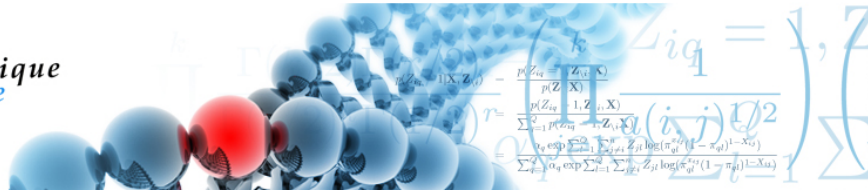


DNA copy number segmentation in cancer studies

Morgane Pierre-Jean
joint work with : Pierre Neuvial, Guillem Rigaill

Centre National de Recherche en Génomique Humaine (CNRGH), Evry

April 3rd, 2019

Outline

Outline

Objectives

Alterations can be observed at several levels

- Gene expression
- Methylation
- DNA structure
- Mutations
- DNA copy number

Why study genetic alterations in cancers ?

- Help to diagnosis
- Identify biomarkers linked to drug resistance
- Personalized treatments

Objectives

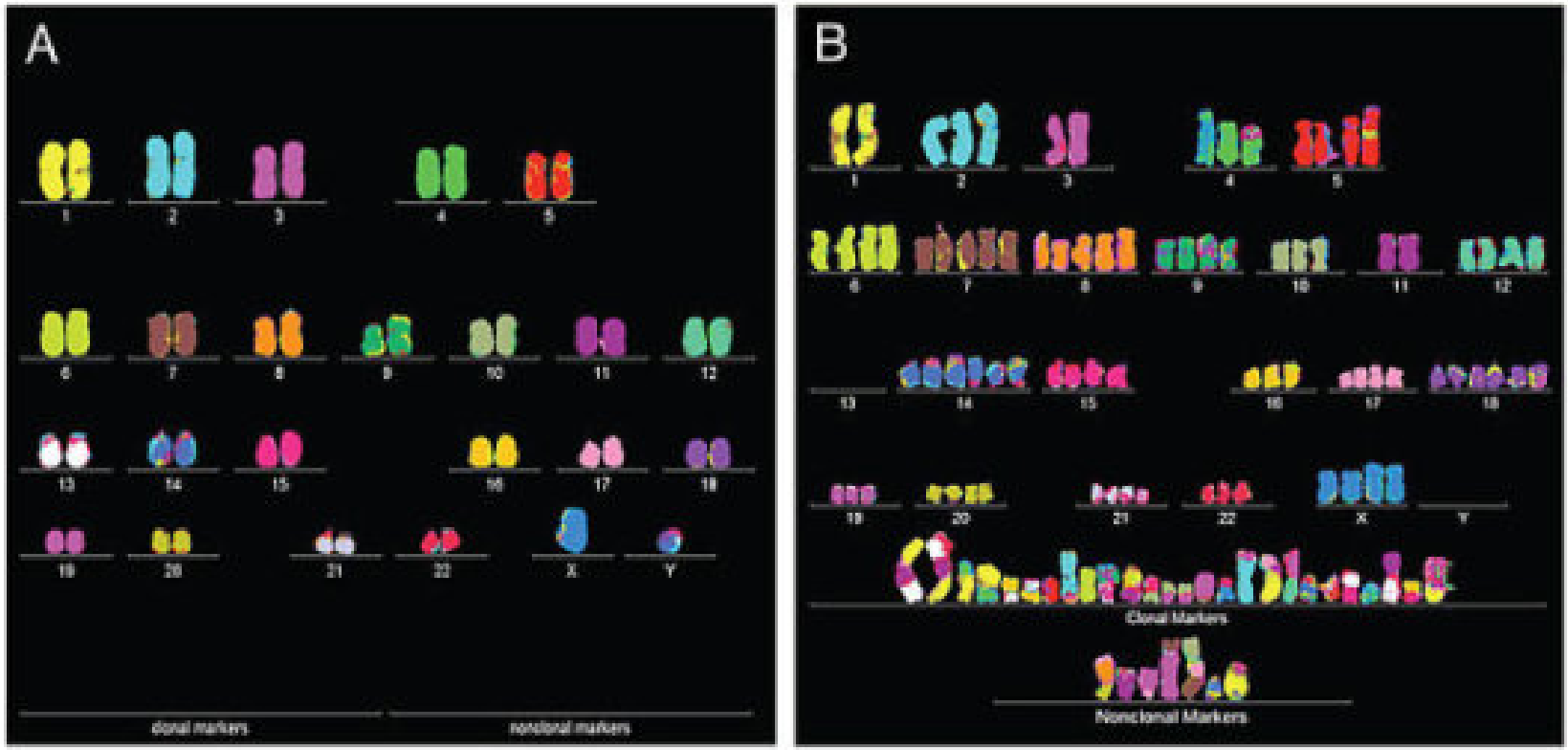
Alterations can be observed at several levels

- Gene expression
- Methylation
- DNA structure
- Mutations
- DNA copy number

Why study genetic alterations in cancers ?

- Help to diagnosis
- Identify biomarkers linked to drug resistance
- Personalized treatments

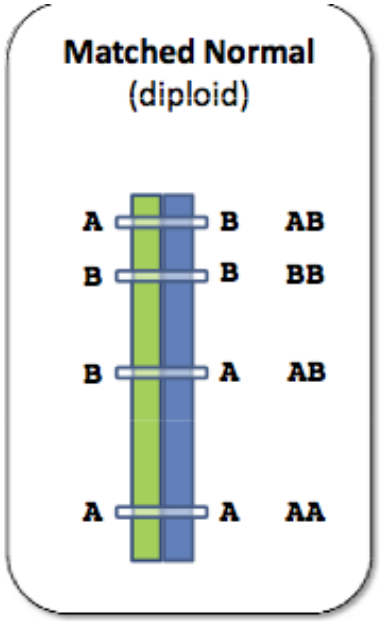
Karyotype



How to measure DNA copy number more precisely ?

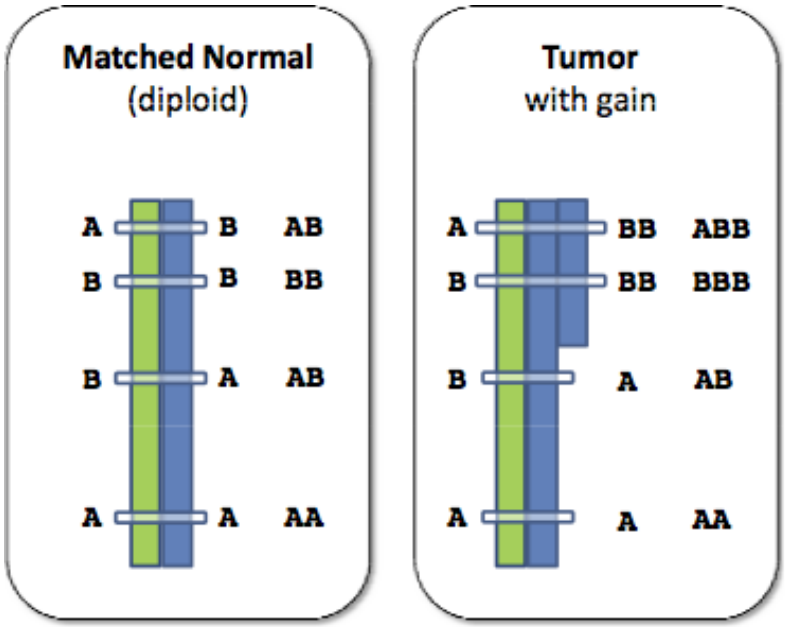
- CGH arrays (measuring total DNA copy number)
- SNP arrays (measuring quantity of alleles for predefined positions)
- Sequencing technologies (WGS or WES)

Illustration of alterations at level of DNA copy number



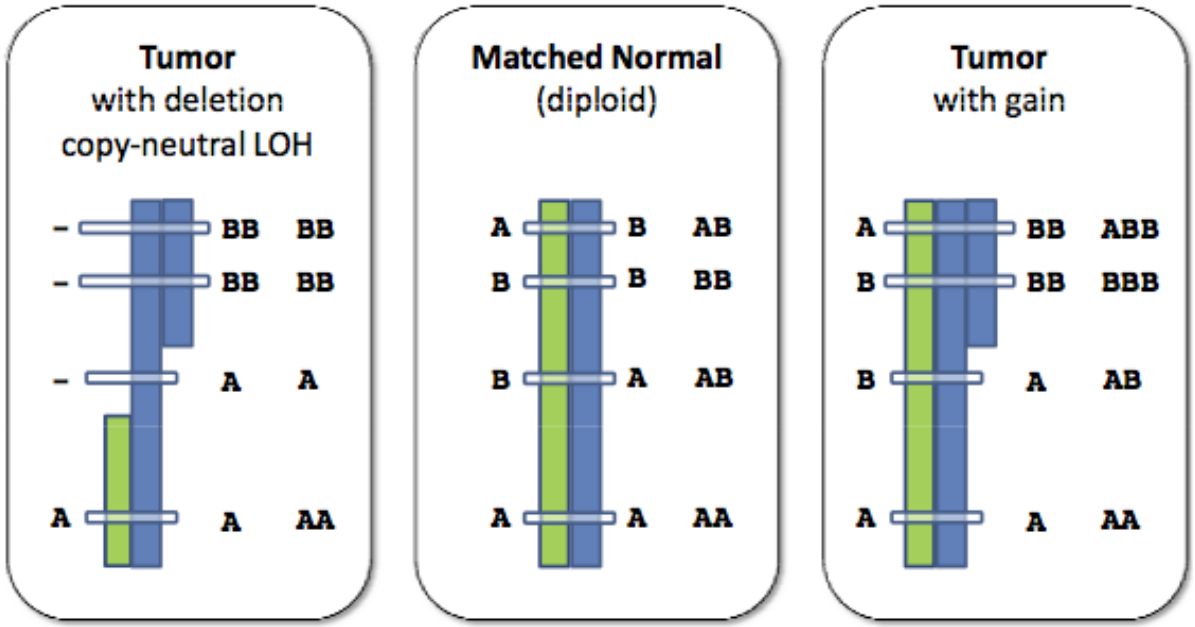
Slide H. Bengtsson

Illustration of alterations at level of DNA copy number



Slide H. Bengtsson

Illustration of alterations at level of DNA copy number

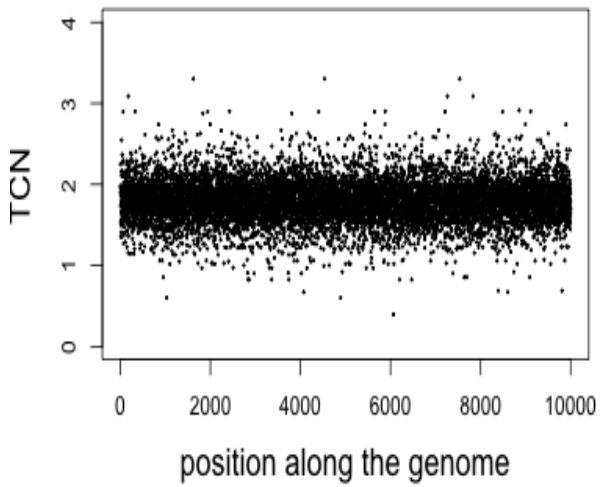


Slide H. Bengtsson

What kind of signals from SNPs arrays ?

