Statistical analysis of DNA copy number data in cancers

Pierre Neuvial

Institut de Mathématiques de Toulouse, Equipe Statistique et Probabilités

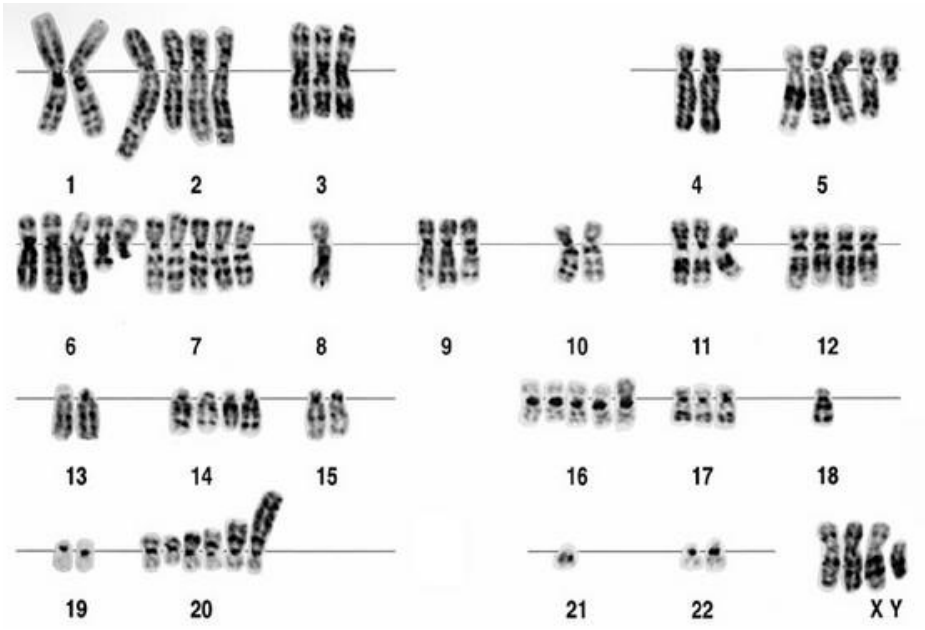
http://www.math-evry.cnrs.fr/members/pneuvial/

Séminaire MIAT, 2/9/2016

Genomic changes at the DNA level are hallmarks of cancer

We inherited 23 paternal and 23 maternal chromosomes, mostly identical

Normal karyotype

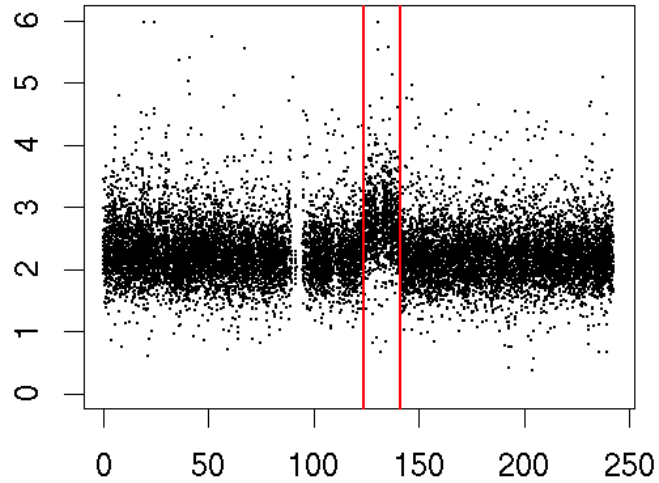


Tumor karyotype

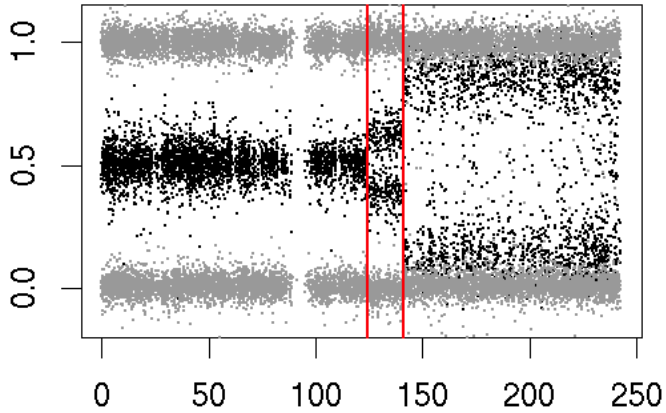
Goal : identify CN changes to improve characterization, classification, and treatment of cancers

SNP array data in cancer samples

Total copy number (c)



Allelic ratio (b)



DNA copy number studies in cancer research

Data types and what information can be retrieved from them

- microarray (CGH arrays, SNP arrays) :
  - gains, losses, copy-neutral LOH
- sequencing (exome, whole genome) :
  - idem + translocations, mutations

Statistical questions tackled here

- identifying breakpoints from DNA copy number data
- performance evaluation in DNA copy number studies
- quantifying tumor heterogeneity

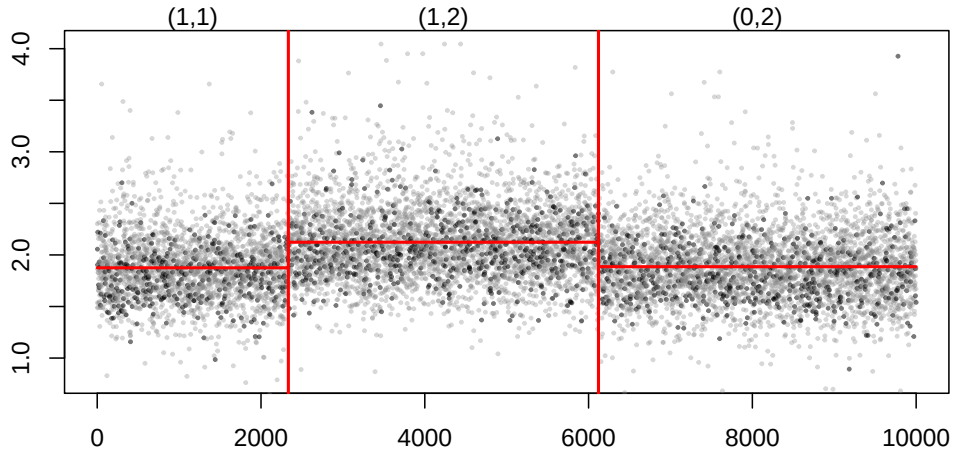
Outline

1. Joint segmentation methods
   - Model and methods
   - Recursive binary segmentation
2. Performance evaluation of copy-number segmentation methods
   - Generating data with known truth
   - Comparing methods for segmenting SNP array data
3. Dissecting tumor heterogeneity from copy number profiles
   - Model and parameter estimation
   - Performance evaluation on synthetic data
   - Very preliminary results on real data

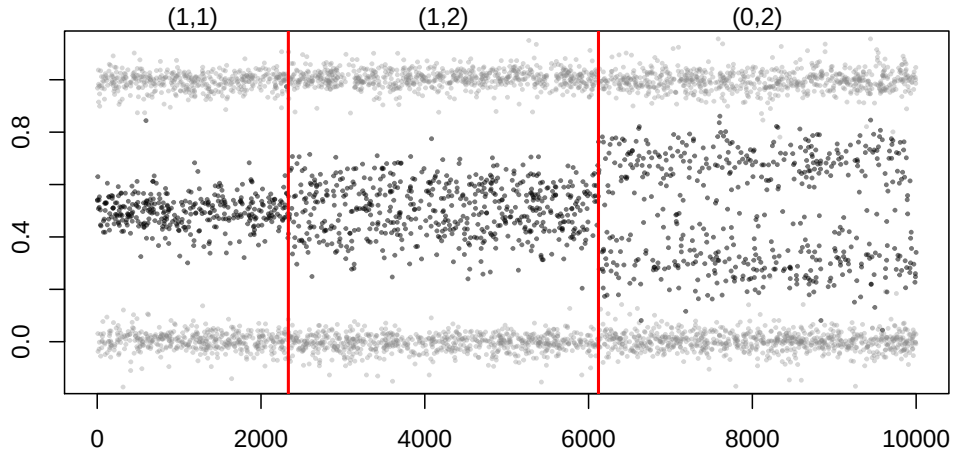
Outline

SNP array data in cancer samples

Total copy number (c)

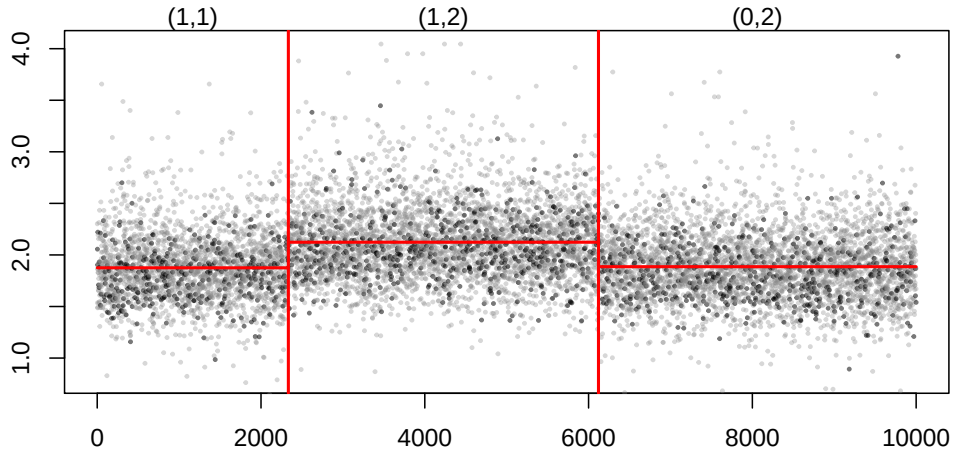


Allelic ratio (b)

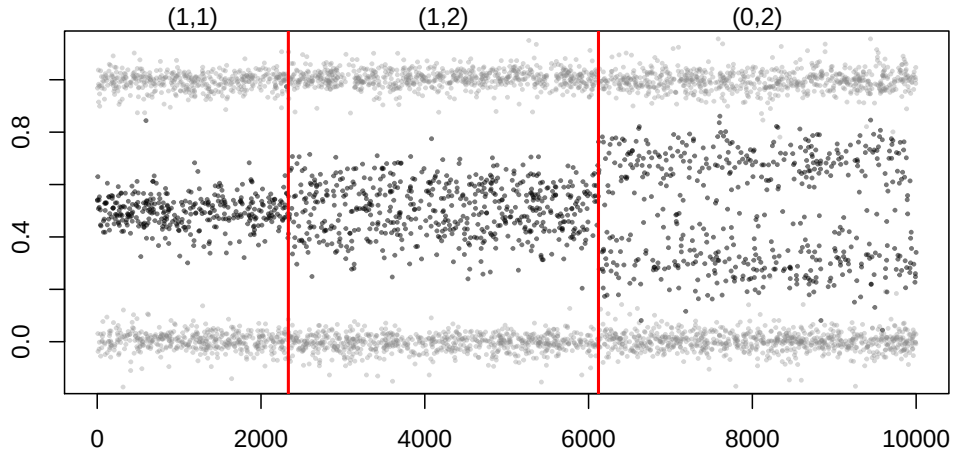


SNP array data in cancer samples

Total copy number (c)



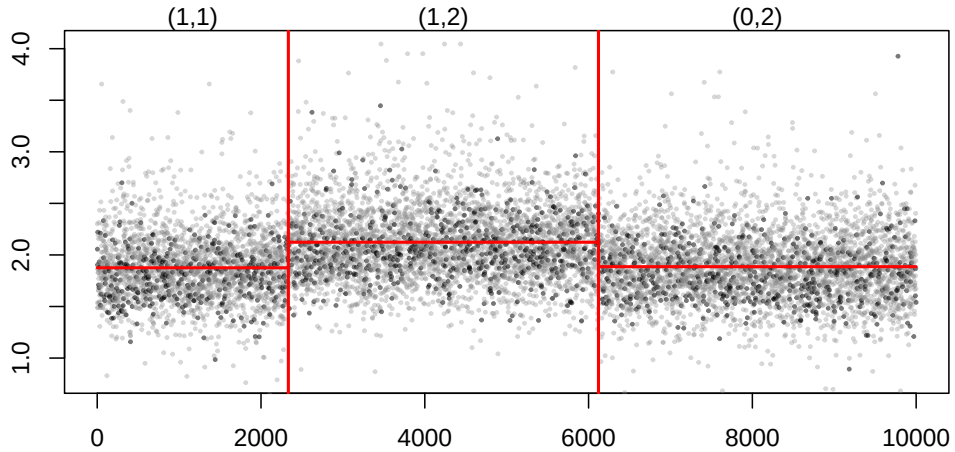
Allelic ratio (b)



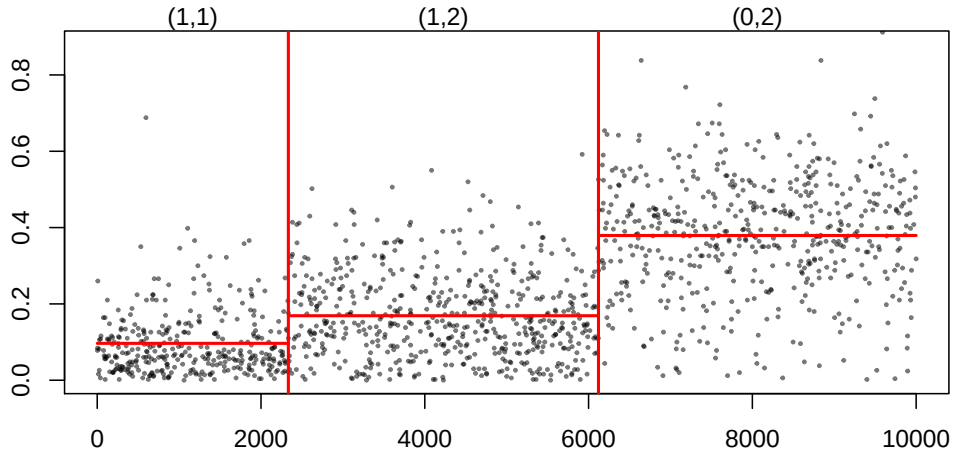
Breakpoints occur at the same position in both dimensions

