations
(CN, SNP-LOH, SV)
Genome
Transcriptome
Proteome
DNA
RNA
mRNA
Post-translation modifications
protein
Transcription
Translation
Intron
Exon
Promoter, regulation, insulation
Nucleus
Splicing, editing
miRNA
Cell Membrane

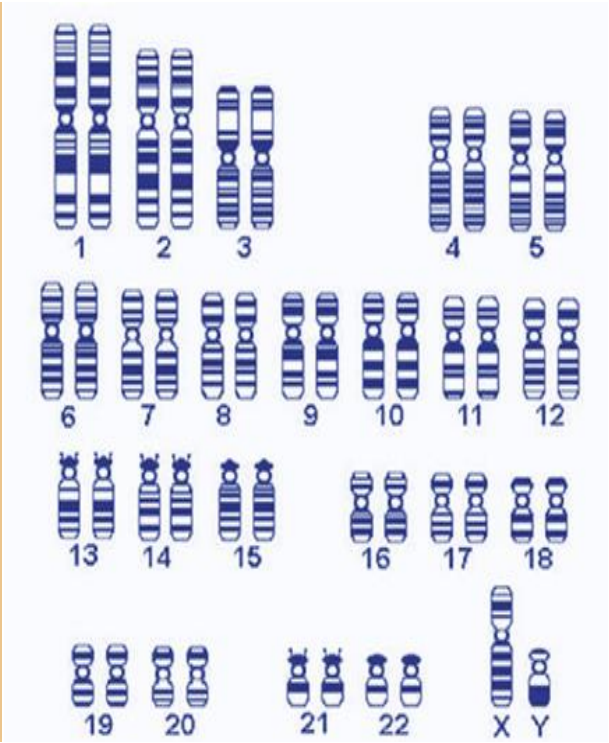
A bit of vocabulary

- "CNV" (Copy Number Variation) includes :
  - "CNA" :
    - Copy Number Anomaly
    - Copy Number Aberration
    - Copy Number Abnormality
  - Polymorphisms (often called … "CNV")

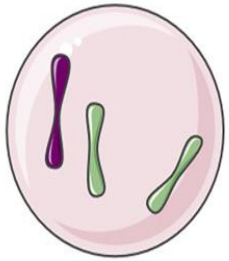
Copy number anomalies and cancer

Kou & al (2020)

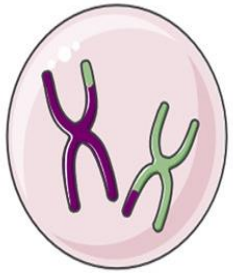
Normal Karyotype



Chromosome Abnormalities



Numerical Chromosome Abnormalities

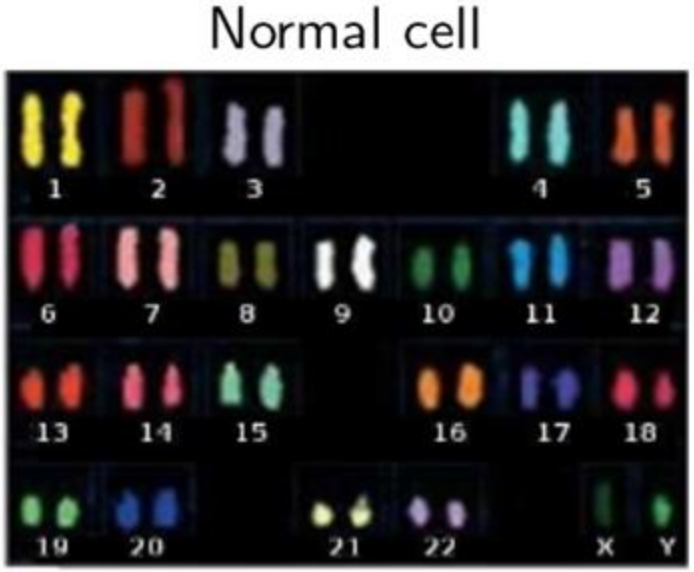


Structural Chromosome Abnormalities

Genomic Instability

Copy number anomalies and cancer

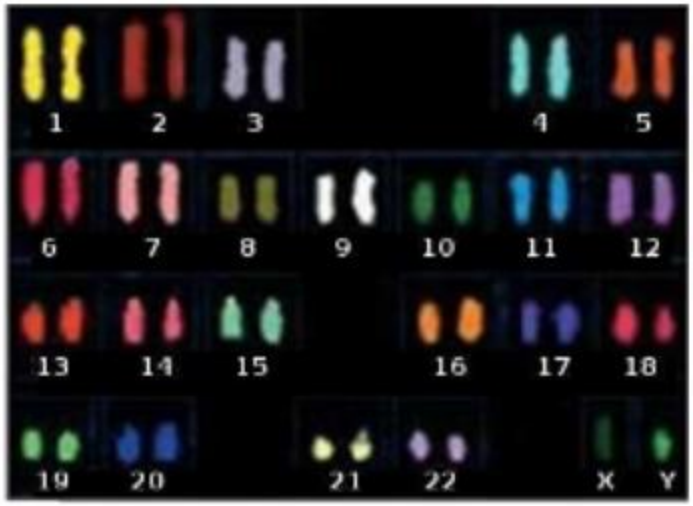
Credit : Philippe HUPPÉ (2008)



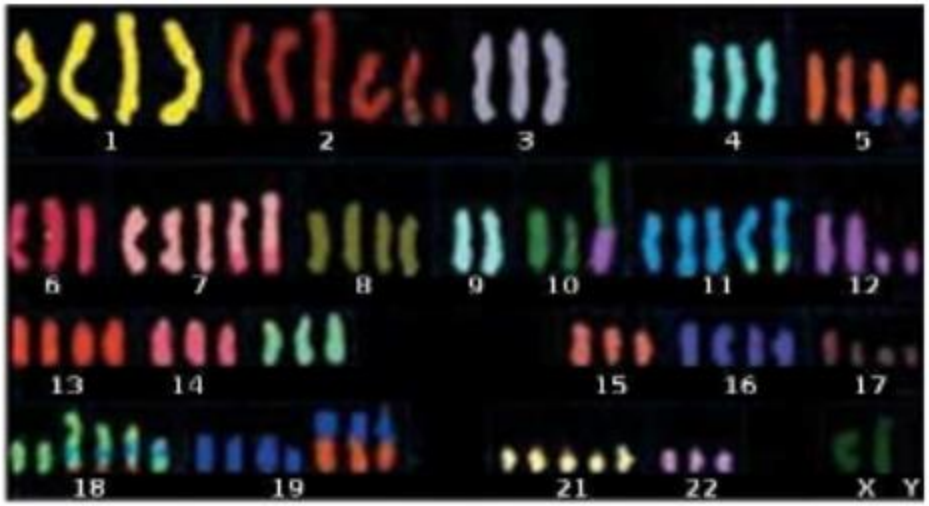
Copy number anomalies and cancer

Credit : Philippe HUPPÉ (2008)

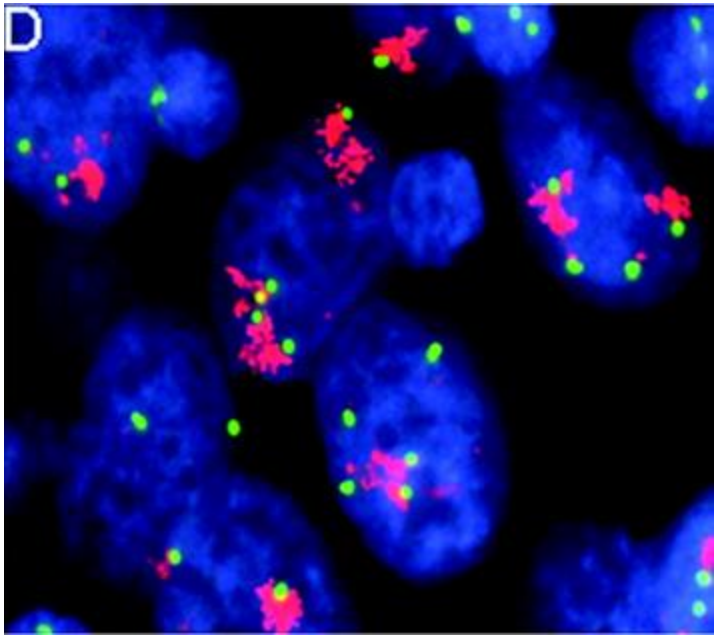
Normal cell



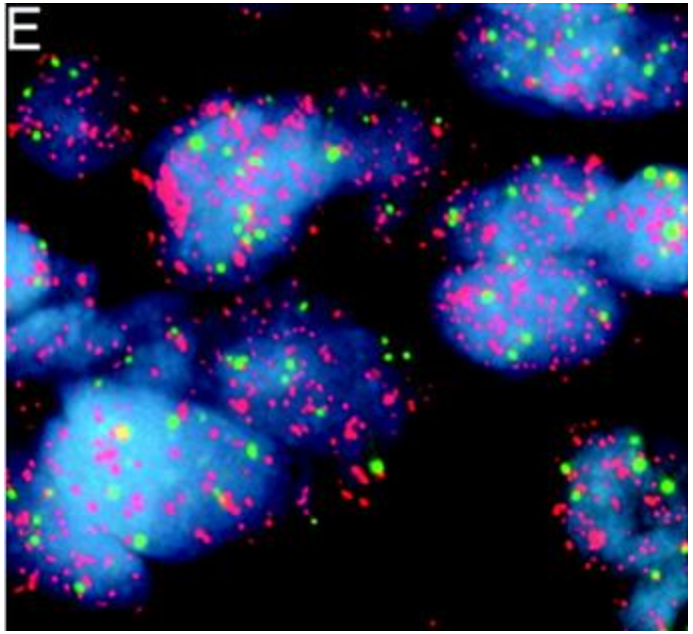
Tumor cell



Chromosomal amplifications



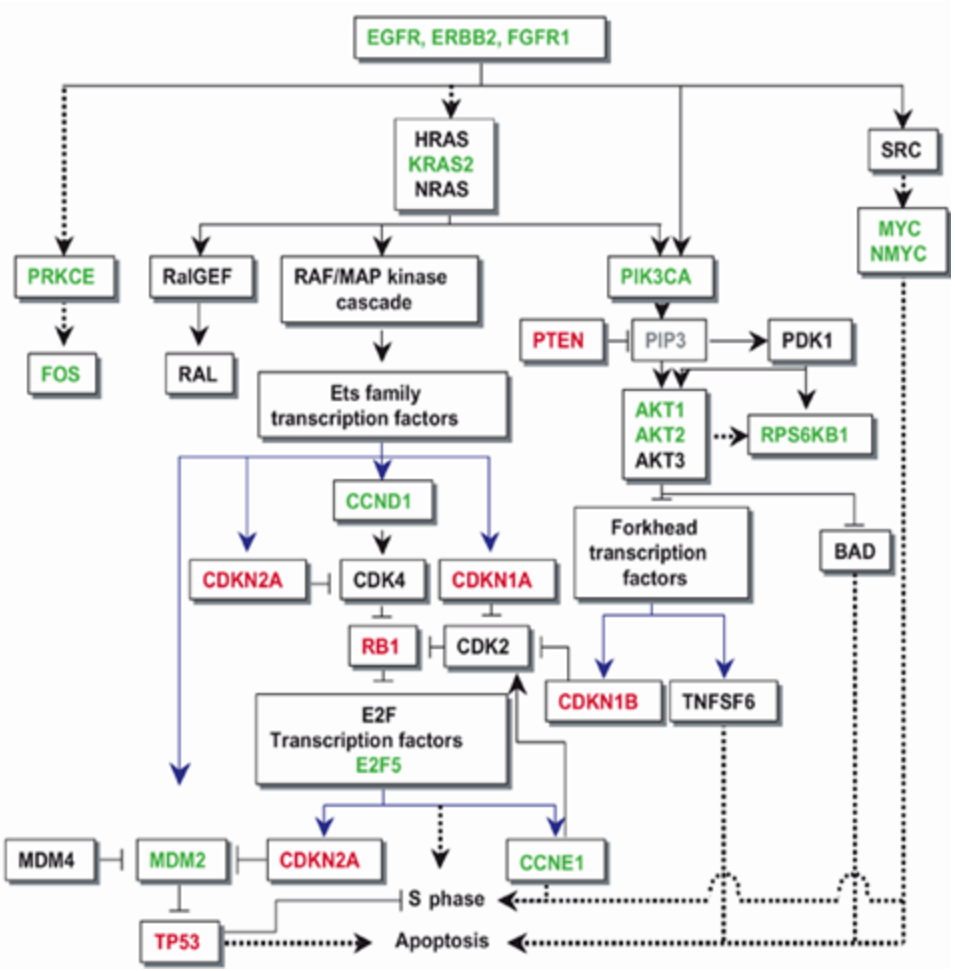
EGFR amplification in lung cancer as **HSR** *(homogeneously stained region)*

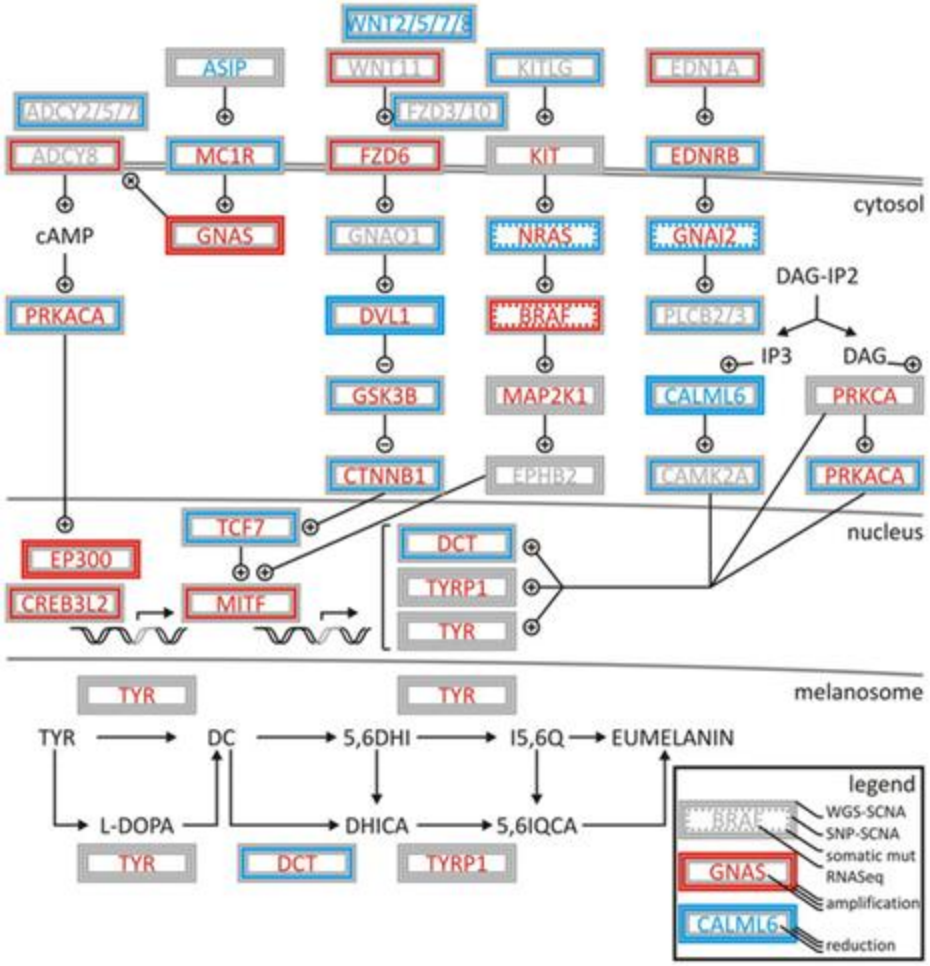


EGFR amplification in lung cancer as **double-minutes**

Varella-Garcia et al, J Clin Pathol 2009

Copy number anomalies and cancer : "Target" genes

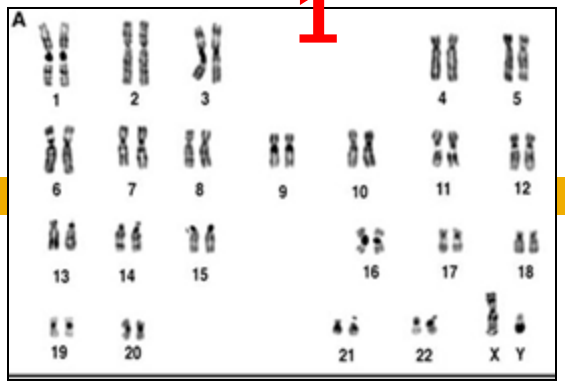




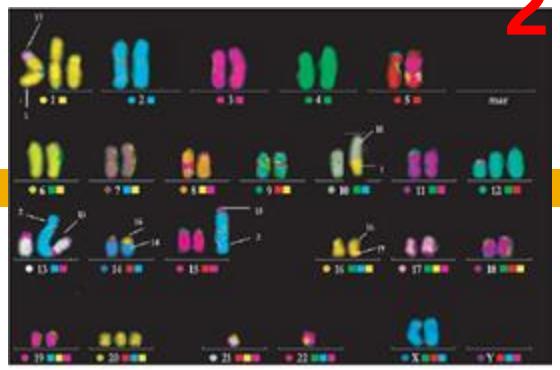
History

A Array

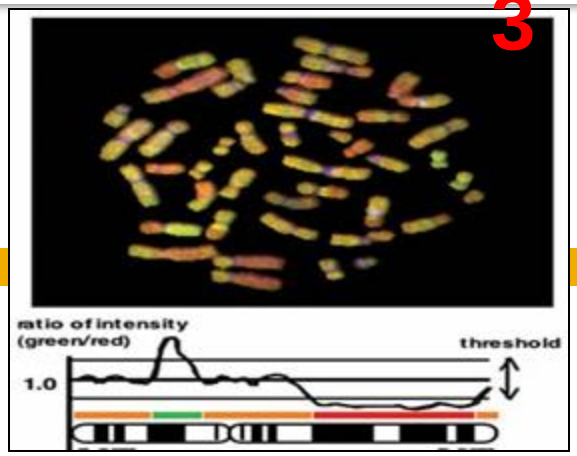
S Sequencing



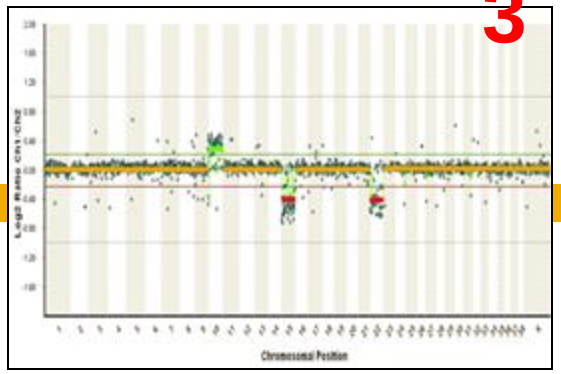
196x : Karyotype



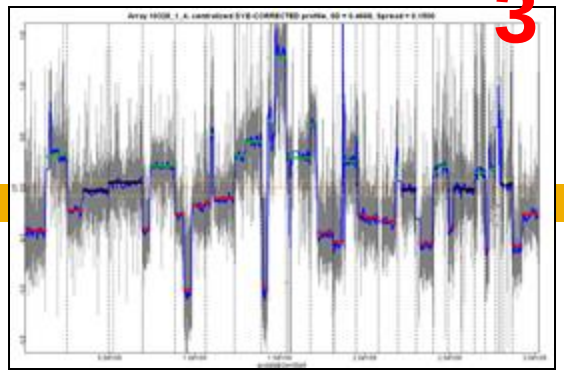
1993 : Spectral karyotyping (SKY)



199x : CGH on chromosomes

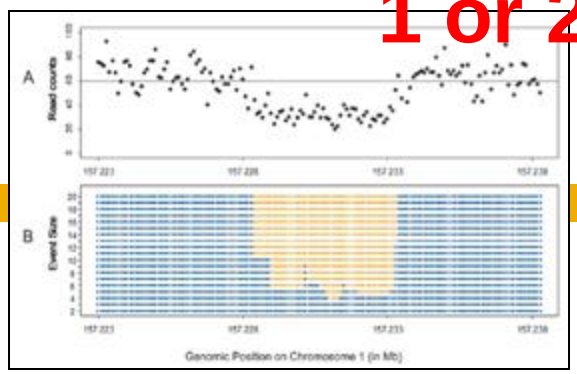


200x : cDNA/BAC-based CGH array



2005 : oligo-based CGH array

A



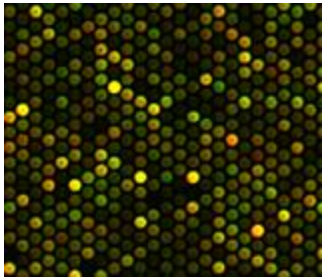
201x : Read-depth from NGS (WGS / WES)