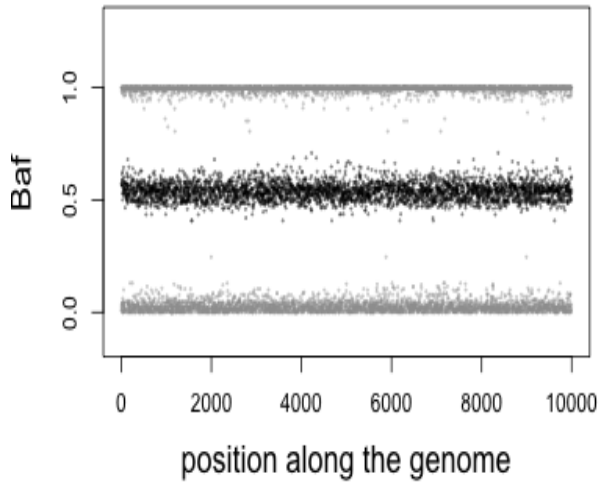
Total copy number

$$c_j = N_j^A + N_j^B$$



B allele fraction

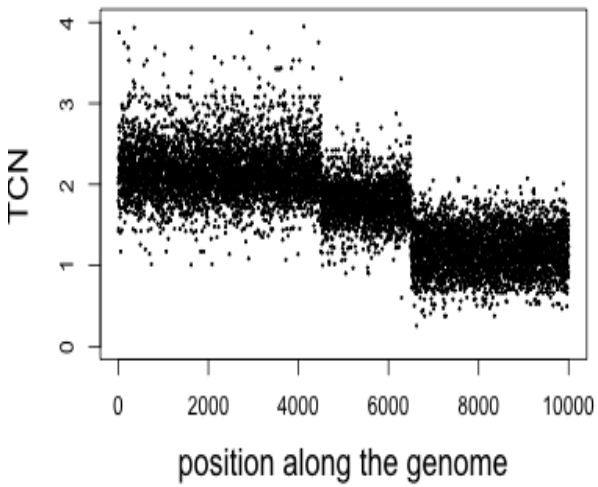
$$b_j = \frac{N_j^B}{c_j}$$



What kind of signals from SNPs arrays ?

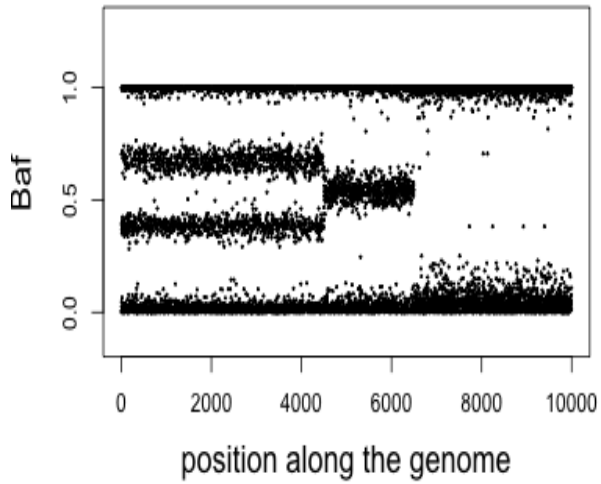
Total copy number

$$c_j = N_j^A + N_j^B$$



B allele fraction

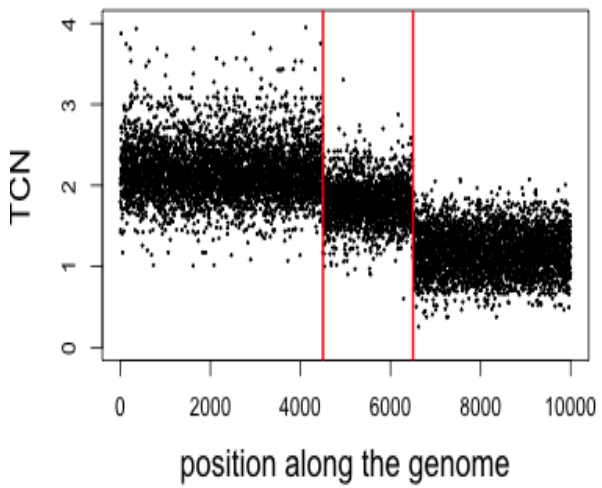
$$b_j = \frac{N_j^B}{c_j}$$



What kind of signals from SNPs arrays ?

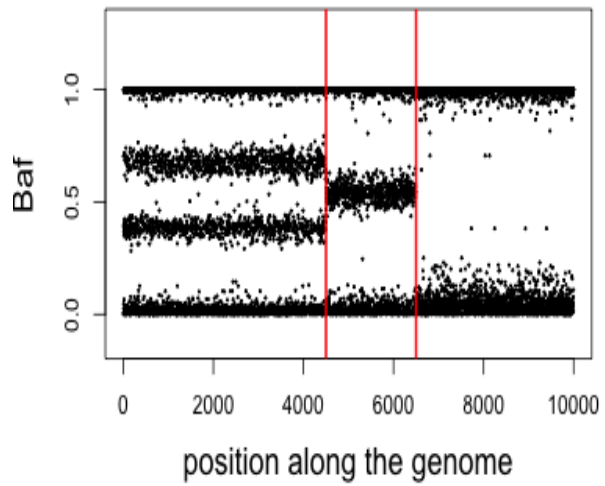
Total copy number

$$c_j = N_j^A + N_j^B$$



B allele fraction

$$b_j = \frac{N_j^B}{c_j}$$



Outline

Binary Segmentation : change-point detection in the mean

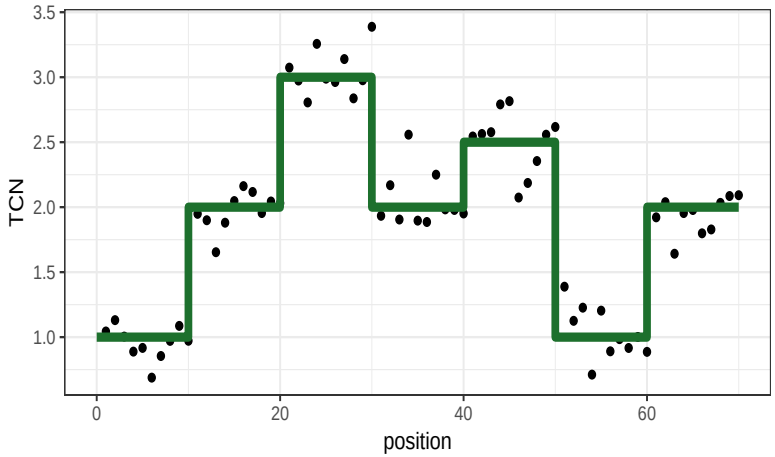
- Take the simple case : dimension is equal to 1 ($d = 1$) :
- Hypothesis : $\mathcal{H}_0$: No breakpoint vs $\mathcal{H}_1$: Exactly one breakpoint.
- The likelihood ratio statistic is given by $\max_{1 \leq i \leq n} |Z_i|$

$$Z_i = \frac{\left(\frac{S_i}{i} - \frac{S_n - S_i}{n-i} \right)}{\sqrt{\frac{1}{i} + \frac{1}{n-i}}}, \tag{1}$$

And $S_i = \sum_{1 \leq l \leq i} y_l$

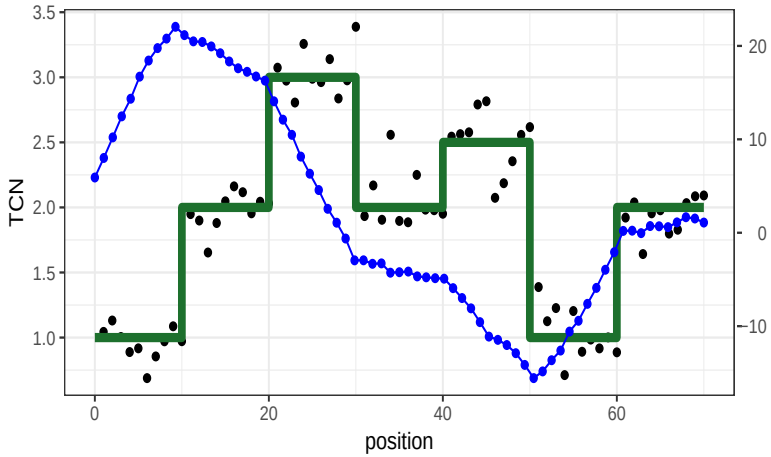
Complexity : $O(dnlog(K))$

- First breakpoint
- For each i : we compute Z_i :
  $b_1 = \arg\max_{1 \leq i \leq n} Z_i$



Complexity : $O(dnlog(K))$

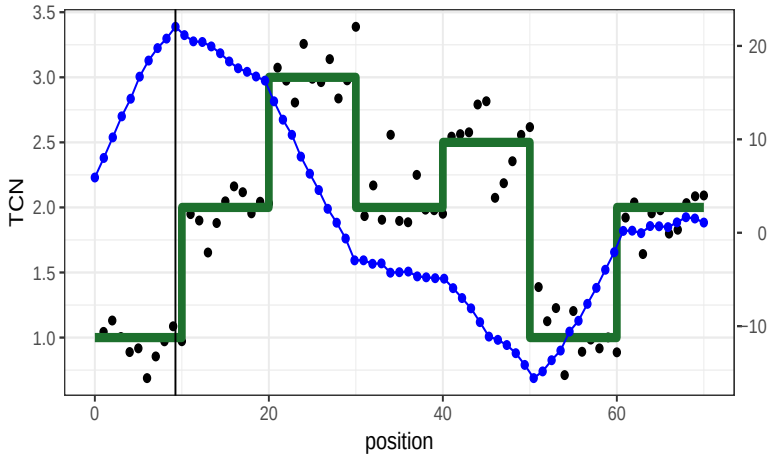
- First breakpoint
- For each i : we compute Z_i :
  $b_1 = \arg\max_{1 \leq i \leq n} Z_i$



Recursive Binary Segmentation (RBS) (Gey and Lebarbier, 2008)

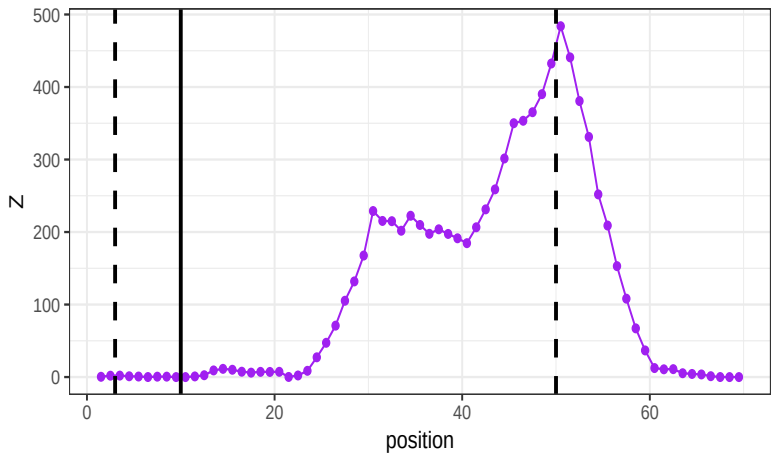
Complexity : $O(dnlog(K))$

- First breakpoint
- For each i : we compute Z_i :
  $b_1 = \arg\max_{1 \leq i \leq n} Z_i$



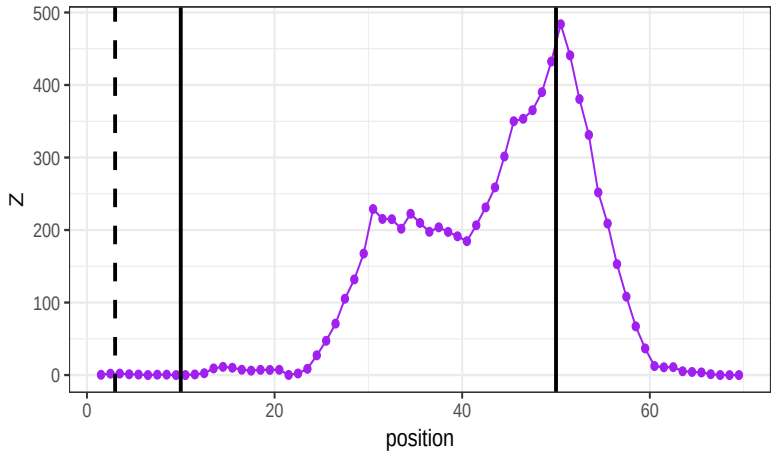
Second step of RBS

- Second breakpoint :
  - $\max_{1 \leq i \leq b_1} Z_i^2$
  - $\max_{b_1 \leq i \leq n} Z_i^2$
- Compute RSE for each segment.
- Keep the RSE which bring the maximum gain
- Add the breakpoint to the active set