SNP array data in cancer samples

Total copy number (c)



Decrease of heterozygosity (d)



$d = 2|b - 1/2|$ (only defined for SNPs heterozygous in the germline)

Model

A change-point model

e.g. [Picard et al. (2005)]

- Biological assumption : DNA copy numbers are piecewise constant
- Statistical model for K change points at $(t_1, \ldots t_K)$:

$$\forall j = 1, \ldots, n \qquad c_j = \gamma_j + \epsilon_j$$

where $\forall k \in \{1, \ldots, K+1\}, \forall j \in [t_{k-1}, t_k[\qquad \gamma_j = \Gamma_k$

Challenges : K and $(t_1, \ldots t_K)$ are unknown

- Choosing K : a model selection problem
- For a fixed K, number of possible partitions $= C_{n-1}^K = \mathcal{O}(n^{K-1})$

Orders of magnitude for SNP arrays : $n \sim 10^4$ to 10^6 and $K \sim 10$ to 100

Model

A change-point model

e.g. [Picard et al. (2005)]

- Biological assumption : DNA copy numbers are piecewise constant
- Statistical model for K change points at $(t_1, \ldots t_K)$:

$$\forall j = 1, \ldots, n \qquad c_j = \gamma_j + \epsilon_j$$

where $\forall k \in \{1, \ldots, K+1\}, \forall j \in [t_{k-1}, t_k[\qquad \gamma_j = \Gamma_k$

Challenges : K and $(t_1, \ldots t_K)$ are unknown

- Choosing K : a model selection problem
- For a fixed K, number of possible partitions $= C_{n-1}^K = \mathcal{O}(n^{K-1})$

Orders of magnitude for SNP arrays : $n \sim 10^4$ to 10^6 and $K \sim 10$ to 100

Need for algorithms of **linear time and space complexity** !

Some (joint) copy number segmentation methods

Method	Time	# dims
Dynamic programming (DP)		
[Rigaill et al.(2010)]	$n \log(n)$	1
[Picard et al. (2005)]	$d \cdot K \cdot n^2$	any
Fused Lasso		
[Harchaoui and Lévy-Leduc(2008)]	$K \cdot n$	1
[Bleakley and Vert (2011)]	$d \cdot K \cdot n$	any
Recursive binary segmentation (RBS/CART)		
[Gey and Lebarbier (2008)]	$dn \log(K)$	any
Circular binary segmentation (CBS)		
[Olshen AB et al. (2004)]	$n \log(n)$	1
[Olshen AB et al. (2011)]	$n \log(n)$	2
[Zhang et al.(2010)]	$d \cdot n^2$	any
Hidden Markov Models (HMM)		
[Lai et al.]	n^2	1
[Chen et al. (2011)]	n^2	2

A two-step approach for joint segmentation : RBS + DP

Strategy proposed by [Gey and Lebarbier (2008)]

1. Run a **fast** but **approximate** segmentation method
2. Prune the obtained candidate breakpoints using dynamic programming (**slower** but **exact**)

Complexity when first step is **Recursive Binary Segmentation**

1. $O(d \cdot n \cdot \log(K))$
2. $O(d \cdot K^2 \cdot K)$

Overall : $O(d \cdot n \cdot \log(K))$

Binary Segmentation

When $d = 1$

- Test $\mathcal{H}_0$: "No breakpoint" vs $\mathcal{H}_1$: "Exactly one breakpoint"
- The likelihood ratio statistic is given by $\max_{1 \leq i \leq n} |Z_i|$

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

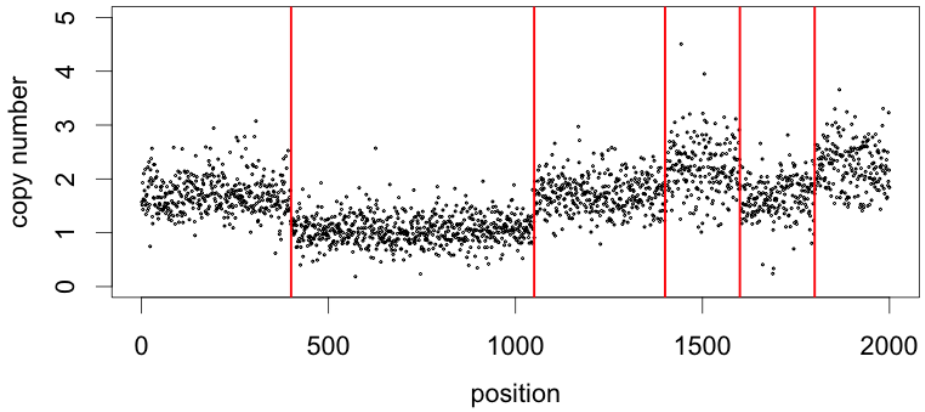
where $S_i = \sum_{1 \leq l \leq i} y_l$.

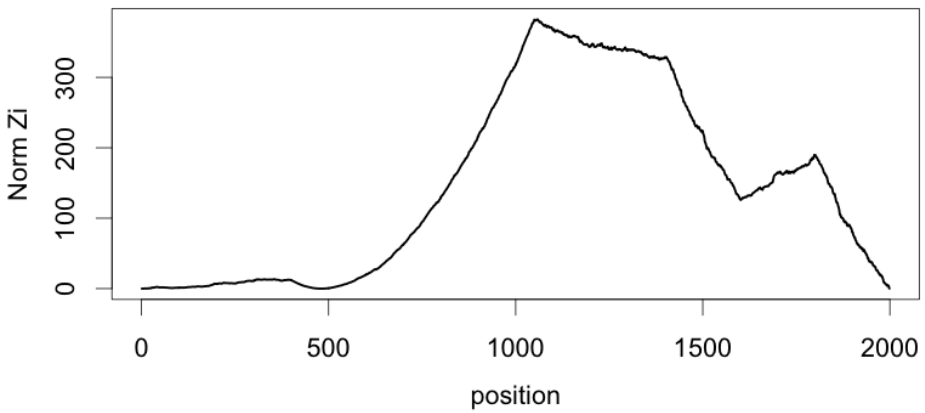
If $d > 1$: the likelihood ratio statistic becomes $\max_{1 \leq i \leq n} \|Z_i\|_2^2$

Recursive Binary Segmentation (RBS)

Complexity : $O(d \cdot n \cdot \log(K))$

- First breakpoint
- For each i : we compute Z_i :
  $b_1 = \arg\max_{1 \leq i \leq n} \|Z_i\|_2^2$

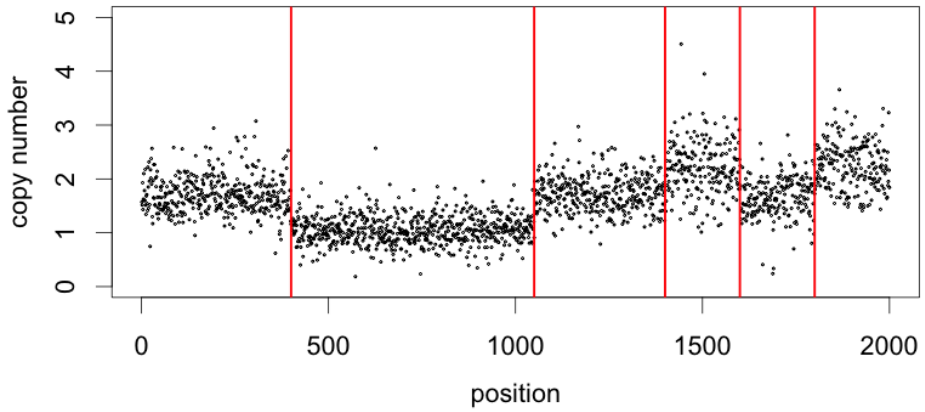


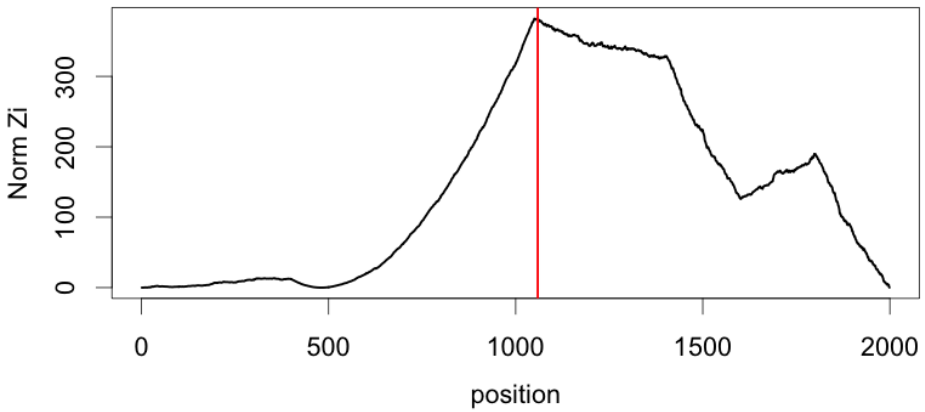


Recursive Binary Segmentation (RBS)

Complexity : $O(d \cdot n \cdot \log(K))$

- First breakpoint
- For each i : we compute Z_i :
  $b_1 = \arg\max_{1 \leq i \leq n} \|Z_i\|_2^2$

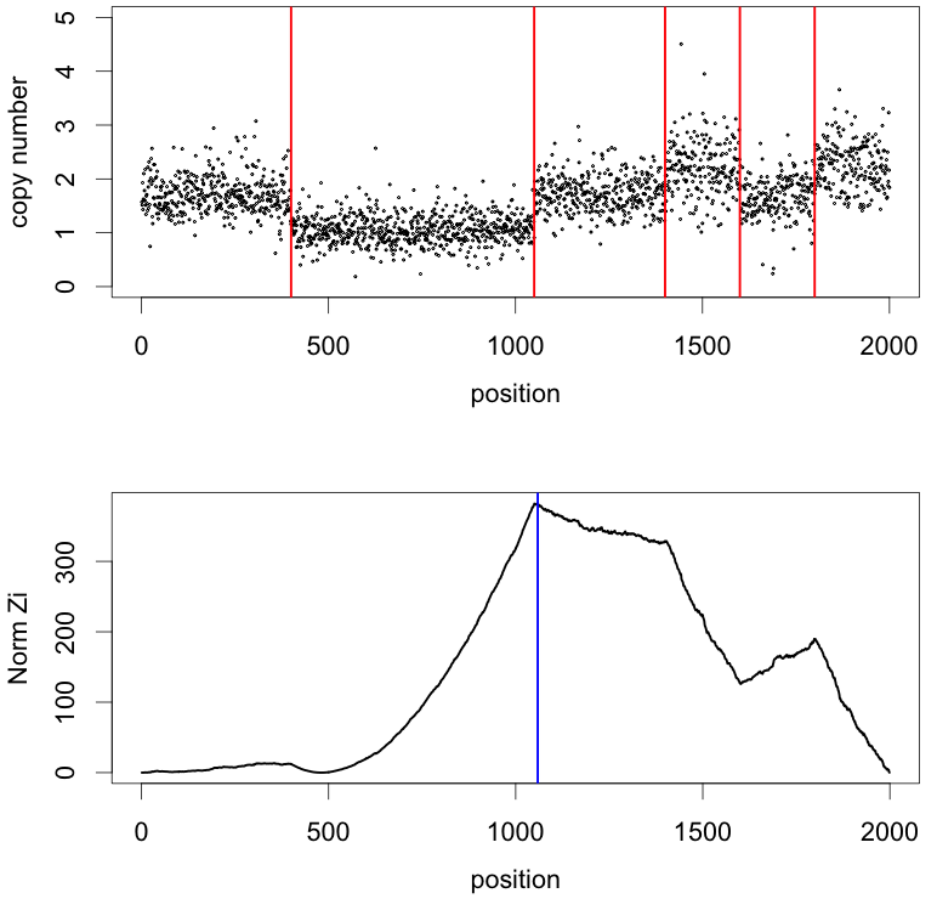




Recursive Binary Segmentation (RBS)