S

A

AGILENT

In-situ printed, Dual color

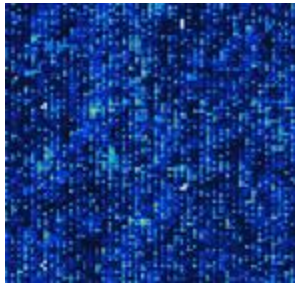


ROCHE-NIMBLEGENE

In-situ printed, Dual color

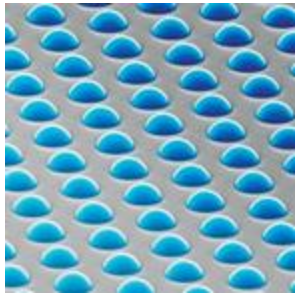
AFFYMETRIX

Photolithographic synthesis, Single color

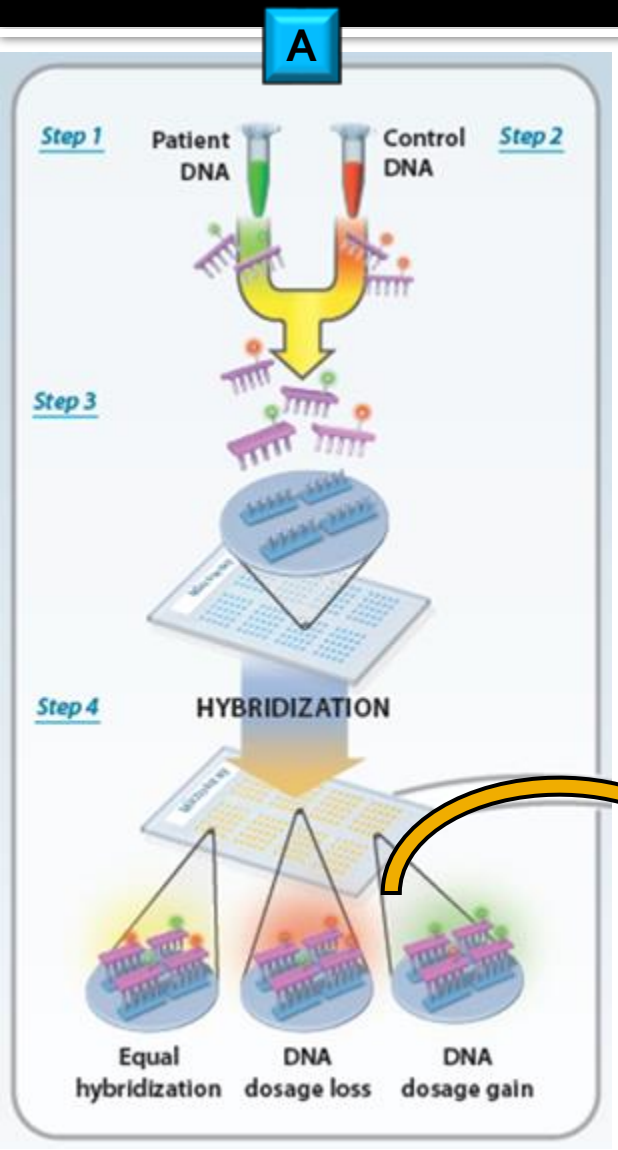


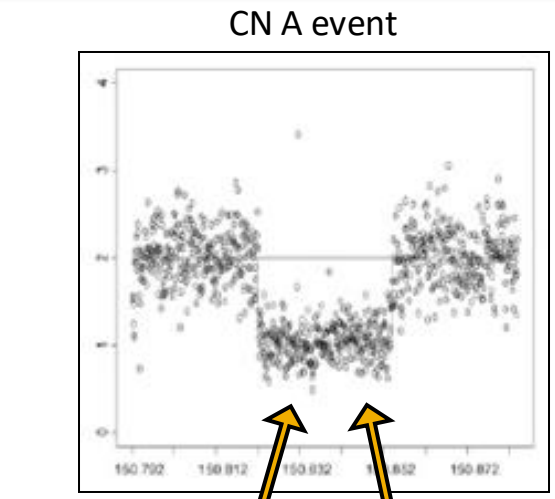
ILLUMINA

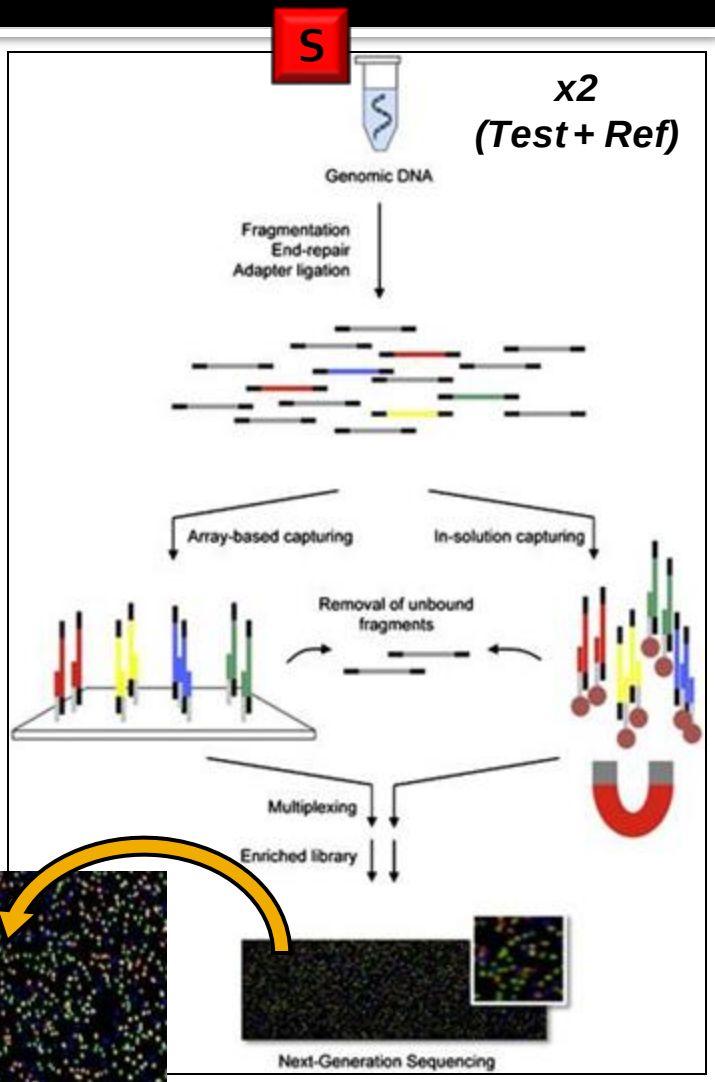
Coated on beads, Dual color



Technical principles

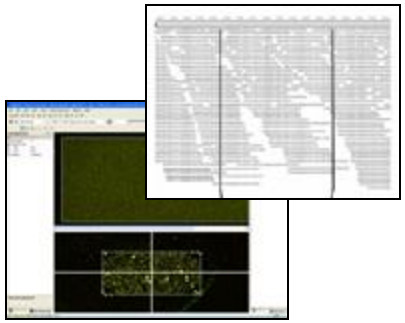




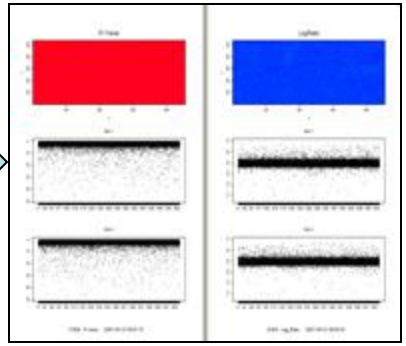


Bioinformatics analysis workflow

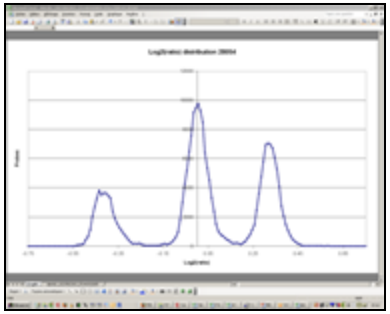
Signals acquisition



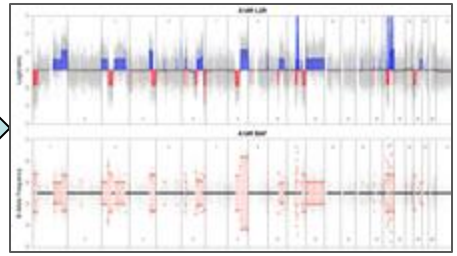
Quality controls



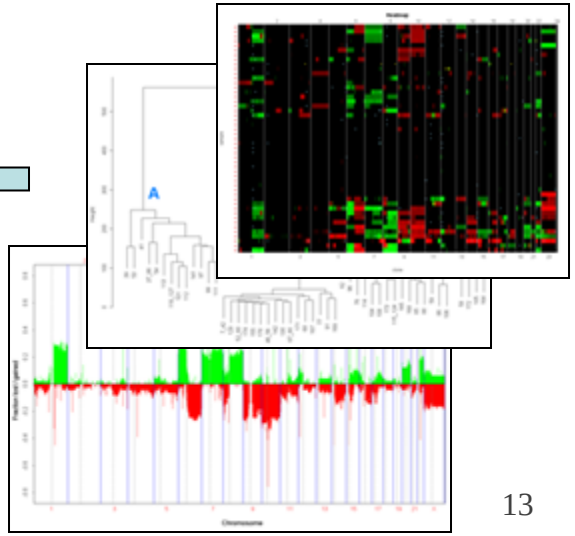
Normalization, centralization



Segmentation, calling

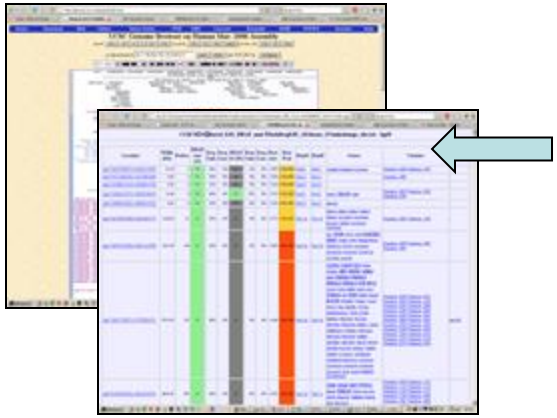


Analysis

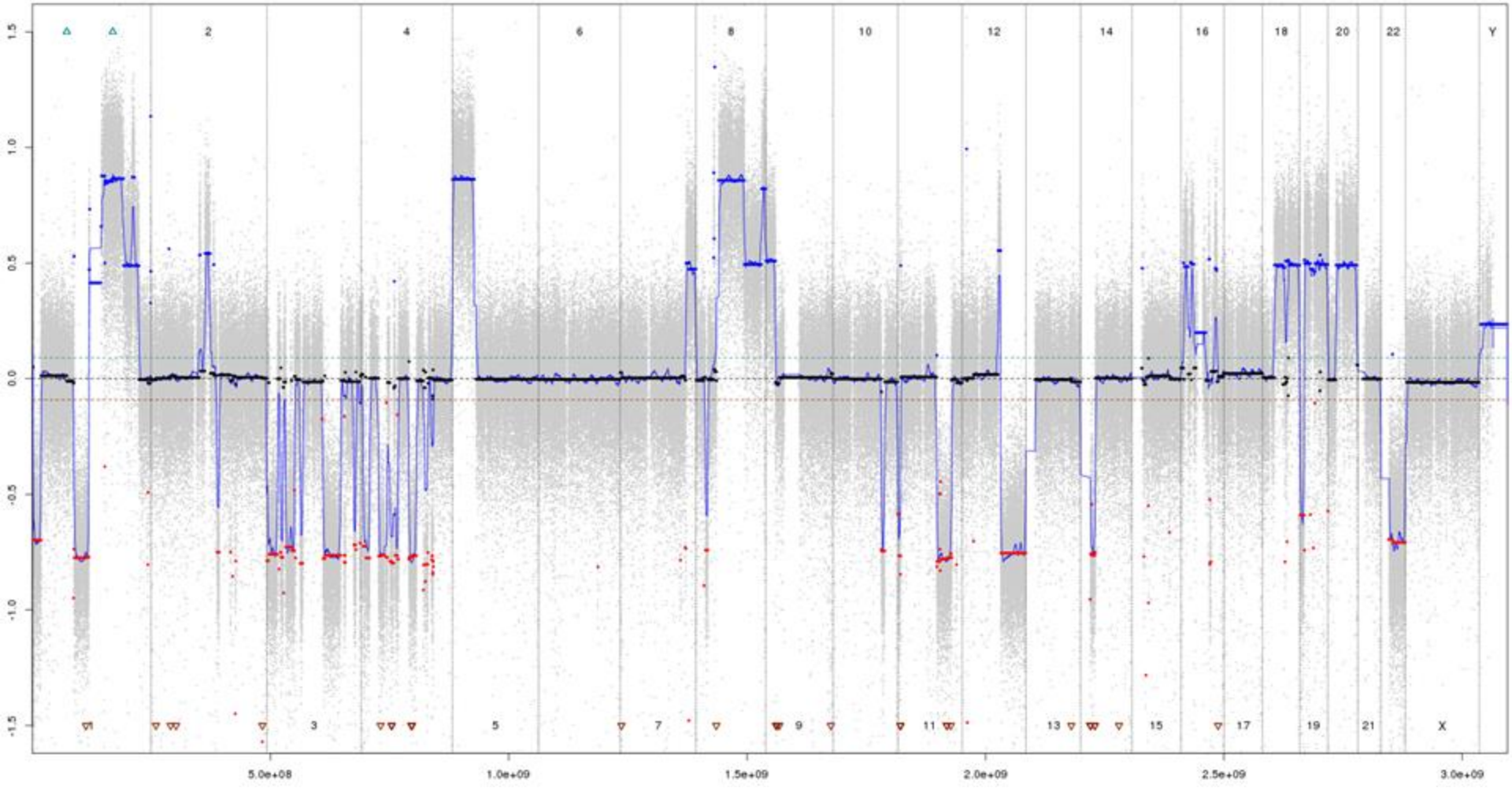


Identification of genomic regions of interest

Annotation



Our aim : a CNA profile



Normalization

Removing / reducing sources of bias

Spatial biases (legacy)

A

S

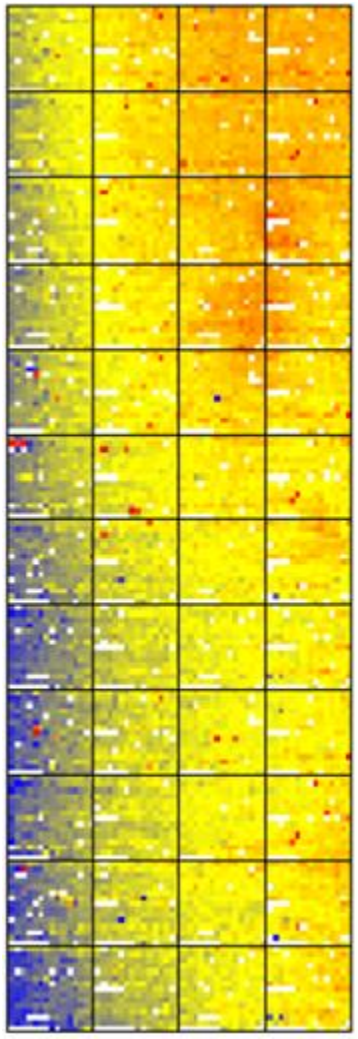
Gradient

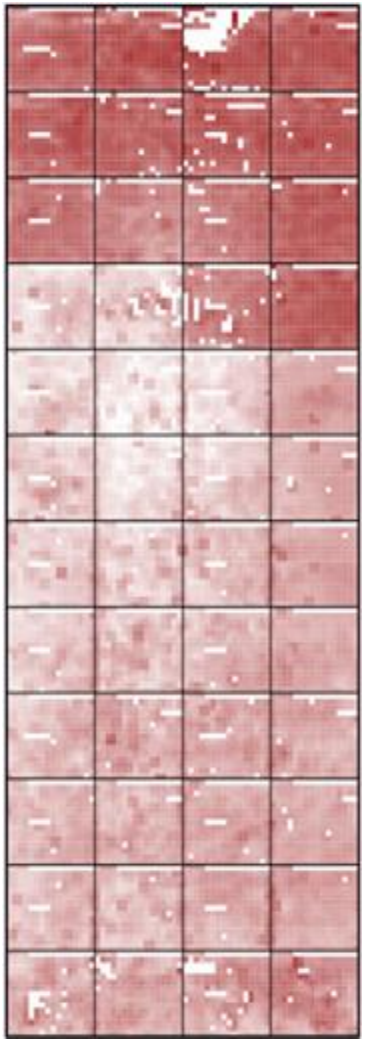
Spotter

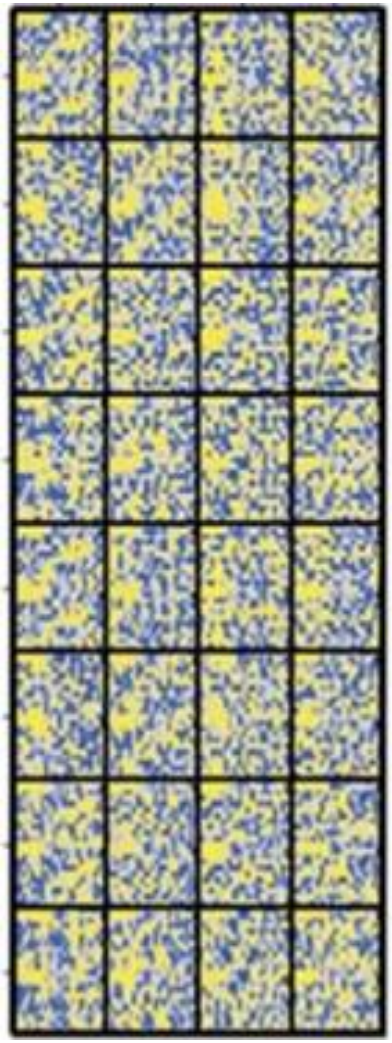
Print-tip

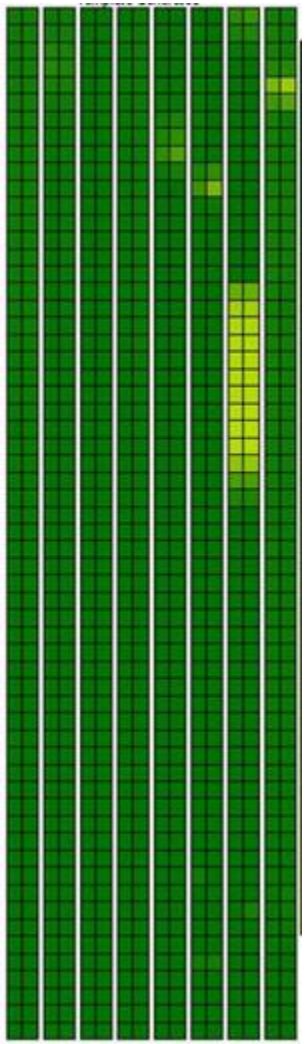
Leak

Density

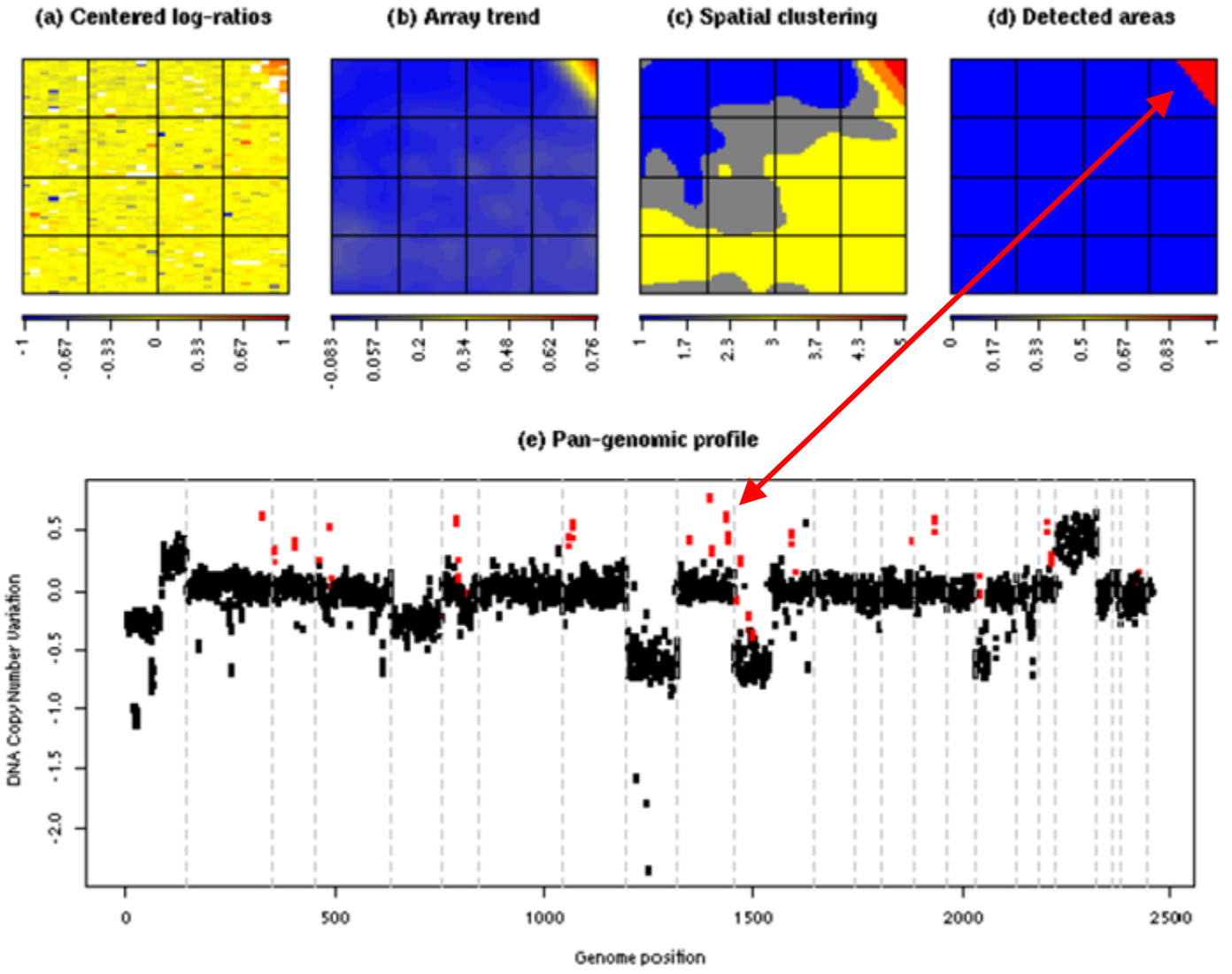




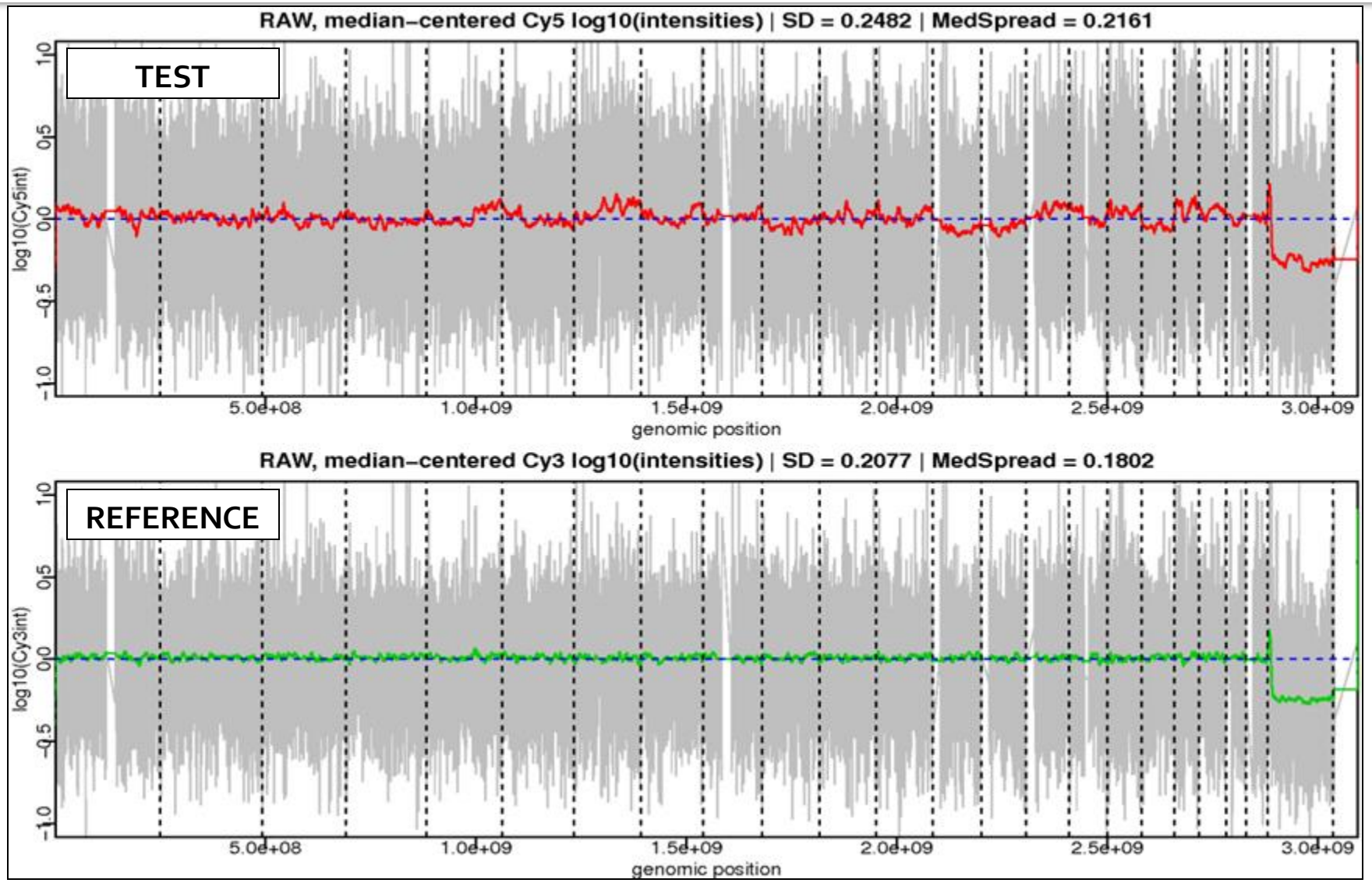




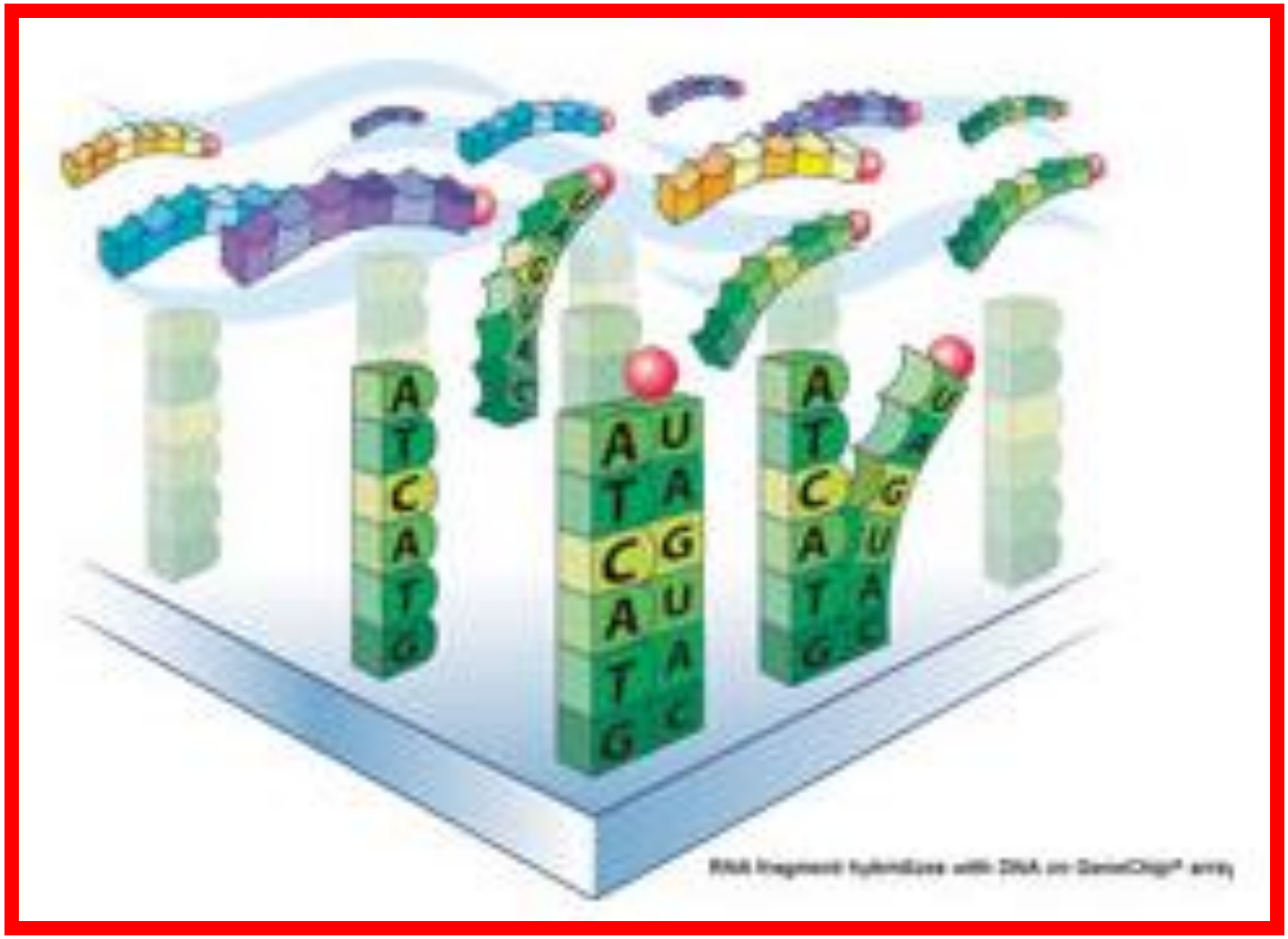
Spatial biases correction



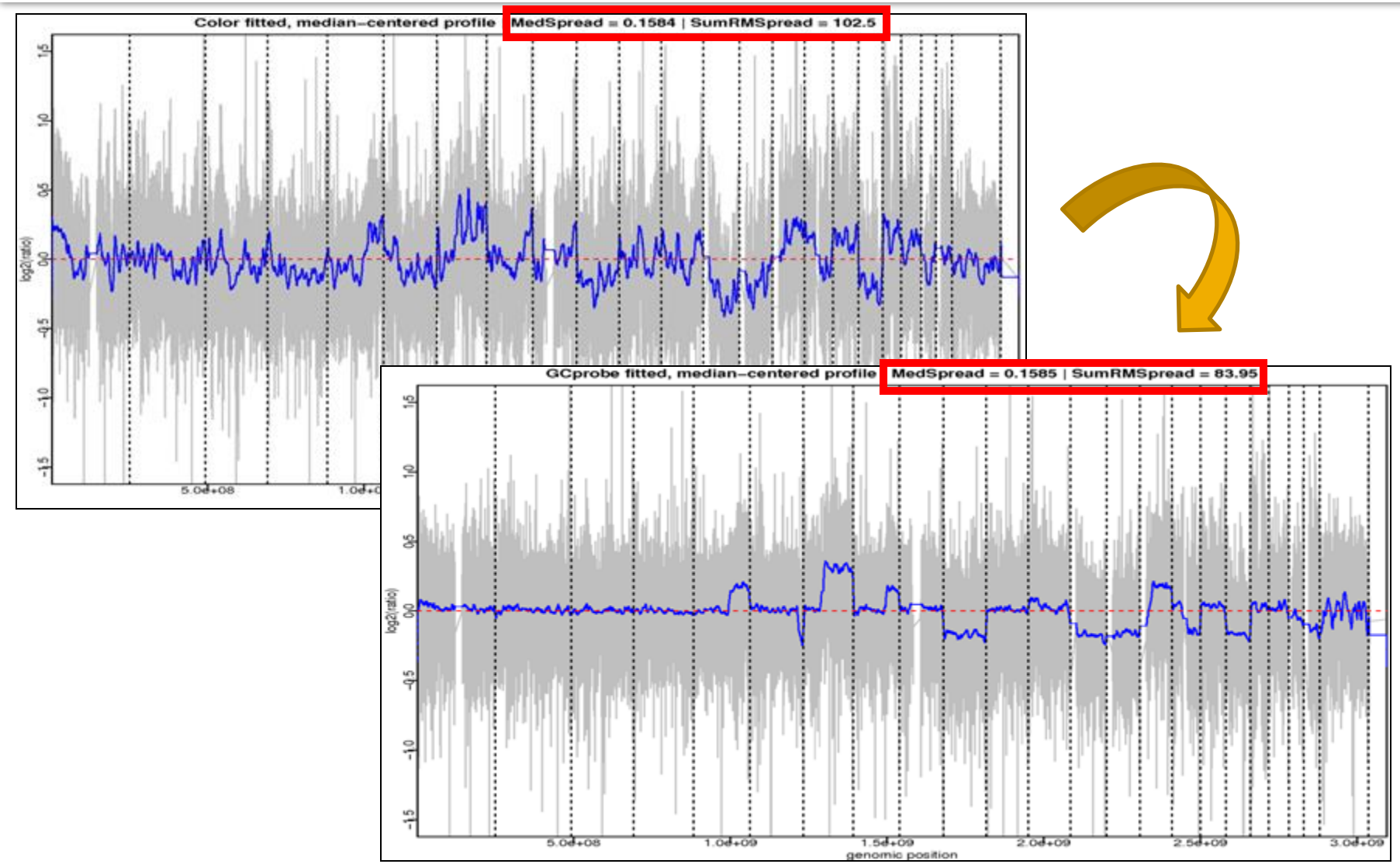
Dye bias / Library bias



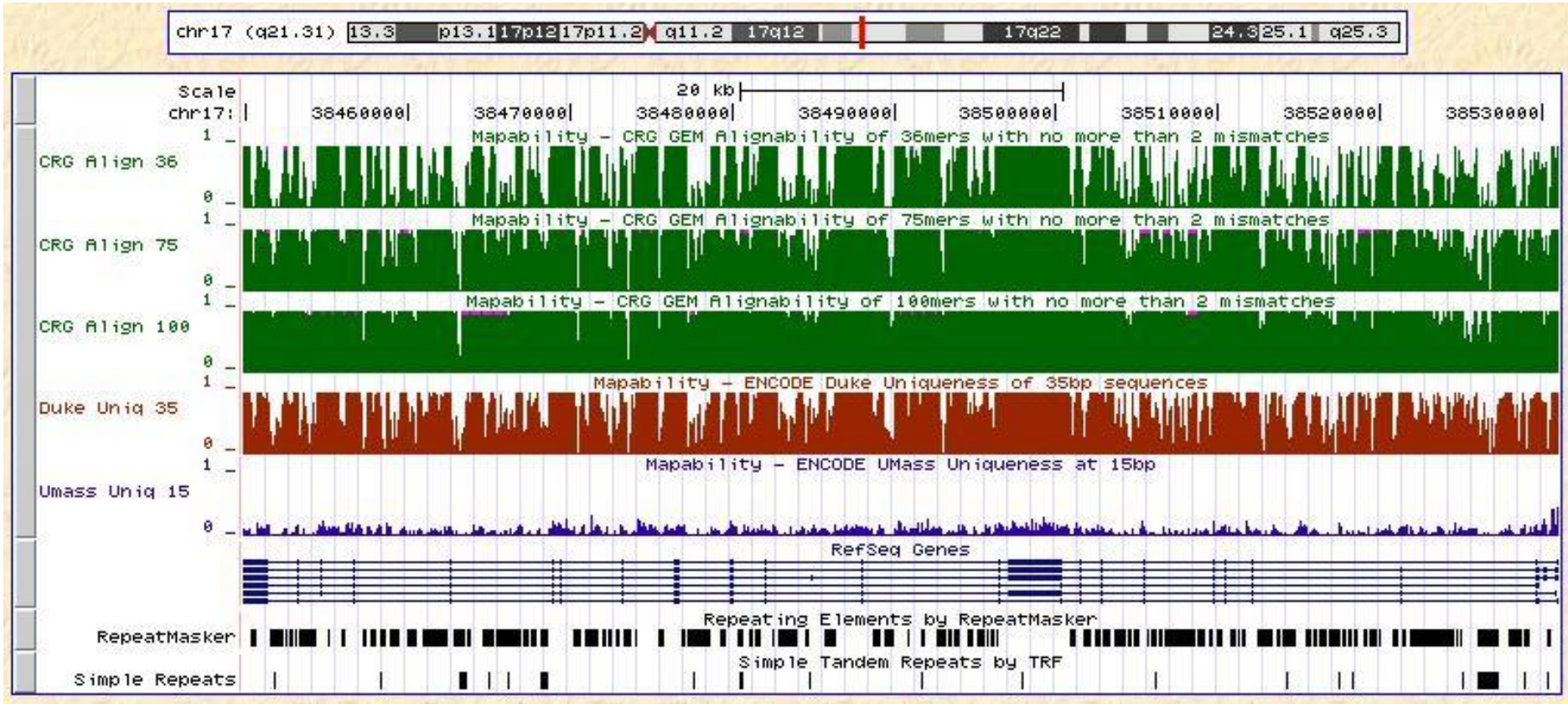
GC-content bias



GC-content bias correction (lowess regression)