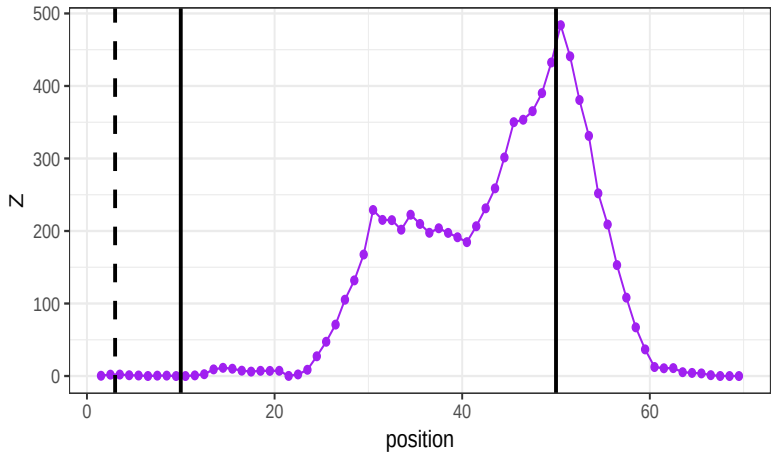
Second step of RBS

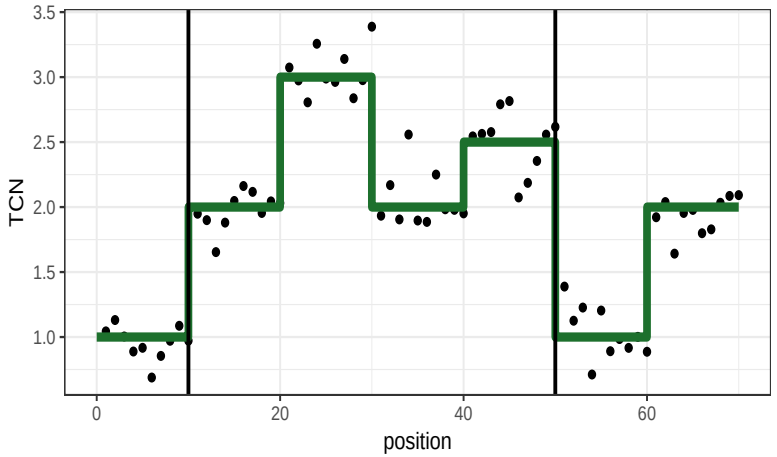
- Second breakpoint :
  - $\max_{1 \leq i \leq b_1} Z_i^2$
  - $\max_{b_1 \leq i \leq n} Z_i^2$
- Compute RSE for each segment.
- Keep the RSE which bring the maximum gain
- Add the breakpoint to the active set



Second step of RBS

- Second breakpoint :
  - $\max_{1 \leq i \leq b_1} Z_i^2$
  - $\max_{b_1 \leq i \leq n} Z_i^2$
- Compute RSE for each segment.
- Keep the RSE which bring the maximum gain
- Add the breakpoint to the active set

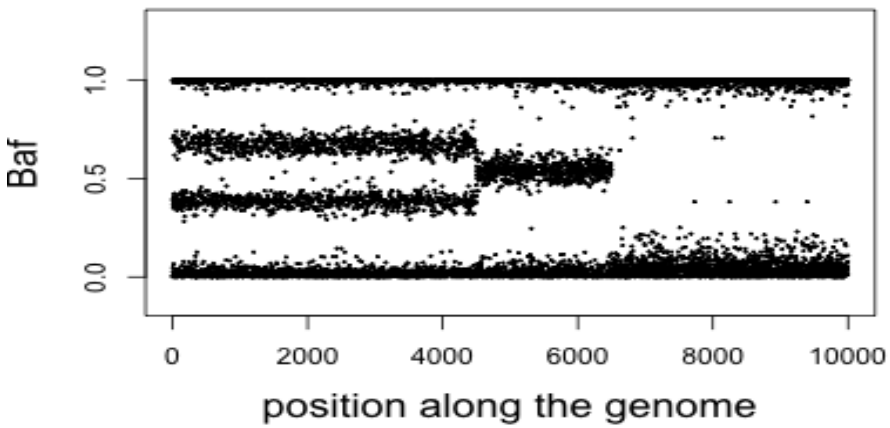


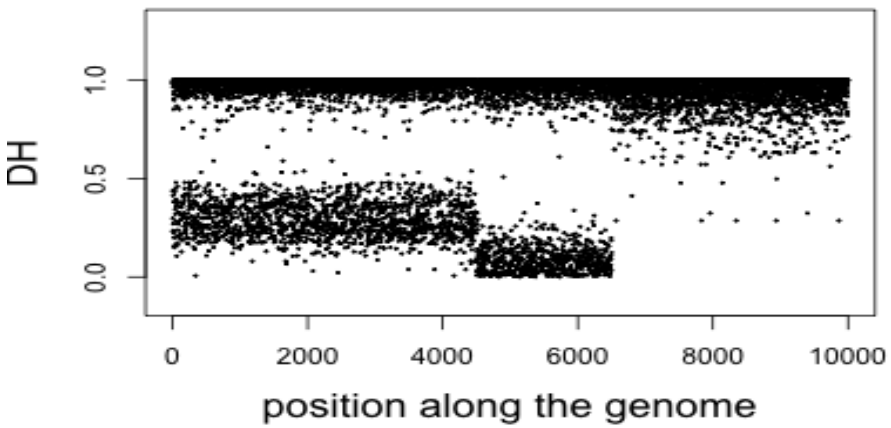


Define decrease of heterozygosity to segment BAF

Decrease of heterozygosity

$$DH = 2 \times |BAF - \tfrac{1}{2}|$$

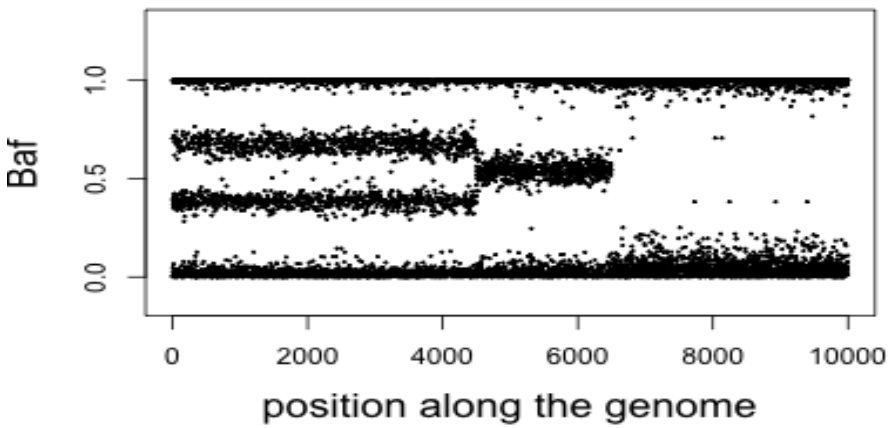


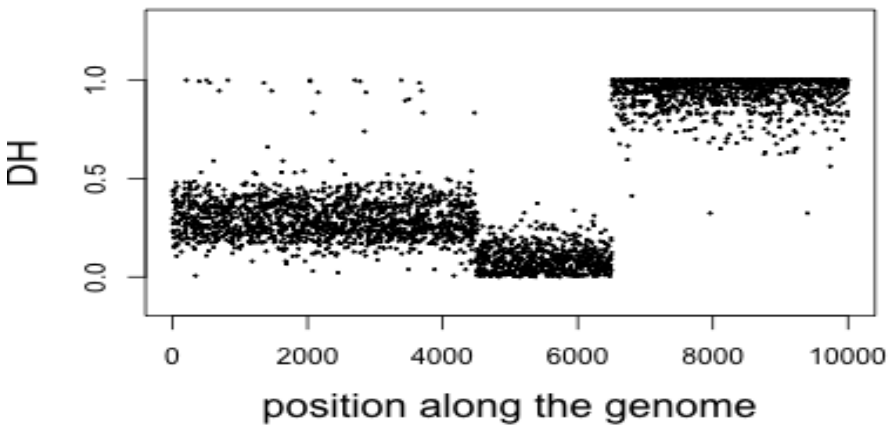


Define decrease of heterozygosity to segment BAF

Decrease of heterozygosity

$$DH = 2 \times |BAF - \tfrac{1}{2}|$$

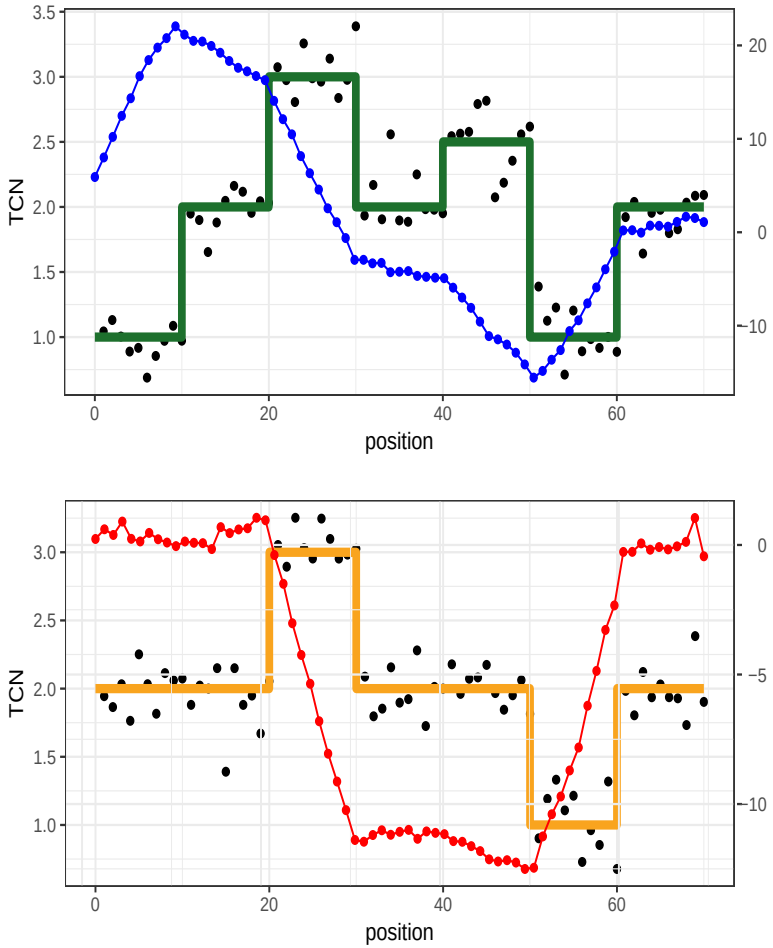




Joint RBS (1)