Complexity : $O(d \cdot n \cdot \log(K))$

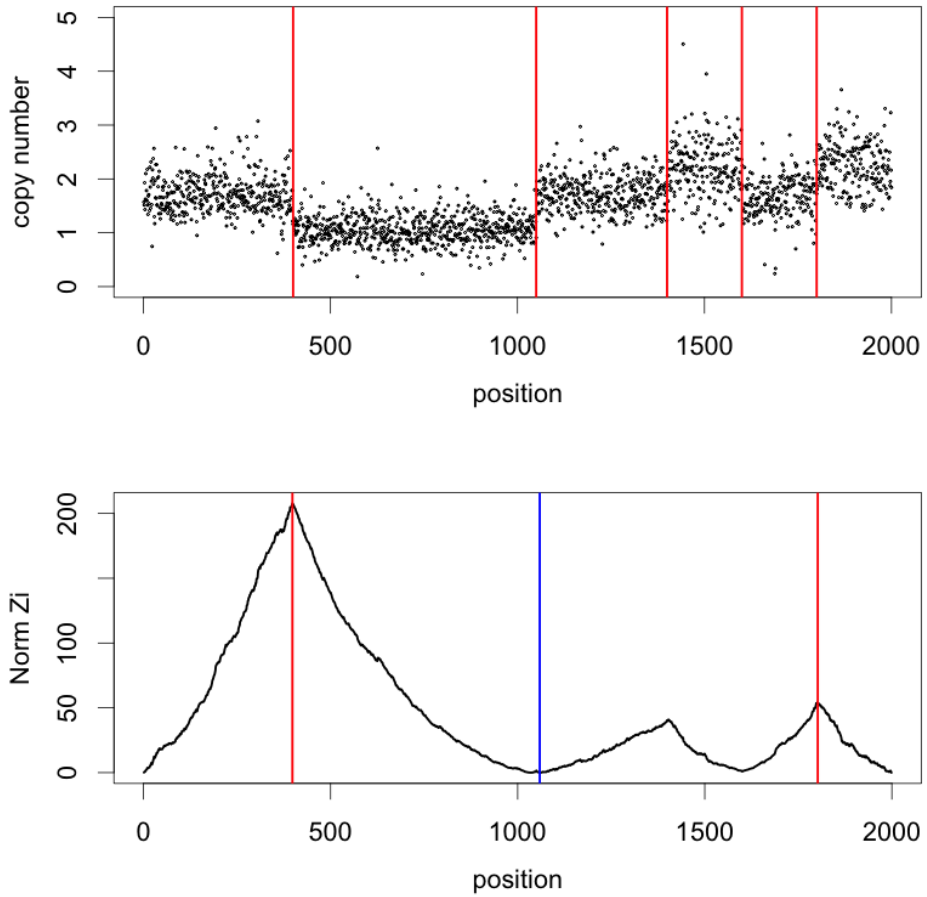
- First breakpoint
- For each i : we compute Z_i :
  $b_1 = \arg\max_{1 \leq i \leq n} \|Z_i\|_2^2$



Recursive Binary Segmentation (RBS)

Complexity : $O(d \cdot n \cdot \log(K))$

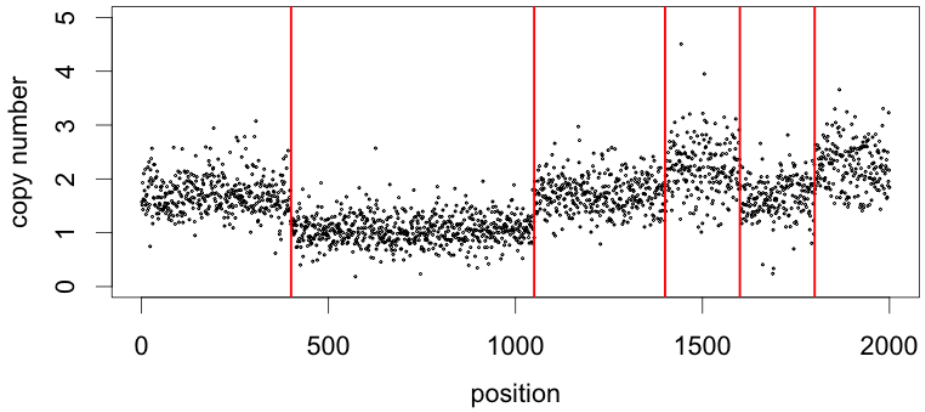
- First breakpoint
- For each i : we compute Z_i :
  $b_1 = \arg\max_{1 \leq i \leq n} \|Z_i\|_2^2$

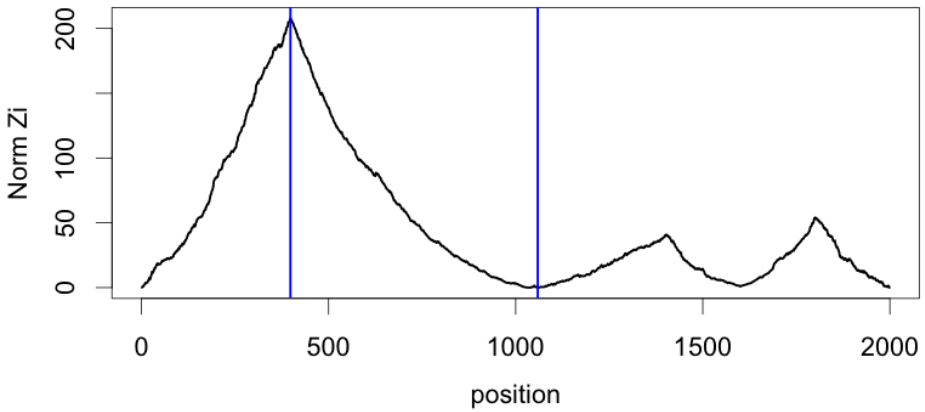


Recursive Binary Segmentation (RBS)

Complexity : $O(d \cdot n \cdot \log(K))$

- First breakpoint
- For each i : we compute Z_i :
  $b_1 = \arg\max_{1 \leq i \leq n} \|Z_i\|_2^2$





Outline

Motivation

Standard approach for developing statistical methods for genomic data :

1. describe a new model/method/learning technique/algorithm
2. show that it performs as expected on simulated data
3. describe a "real data application" with limited ground truth

$\Rightarrow$ Can we design **more convincing performance assessment frameworks** ?

Contribution

A performance assessment framework tailored to a specific application

- Pierre-Jean, Rigaill and Neuvial, *Brief. in Bioinformatics* (2015)
- Implementation : R packages `acnr` and `jointseg` available from github

Back to motivation

Questions of interest

- Are 2d (i. e., joint) methods always better than 1d methods ?
- Is dynamic programming always the best ?

Under Gaussian simulations, the answers are obvious. In practice ?

Contributions

- An evaluation framework allowing to address the above questions
- Identification of biological parameters that drive the methods' performance

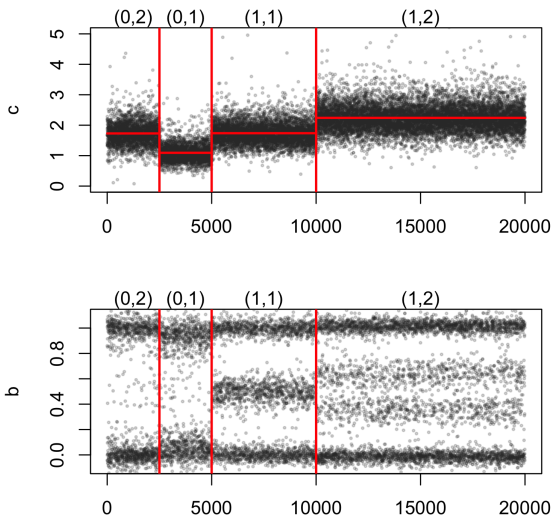
Proposed approach

Limitations of existing approaches

- simulation models : hard to get biological insight
- dilution series [Staaf et al. (2008)] : few regions
- automatically annotated data sets [Willenbrock & Fridlyand (2004)] : depend on a segmentation method
- manually annotated data sets [Hocking et al. (2013)] : SNR cannot be tuned

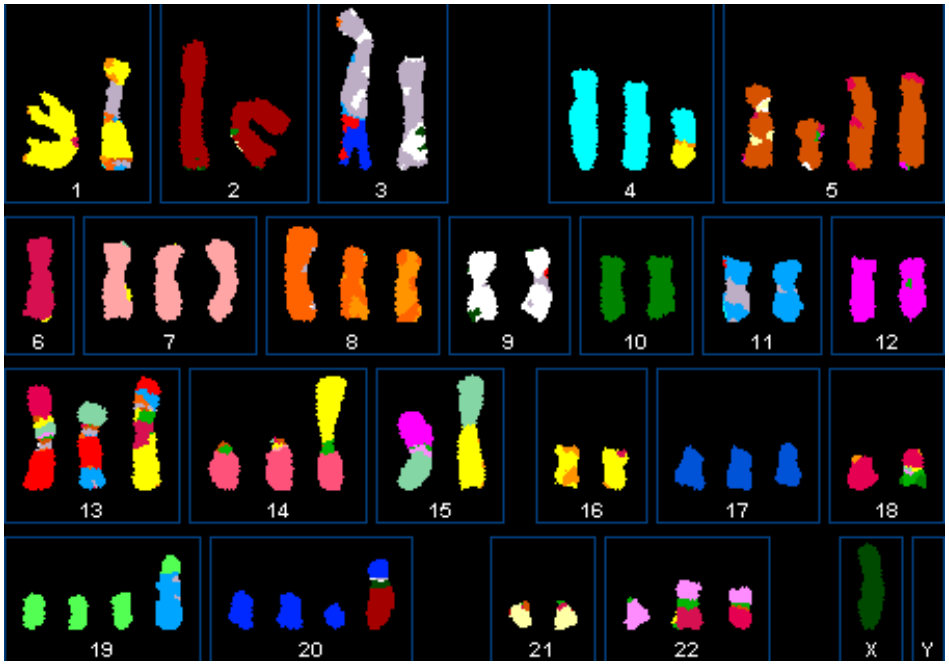
Ingredients for the proposed approach

1. breakpoint positions : $(t_k)_{k=1\cdots K}$
2. copy-number state labels : $(\Gamma_k)_{k=1\cdots K+1}$
3. signal : resampled from real data



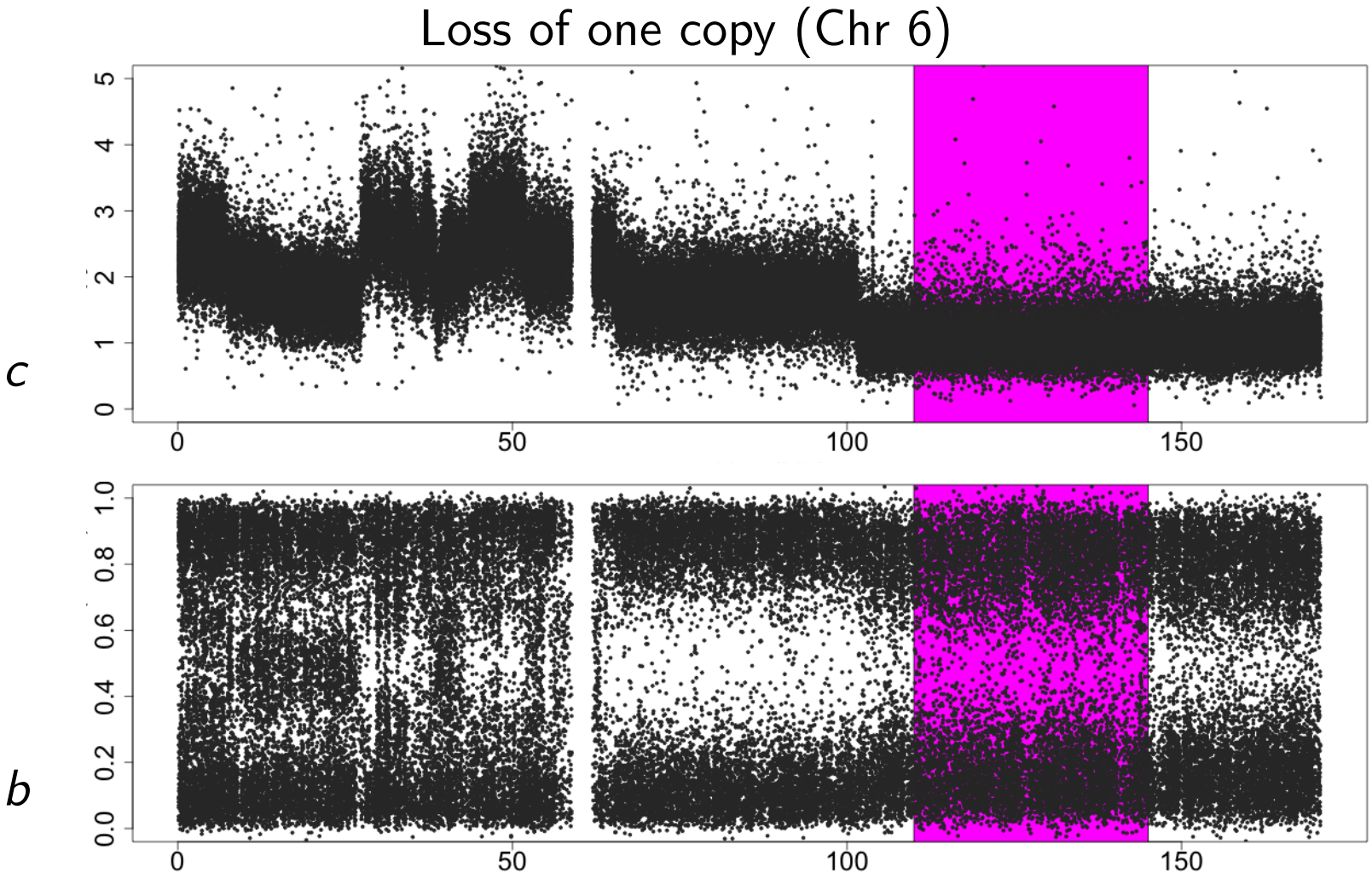
This requires real data with known "truth"

Lung cancer cell line NCI-H1395



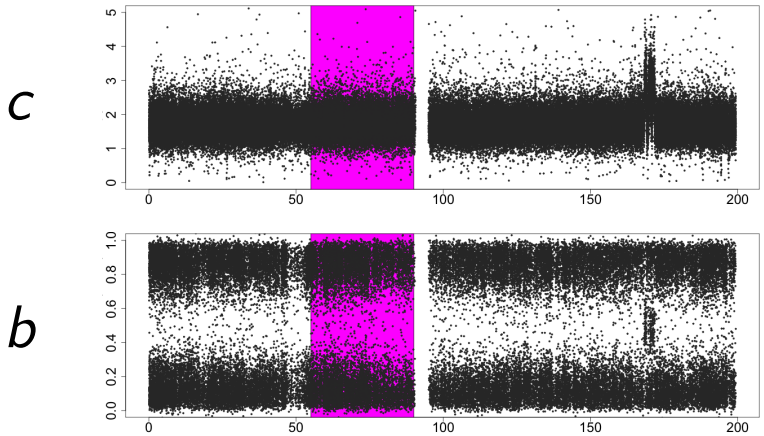
from :
http://www.path.cam.ac.uk/~pawefish/LungCellLineDescriptions/NCI-H1395.html