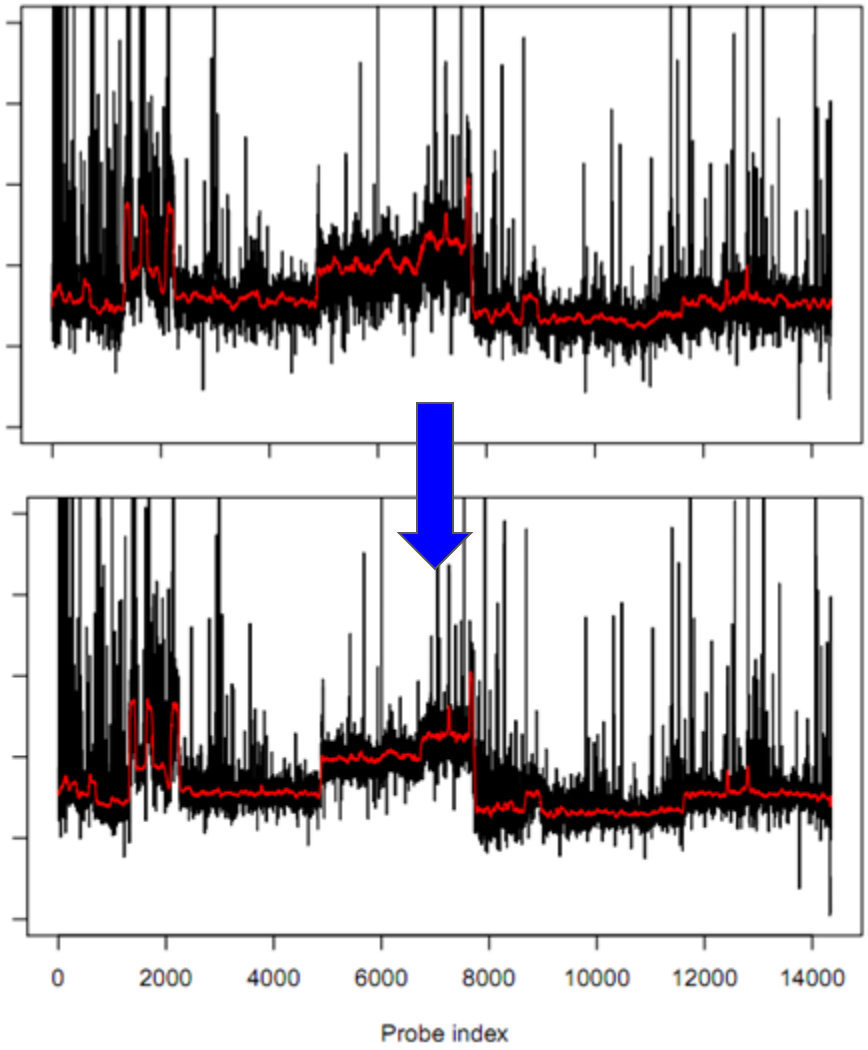


Reads mappability



Ribeca et al., PLoS One, 2012

Residual bias / Wave effect



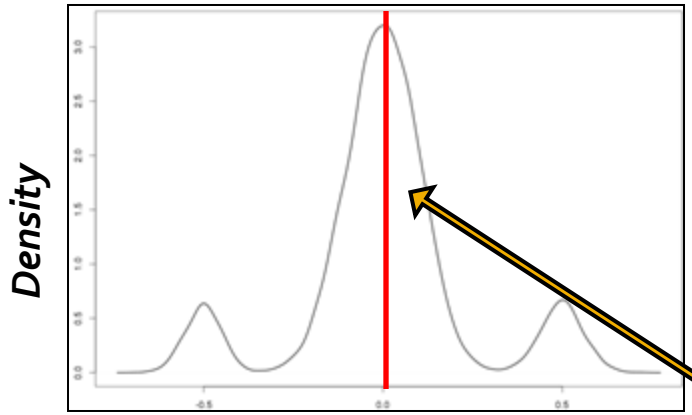
- *Joint* normalization
- Spline smoothing using spatial correlation
- *cghseg* R package
- Efficient with few samples (~5 - 10)
- (very) Slow…

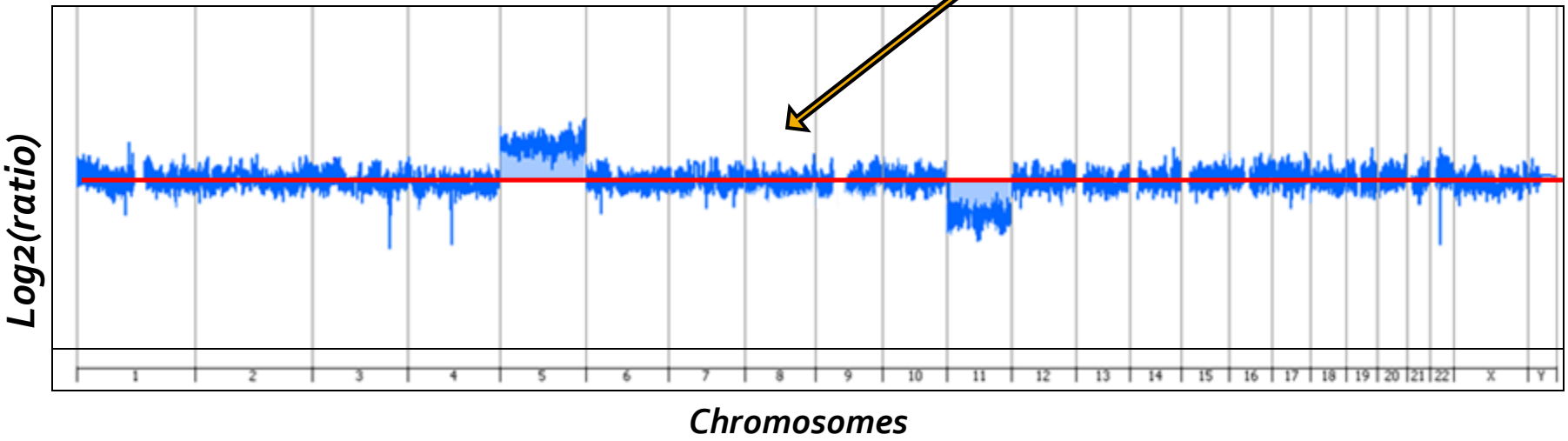
Picard et al., Biostatistics, 2011

Centralization

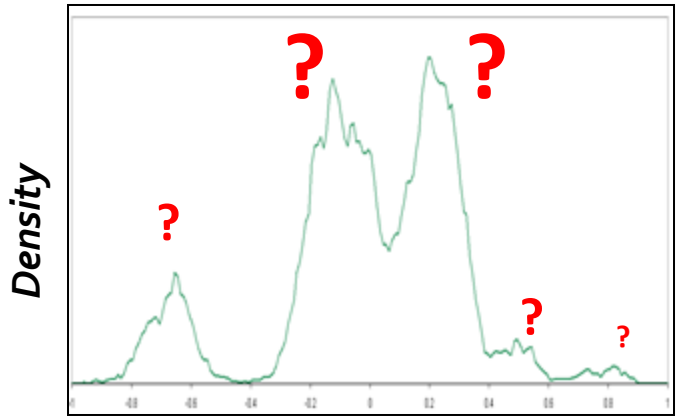
Finding a basis

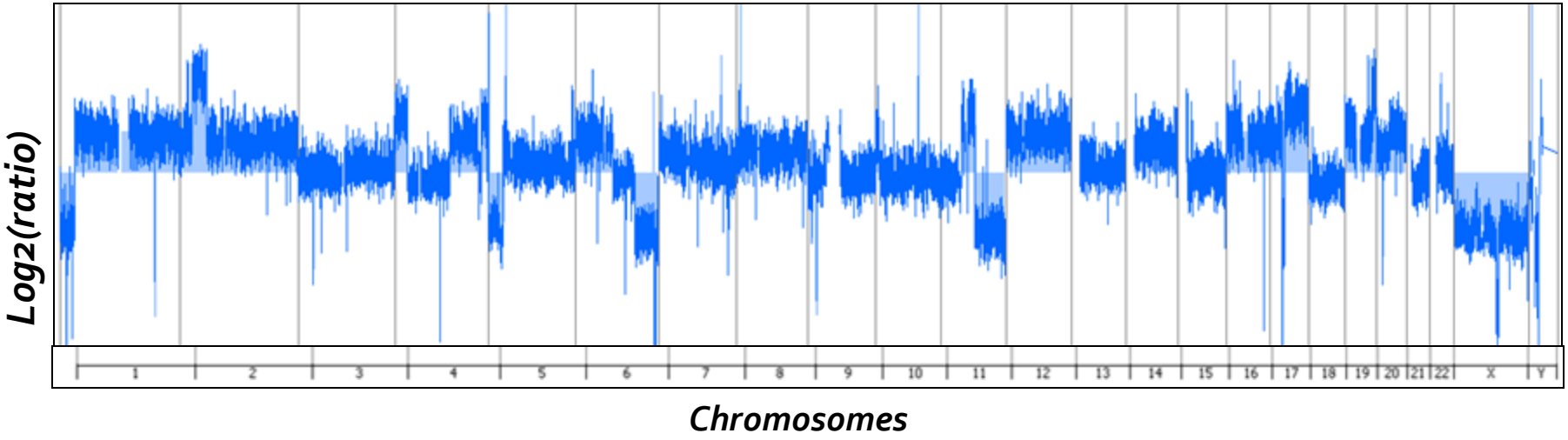
Centralization :
An (synthetic) obvious example



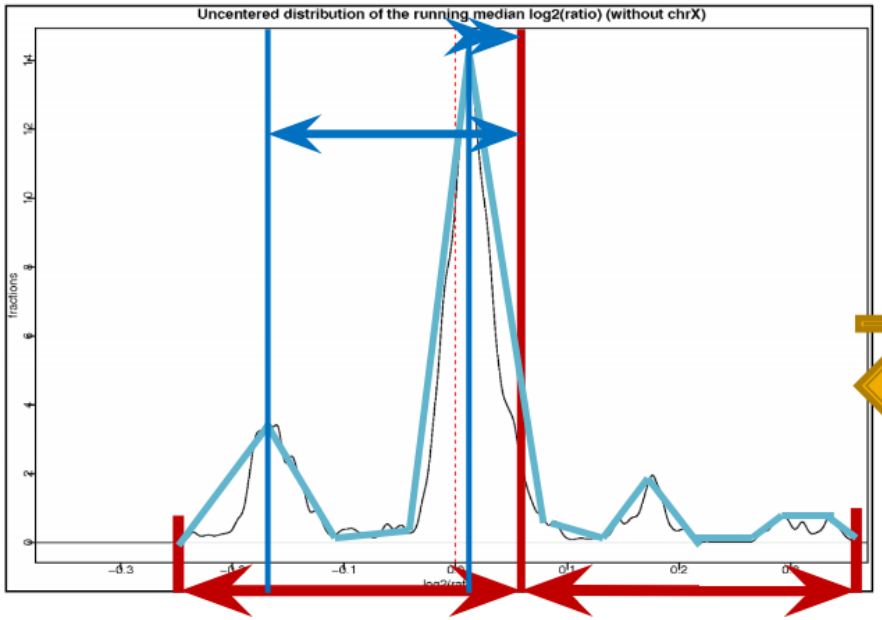


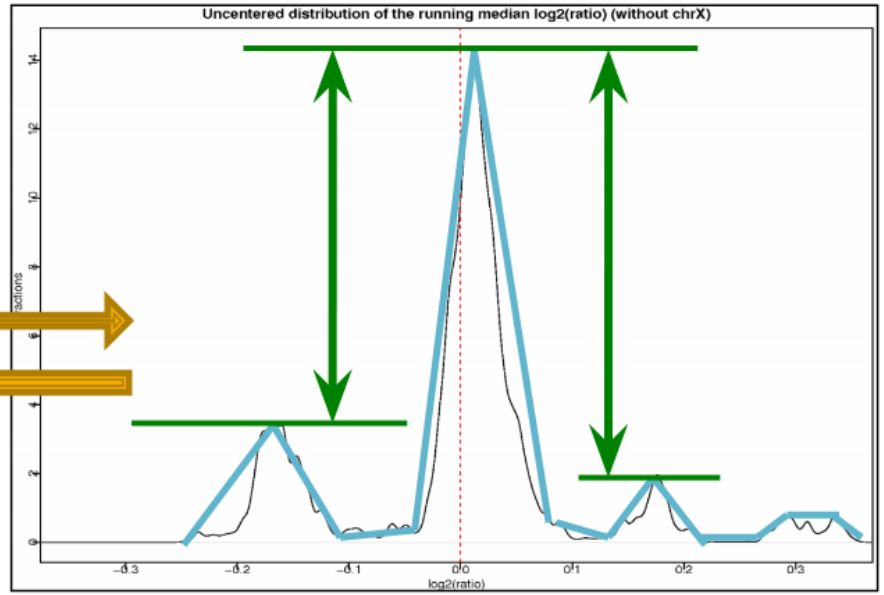
Centralization : A typical cancer example





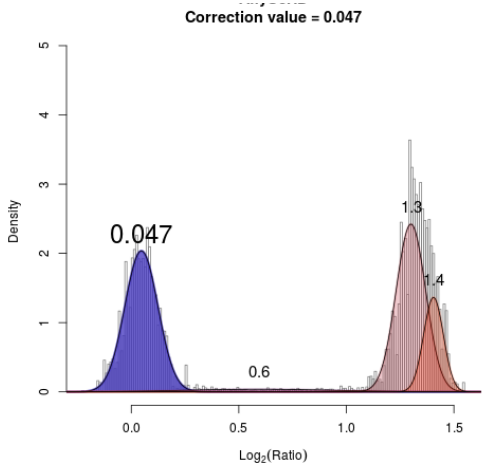
Centralization : Basic considerations





Centrality

Population



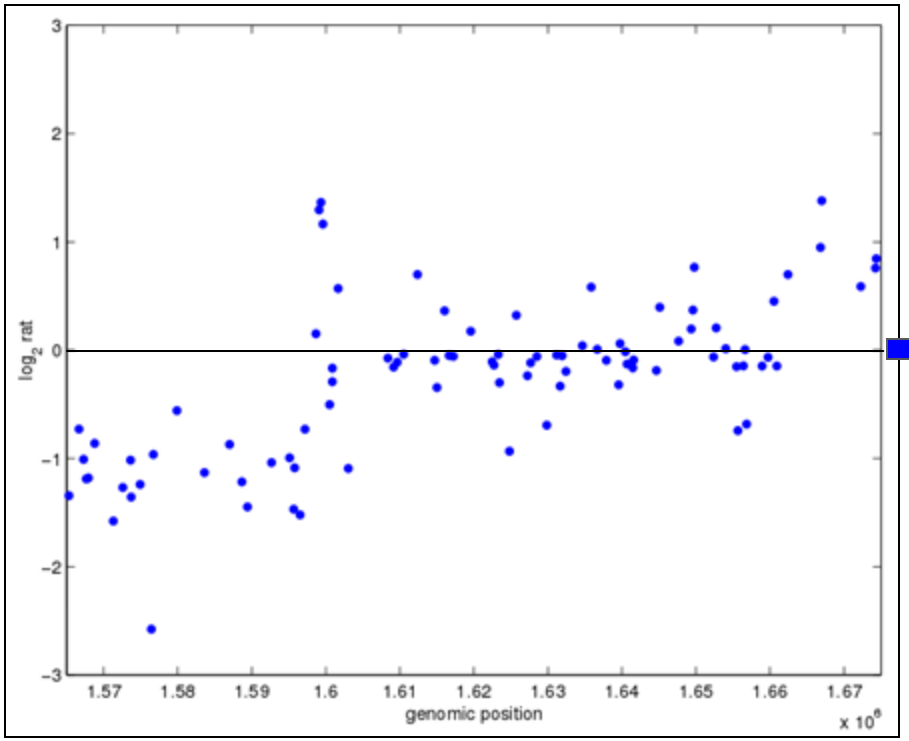
Segmentation

Data reduction

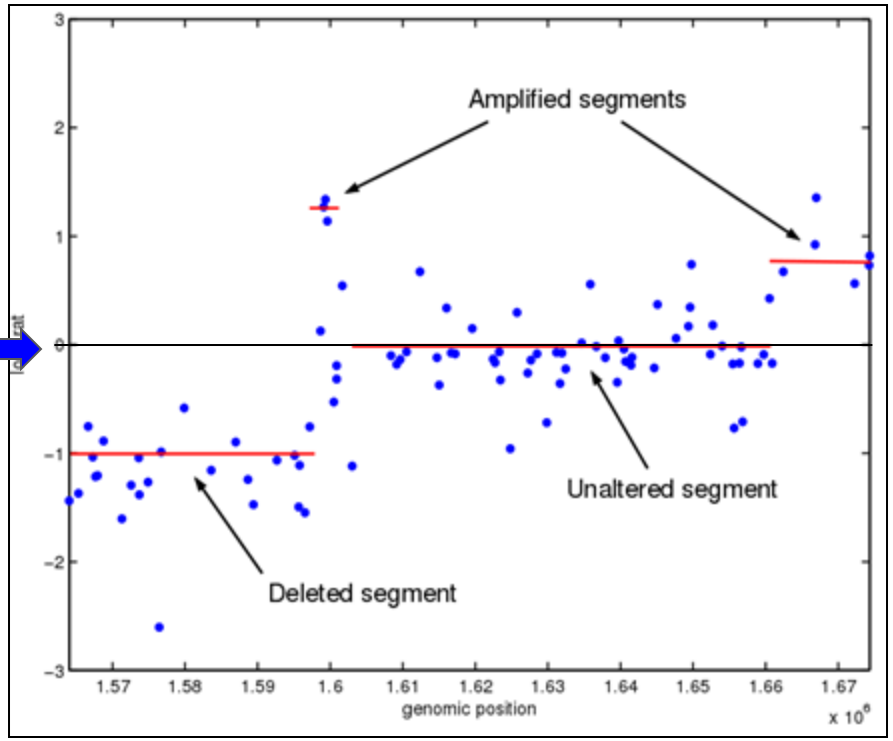
Segmentation : Data reduction

Credit : Stéphane ROBIN

Raw profile



Segmented profile



Numerous, noisy positional measurements

N = 100

Reduced, denoised genomic intervals

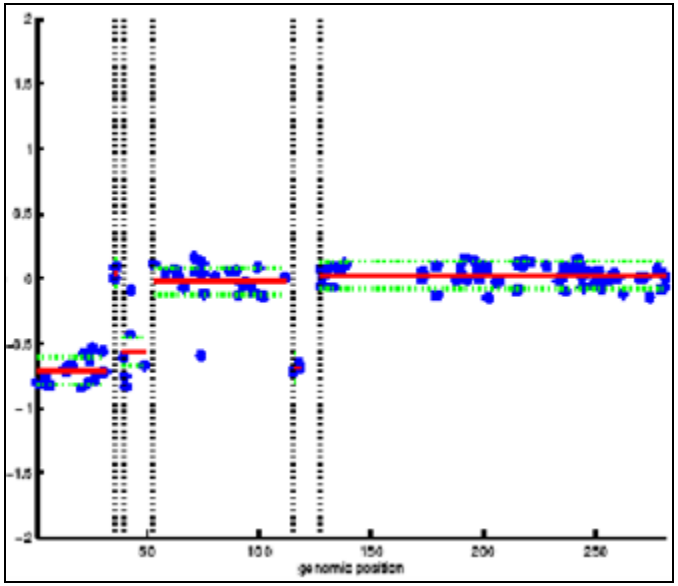
S = 4

Segmentation : Challenging breakpoints detection

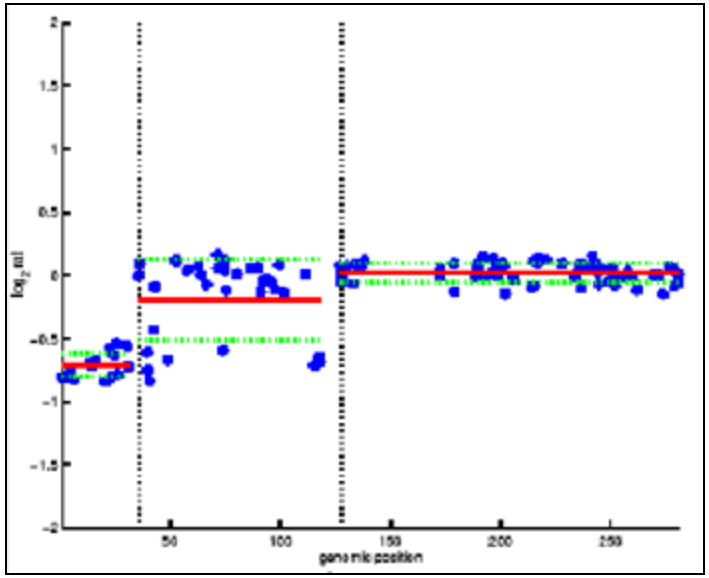
- **Two unknowns for breakpoints :**
  - **Localization**
  - **Quantity**

- **Three families of algorithms :**
  - **Smoothers** (wavelet)
  - **Change-point**
    - **Binary segmentation** (CBS)
    - **Optimal partitionning** (PELT)
  - **HMM modeling** (bioHMM)

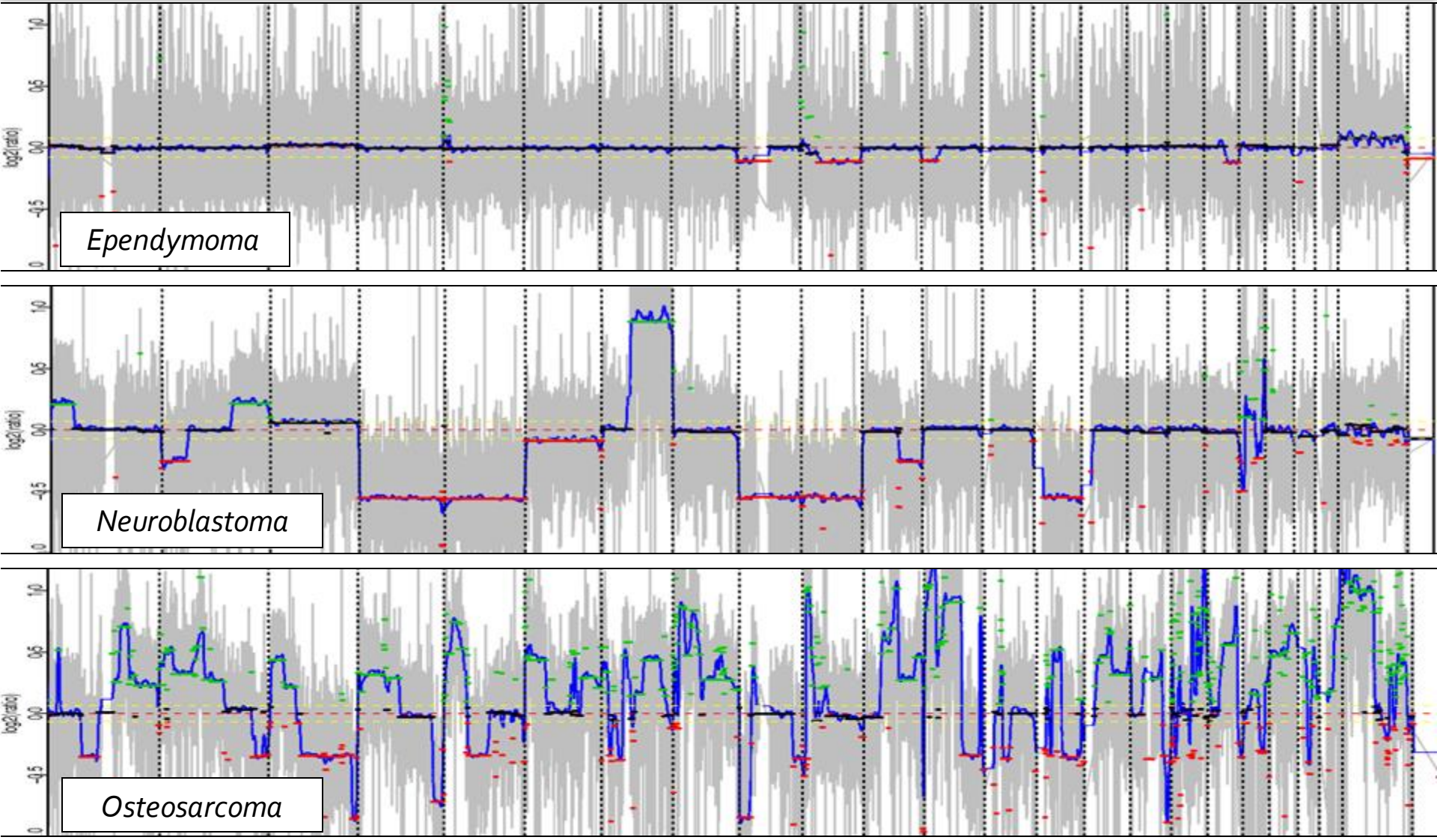
Homoscedastic (m)



Heteroscedastic (m, V)