Complexity : $O(dnlog(K))$

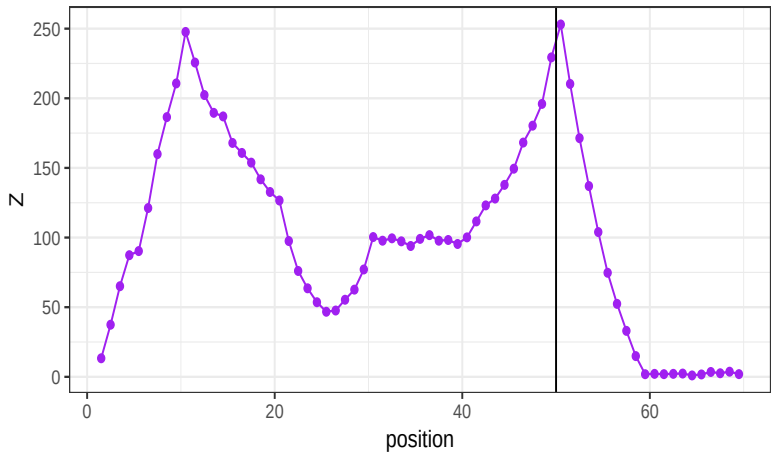
- First breakpoint
- For each i : we compute $Z_i = (Z_1, Z_2)_i : t_1 = \arg\max_{1 \leq i \leq n} \|Z_i\|_2^2$



Joint RBS (1)

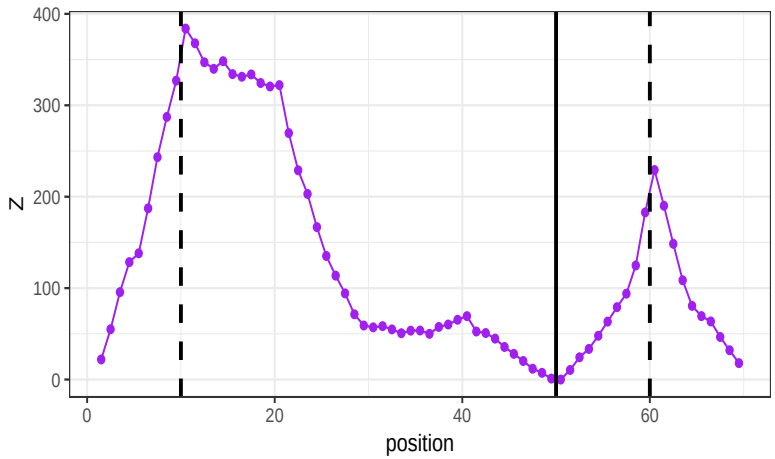
Complexity : $O(dnlog(K))$

- First breakpoint
- For each i : we compute $Z_i = (Z_1, Z_2)_i : t_1 = \arg\max_{1 \leq i \leq n} \|Z_i\|_2^2$



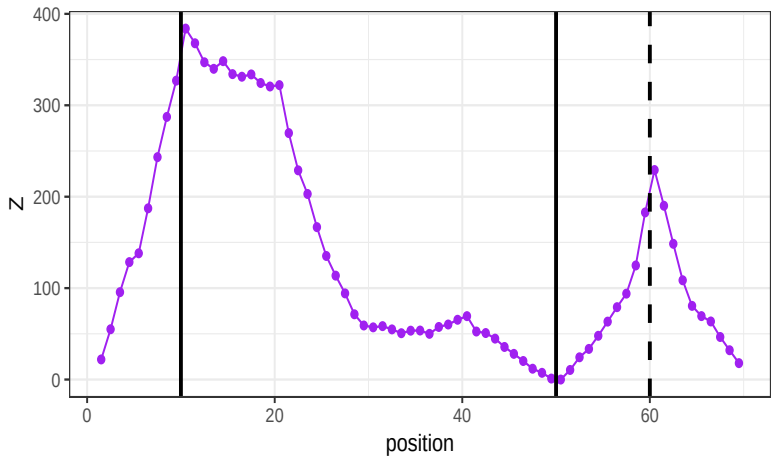
Joint RBS (2)

- Second breakpoint :
  - $\max_{1 \leq i \leq t_1} \|Z_i\|_2^2$
  - $\max_{t_1 < i \leq n} \|Z_i\|_2^2$
- Compute RSE for each segment.
- Keep the RSE which bring the maximum gain
- Add the breakpoint to the active set



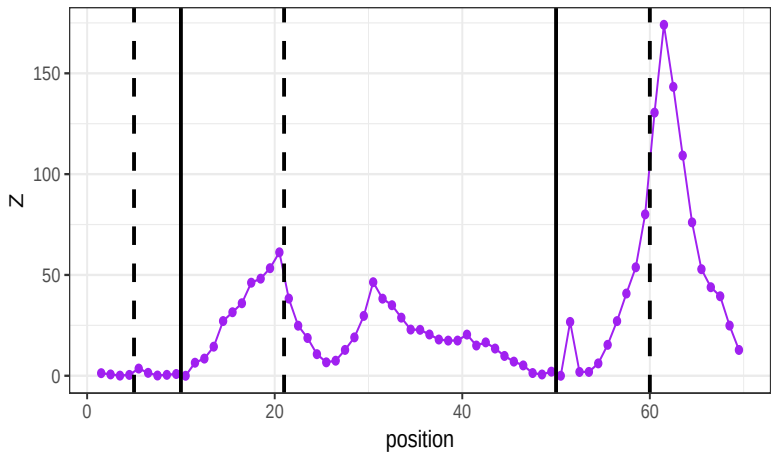
Joint RBS (2)

- Second breakpoint :
  - $\max_{1 \leq i \leq t_1} \|Z_i\|_2^2$
  - $\max_{t_1 < i \leq n} \|Z_i\|_2^2$
- Compute RSE for each segment.
- Keep the RSE which bring the maximum gain
- Add the breakpoint to the active set



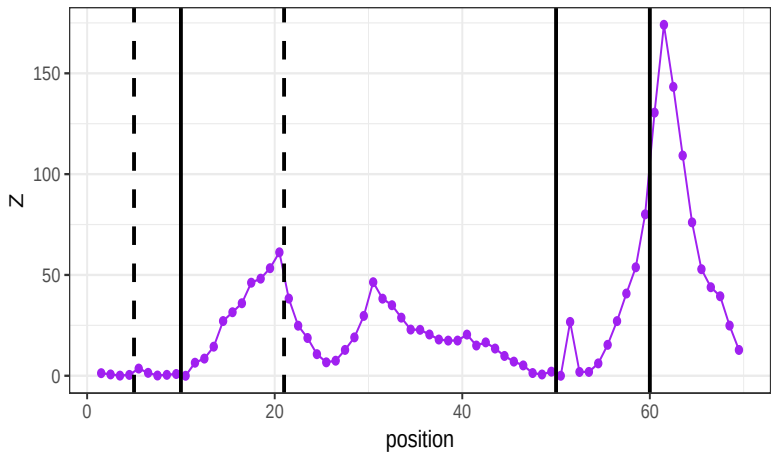
Joint RBS (3)

- Third breakpoint :
  - $\max_{1 \leq i \leq t_1} \|Z_i\|_2^2$
  - $\max_{t_1 < i \leq t_2} \|Z_i\|_2^2$
  - $\max_{t_2 < i \leq n} \|Z_i\|_2^2$
- Compute RSE for each segment.
- Keep the RSE which bring the maximum gain
- Add the breakpoint to the active set



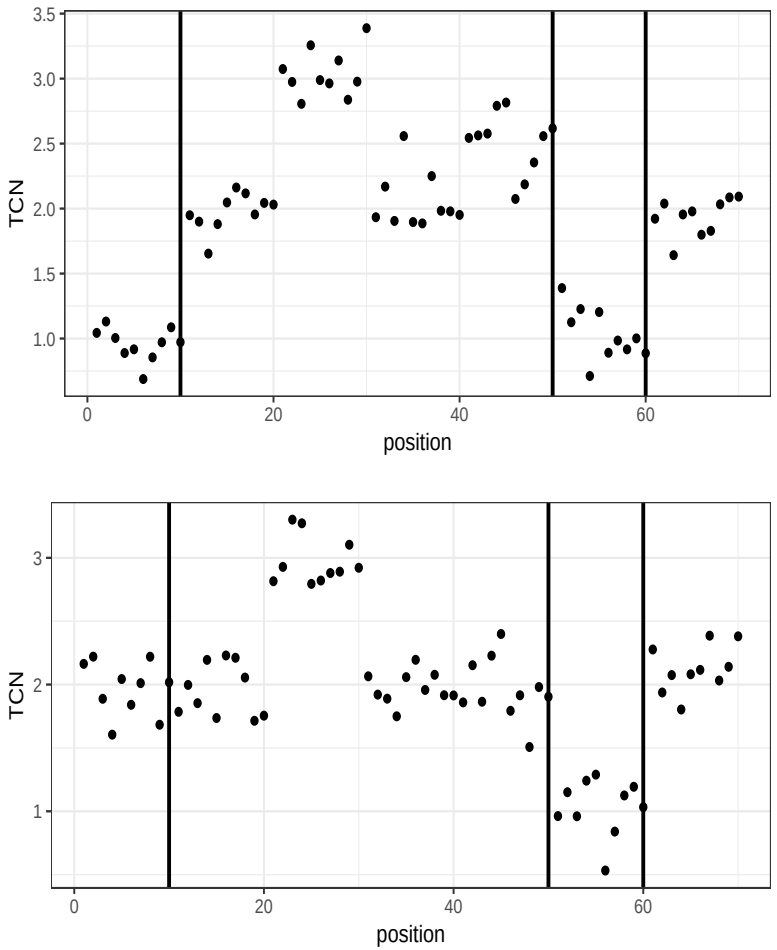
Joint RBS (3)

- Third breakpoint :
  - $\max_{1 \leq i \leq t_1} \|Z_i\|_2^2$
  - $\max_{t_1 < i \leq t_2} \|Z_i\|_2^2$
  - $\max_{t_2 < i \leq n} \|Z_i\|_2^2$
- Compute RSE for each segment.
- Keep the RSE which bring the maximum gain
- Add the breakpoint to the active set



Joint RBS (3)

- Third breakpoint :
  - $\max_{1 \leq i \leq t_1} \|Z_i\|_2^2$
  - $\max_{t_1 < i \leq t_2} \|Z_i\|_2^2$
  - $\max_{t_2 < i \leq n} \|Z_i\|_2^2$
- Compute RSE for each segment.
- Keep the RSE which bring the maximum gain
- Add the breakpoint to the active set



Outline

How generate realistic data ?

1. Manual annotation

How generate realistic data ?

1. Manual annotation
2. Difficulty is controlled with the proportion of tumor cells

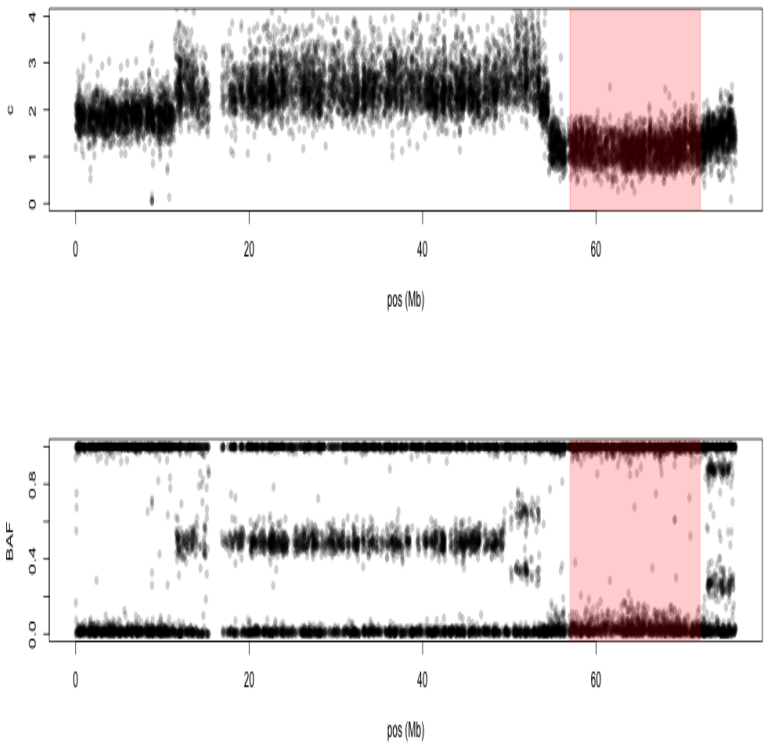
How generate realistic data ?