Real data annotation : NCI-H1395, chr 6

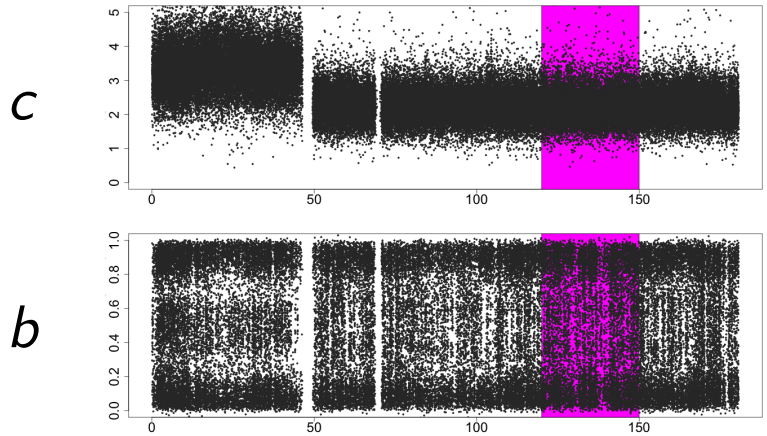


Real data annotation : NCI-H1395

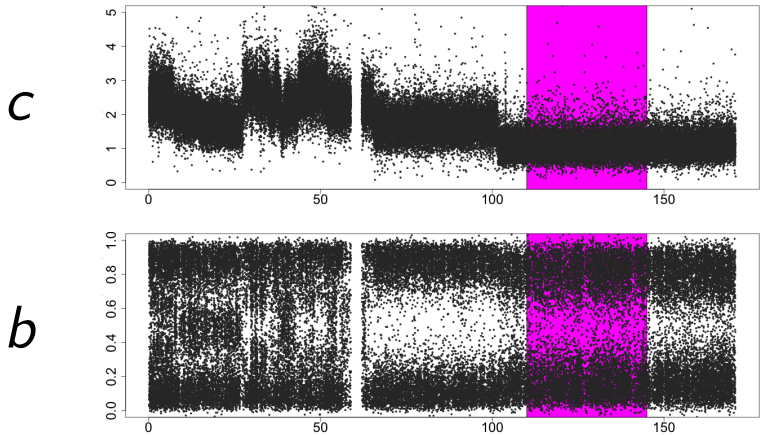
Copy-neutral LOH (Chr 3)



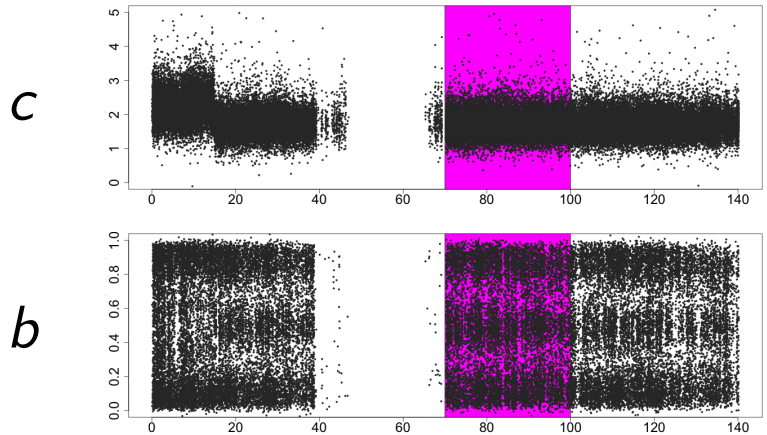
Gain of one copy (Chr 5)



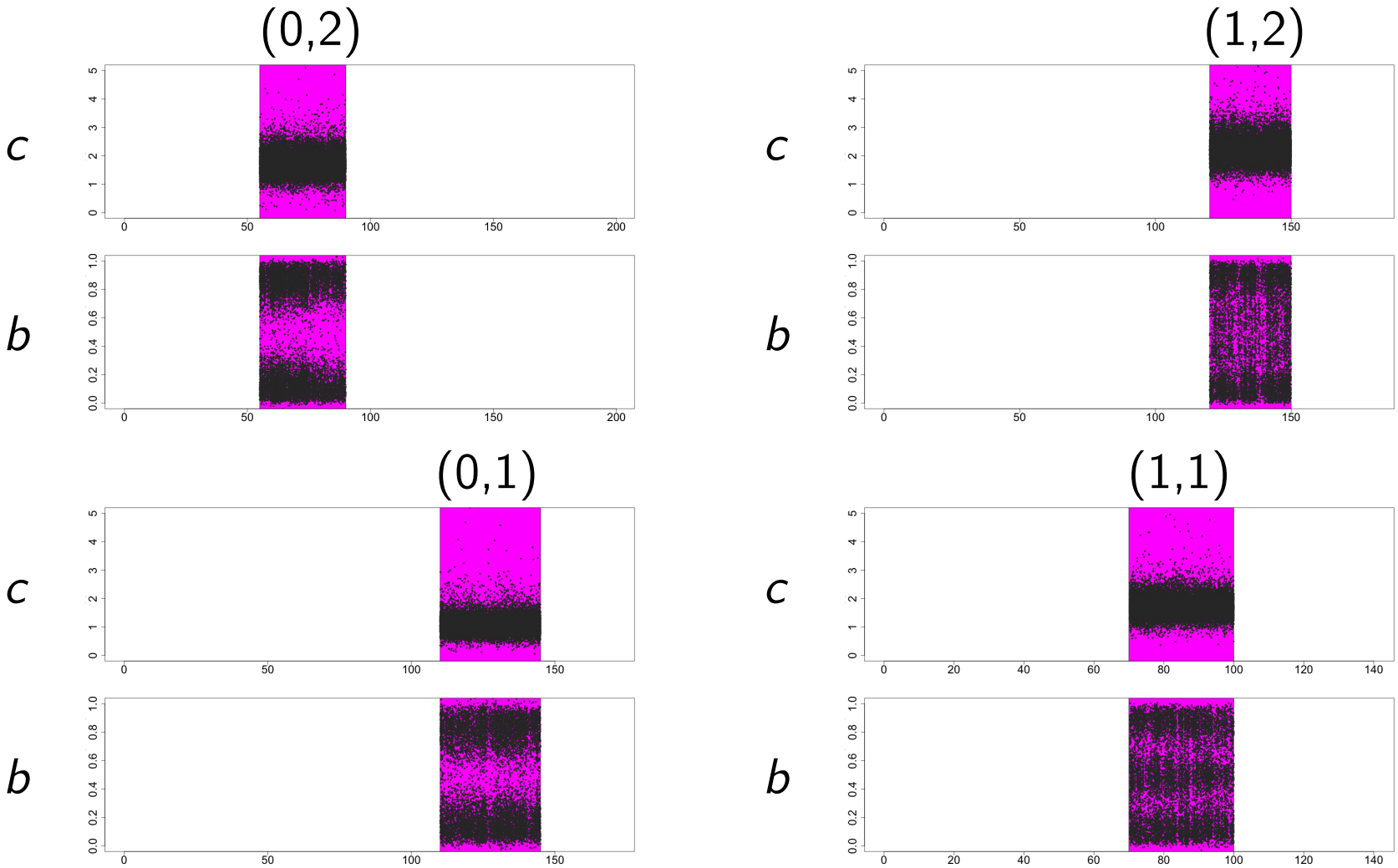
Loss of one copy (Chr 6)



Normal (Chr 9)

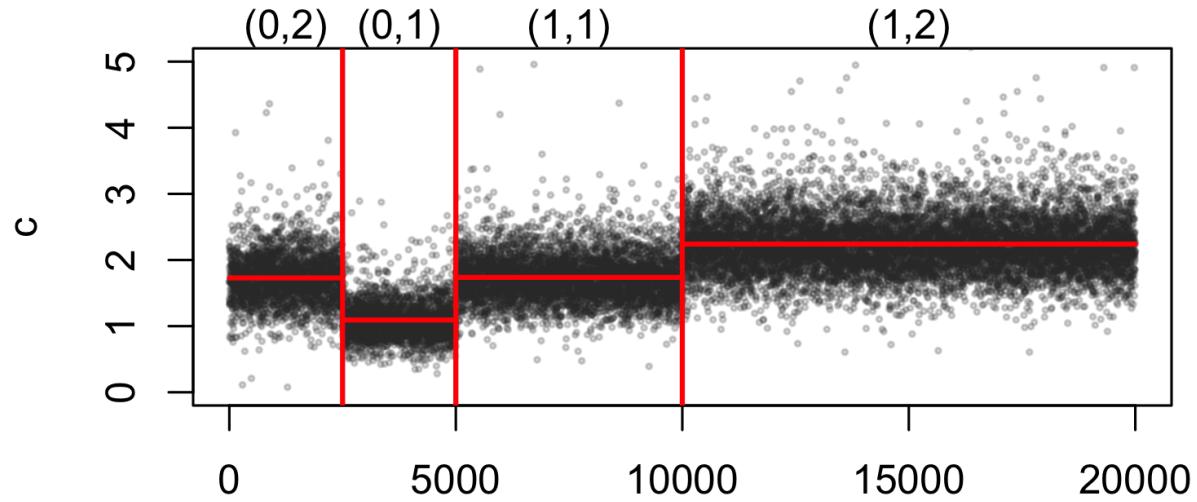


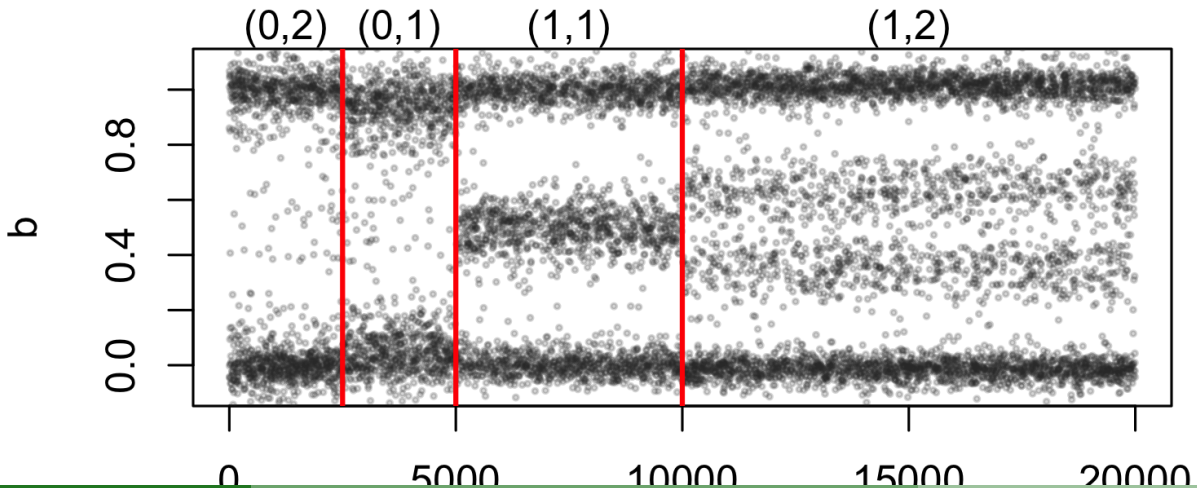
Real data annotation : NCI-H1395



Synthetic data generation

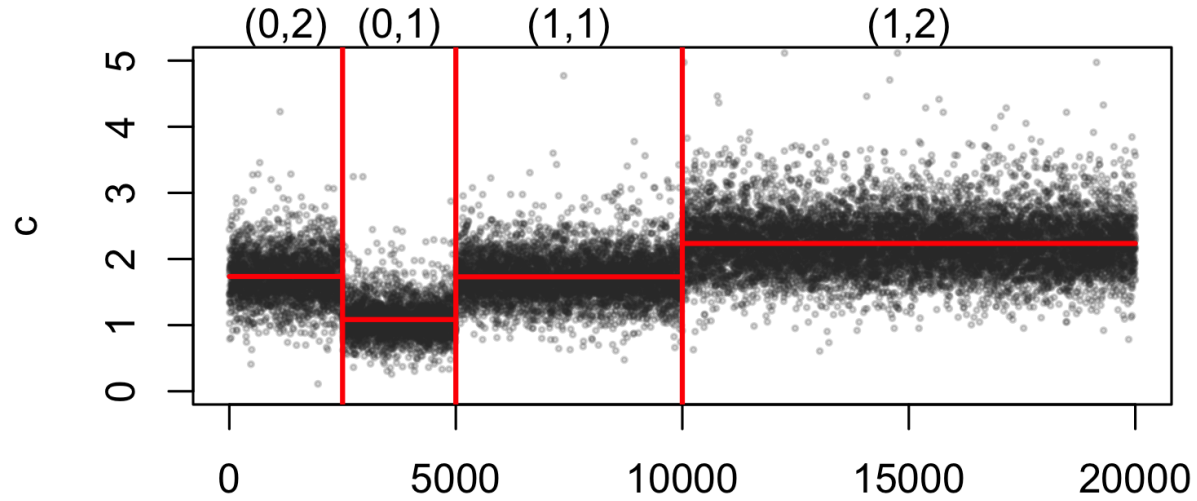
Example : data set 1, 100% tumor cells

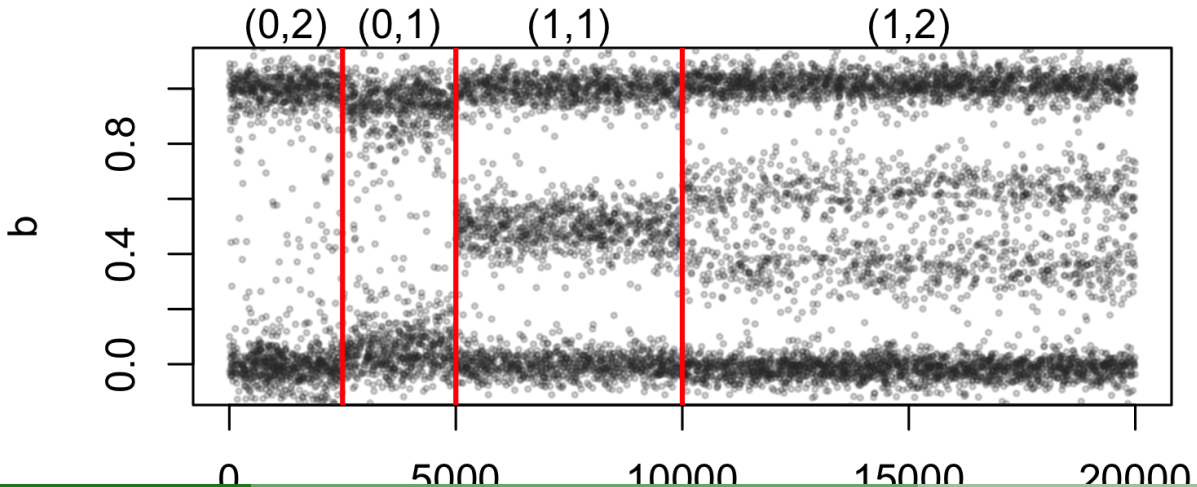




Synthetic data generation

Example : data set 1, 100% tumor cells (same "truth")