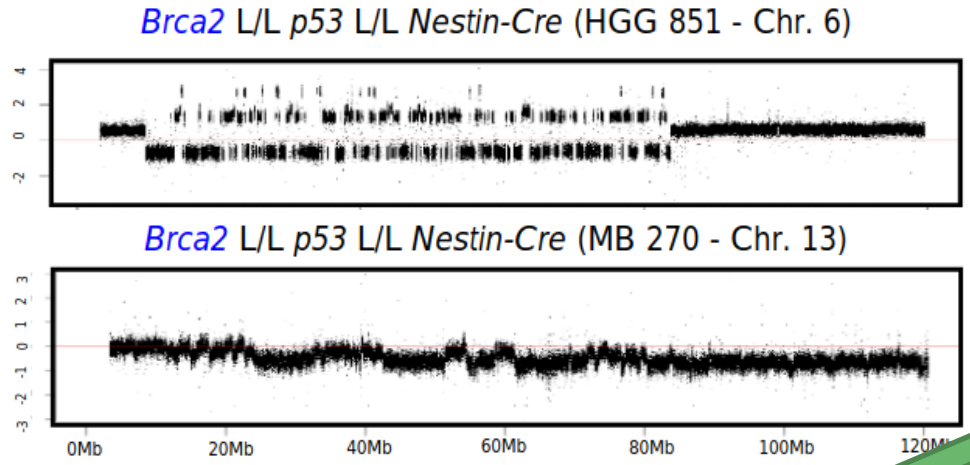


Segmentation :
Tumor profile complexity

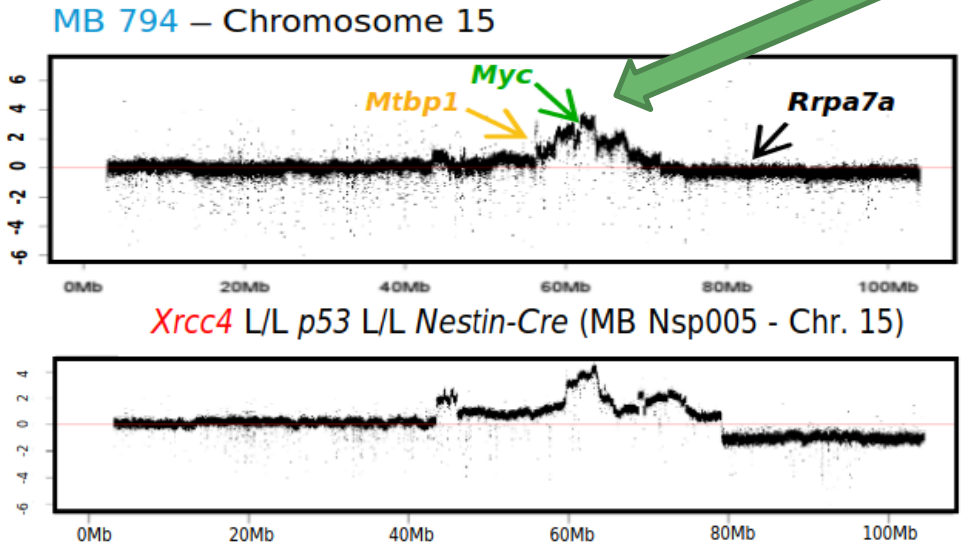


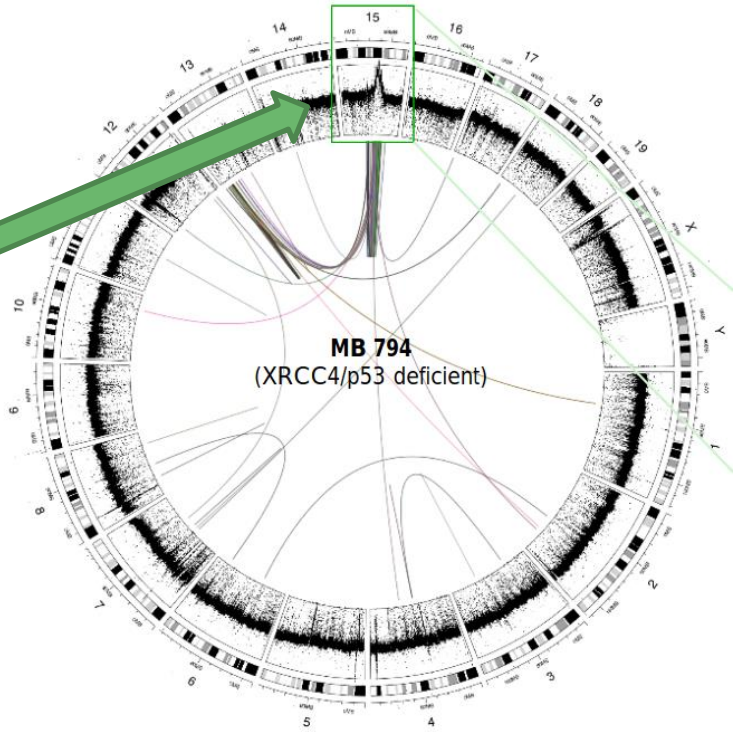
Segmentation : Extreme events *(chromothripsis, chromoanasynthesis)*

CHROMOTHRIPSIS



CHROMOANASYNTHESIS

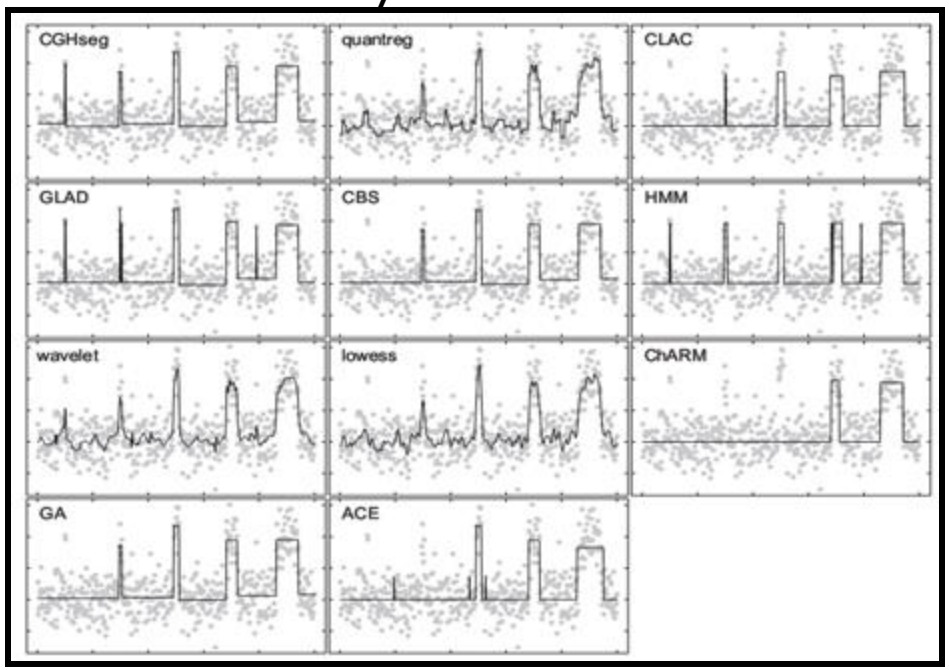




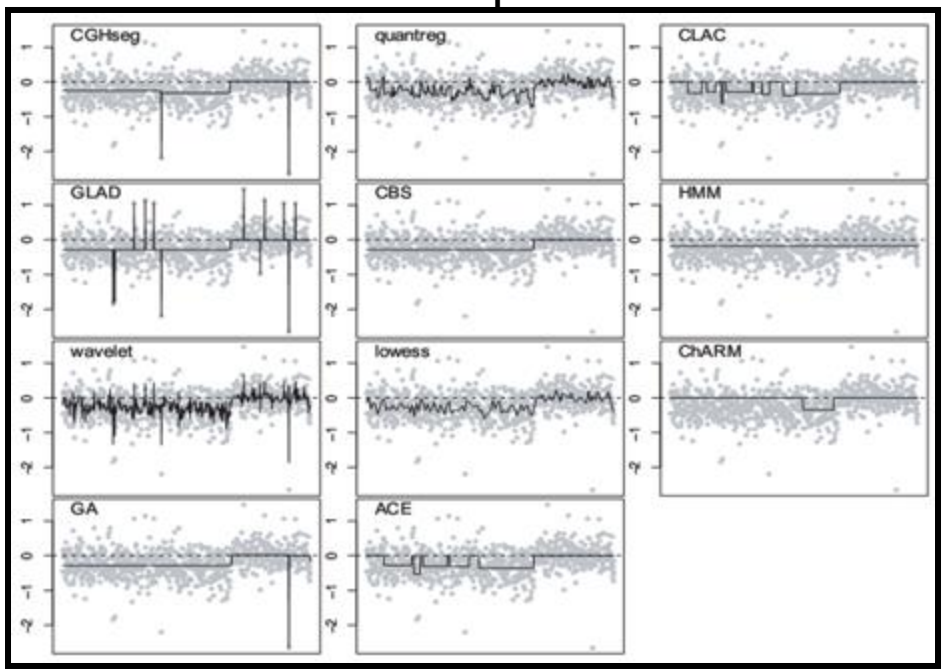
Segmentation : Several methods available

Lai et al. 2005

Synthetic data



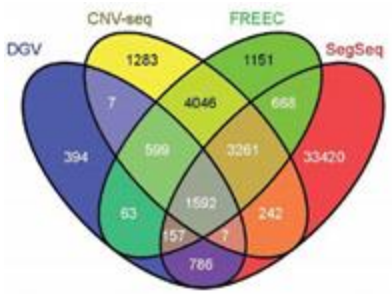
Tumoral profile

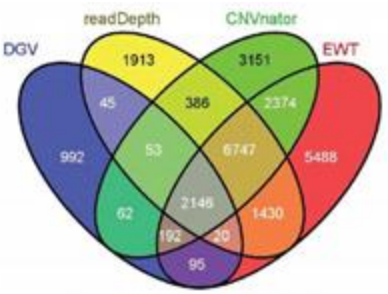


Algorithm	Tool	Autor	Publication Year	Model type	Criterion	K choice
Dynamic programming	CGH-plotter	Autio	2003	Other	Least-square	Ad hoc
Genetic algorithm	*NA*	Jong	2003	Heteroscedastic	Max likelihood	Penalized
EM	aCGH-HMM	Fridlyand	2004	HMM	Max likelihood	Penalized
Adaptative smoothing	GLAD	Hupé	2004	Homoscedastic	Max likelihood	Penalized
CBS	DNAcopy	Olshen	2004	Homoscedastic	Partial sums / perm°	Ad hoc
Dynamic programming	CGHseg	Picard	2005	Homoscedastic	Max likelihood	Penalized

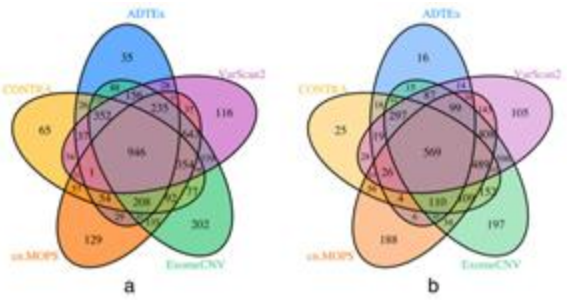
Segmentation : Several methods available

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

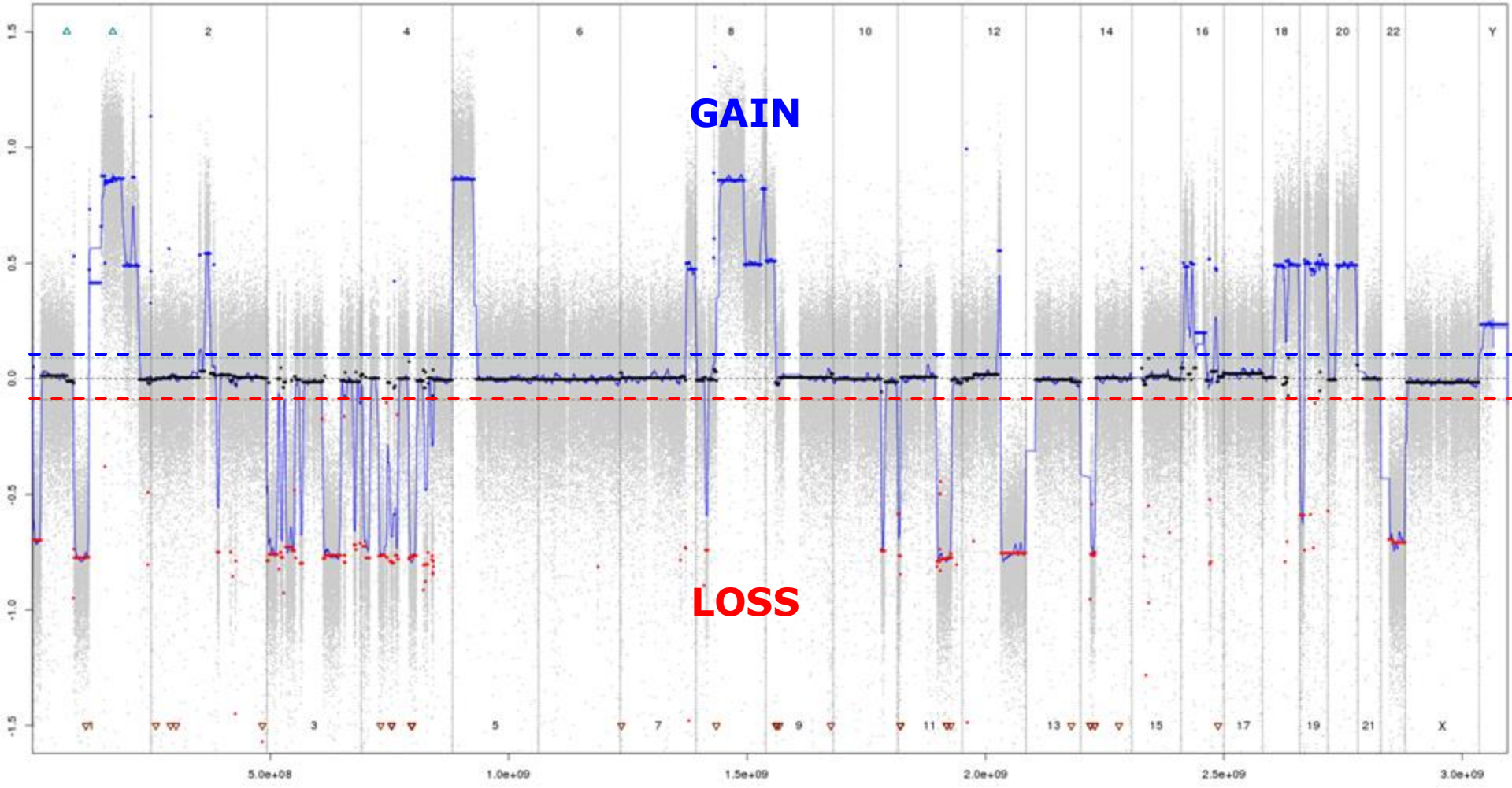


Zare et al. BMC Bioinformatics 2017

Calling

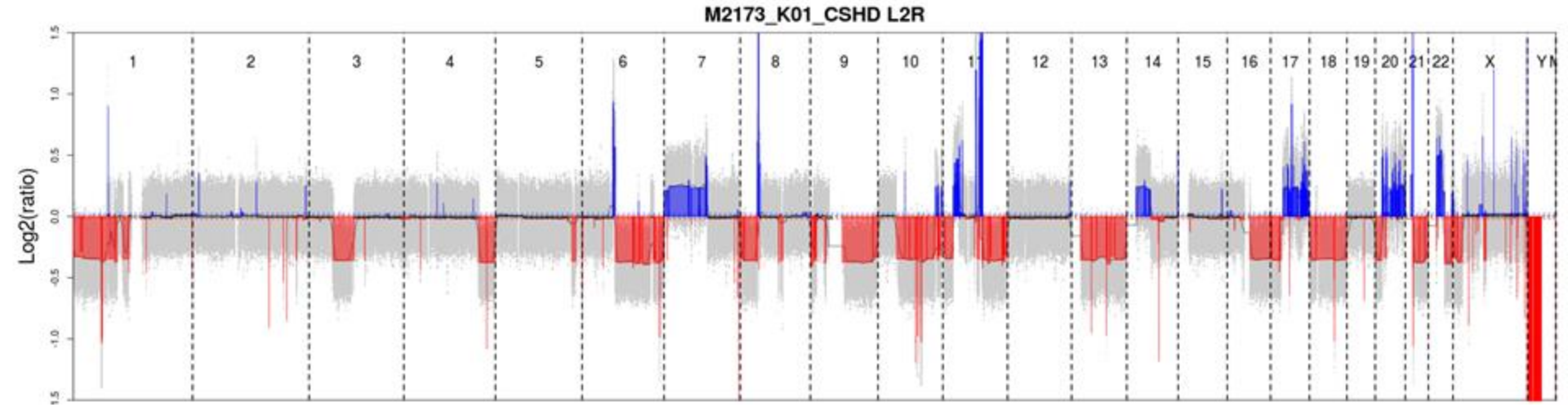
Who's who

Sample profile

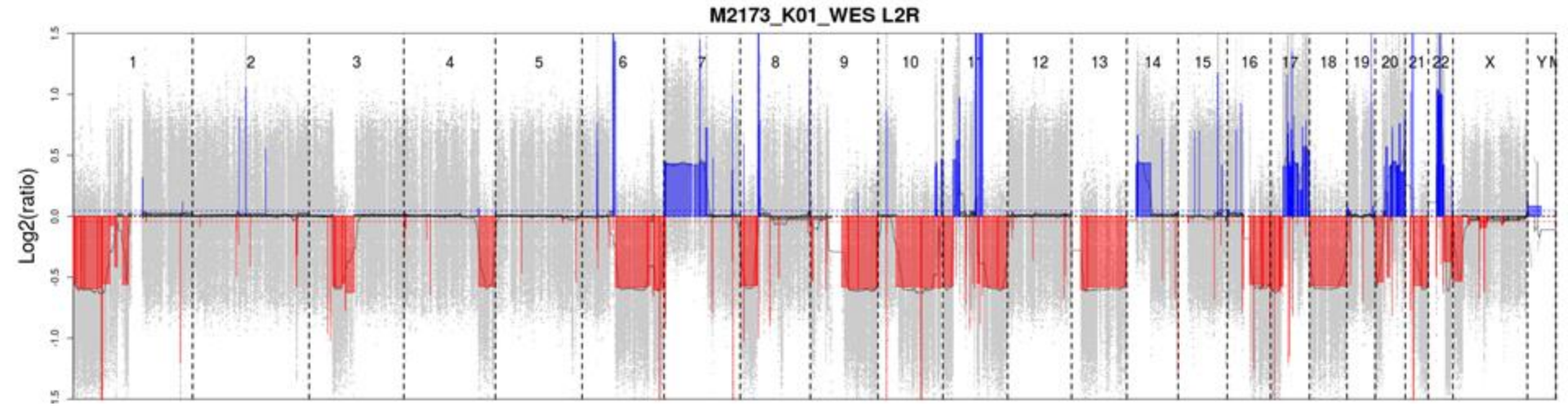


NGS or microarray ?

NGS versus microarrays



MICROARRAY (Affymetrix CytoScan HD)



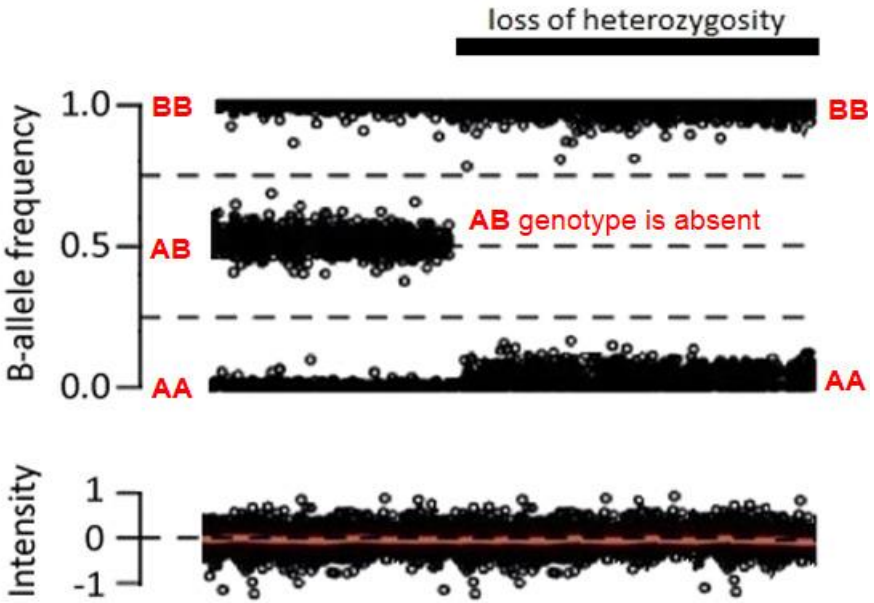
WES (Agilent SureSelect v5 capture sequenced on Illumina HiSeq)

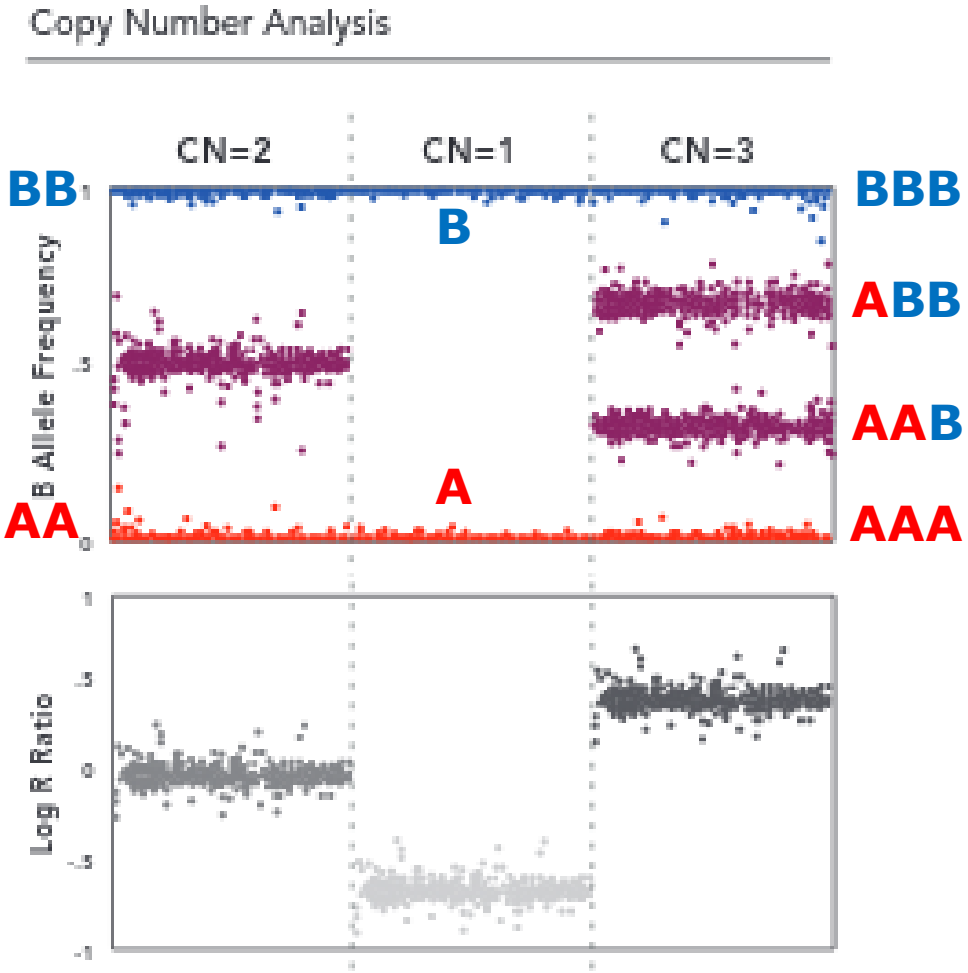
Beyond canonical CGH

From relativity to absoluteness

Adding the B-Allele Frequency (BAF)

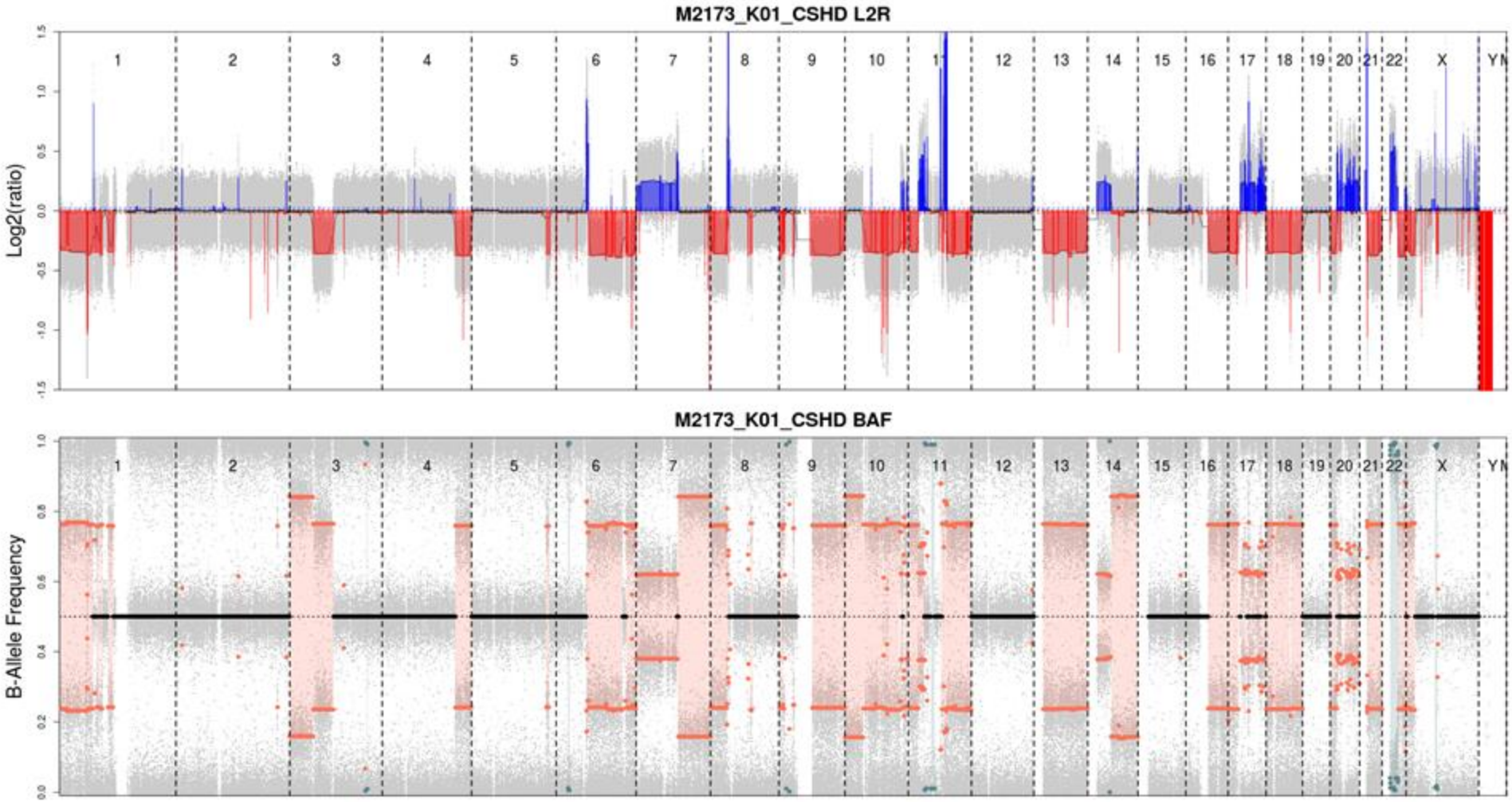
REFERENCE	A T C C G T G C C A A T G T C
READS	A T C C G T G C C A A T G T C A T C C G T G C C G A T G T C A T C C G T G C C A A T G T C A T C C G T G C C G A T G T C A T C C G T G C C G A T G T C A T C C G T G C C A A T G T C
"B" COUNTS	3
TOTAL COUNTS	6
BAF	0.5