1. Manual annotation
2. Difficulty is controlled with the proportion of tumor cells
3. Resampling

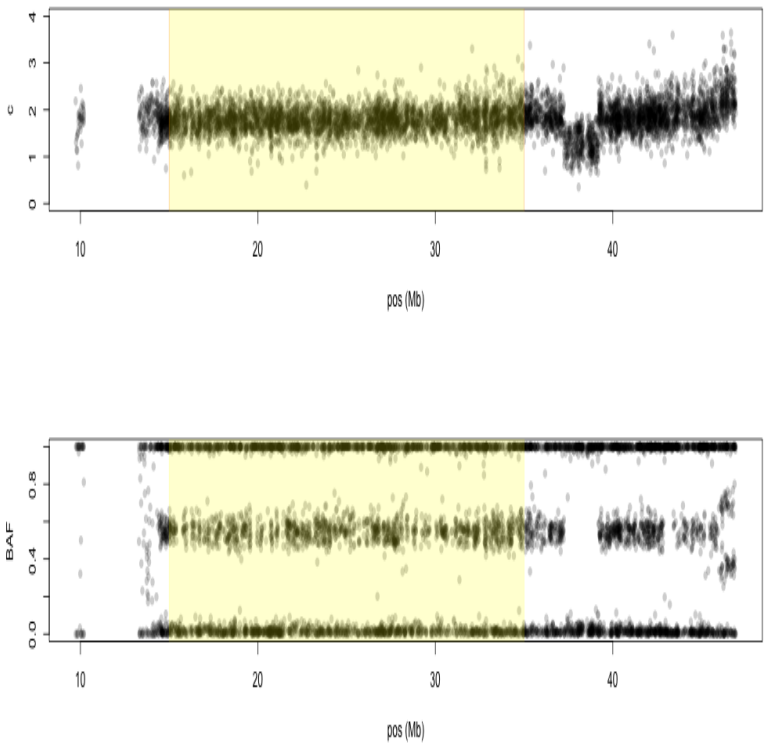
Real data annotation

Step 1- Annotate a real data set

Loss of one copy (Chr18)



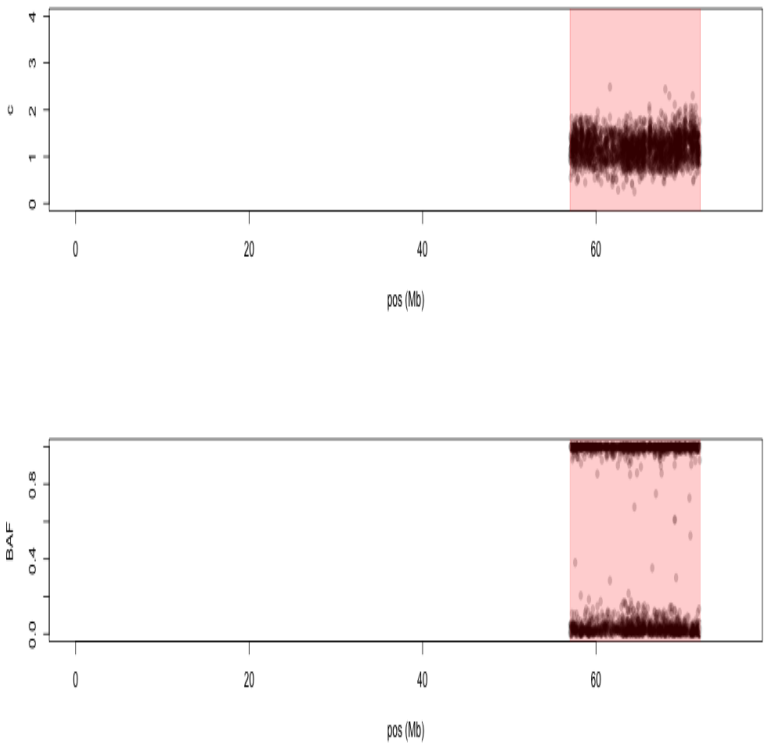
Normal region (Chr21)



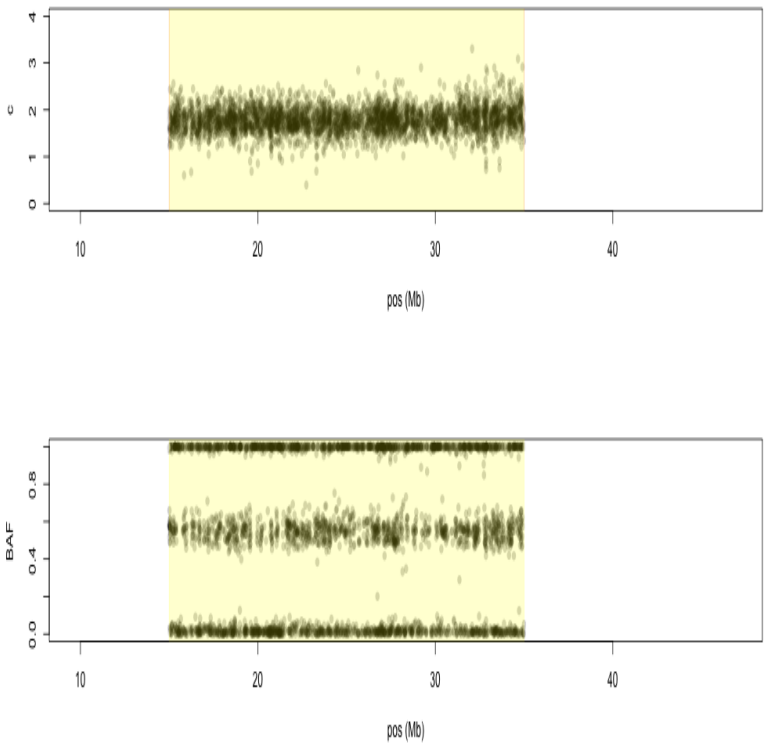
Real data annotation

Step 1- Annotate a real data set

Loss of one copy (Chr18)

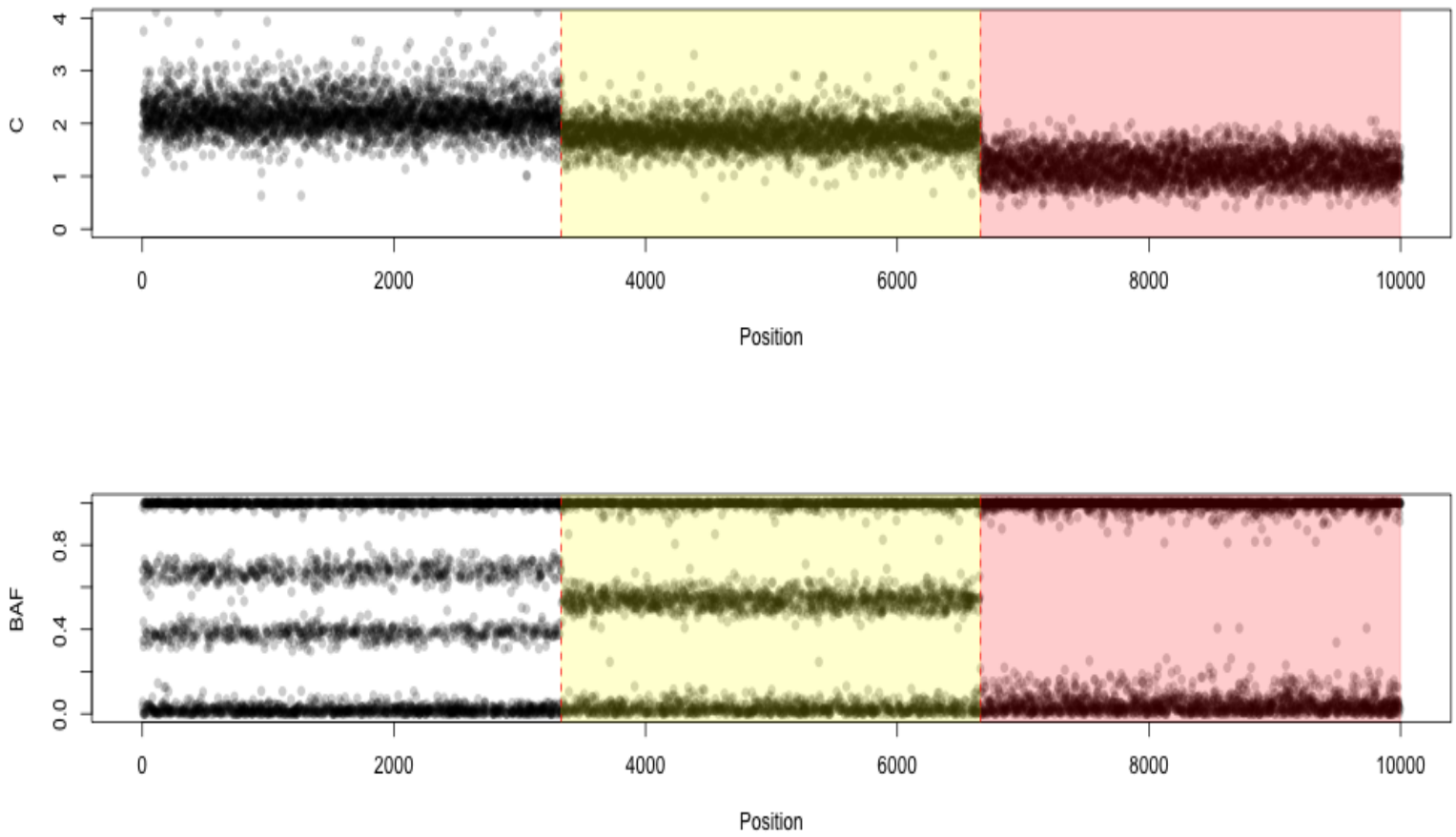


Normal region (Chr21)