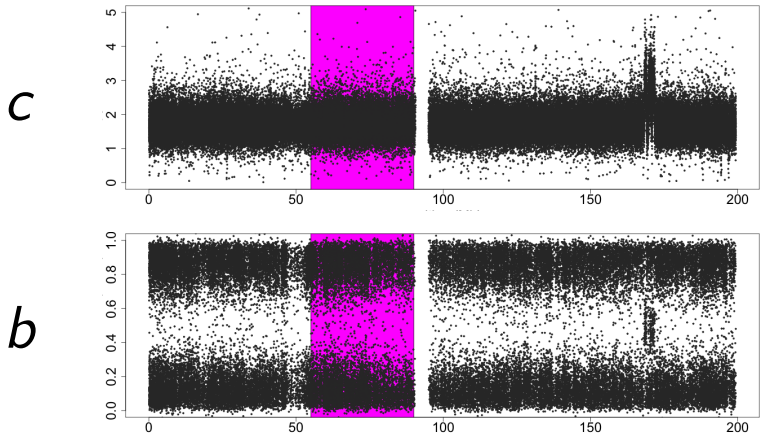




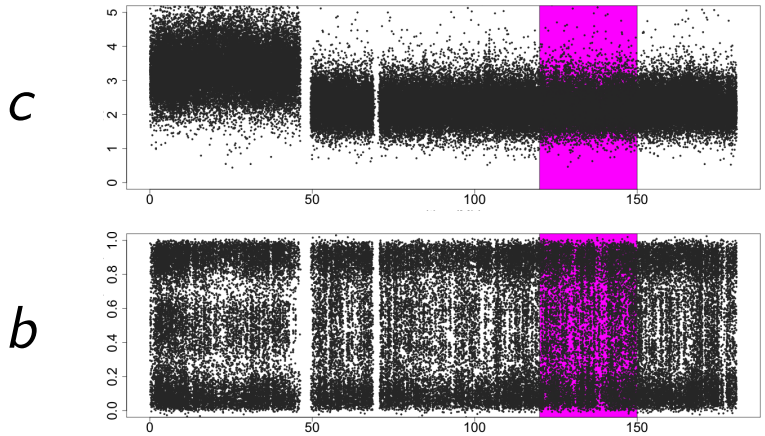
Real data annotation : NCI-H1395

100% tumor cells

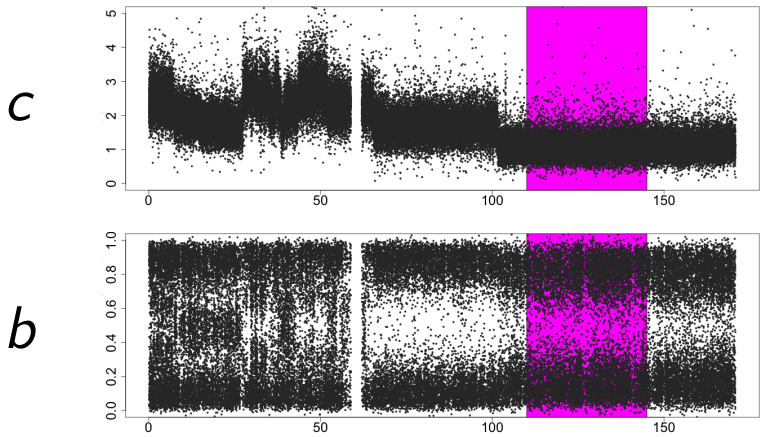
Copy-neutral LOH (Chr 3)



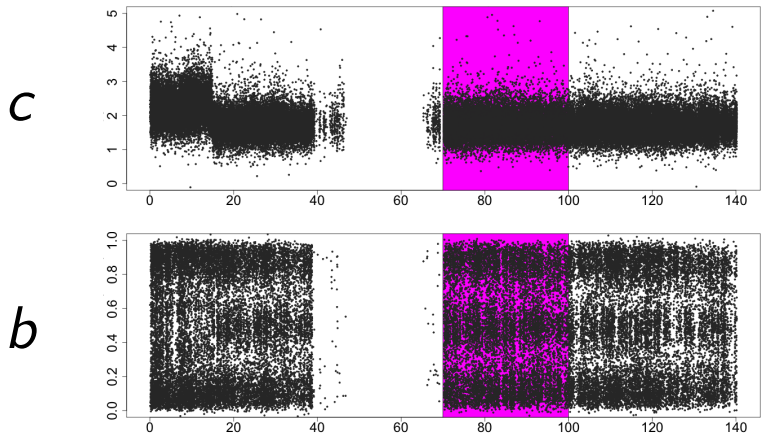
Gain of one copy (Chr 5)



Loss of one copy (Chr 6)



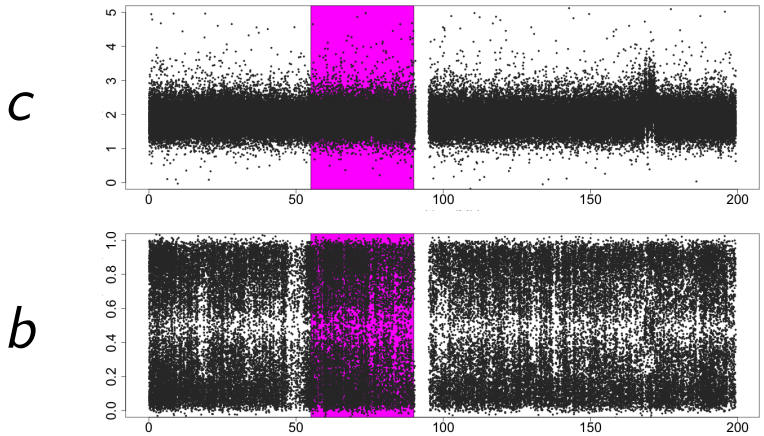
Normal (Chr 9)



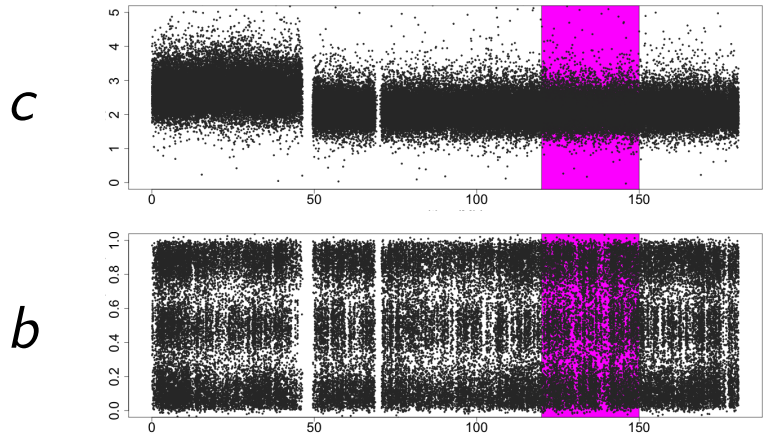
Real data annotation : NCI-H1395

70% tumor cells (using annotation from the 100% data set !)

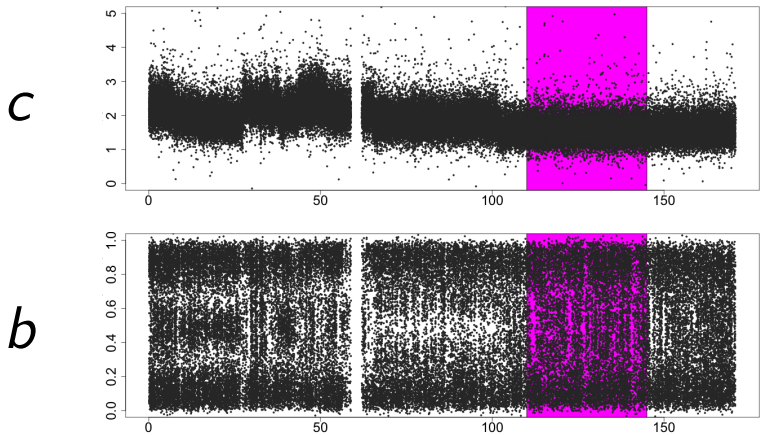
Copy-neutral LOH (Chr 3)



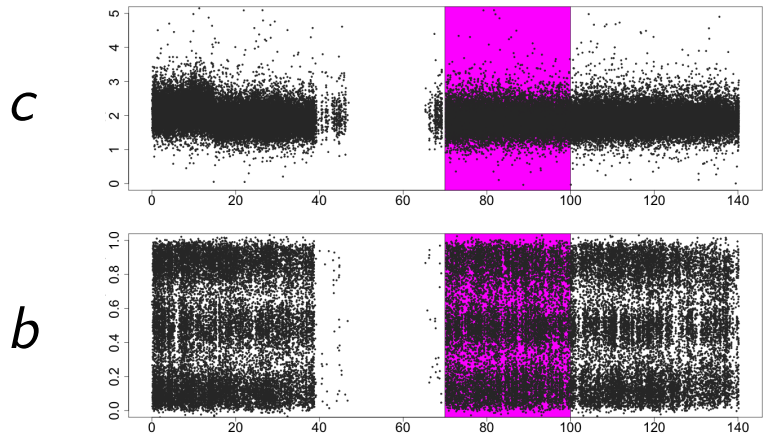
Gain of one copy (Chr 5)



Loss of one copy (Chr 6)



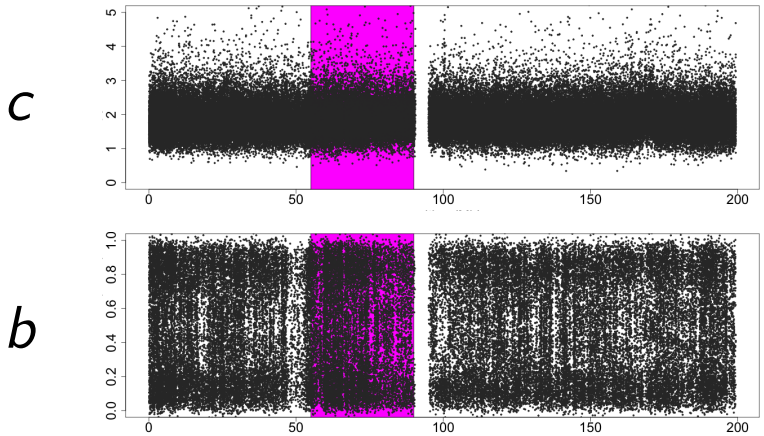
Normal (Chr 9)



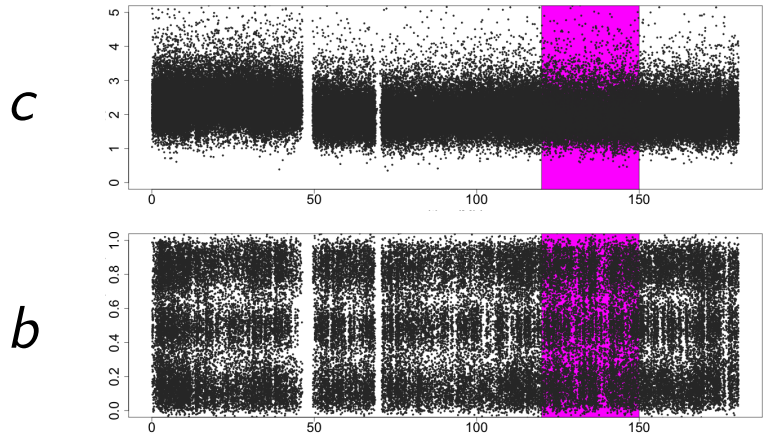
Real data annotation : NCI-H1395

50% tumor cells (using annotation from the 100% data set !)