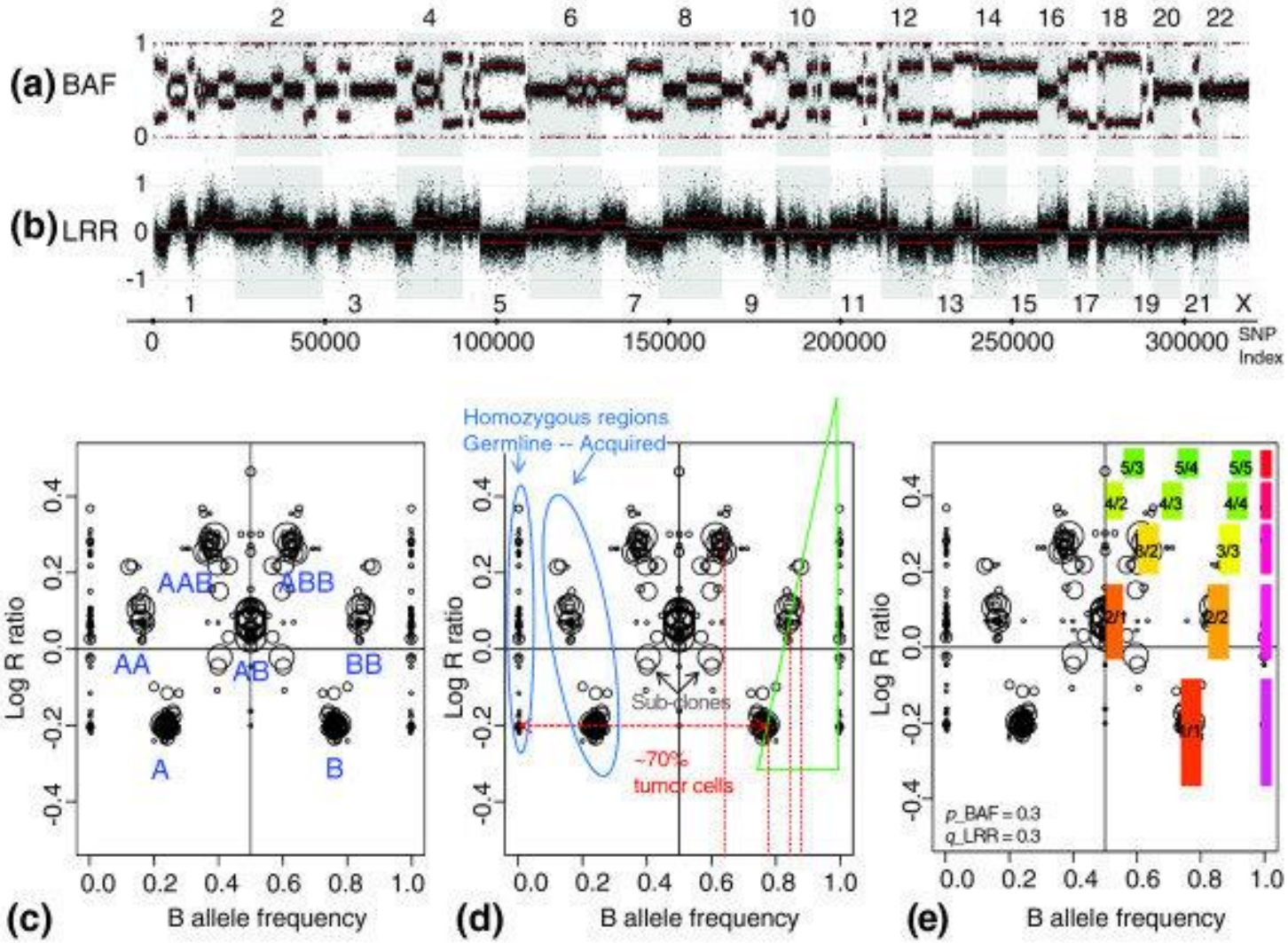
Up to TCN & ASCN : L2R + BAF

TCN : Total copy cumber ; ASCN : Allele-specific copy number



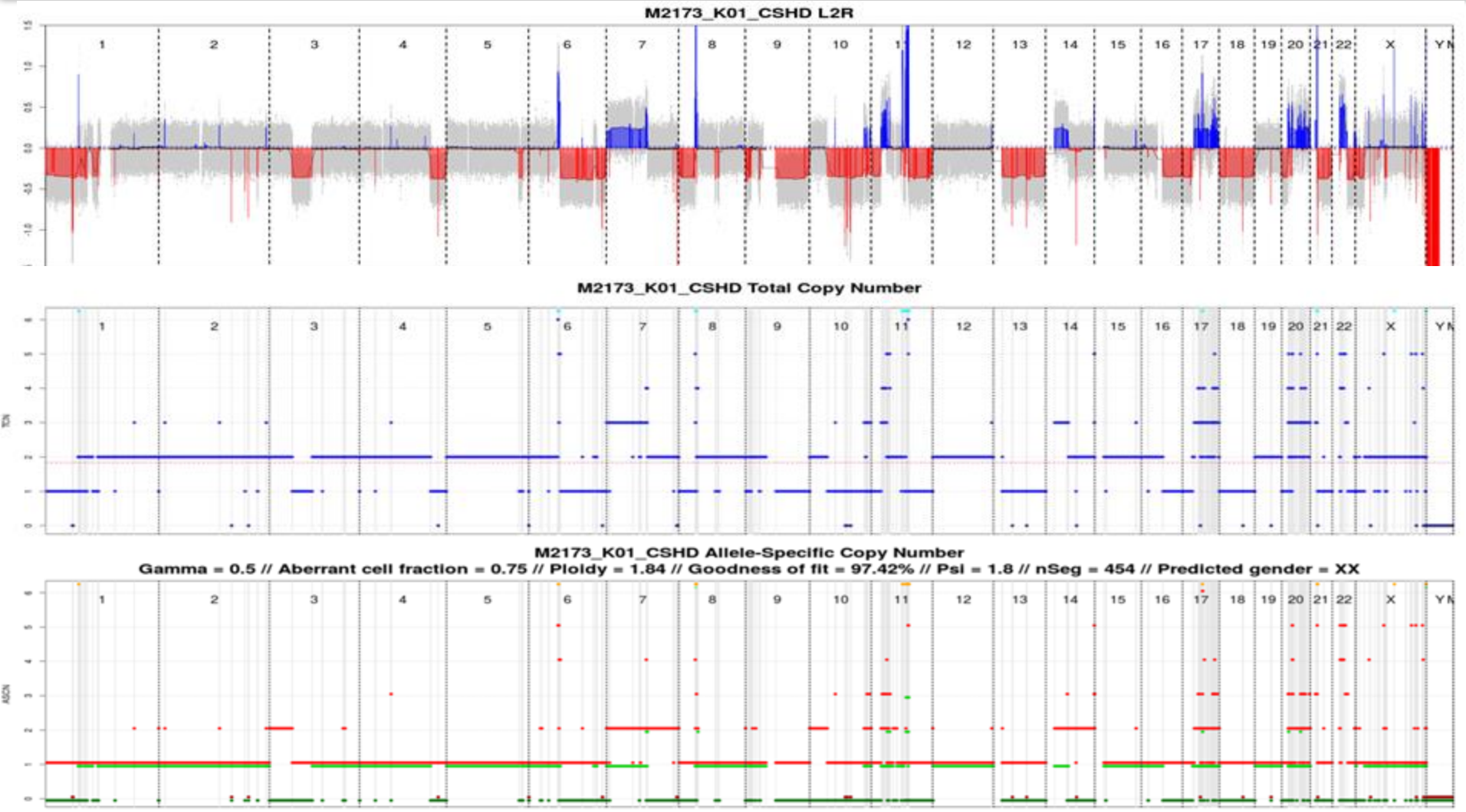
Up to TCN & ASCN : Modeling

GAP (Popova et al, Genome Biol, 2009)



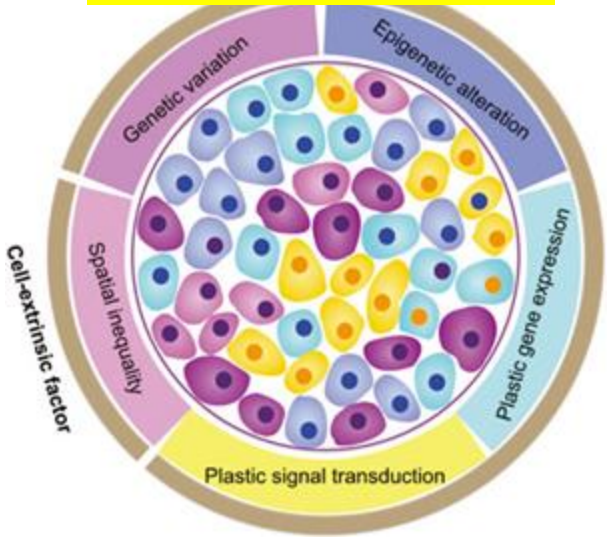
- Functional combination of L2R & BAF
- Also modeling **global ploidy**
- Derives putative **tumor cellularity**
- Tools : ASCAT, facets, sequenza, ControlFREEC, GAP …

Up to TCN & ASCN : Results

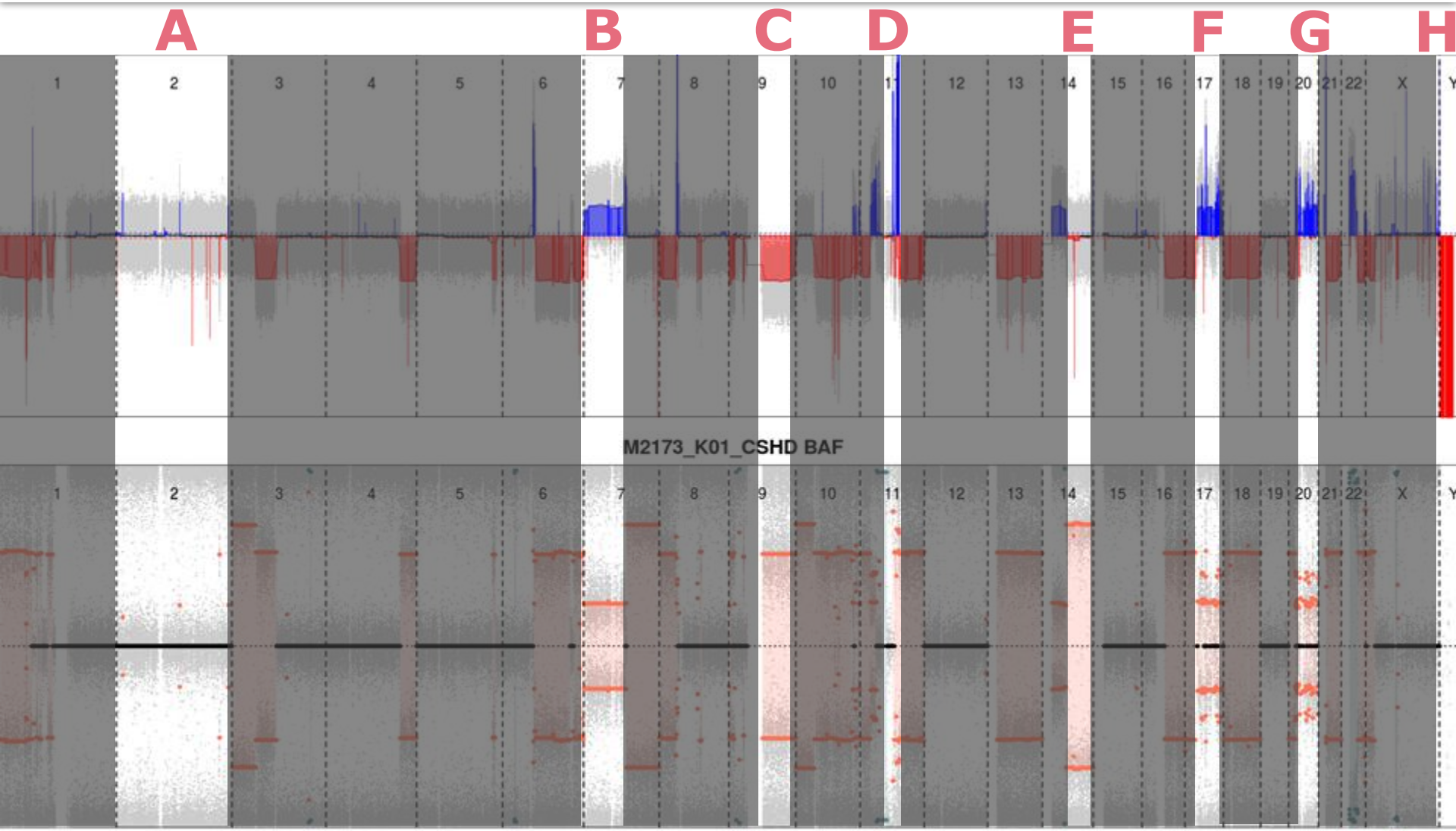


Up to TCN & ASCN : Clonality

- Bulk = mix of several (different ?) cells
- Current algorithm infer a **single, major** clone
- In case of polyclonality, risks are :
  - ***Over***complexification (ploidy doubled, tripled, +)
  - ***Under***complexification (true aberrations ignored)
- Polyclonality **deconvolution** algorithms exist (*Battenberg*, *cloneCNA*, …) but complex (many parameters) and with low efficiency
- Long live **single cell** !



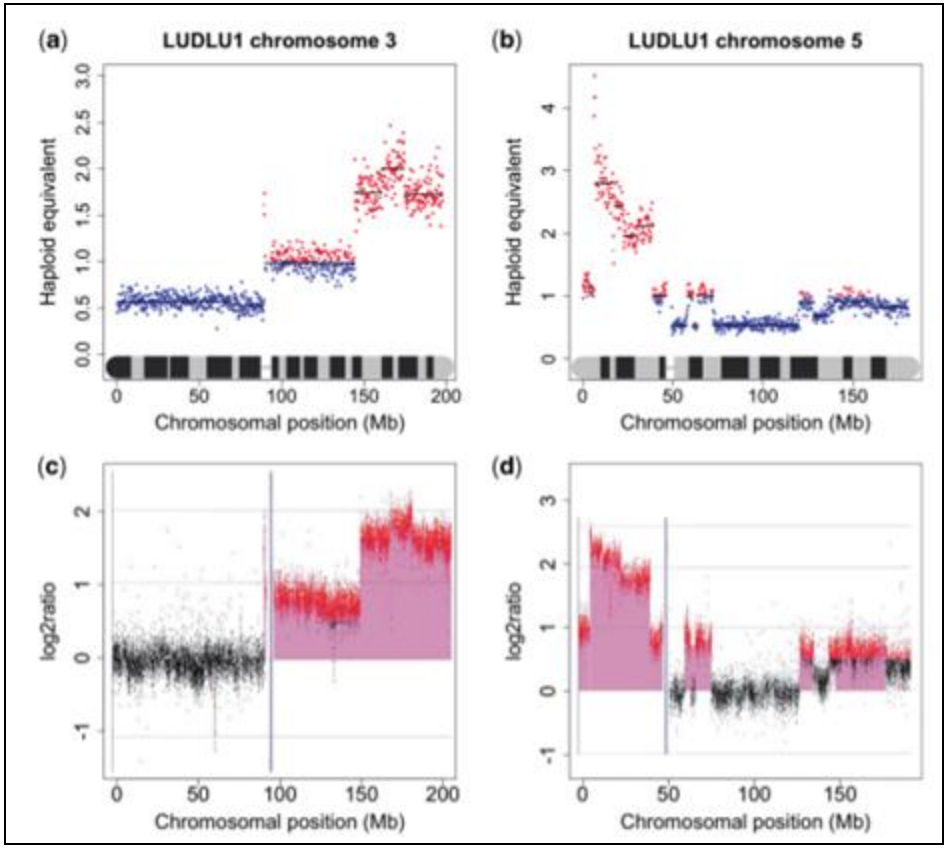
Quiz : name these putative event cases !



NGS : beyond microarrays

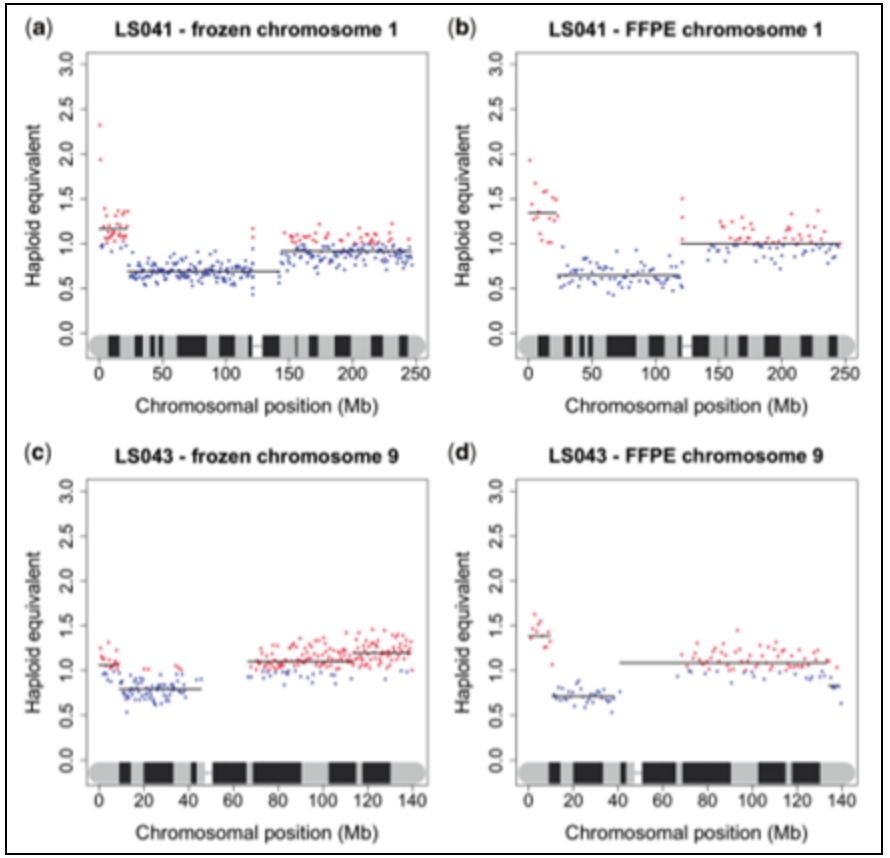
NGS : low input, FFPE

Wood, 2010



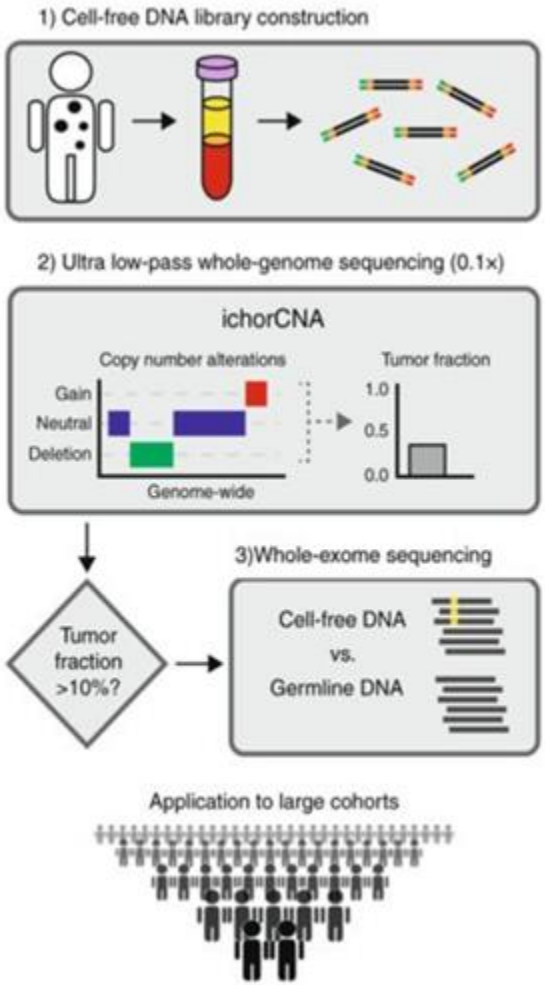
- **2 to 5 ng** of DNA

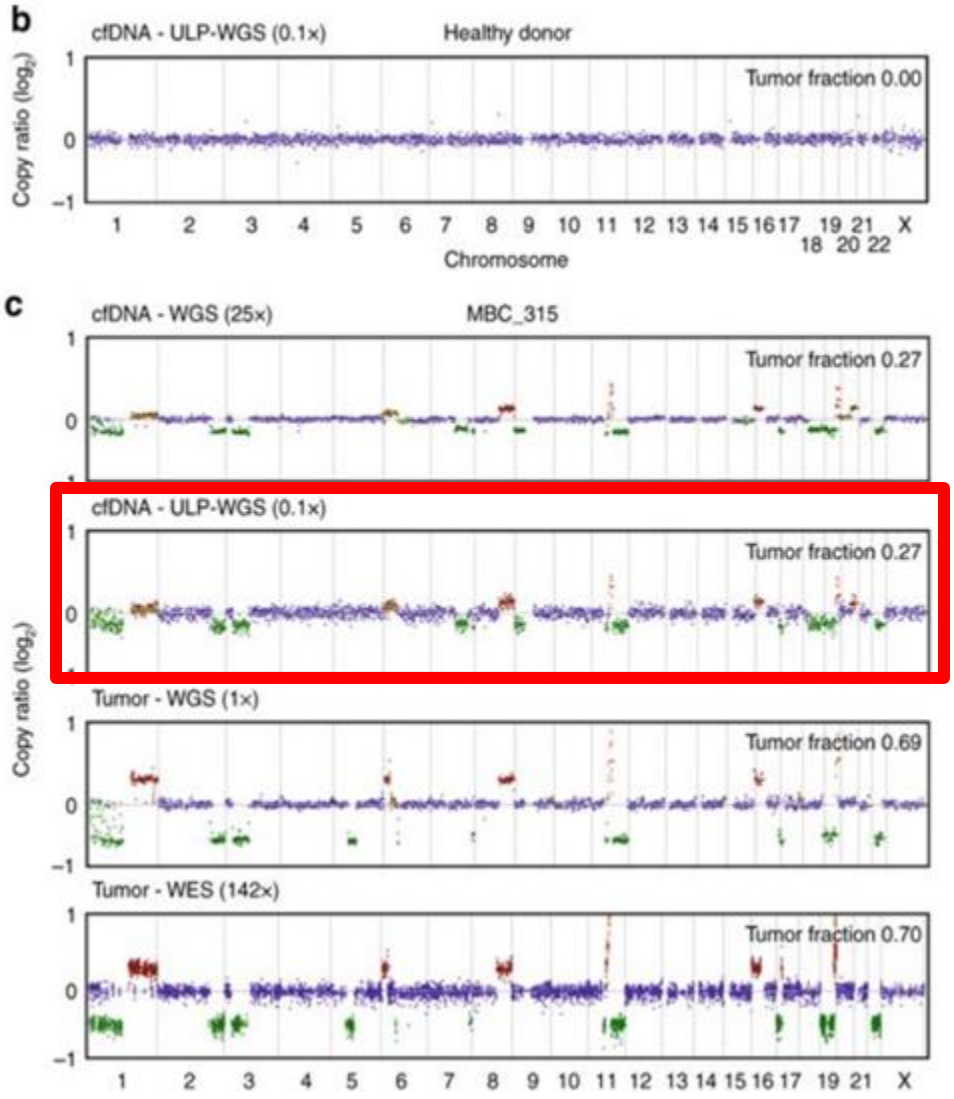
- **FFPE** (Formalin-fixed paraffin-embedded) samples



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)

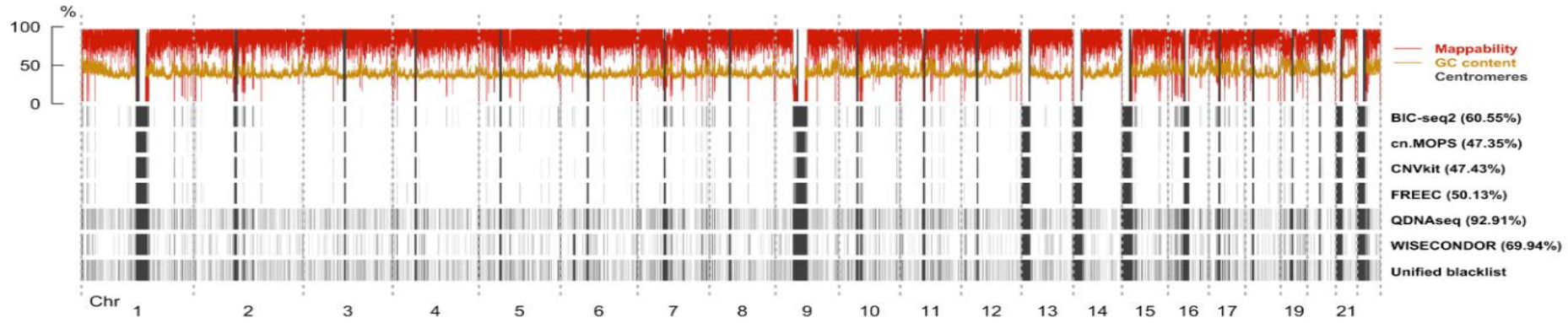


Figure S3. **Representation of the blacklists across the considered tools at 30 kb.**

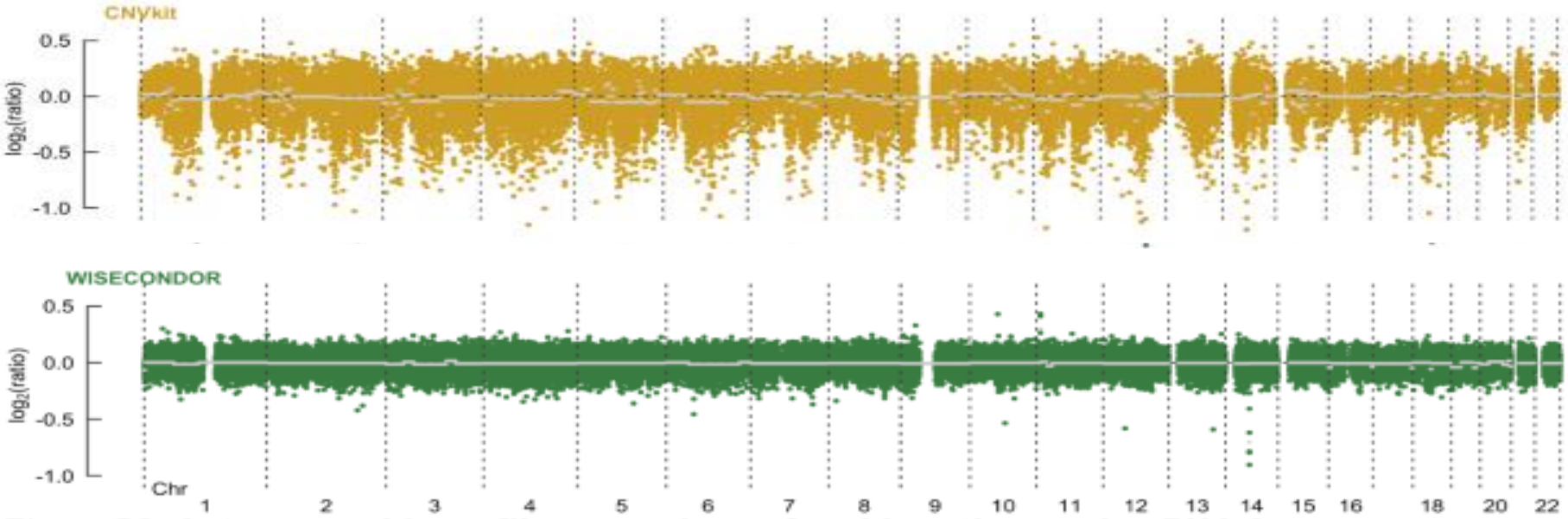
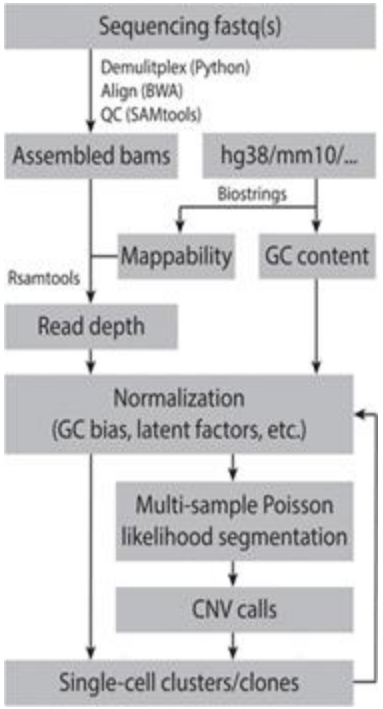
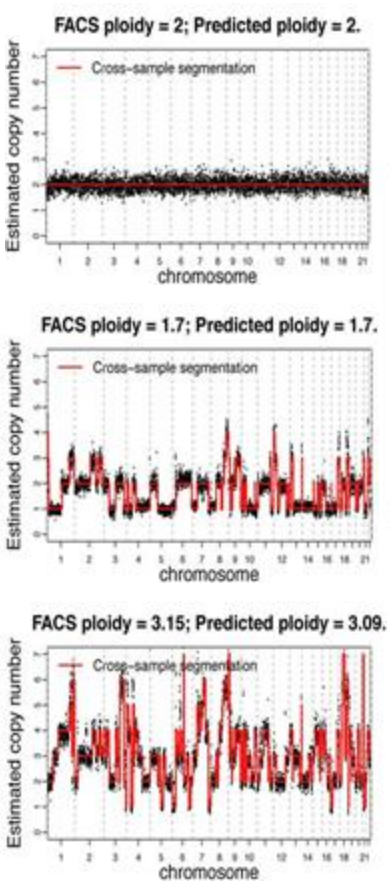
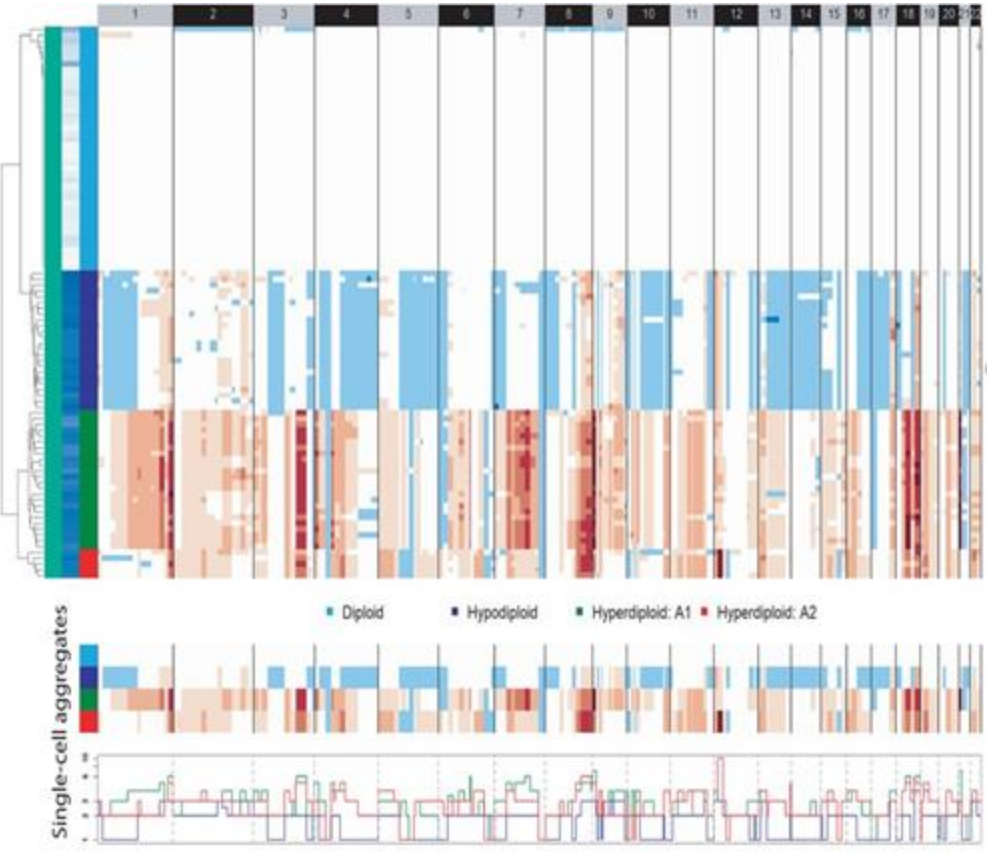