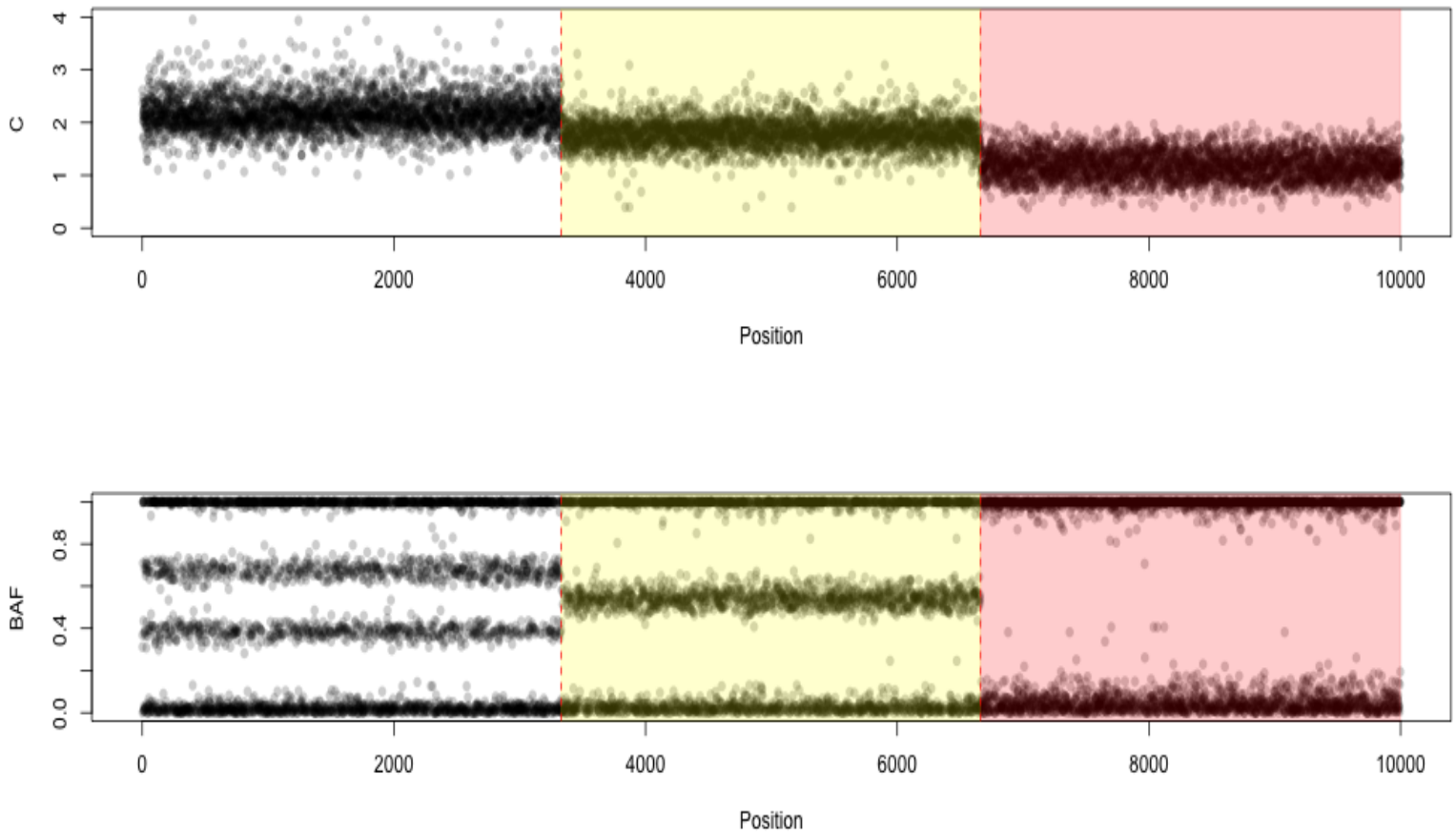
Data generation by resampling

Step 2 - Synthetic data generation by resampling
100% tumor cells



Data generation by resampling

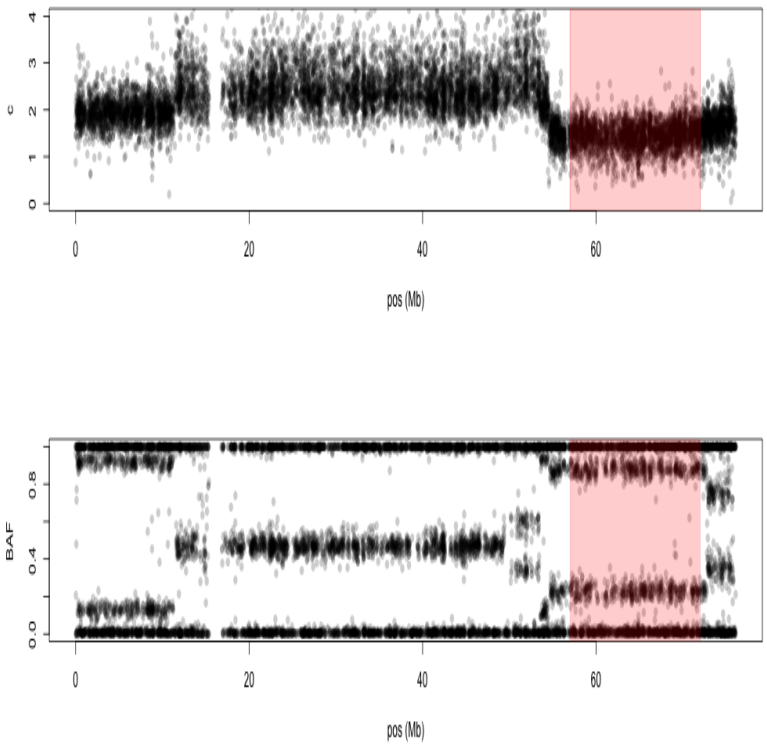
Step 2 - Synthetic data generation by resampling
100% tumor cells - same truth



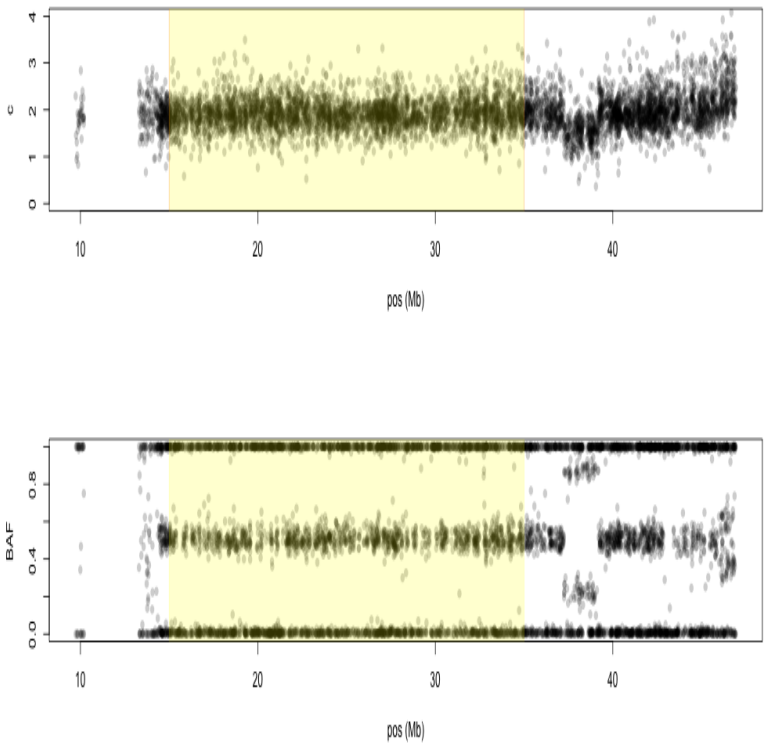
Real data annotation

70%- Using annotation from 100% data set

Loss of one copy (Chr18)



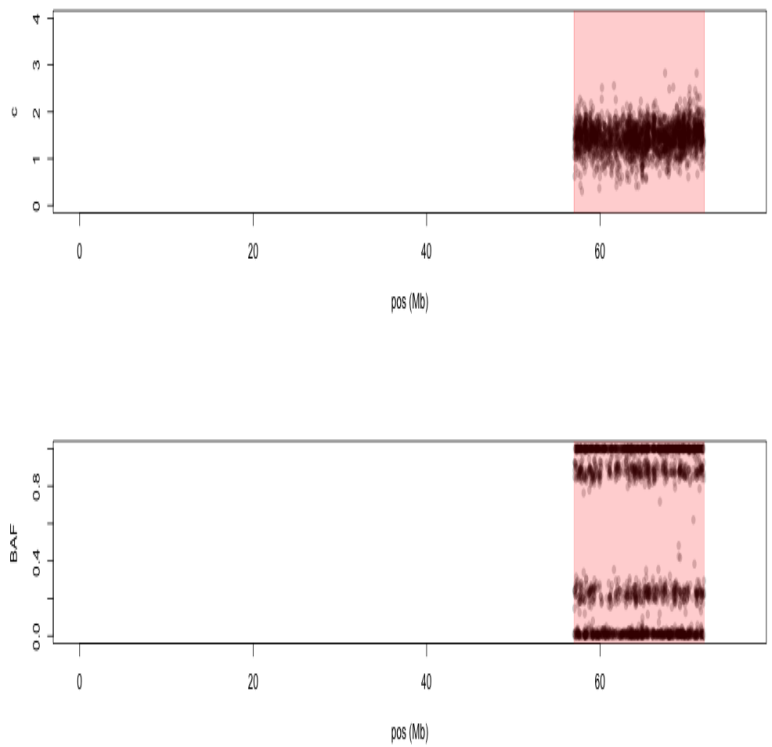
Normal region (Chr21)



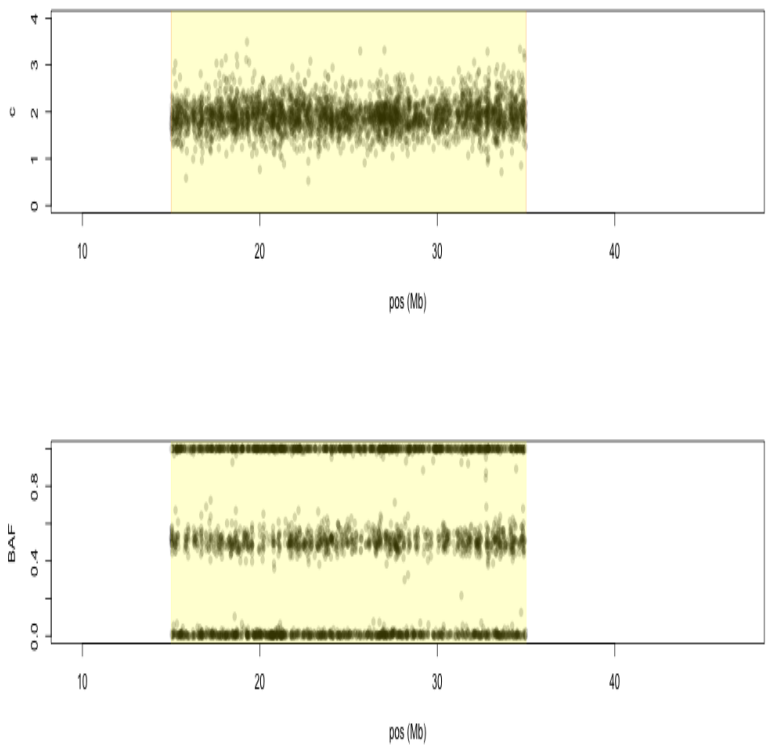
Real data annotation

70%- Using annotation from 100% data set

Loss of one copy (Chr18)



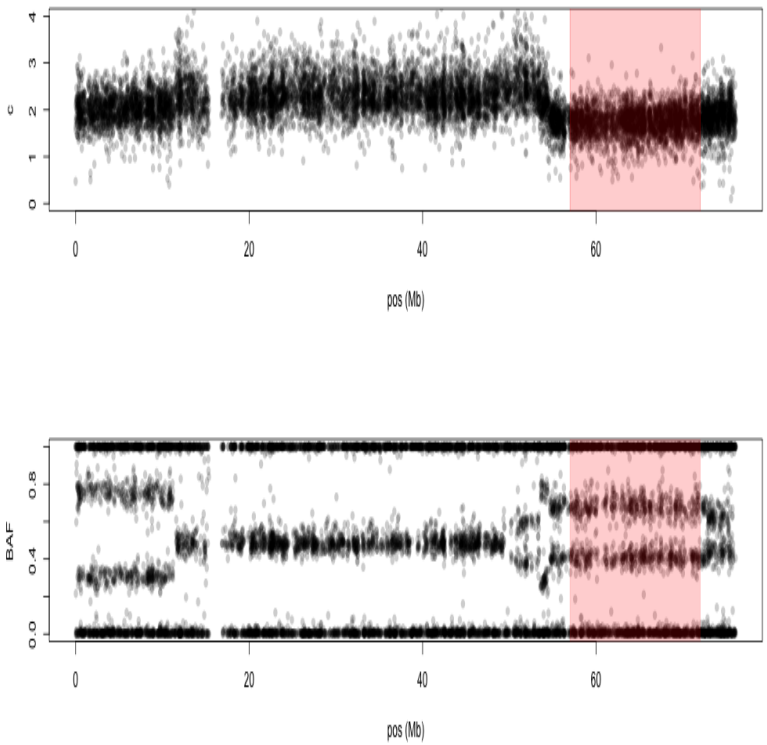
Normal region (Chr21)