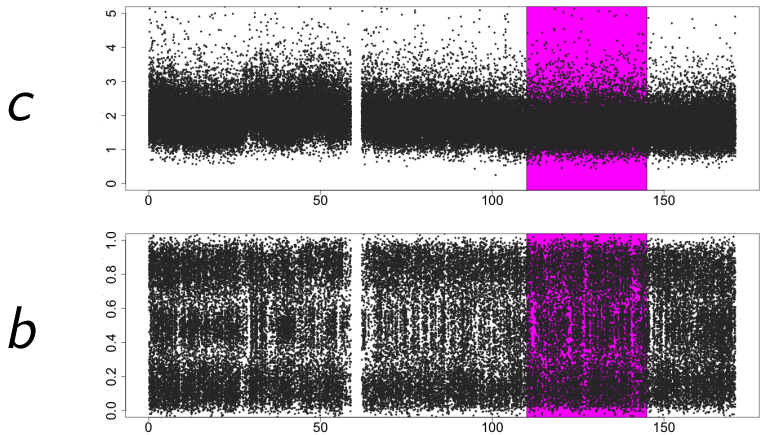
Copy-neutral LOH (Chr 3)



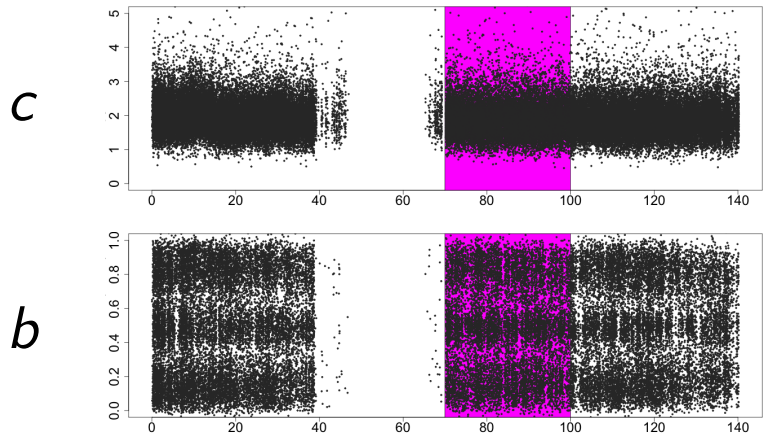
Gain of one copy (Chr 5)



Loss of one copy (Chr 6)



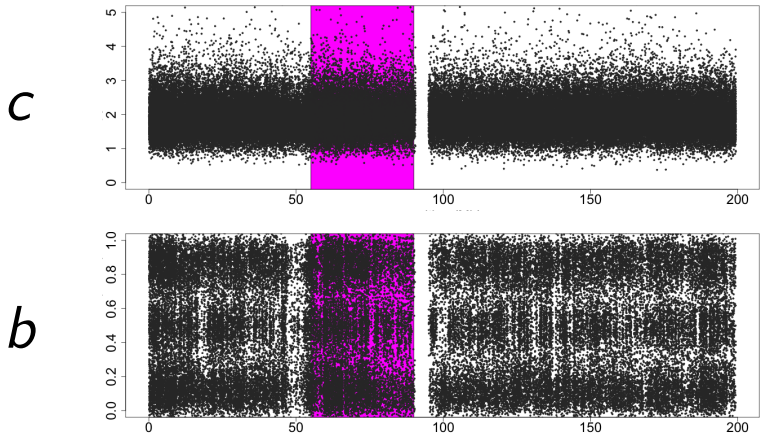
Normal (Chr 9)



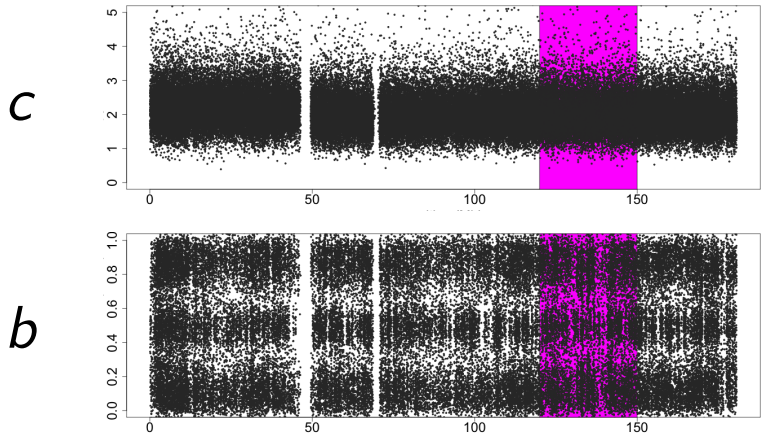
Real data annotation : NCI-H1395

30% tumor cells (using annotation from the 100% data set !)

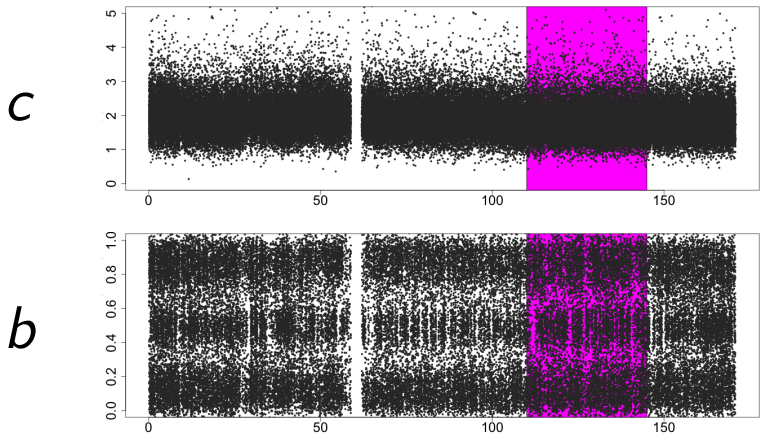
Copy-neutral LOH (Chr 3)



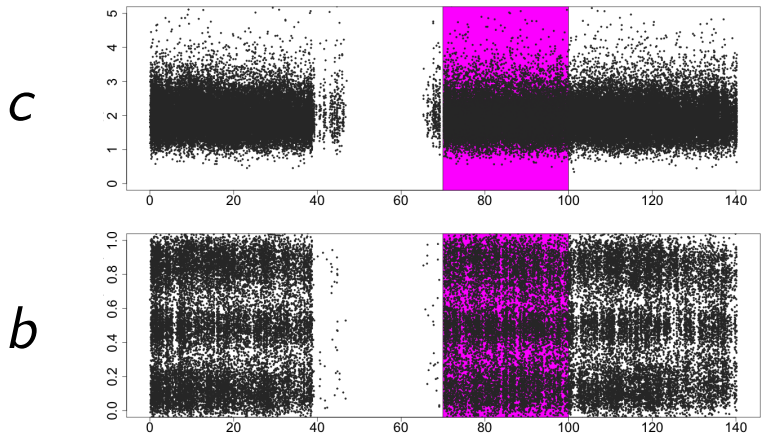
Gain of one copy (Chr 5)



Loss of one copy (Chr 6)

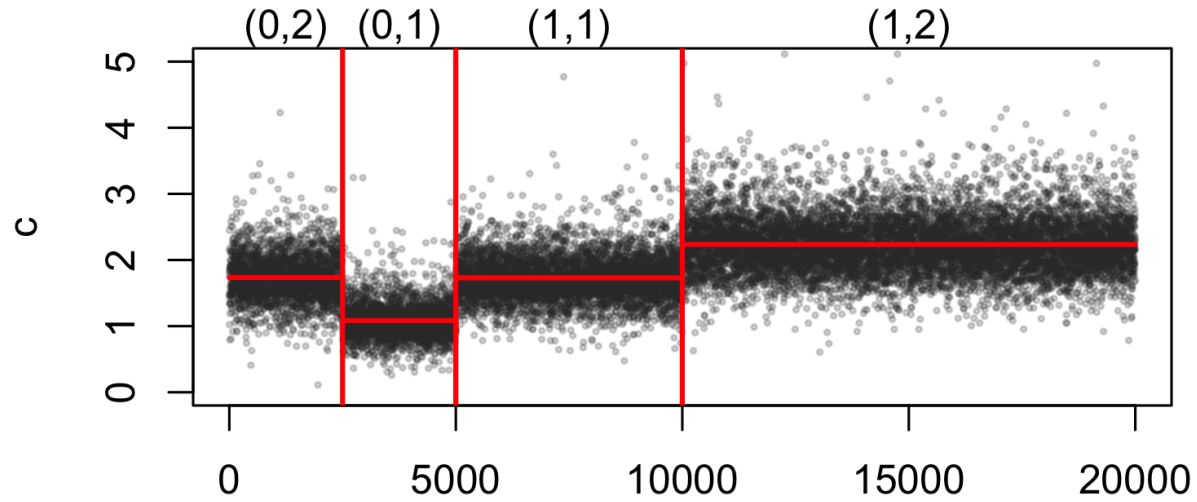


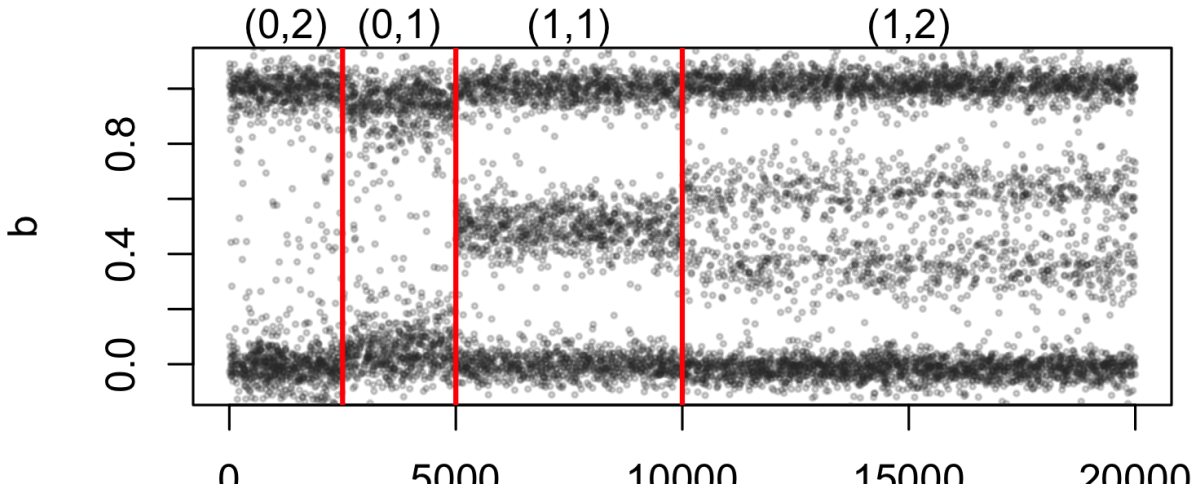
Normal (Chr 9)