Figure S8. **Autosome-wide profile comparison of problematic sample gDNA-3.**

NGS : Single Cell CNA (SCOPE)







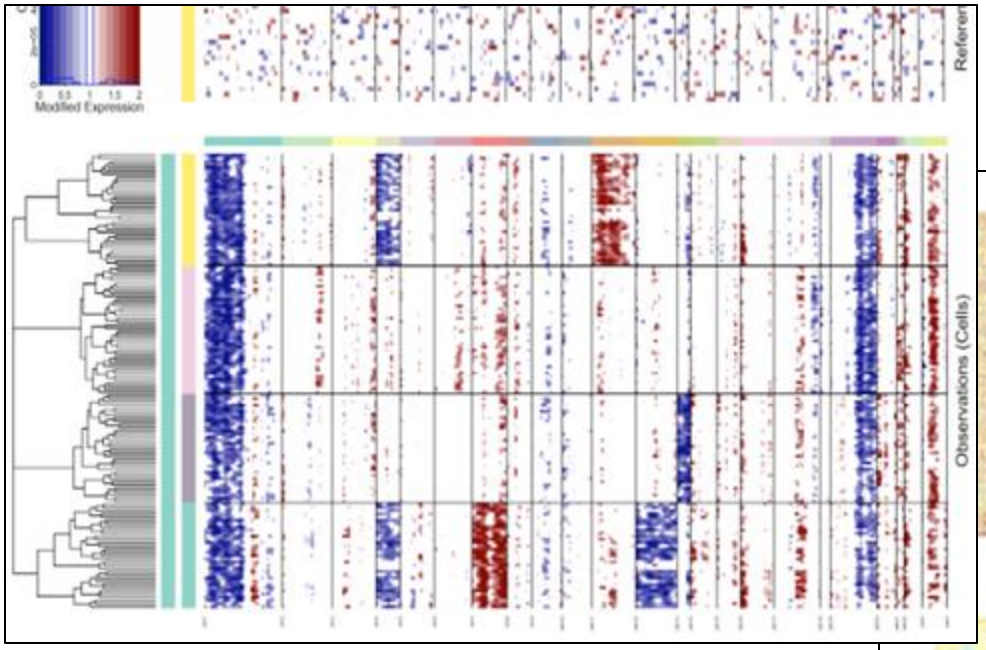
SCOPE Rujn Wang et al., bioRxiv, 2019/03

WARNING :

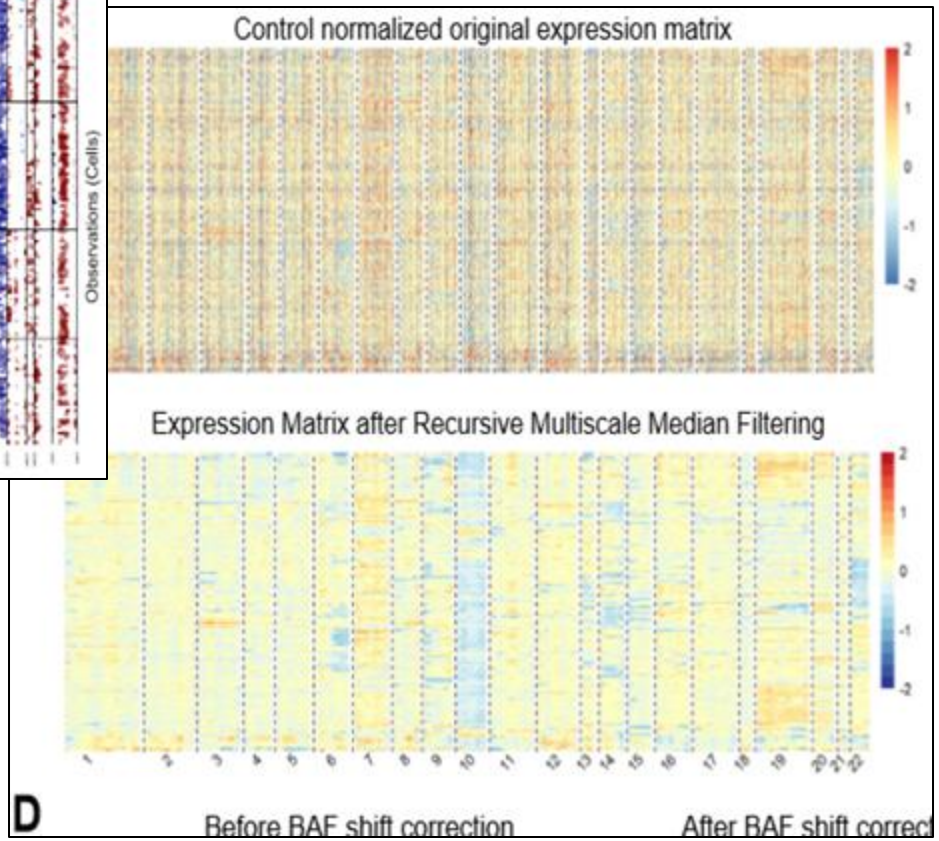
- Limited resolution : > 2 Mb (binning)
- Requires > 750,000 reads / cell

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

InferCNV (Broad Institute)



CaSpER (Armanci et al, BioRxiv 2019)

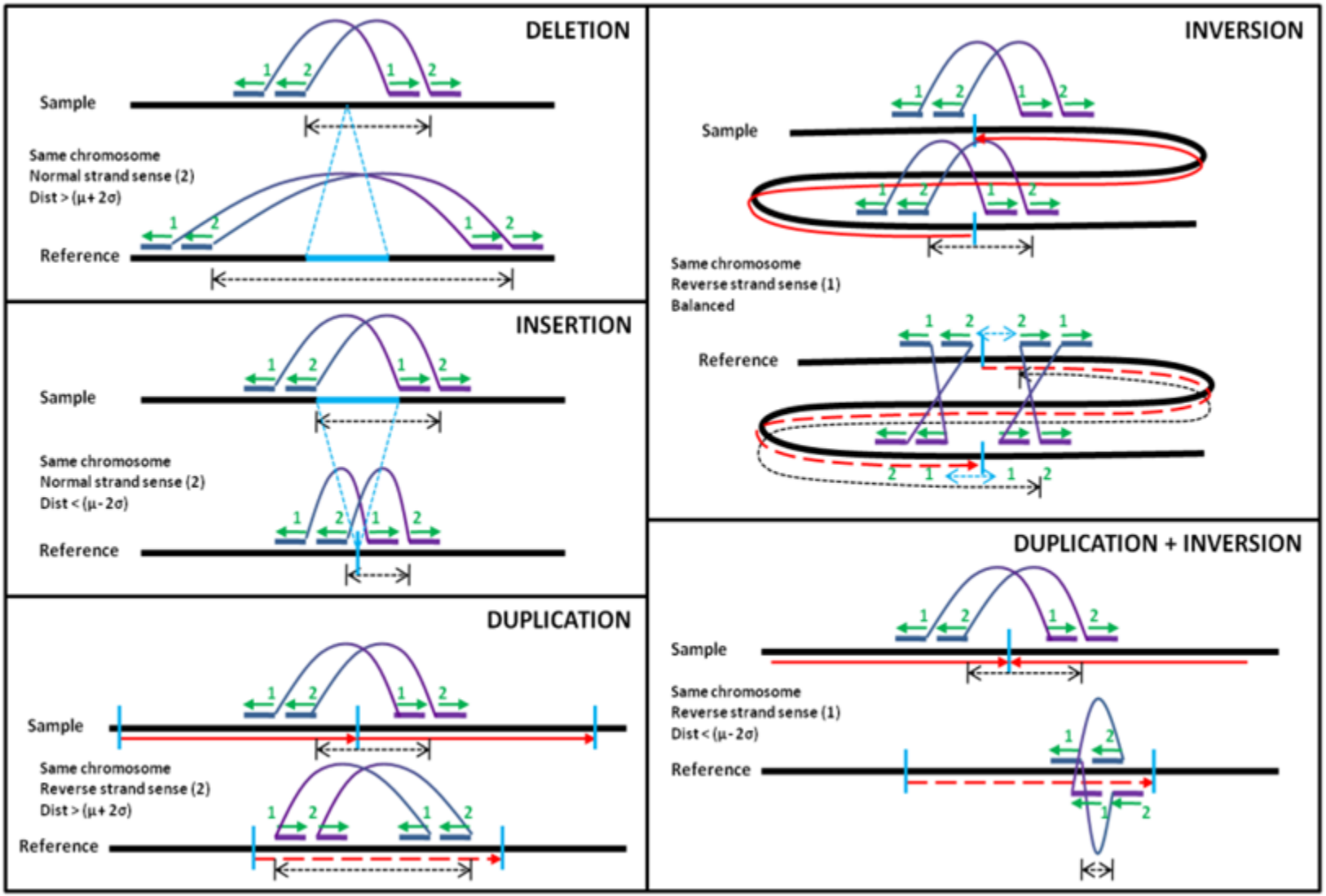


WARNING :

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

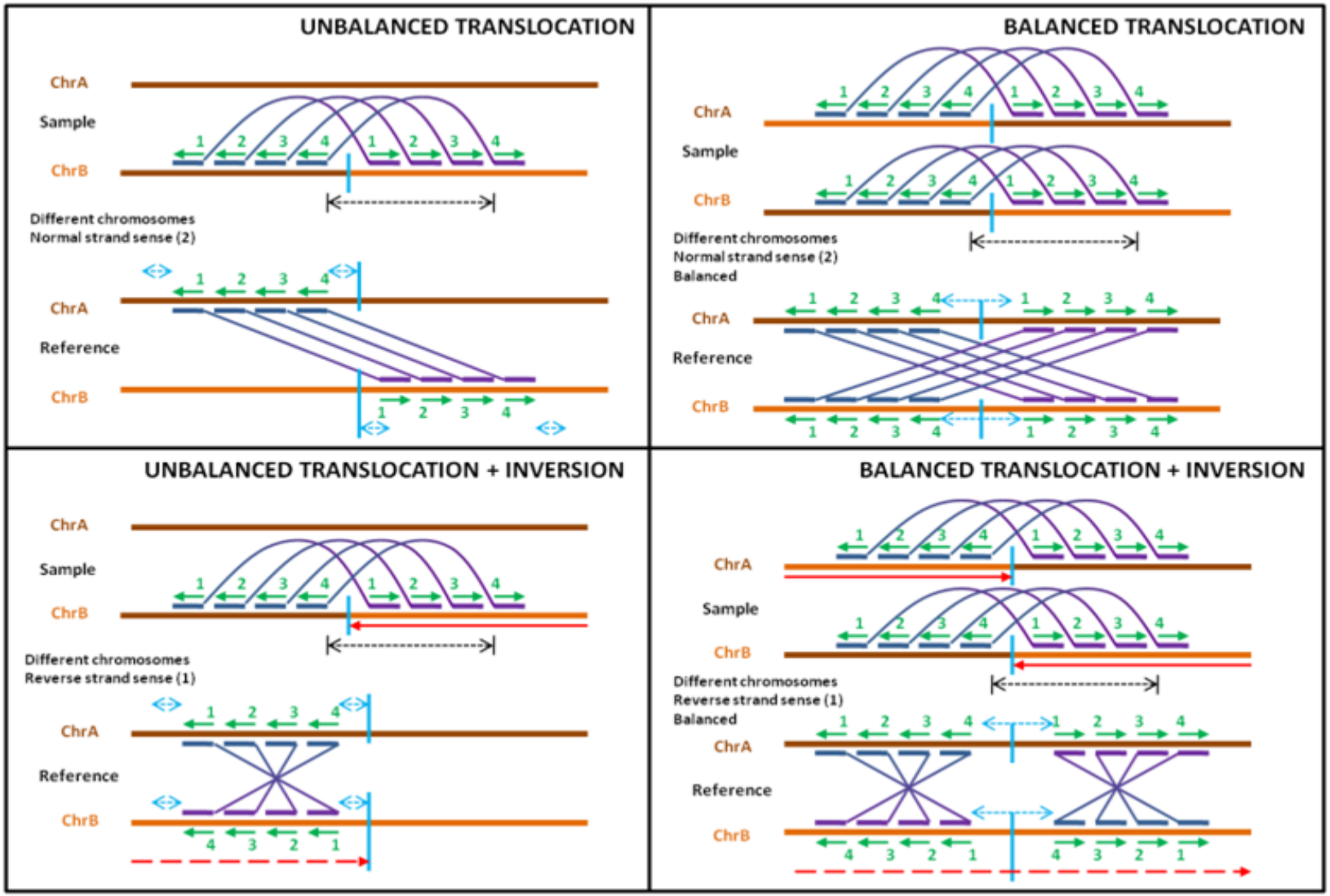
WGS : Intra-chromosomal structural variations

Courtesy of Bruno Zeitouni

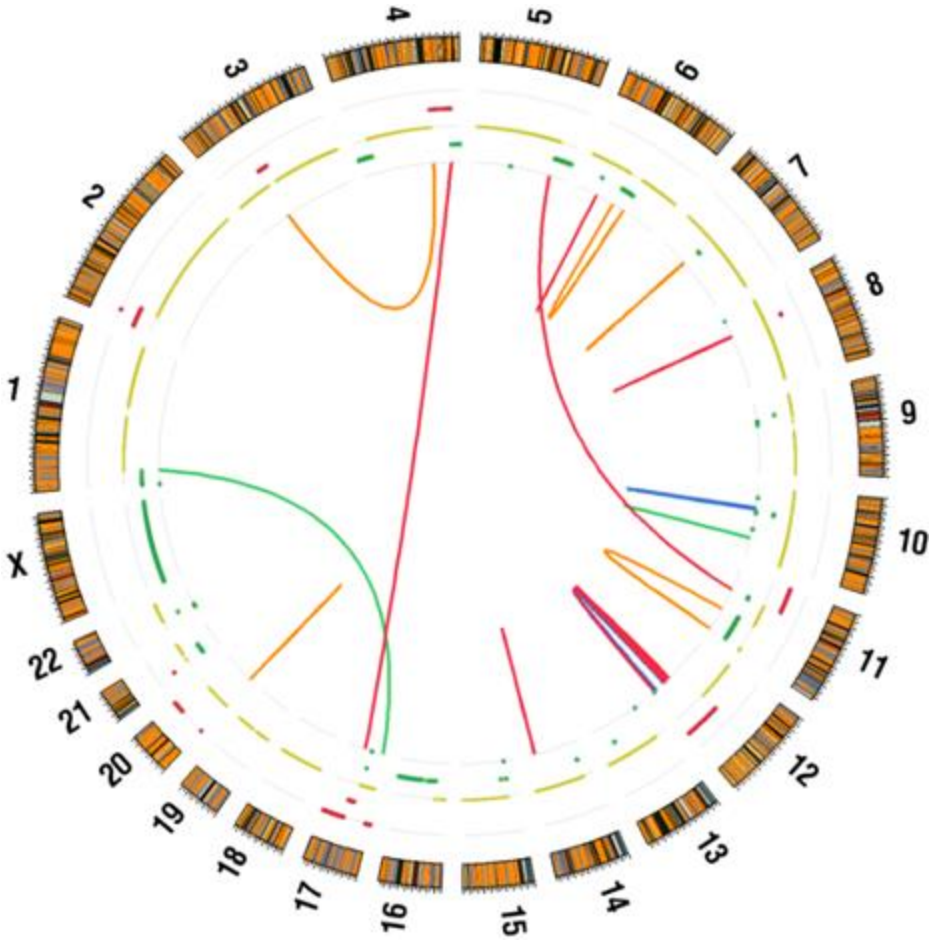


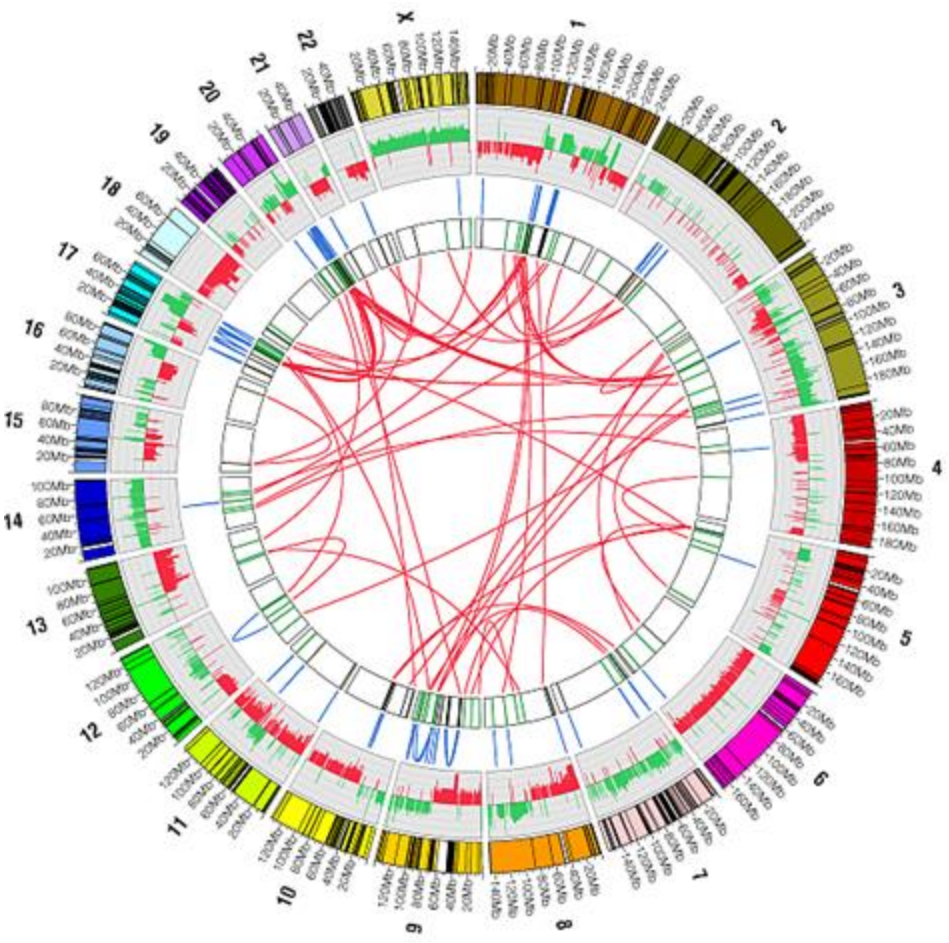
WGS : Inter-chromosomal structural variations

Courtesy of Bruno Zeitouni



Visualizing structural variations



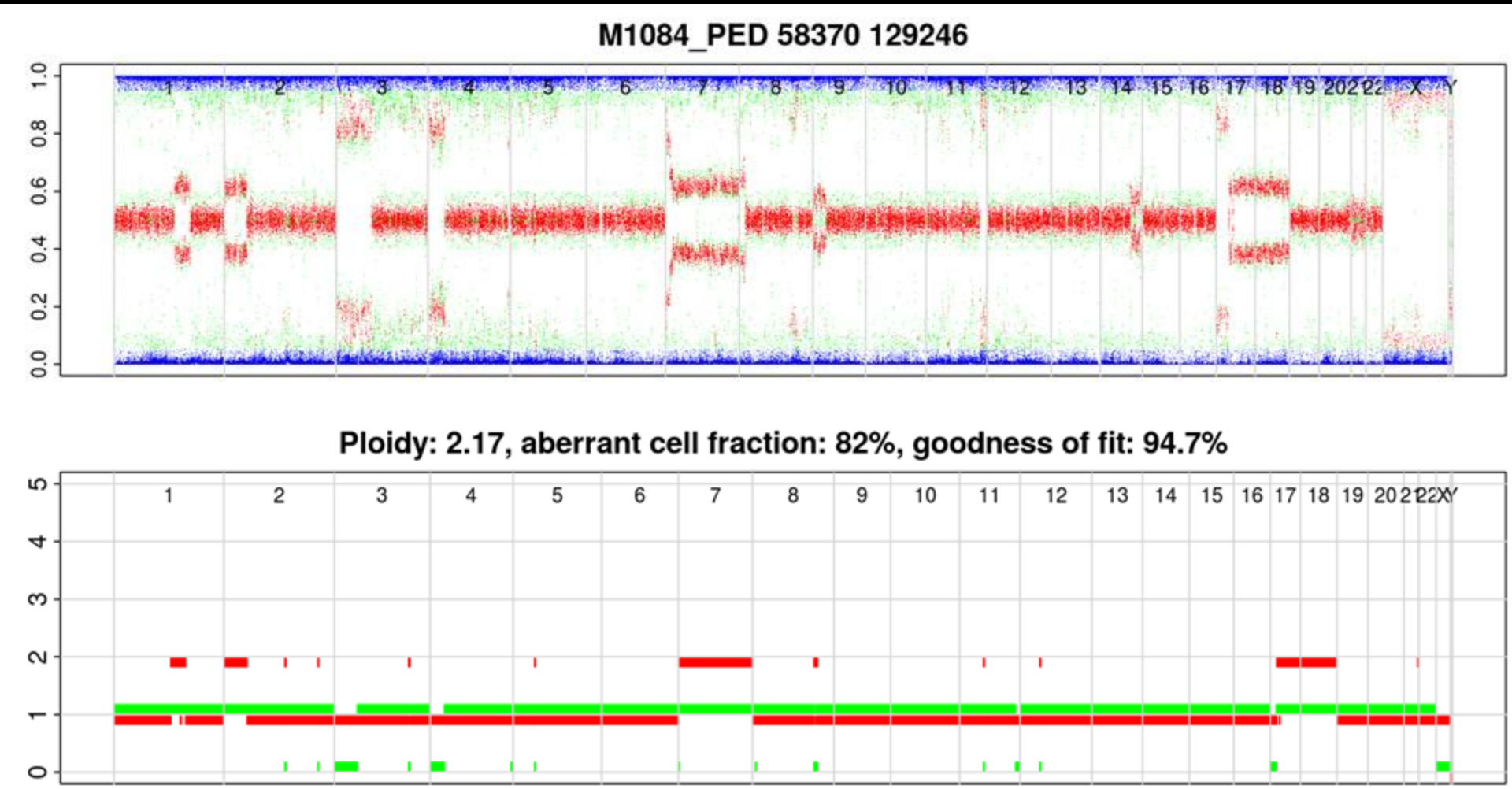


NGS versus microarrays

	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**

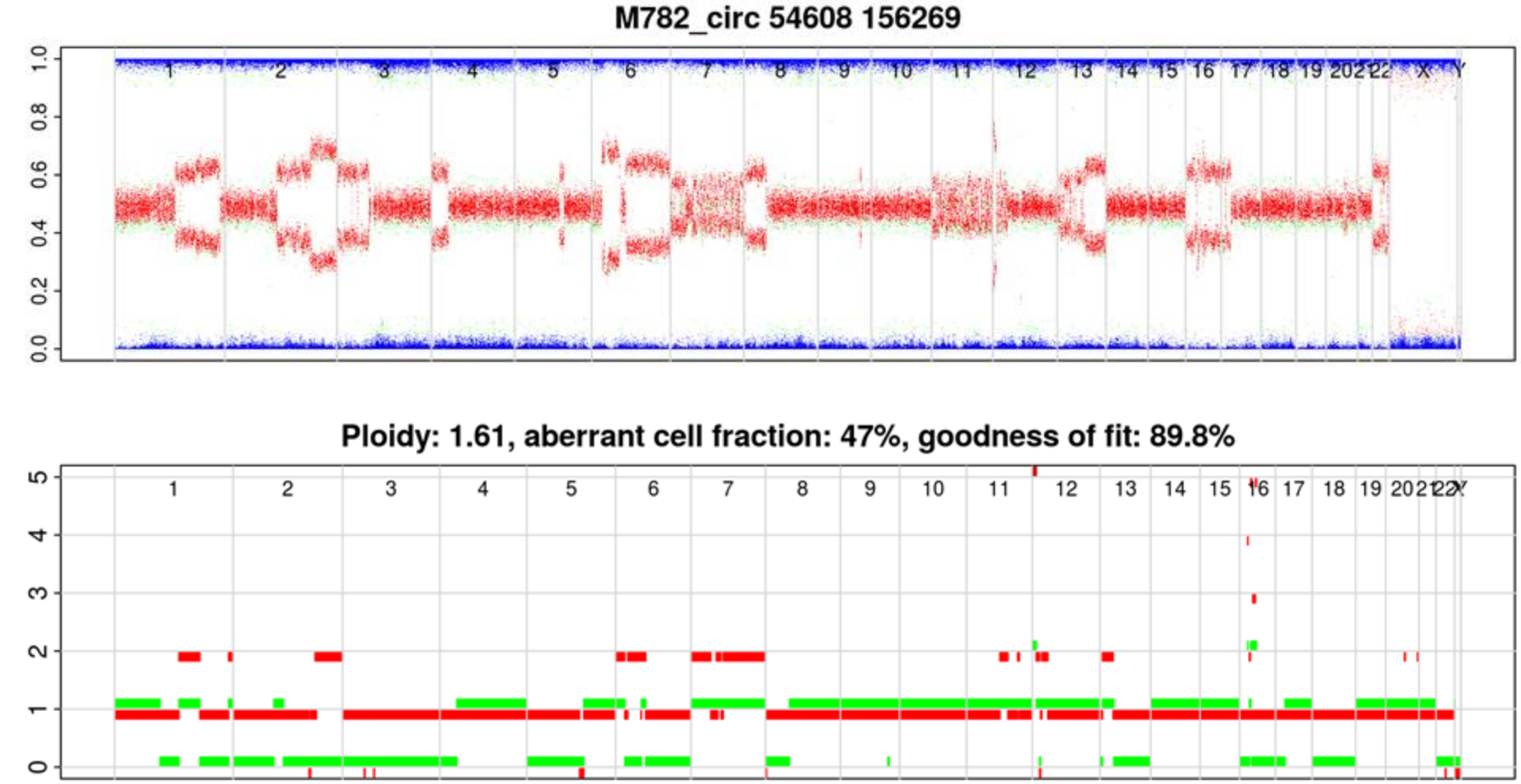
Microarrays : still alive !