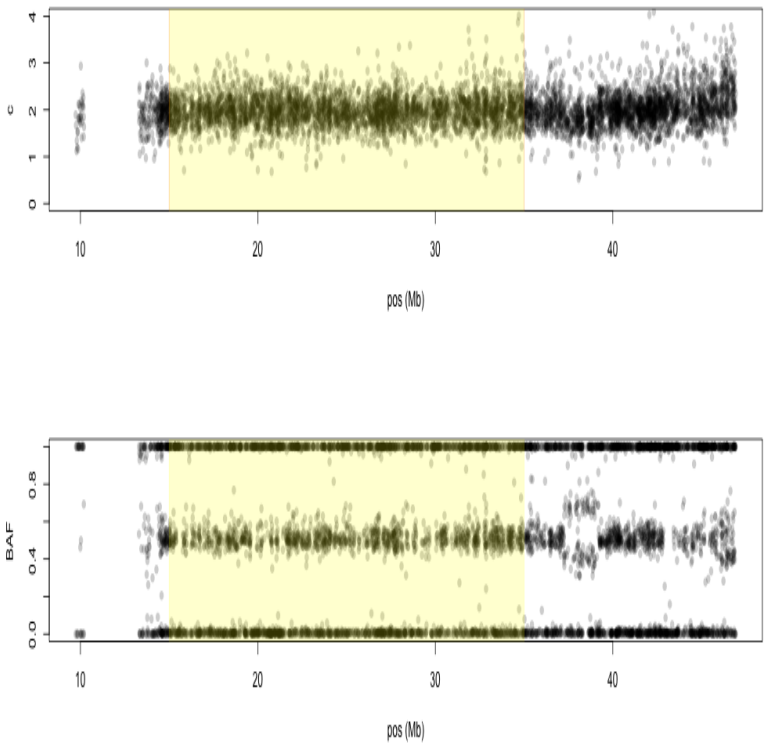
Real data annotation

50%- Using annotation from 100% data set

Loss of one copy (Chr18)



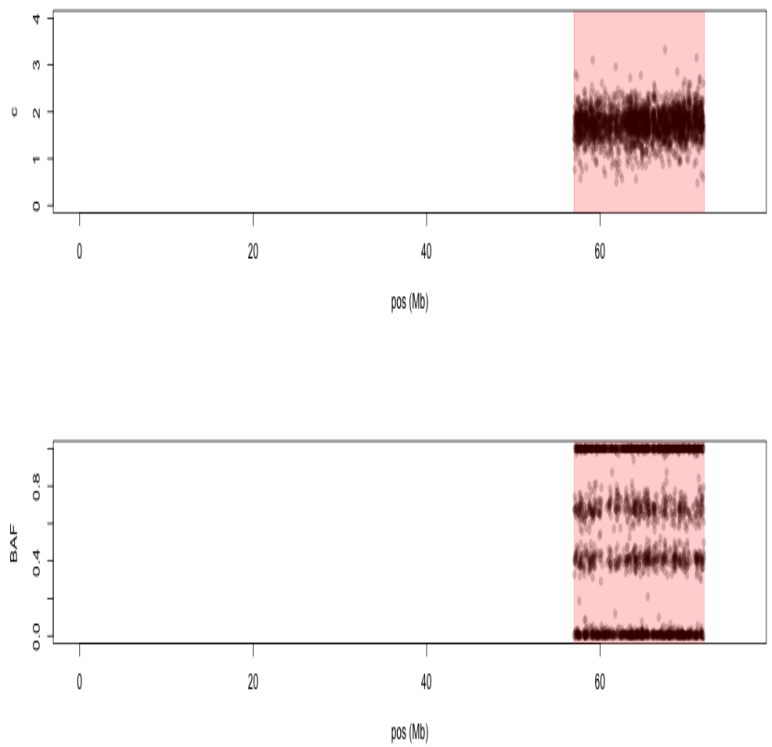
Normal region (Chr21)



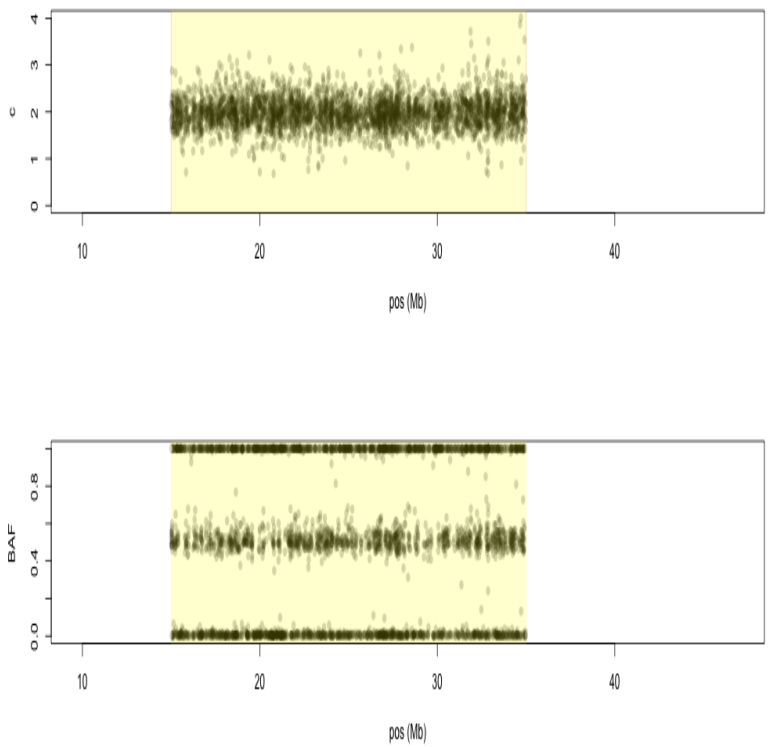
Real data annotation

50%- Using annotation from 100% data set

Loss of one copy (Chr18)

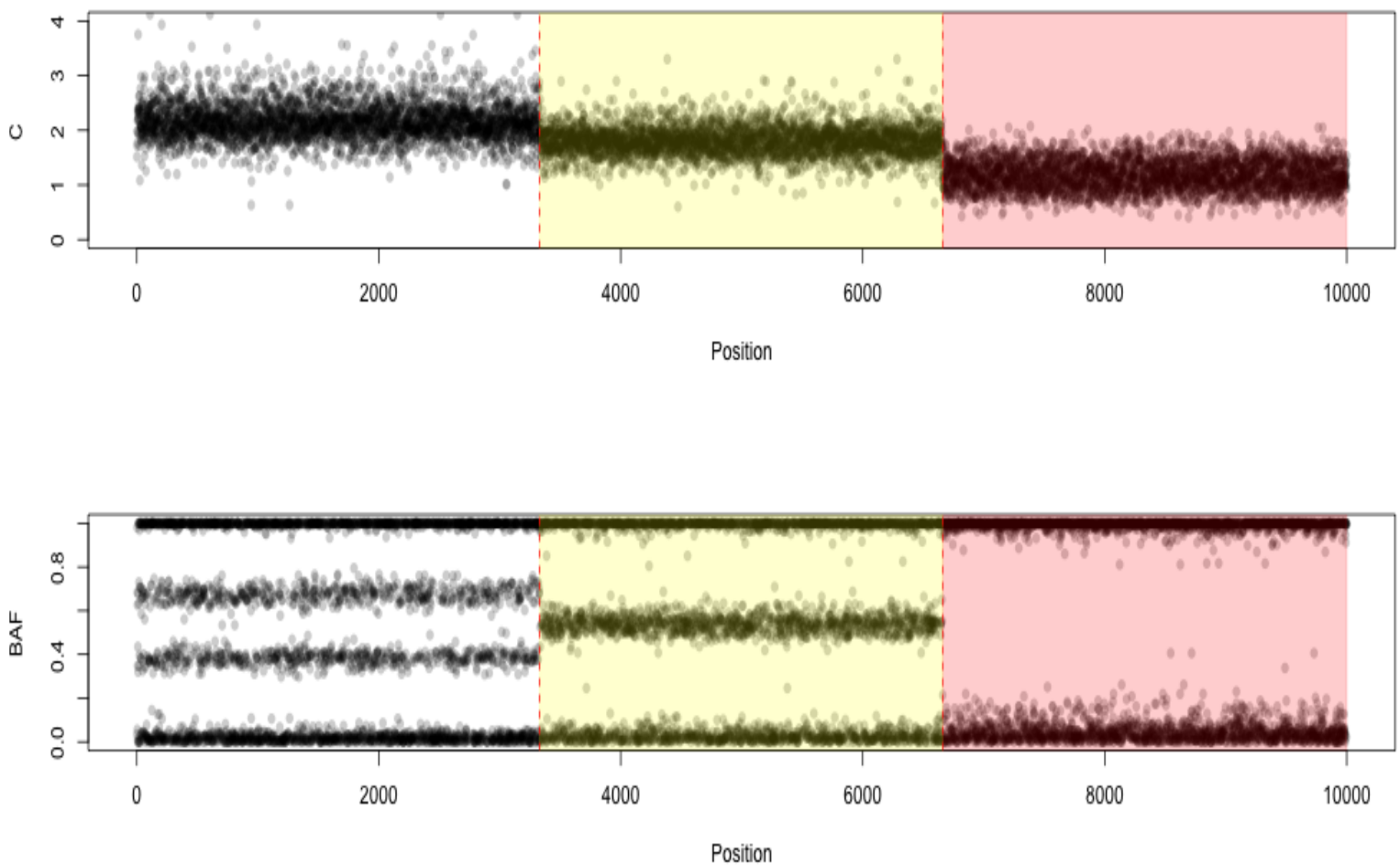


Normal region (Chr21)