Signal-to-noise ratio can be controlled

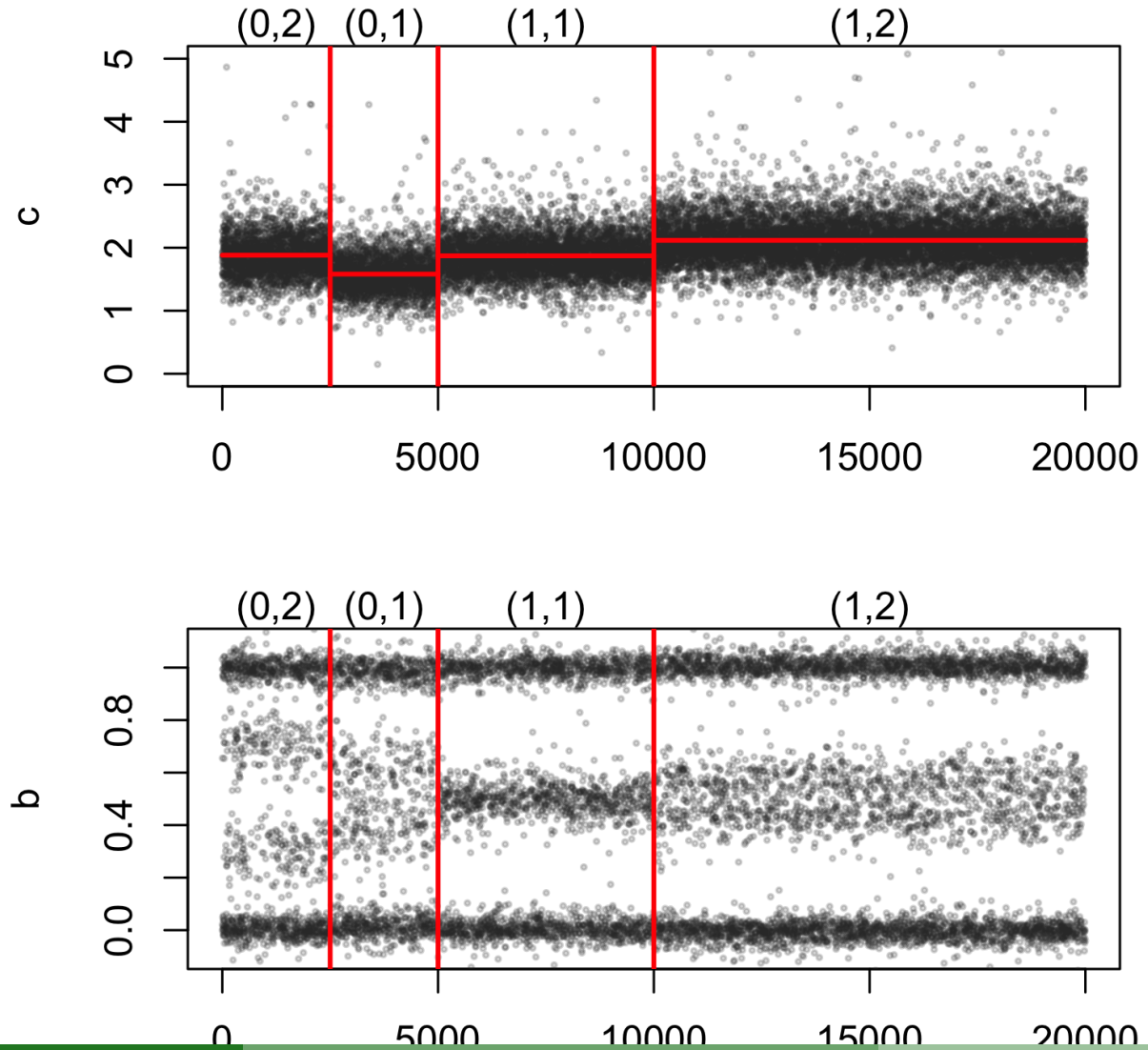
Example : data set 1, 100% tumor cells





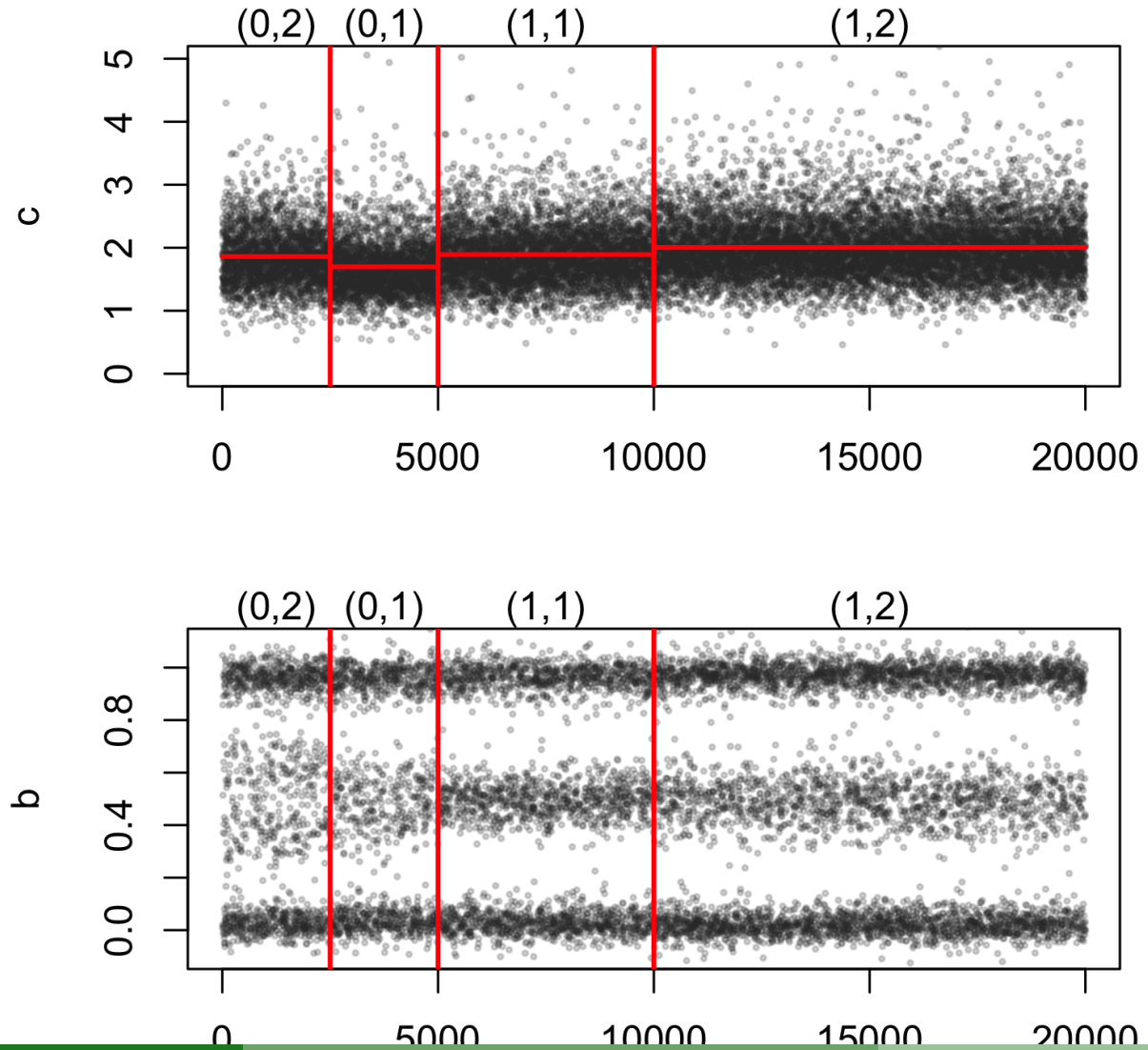
Signal-to-noise ratio can be controlled

Example : data set 1, 70% tumor cells (same "truth")



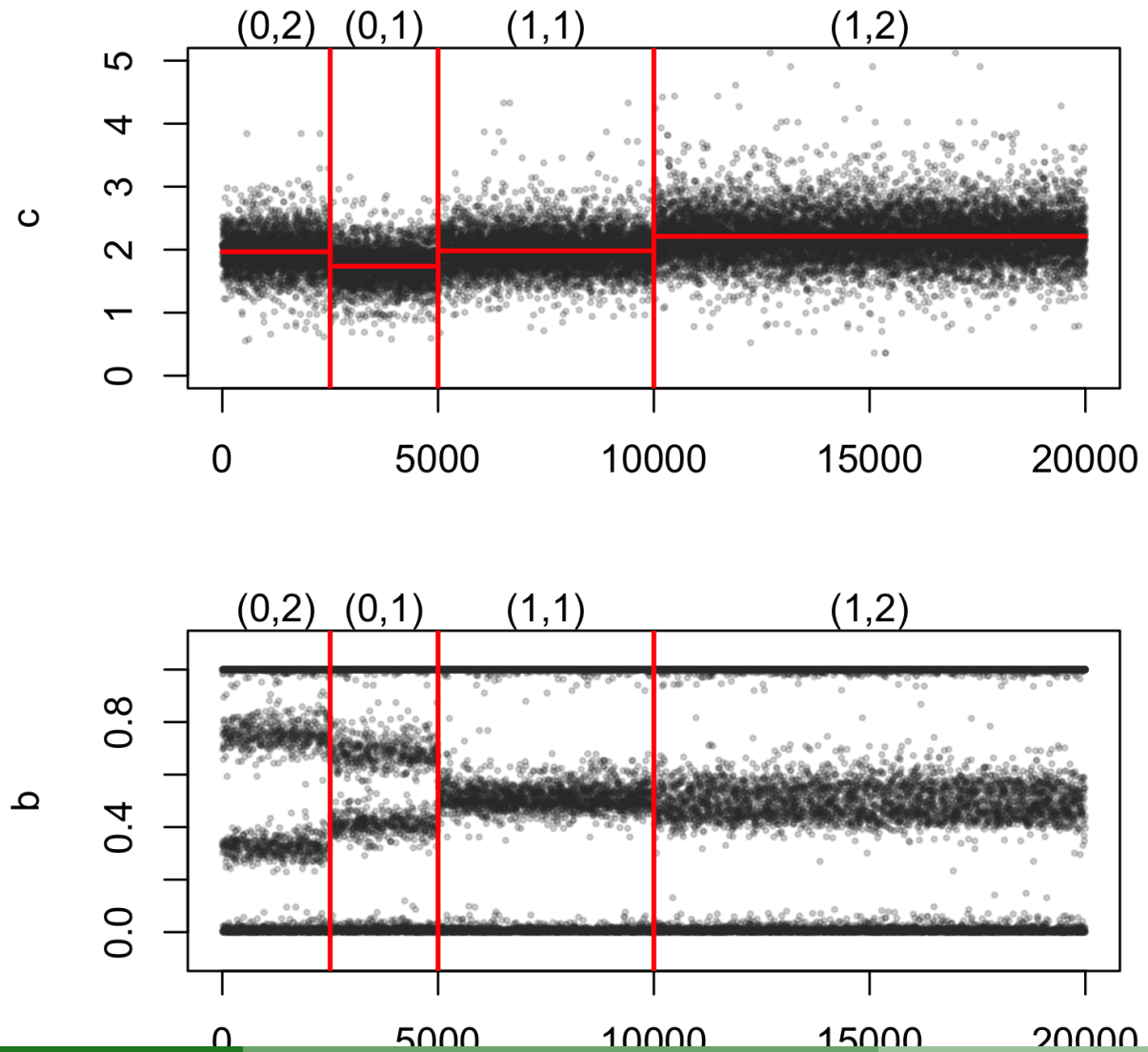
Signal-to-noise ratio can be controlled

Example : data set 1, 50% tumor cells (same "truth")



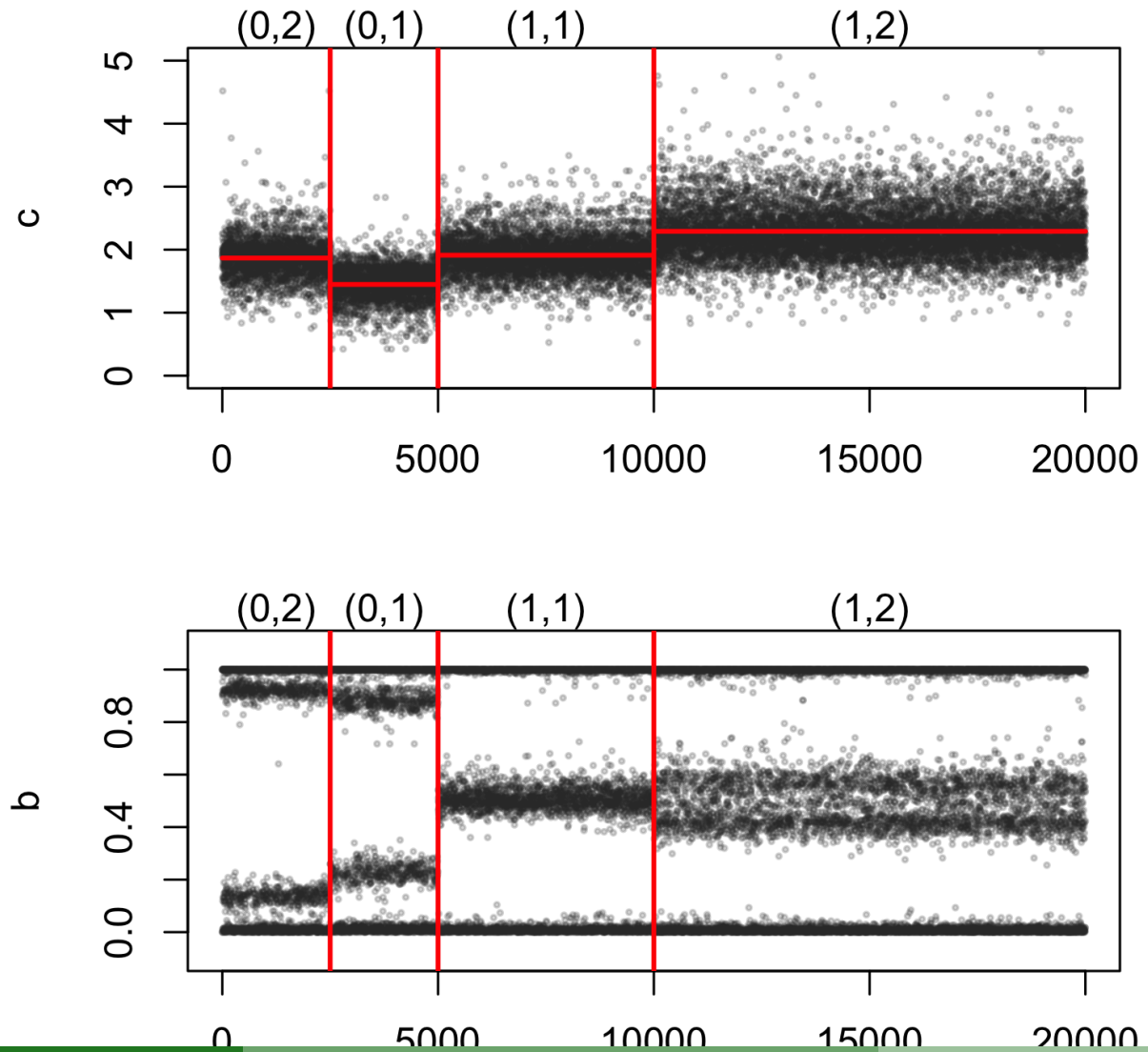
Signal-to-noise ratio can be controlled

Example : data set 2, 50% tumor cells (same "truth")