FFPE samples

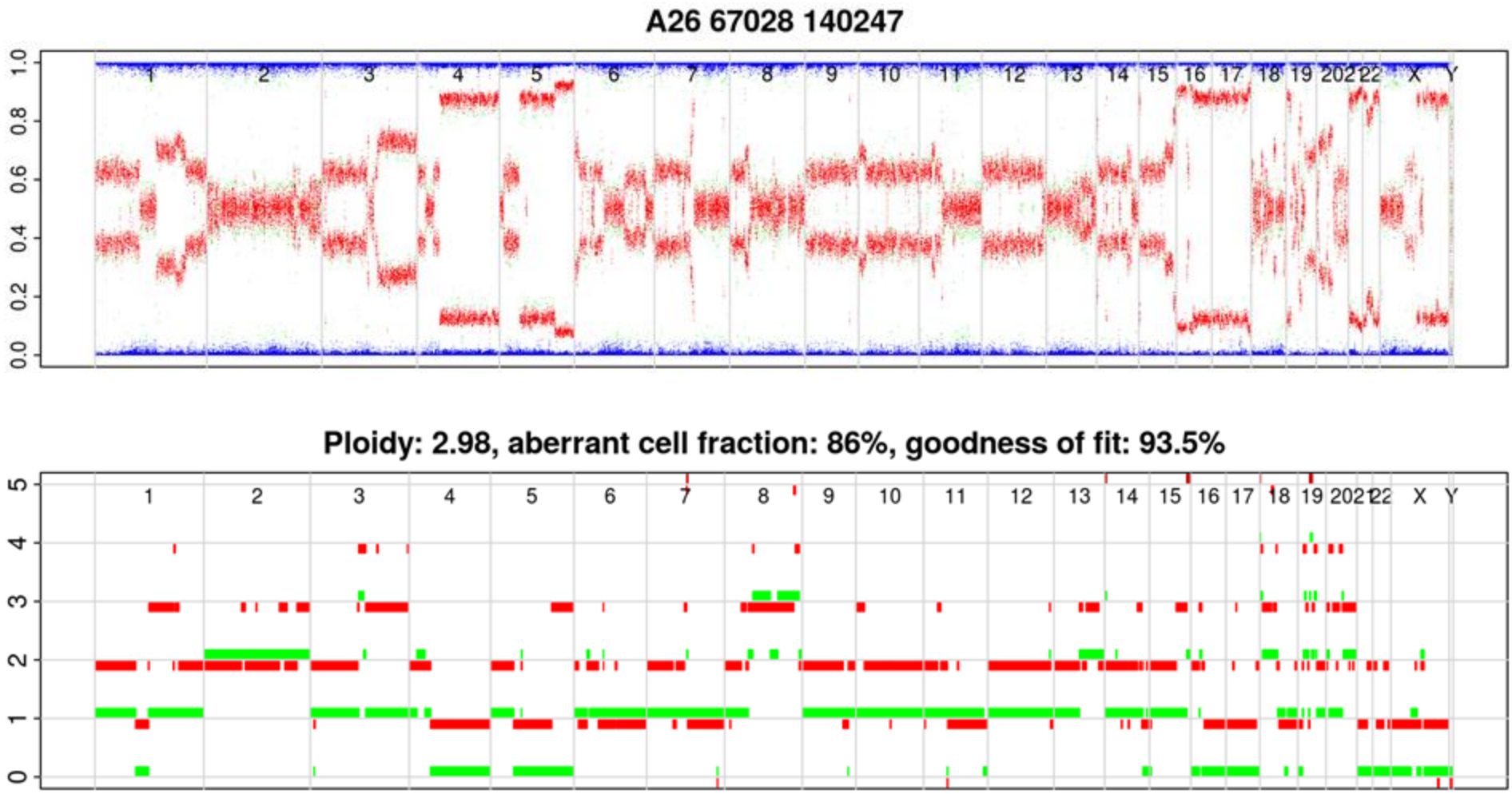


Affymetrix OncoScan

ctDNA

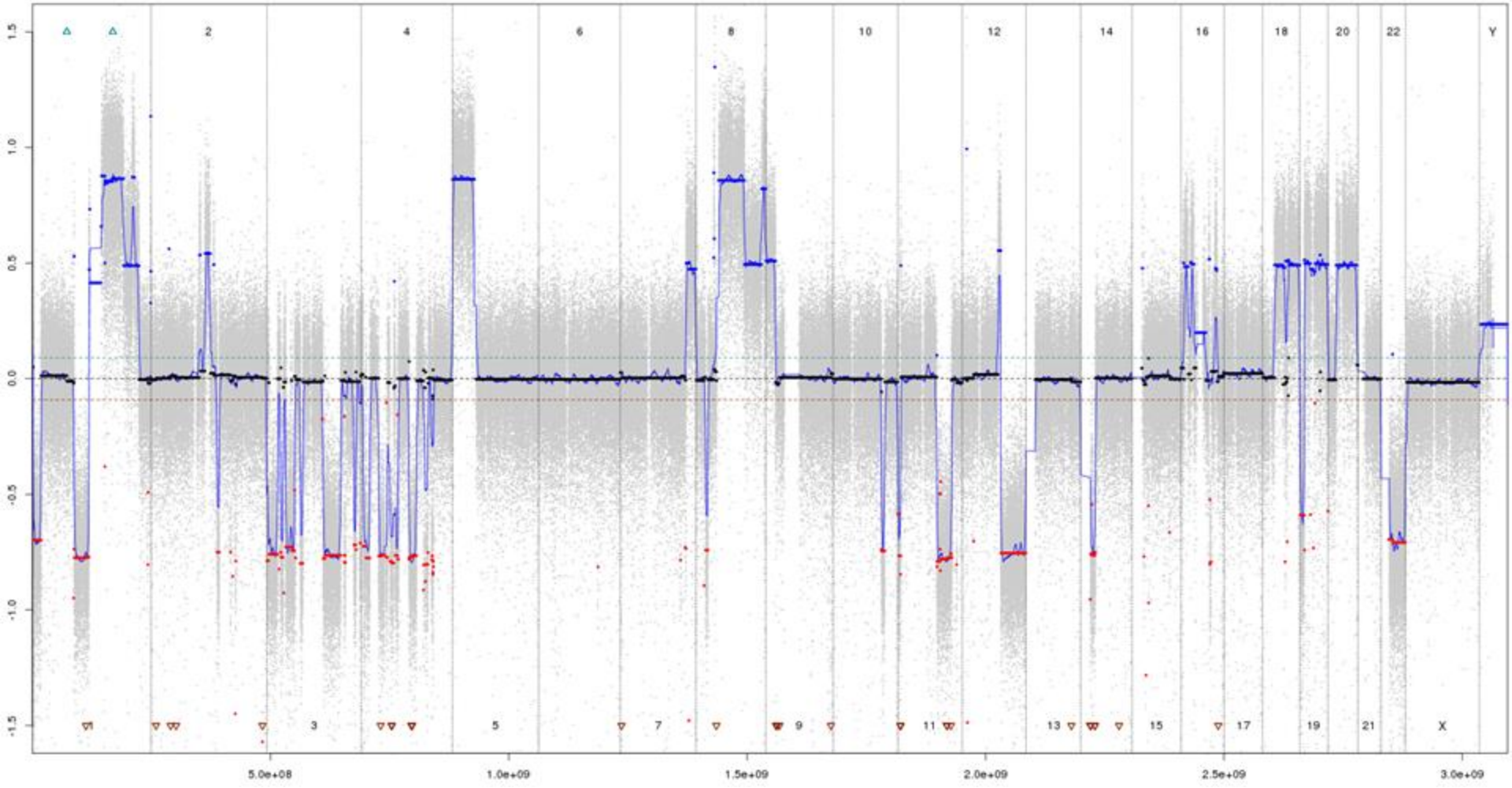

M782_circ 54608 156269
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y
0.0 0.2 0.4 0.6 0.8 1.0
Ploidy: 1.61, aberrant cell fraction: 47%, goodness of fit: 89.8%
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X
0 1 2 3 4 5

ascites DNA

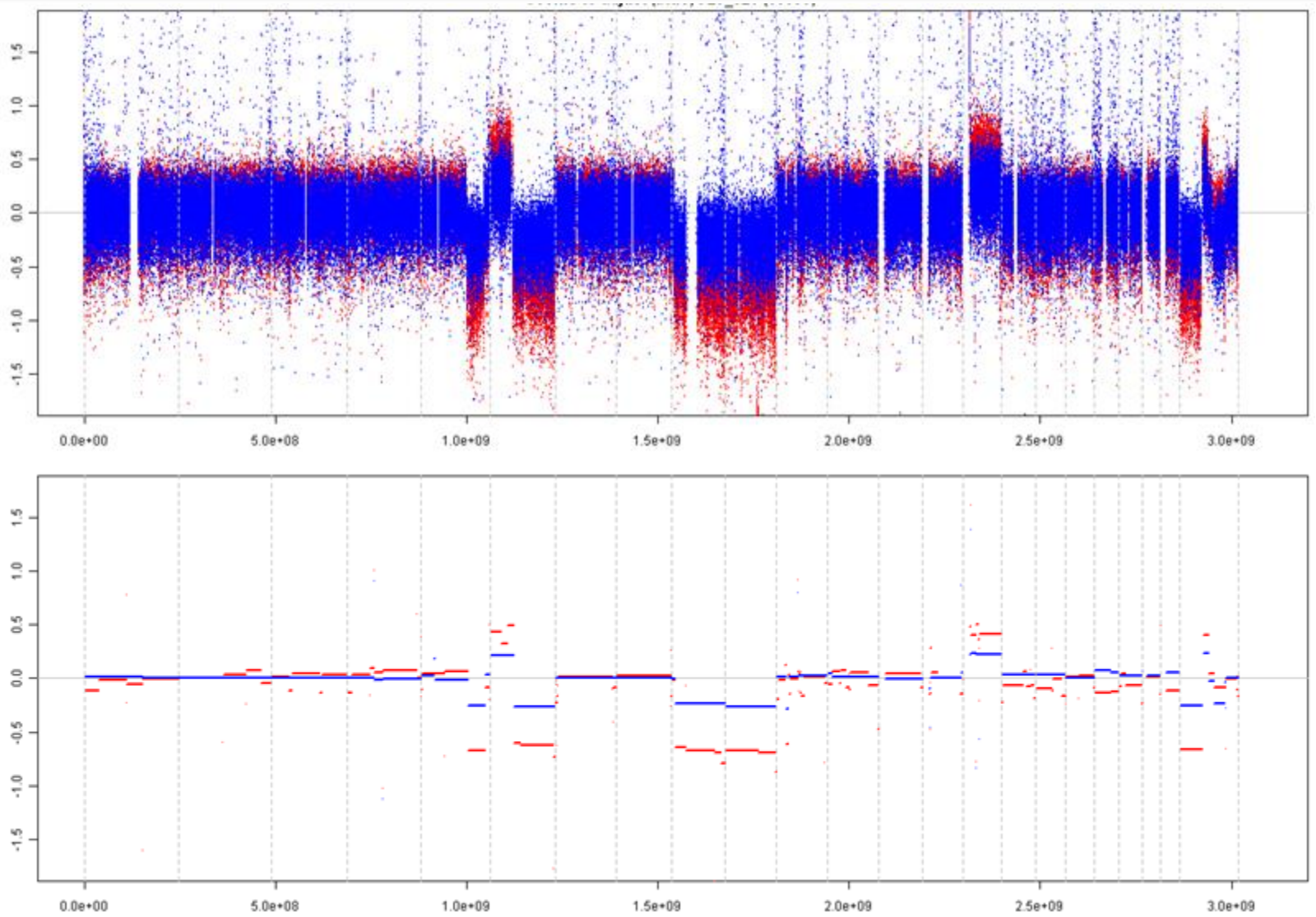

A26 67028 140247
1.0
0.8
0.6
0.4
0.2
0.0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 22 X Y
Ploidy: 2.98, aberrant cell fraction: 86%, goodness of fit: 93.5%
5
4
3
2
1
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 22 X Y

Further analyses

Sample profile



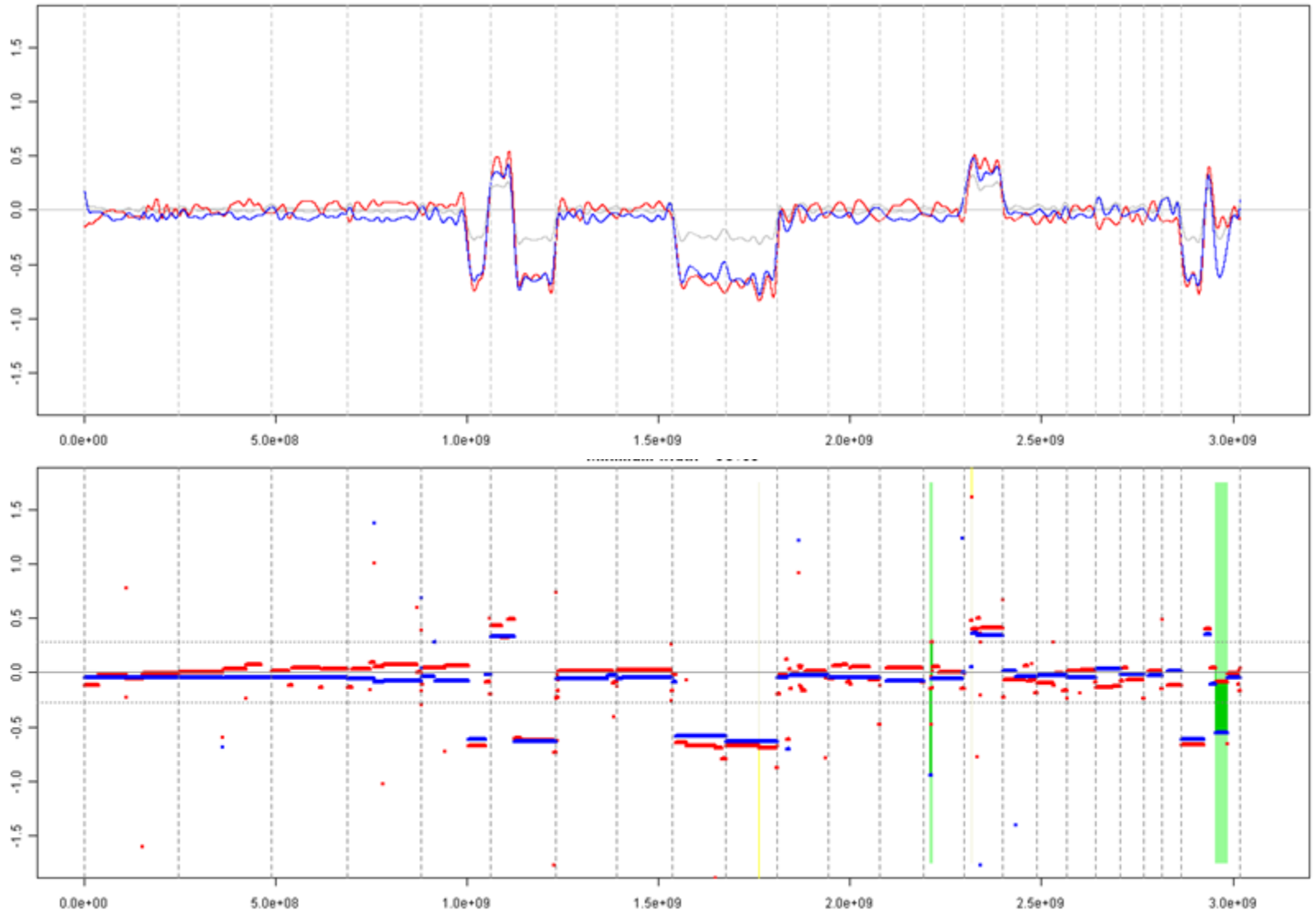
Comparison of pairs (scaling)



d0

d21

Comparison of pairs (calling)

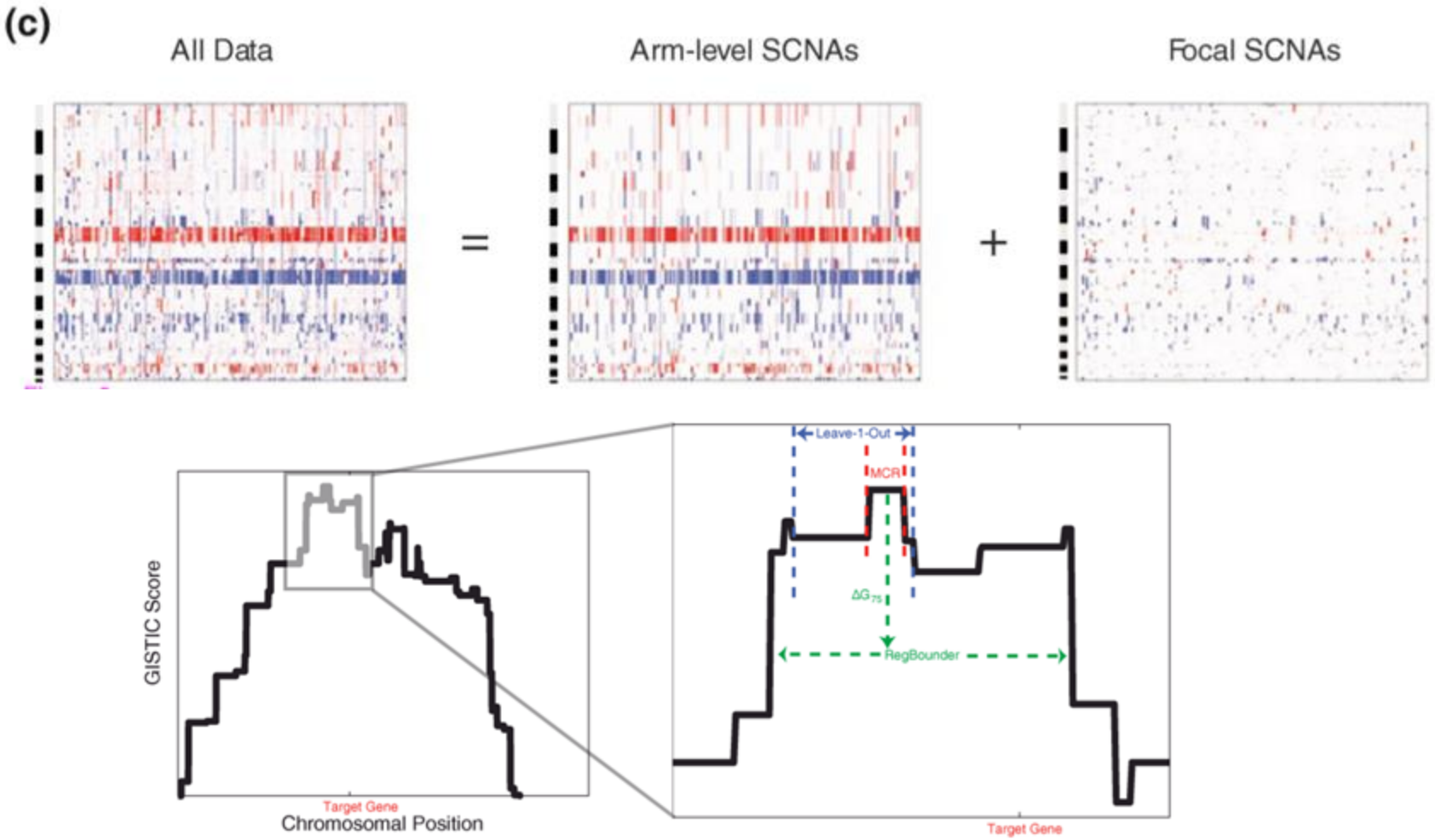


d0

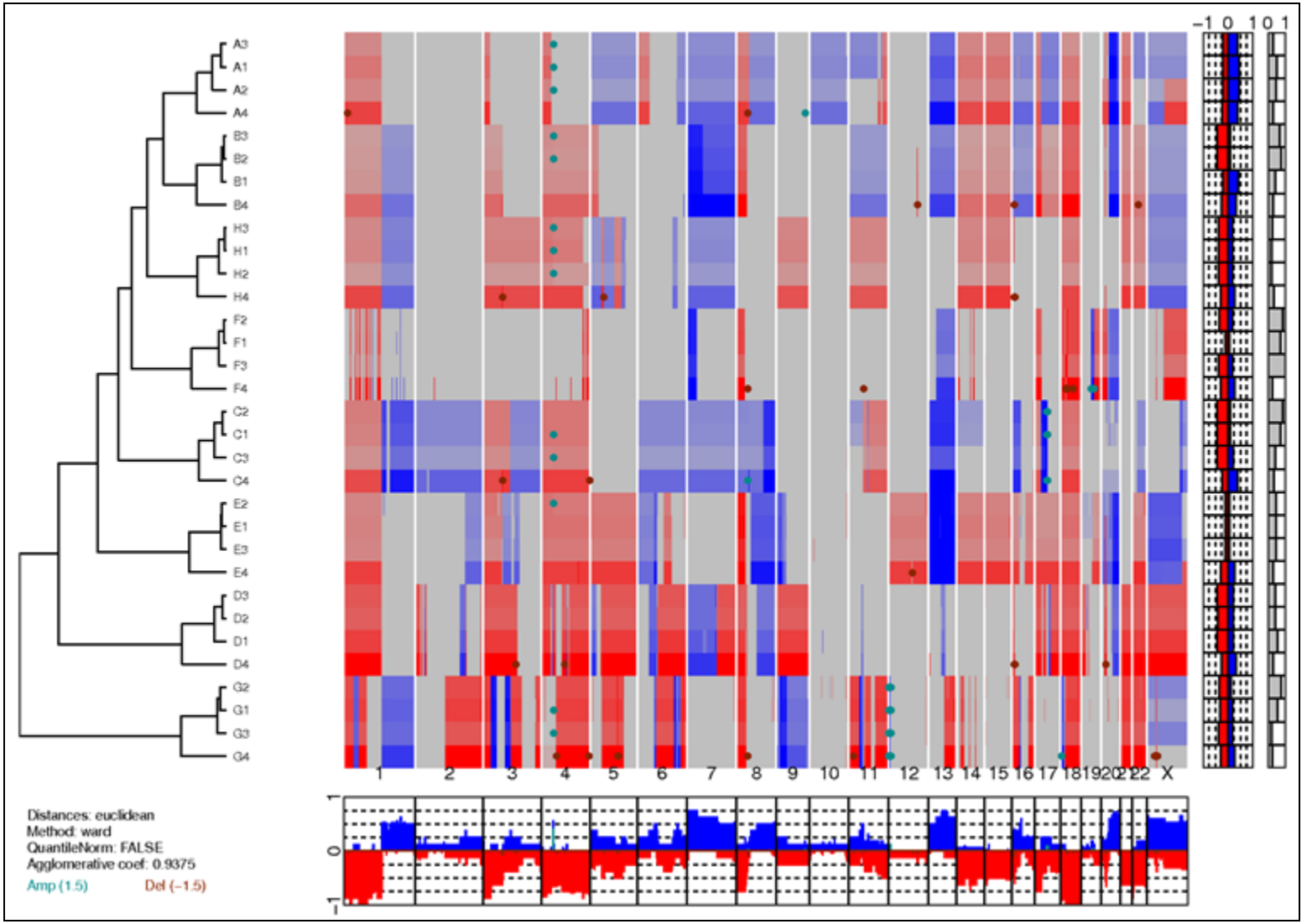
d21

Minimal common regions (GISTIC2)