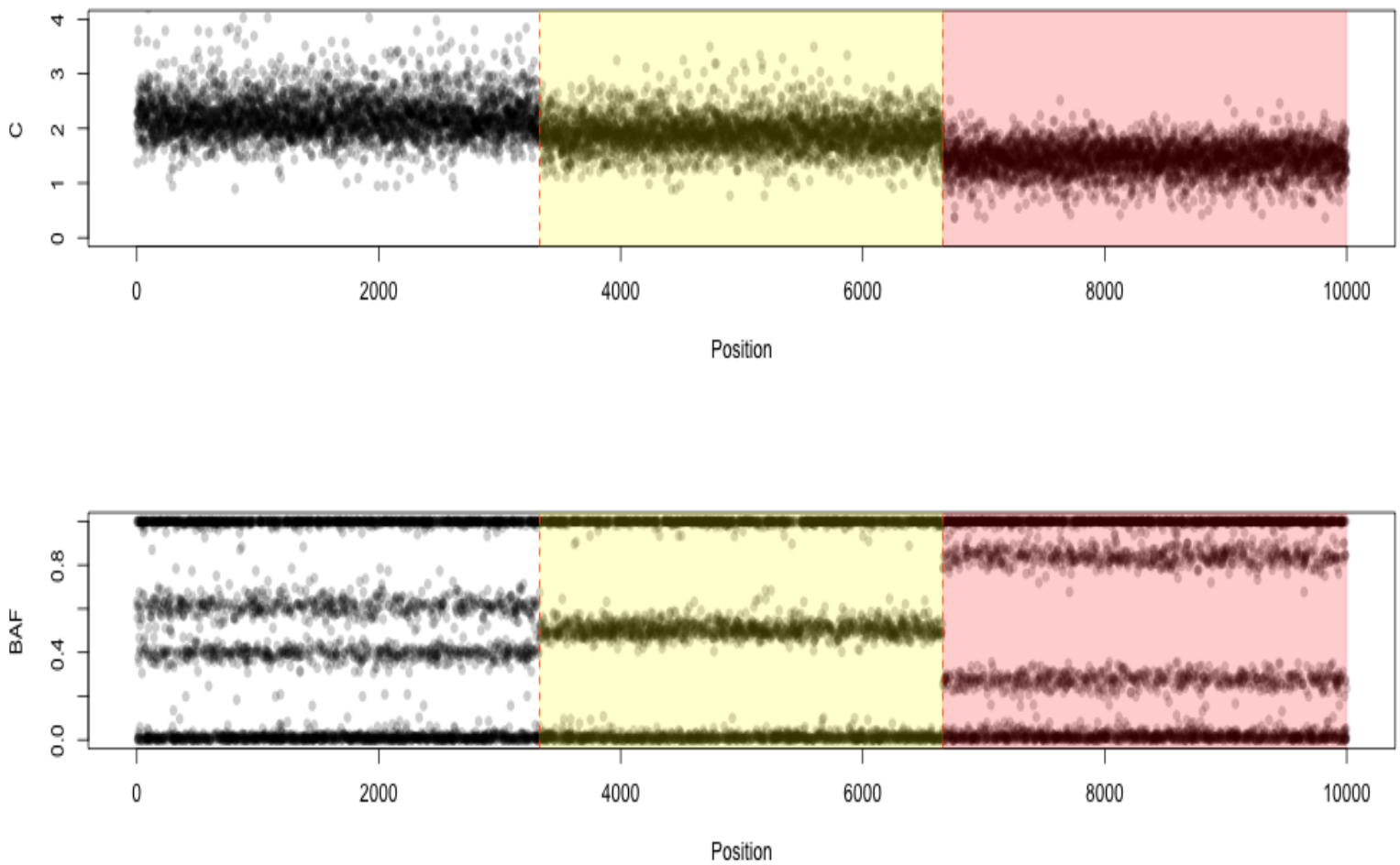
Signal-to-noise can be controlled

data set 100% tumor cells



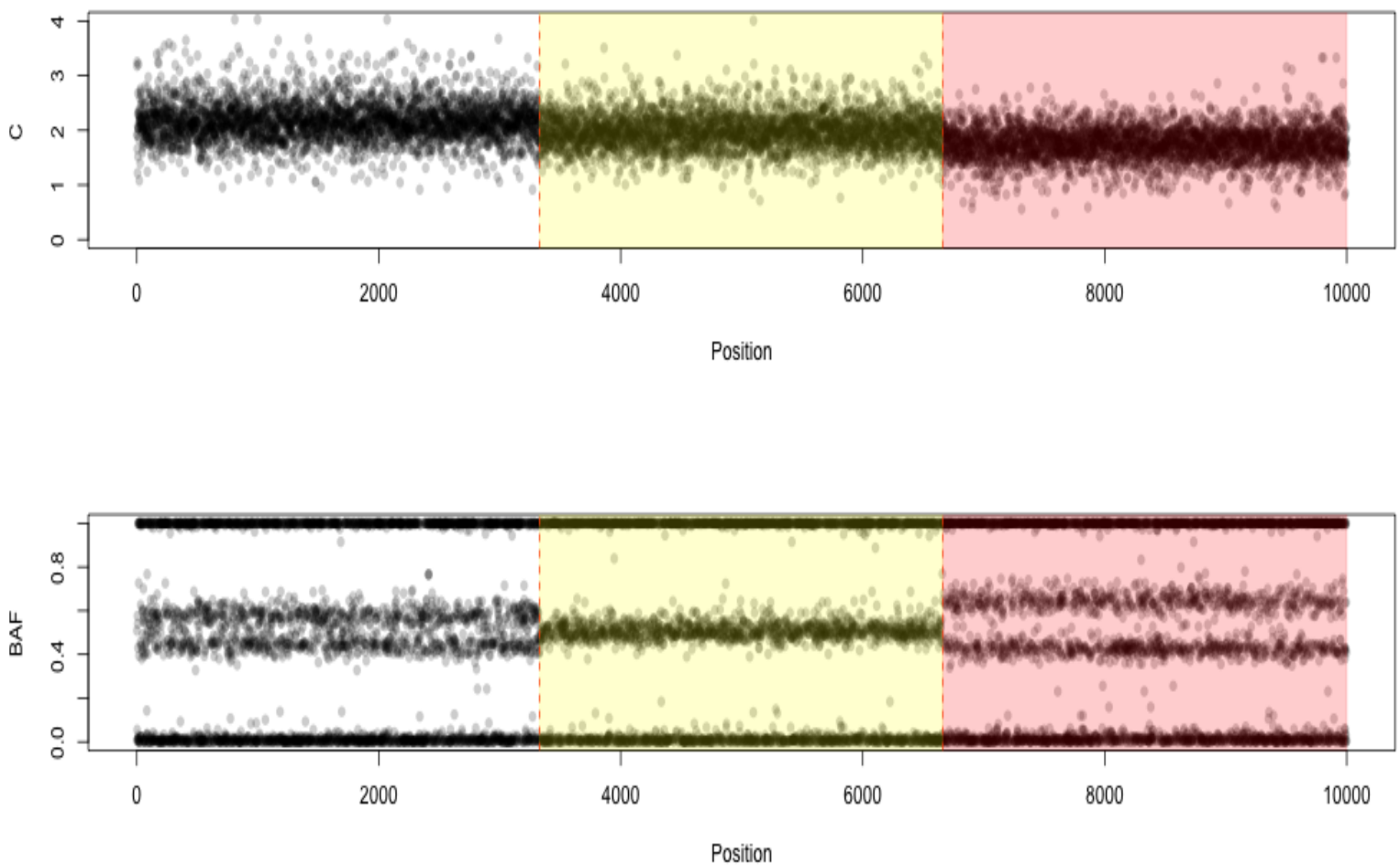
Signal-to-noise can be controlled

data set 79% tumor cells - same truth



Signal-to-noise can be controlled

data set 50% tumor cells - same truth



Summary of the proposed approach

Features

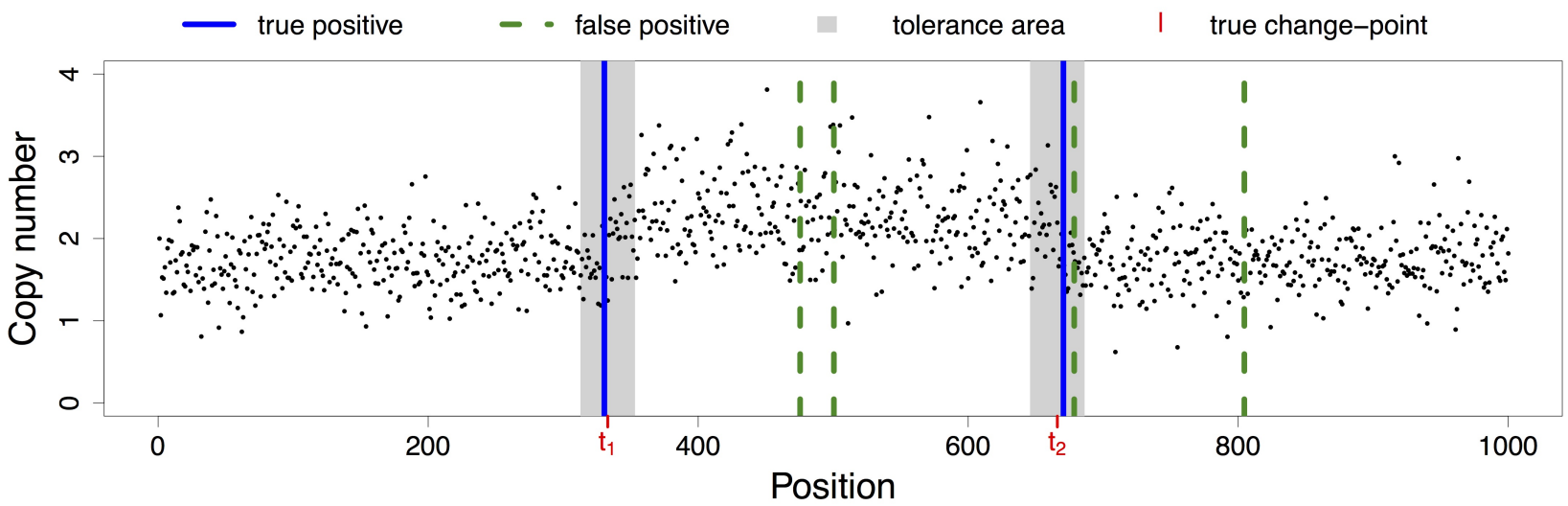
- based on real copy-number data
- SNR driven by biological parameters
- allows for synthetic data generation

A resampling-based data generation framework

- truth (specified by user or automatic)
  - K breakpoint positions
  - K+1 copy-number state labels
- signal (generated from two public SNP array dillution series)
  - GSE11976 (Illumina, HCC1395) : 34, 50, 79 and 100% of tumor cells
  - GSE29172 (Affymetrix, NCI-H1395) : 30, 50, 70 and 100% of tumor cells

Outline

Define TPR and FPR



Performance evaluation : pAUC

Statistic	Method	Data set 1			Data set 2		
		100%	70%	50%	100%	79%	50%
(c, d)	PSCBS (Olshen et al., 2011)	0.89	0.60	0.16	0.97	0.88	0.51
	GFLars (Vert and Bleakley, 2010)	0.60	0.42	0.14	0.97	0.91	0.60
	RBS (Lebarbier, 2005)	0.93	0.63	0.22	0.97	0.95	0.75
(c)	CBS(Olshen et al., 2004)	0.92	0.59	0.16	0.91	0.84	0.45
	GFLars (Vert and Bleakley, 2010)	0.94	0.64	0.18	0.96	0.89	0.49
	RBS (Lebarbier, 2005)	0.91	0.62	0.17	0.90	0.84	0.48
	cghseg (Rigaill, 2010)	0.93	0.61	0.18	0.95	0.88	0.49
(d)	CBS (Olshen et al., 2004)	0.35	0.17	0.10	0.71	0.83	0.64
	GFLars (Vert and Bleakley, 2010)	0.35	0.18	0.10	0.71	0.84	0.66
	RBS (Lebarbier, 2005)	0.34	0.17	0.09	0.69	0.83	0.65
	cghseg (Rigaill, 2010)	0.35	0.18	0.10	0.70	0.84	0.67

Outline

Conclusion

Conclusions

- Perform joint segmentation is better
- `acnr` and `jointseg` are available on cran
- Work published (Pierre-Jean et al., 2015)
- How to detect change-point in the whole distribution ?
  A. Célisse's talk

Future works at CNRGH

- Florence Mauger, PhD and Nouara Oussada, M2 Genhiome
- circulating DNA (plasma)
- Project on WGS data
- Context : tumoral cells detection
- Performance evaluation of 3 pipelines :
  ichorCNA, QDNAseq and CNAnorm

Thank you for your attention

Thanks to Pierre Neuvial, Guillem Rigaill

Gey, S. and Lebarbier, E. (2008). Using CART to detect multiple change points in the mean for large sample. Technical report, Statistics for Systems Biology research group.

Lebarbier, E. (2005). Detecting multiple change-points in the mean of gaussian process by model selection. *Signal processing*, 85(4) :717–736.

Olshen, A. B., Bengtsson, H., Neuvial, P., Spellman, P. T., Olshen, R. A., and Seshan, V. E. (2011). Parent-specific copy number in paired tumor-normal studies using circular binary segmentation. *Bioinformatics*, 27(15) :2038–2046.

Olshen, A. B., Venkatraman, E. S., Lucito, R., and Wigler, M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, 5(4) :557–572.

Pierre-Jean, M., Rigaill, G., and Neuvial, P. (2015). Performance evaluation of DNA copy number segmentation methods. *Briefings in Bioinformatics*, 16(4) :600–615.

Rigaill, G. (2010). Pruned dynamic programming for optimal multiple change-point detection. Technical report, http ://arXiv.org/abs/1004.0887.

Vert, J.-P. and Bleakley, K. (2010). Fast detection of multiple change-points shared by many signals using group LARS. *Advances in Neural Information Processing Systems*, 23 :2343–2351.