Mermel et al, Genome Biology, 2011

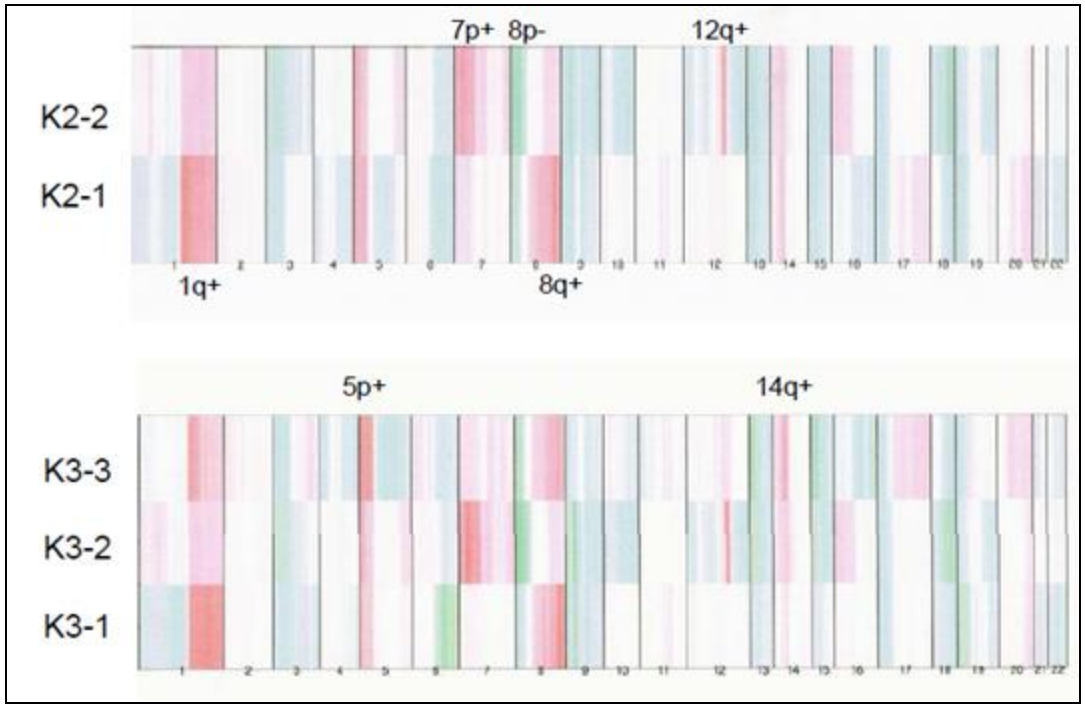


Hierarchical clustering, heatmap frequency of aberrations, genomic instability

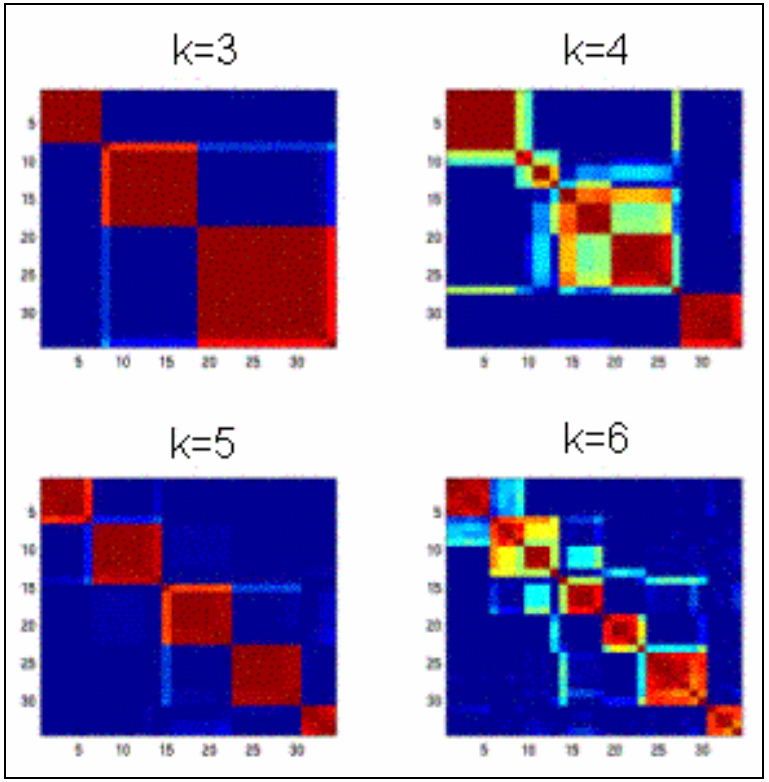


Other clustering methods

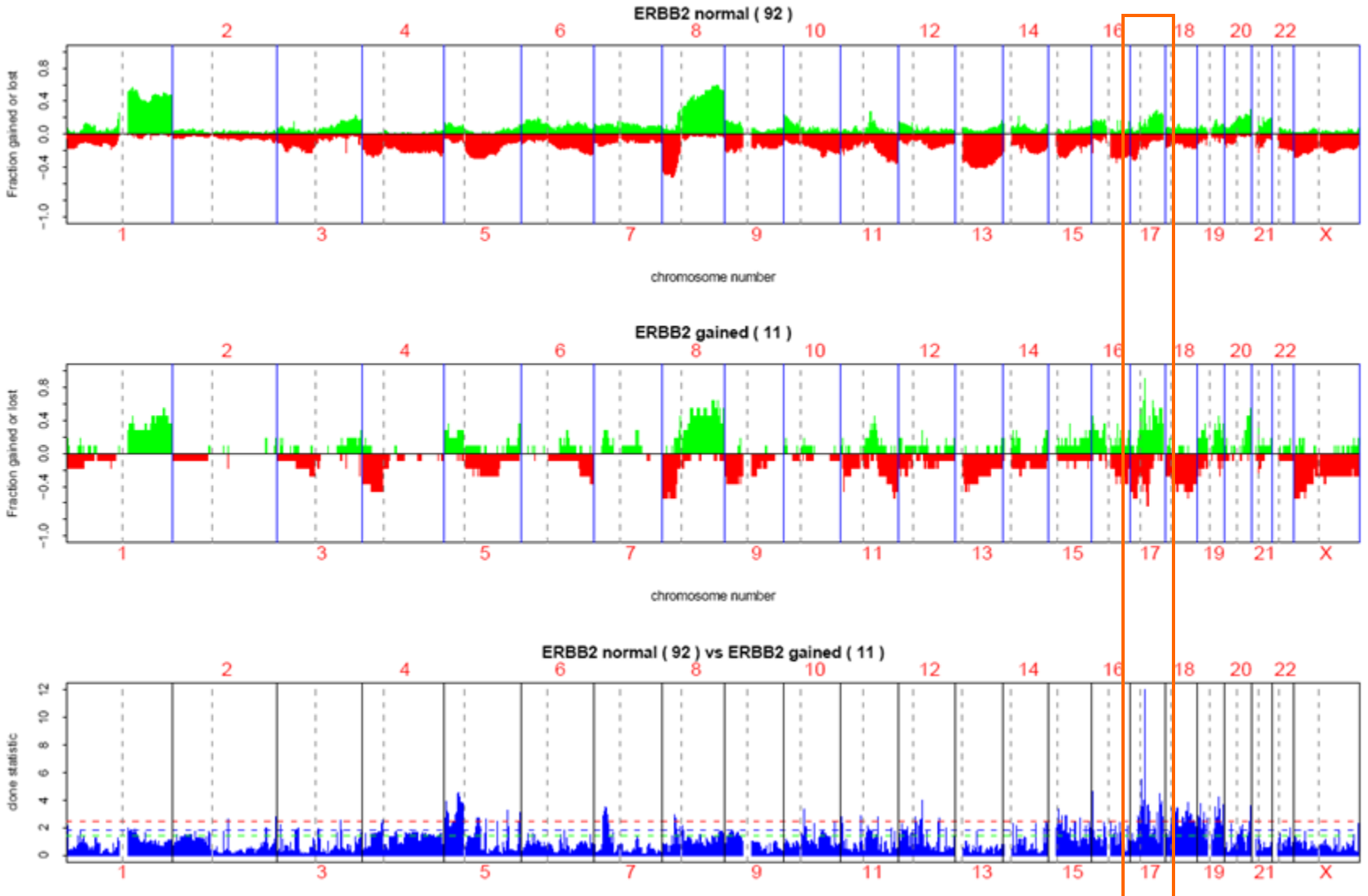
K-means



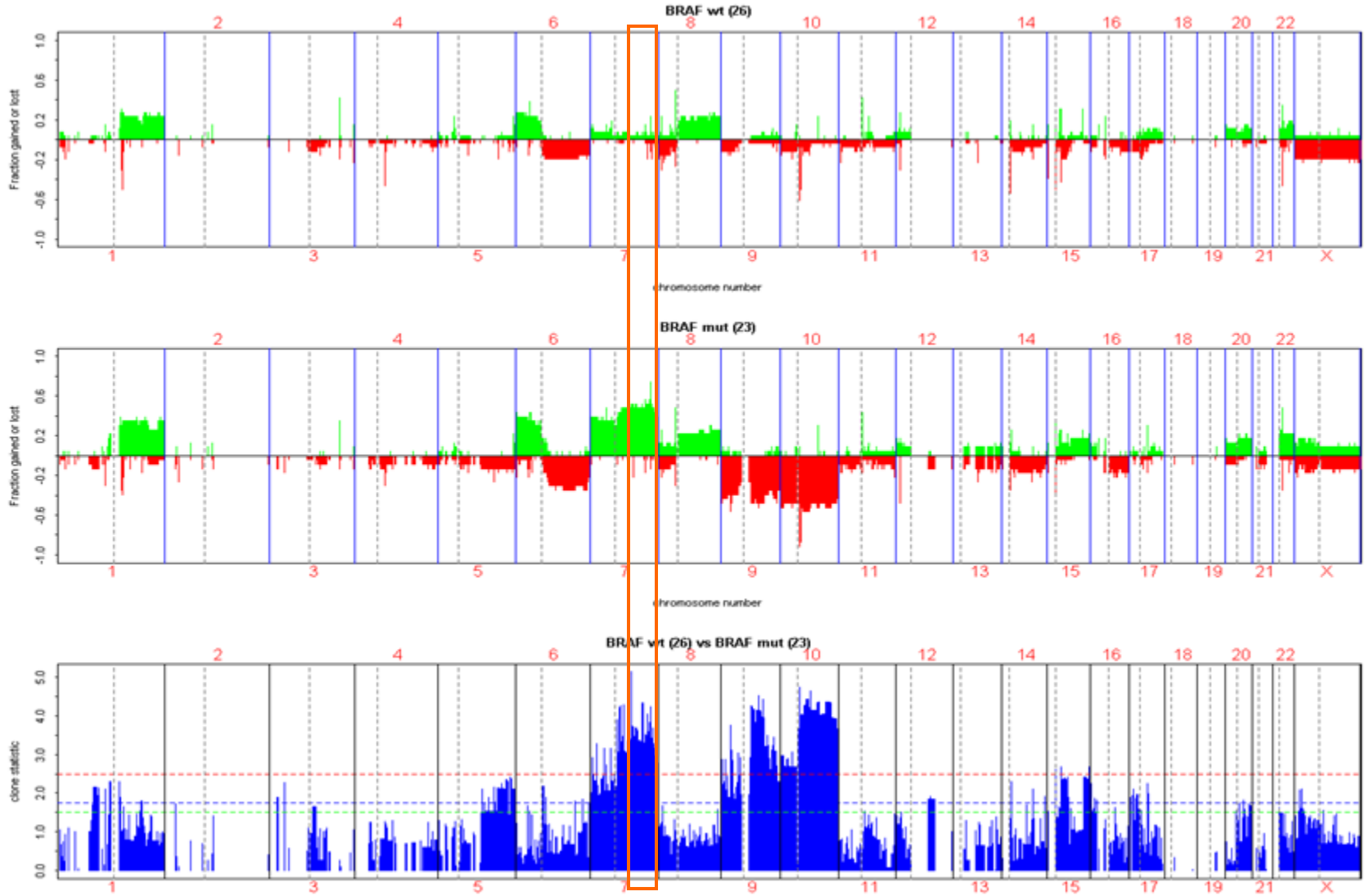
Non-negative Matrix Factorization



Subpopulations and annotations (Continuous data)



Subpopulations and annotations (Continuous data)



Genomic regions annotations

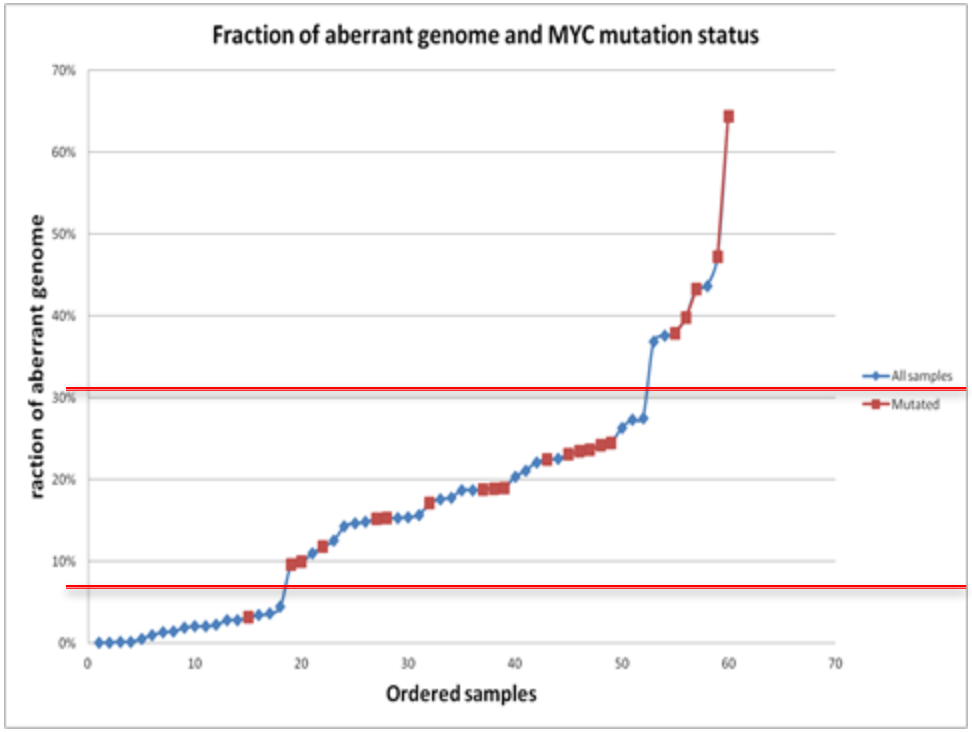
Loc	Width	Band1	Band2	Num.probes	Status	log2(ratio)	Ratio	Genes	CNV	miRNA	CpGisl
8:161471-6914076	6.75Mb	8p23.3	8p23.1	409	L	-1.13	0.46	2 12 32	1949	15	120
8:6939250-7786708	847.46Kb	8p23.1	8p23.1	10	L	-0.34	0.79	3 23	135	-	11
8:8100383-22878739	14.78Mb	8p23.1	8p21.3	879	L	-1.09	0.47	1 15 39 109	1560	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
8:27015529-27667961	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	511.36Kb	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 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	-

Subpopulations and annotations (Discrete data)

	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%

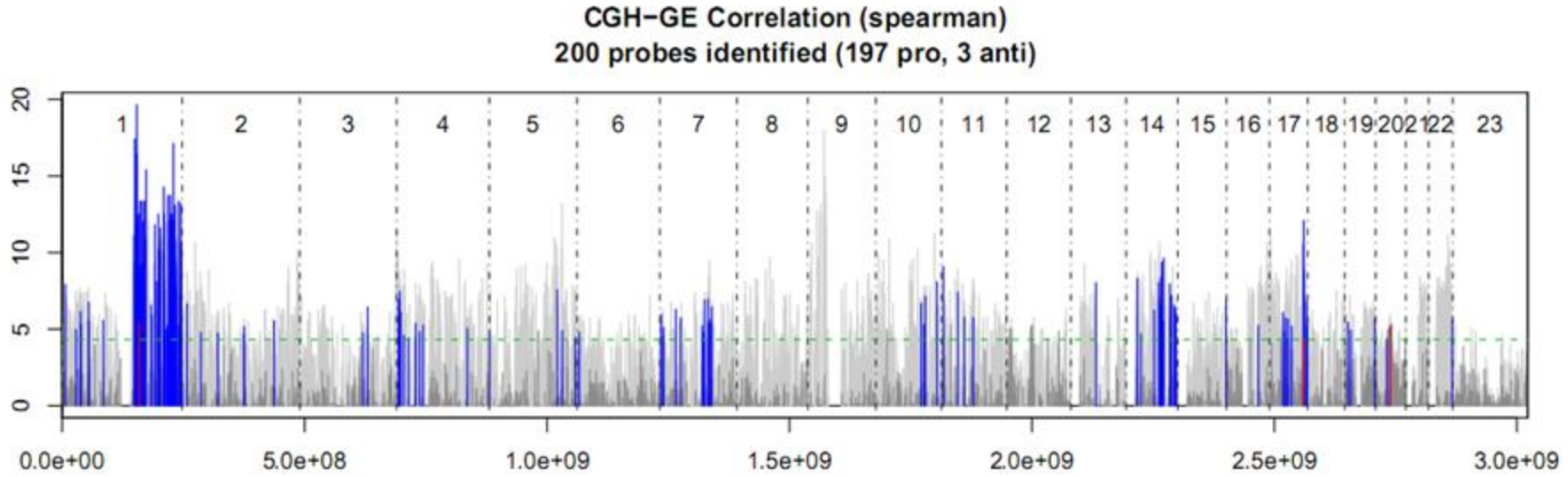
Subpopulations and annotations (Discrete data)

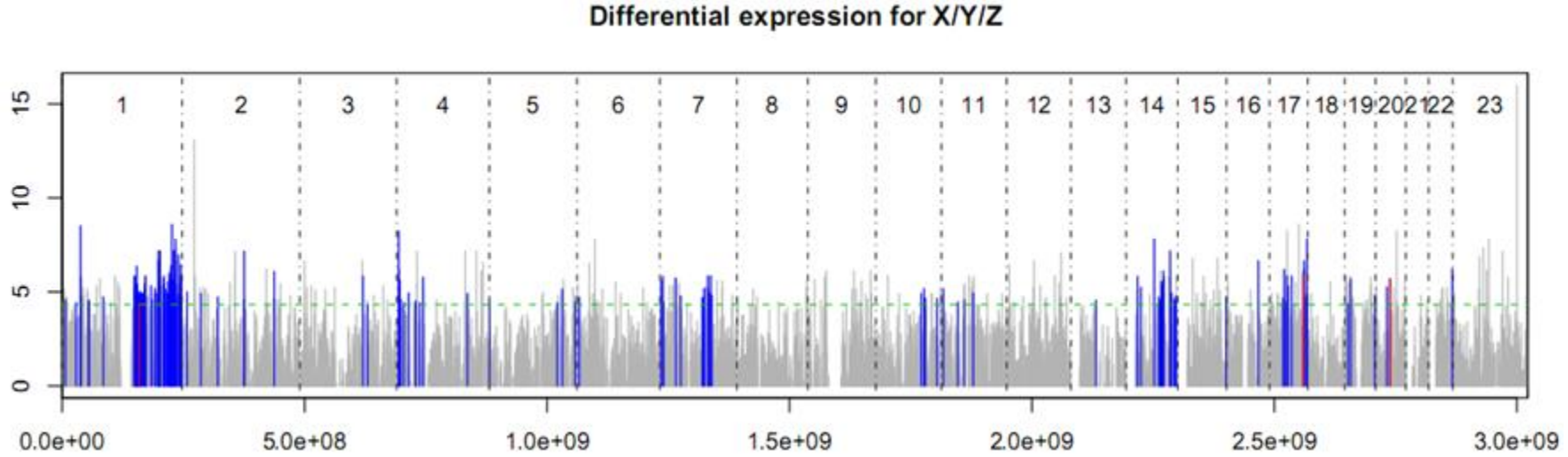


Fisher's exact test, adjusted P-values

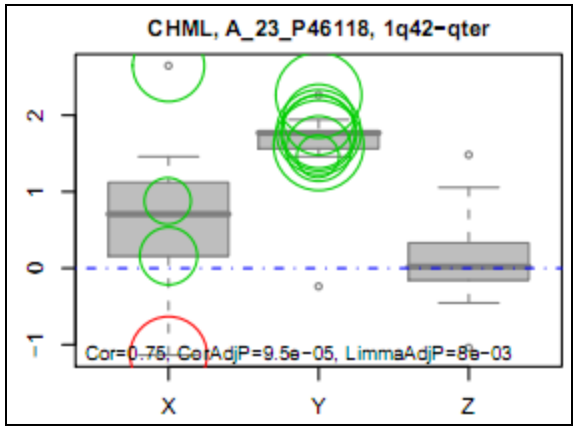
F-tests	Aberrant_Genome	Hclust_2branches
Age_median_68.5	8.19E-01	1.00E+00
Age_mean_67.45	4.11E-01	7.97E-01
Sex	1.06E-01	4.32E-01
TNM_Simplified	4.32E-01	3.75E-01
TNM_1	7.11E-01	1.00E+00
TNM_2	2.48E-01	3.29E-01
TNM_3+4	9.31E-01	5.92E-01
Differentiation	4.46E-01	3.44E-01
Fibrosis	2.37E-01	2.59E-01
Tissue_Invasion	5.88E-01	7.81E-01
Mutation_EGFR	9.25E-06	8.60E-05
Mutation_KRAS	2.44E-01	8.30E-02
EGFRwt_KRASwt	2.59E-05	6.87E-03

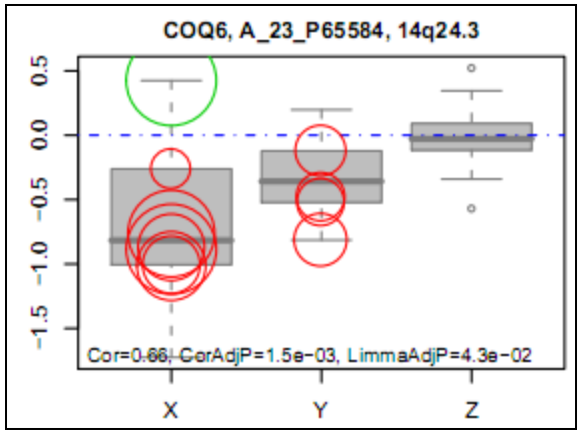
CNA+GEX integration : Correlation analysis

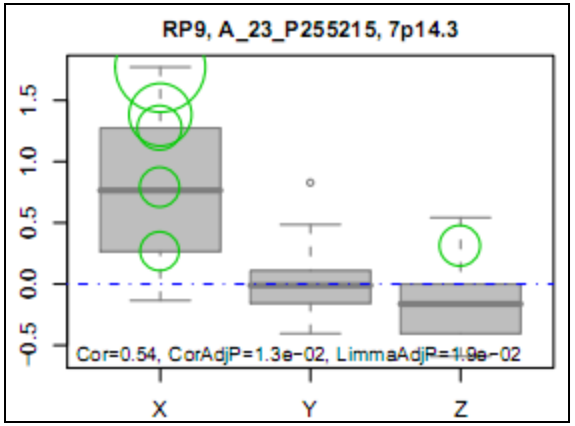




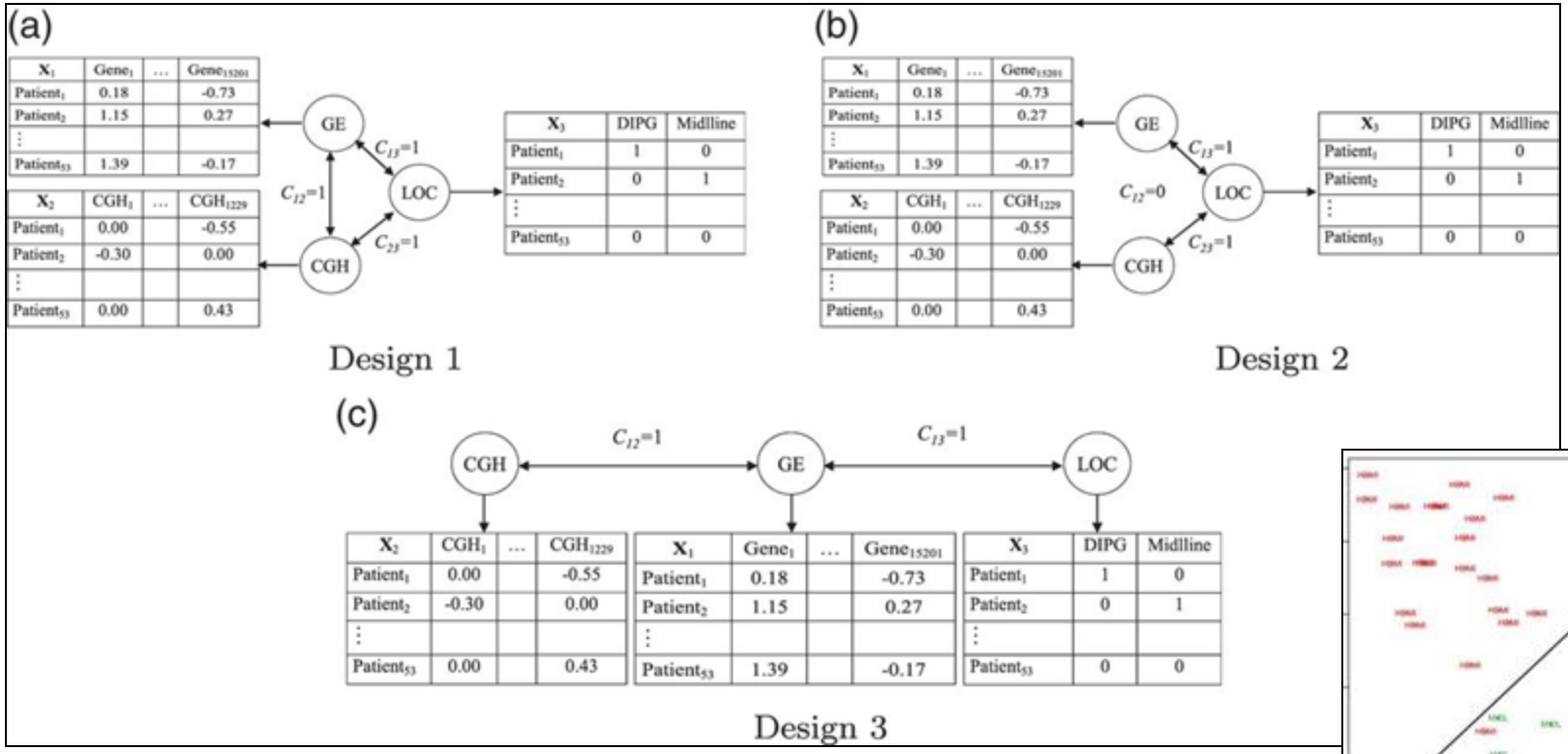
CNA+GEX integration : Correlation analysis

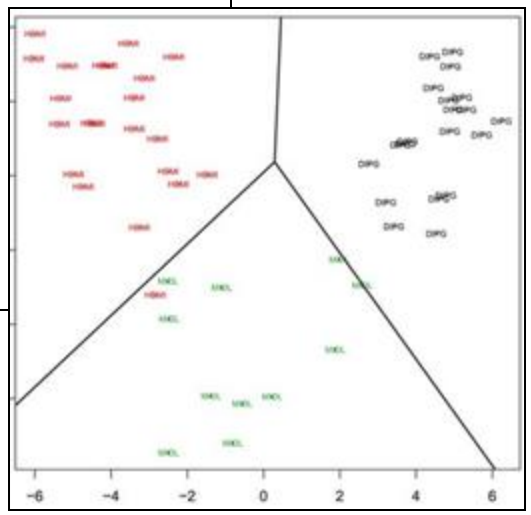






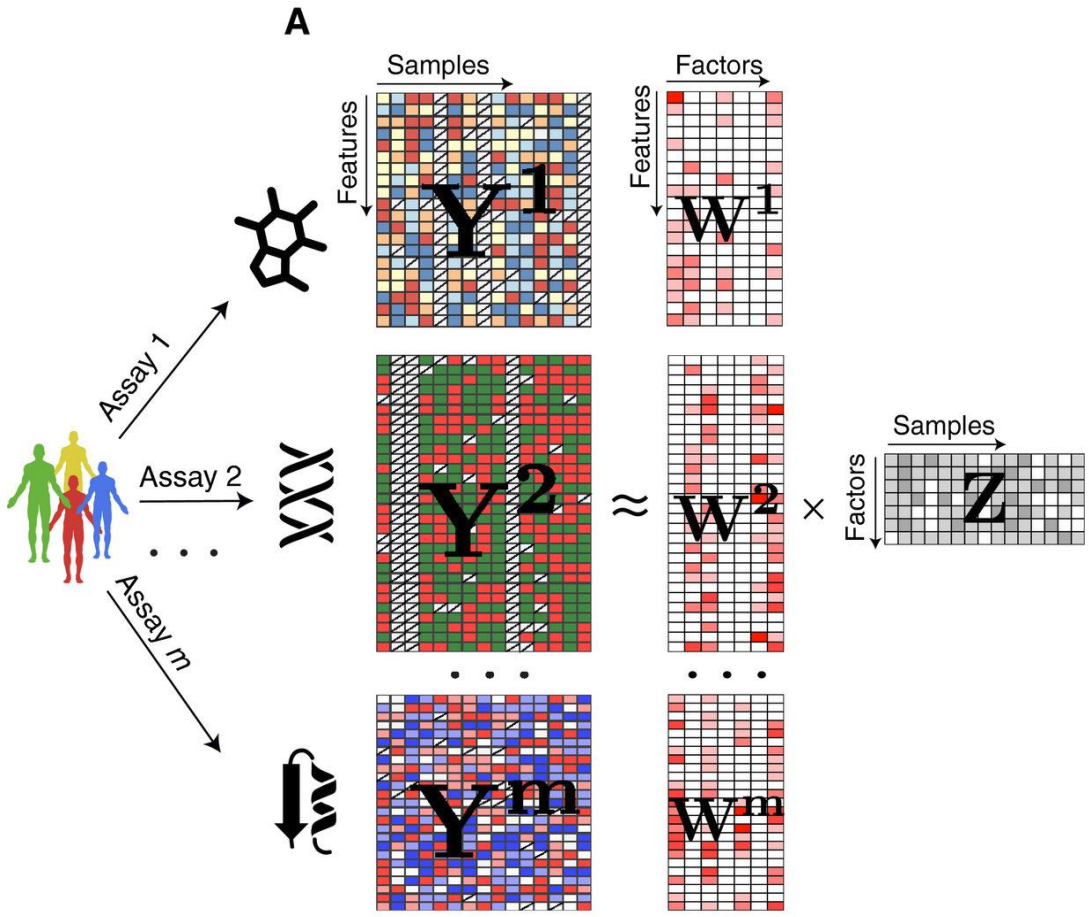
CNA+GEX integration : Multiblock analysis (RGCCA, SGCCA)

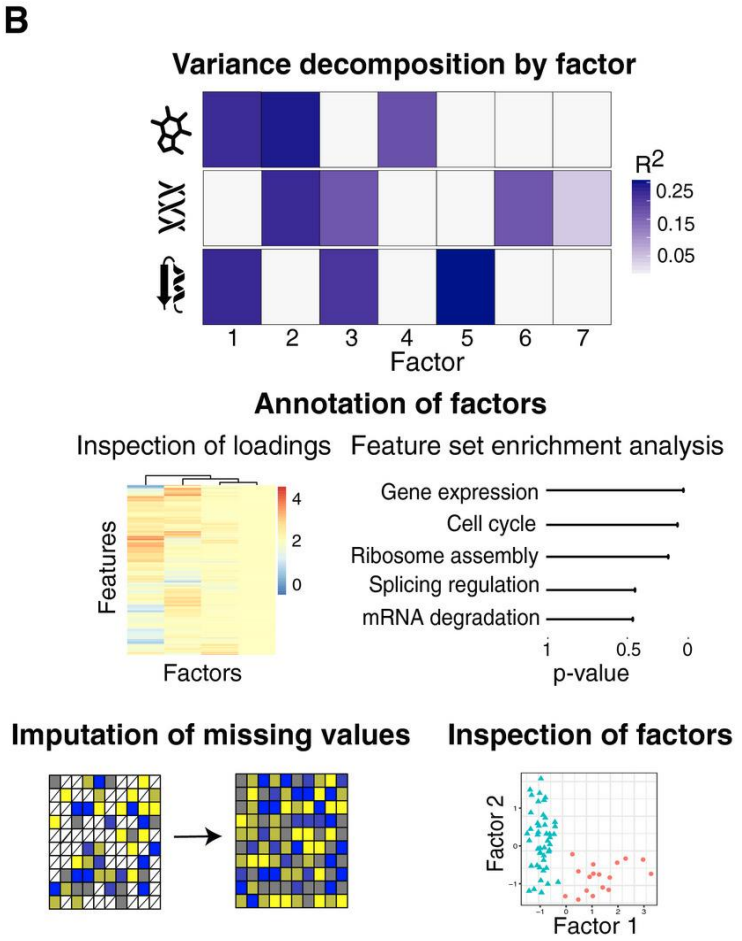




Philippe et Tenenhaus, Biostatistics, 2014

CNA+GEX integration : Factor analysis (MOFA, LIGER)





Arguelaget et al. 2018