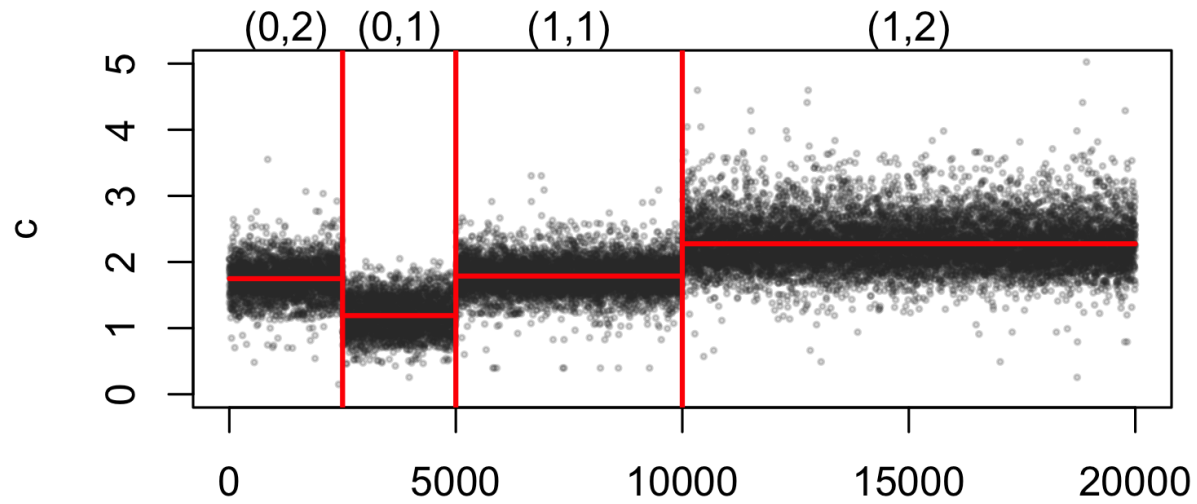
Signal-to-noise ratio can be controlled

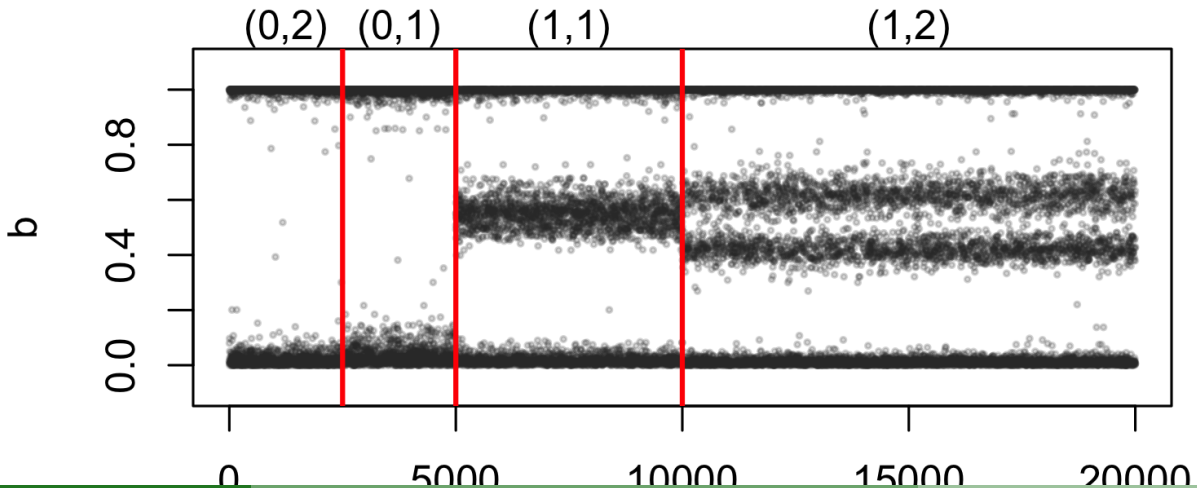
Example : data set 2, 79% tumor cells (same "truth")



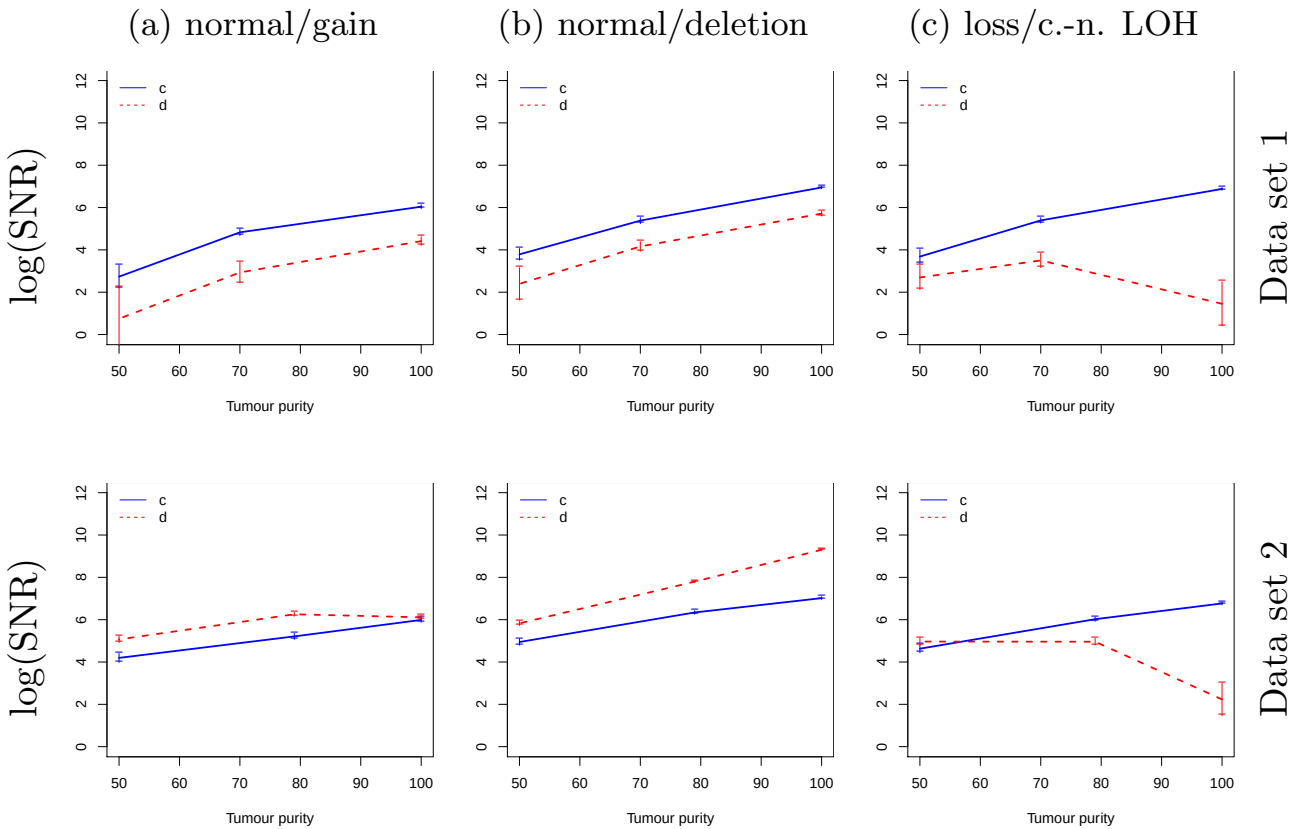
Signal-to-noise ratio can be controlled

Example : data set 2, 100% tumor cells (same "truth")





Signal depends heavily on the type of breakpoint



- difficulty generally increases with normal contamination
- SNR levels depend on the type of copy number transition
- neither *c* or *d* is always the best statistic

Summary of the proposed approach

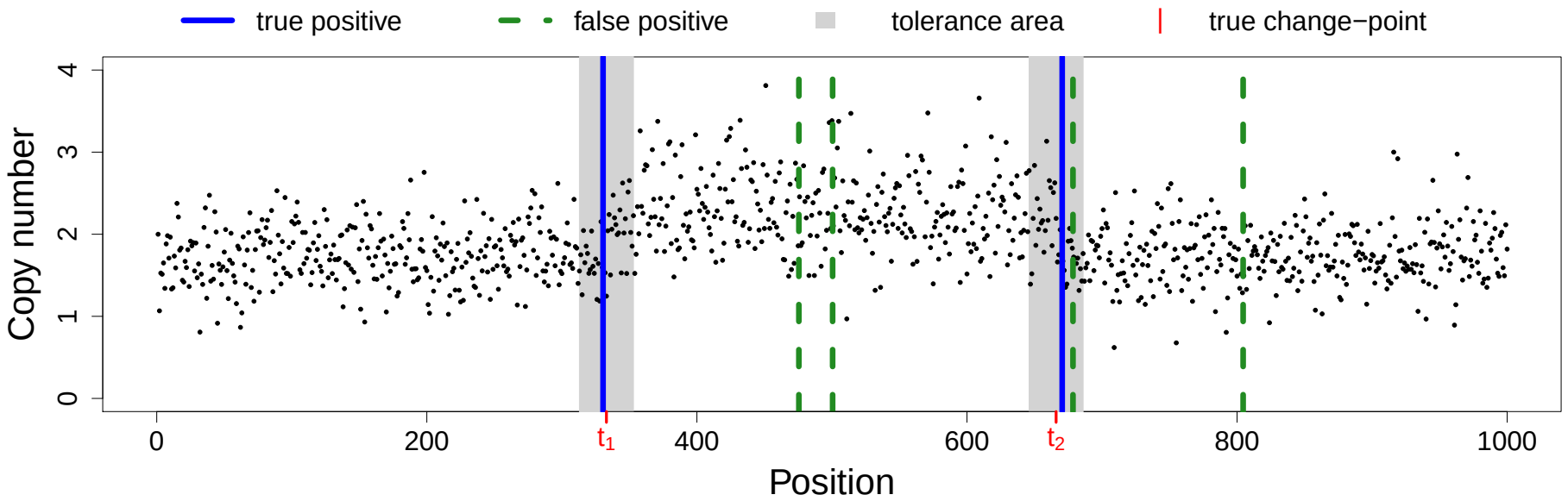
Features

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

A resampling-based data generation framework

1. truth (either user-specified or automatically generated)
   - K breakpoint positions
   - $K + 1$ copy-number state labels
2. signal (generated from two public SNP array dilution series)
   - GSE11976 (Illumina, HCC1395) : 34, 50, 79 and 100% of tumor cells
   - GSE29172 (Affy., NCI-H1395) : 30, 50, 70 and 100% of tumor cells.

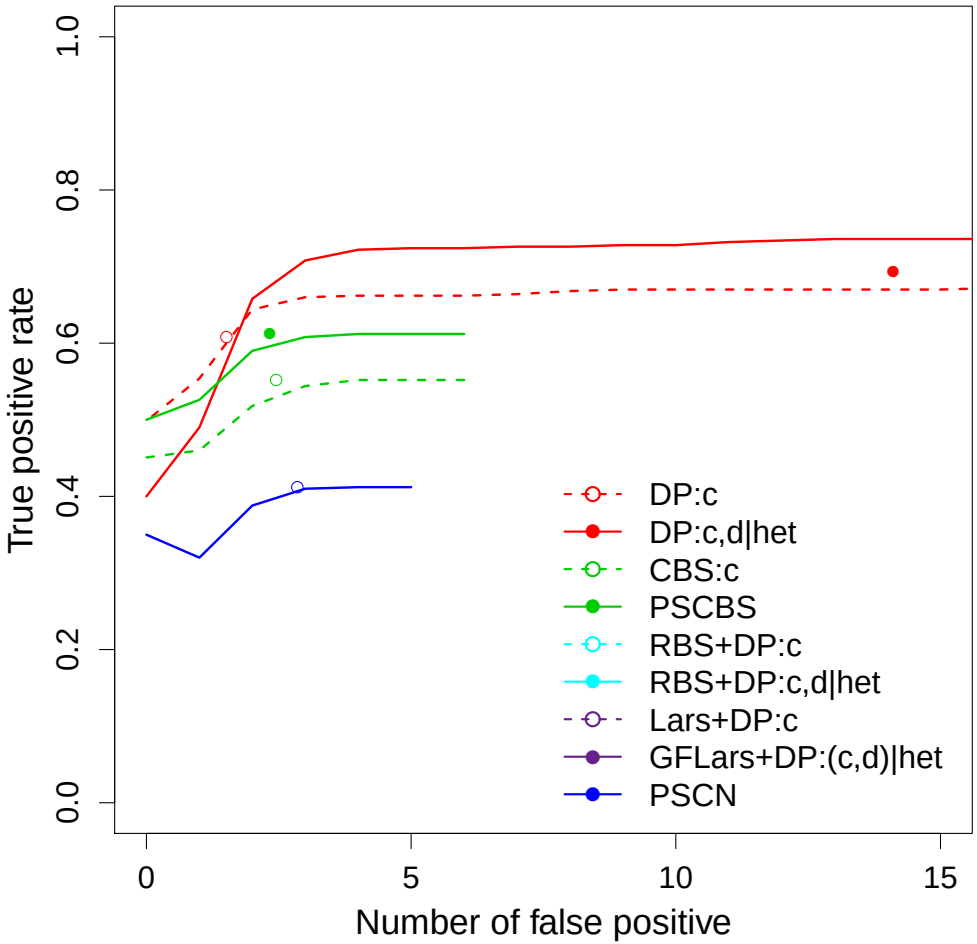
Defining true and false positives

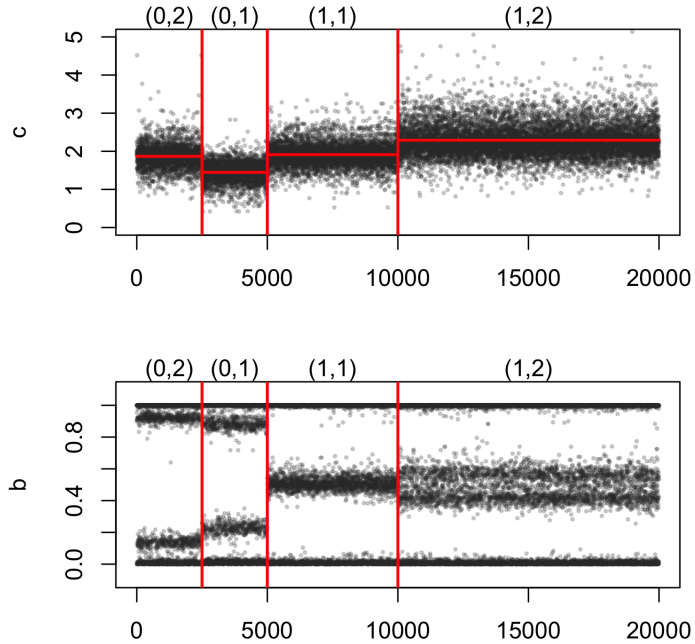


- two breakpoints at t_1 and t_2
- TP=2, FP=4

Taking both dimensions into account helps

100 profiles, $n = 5000$, $K = 5$, purity $= 79\%$, precision $= 1$

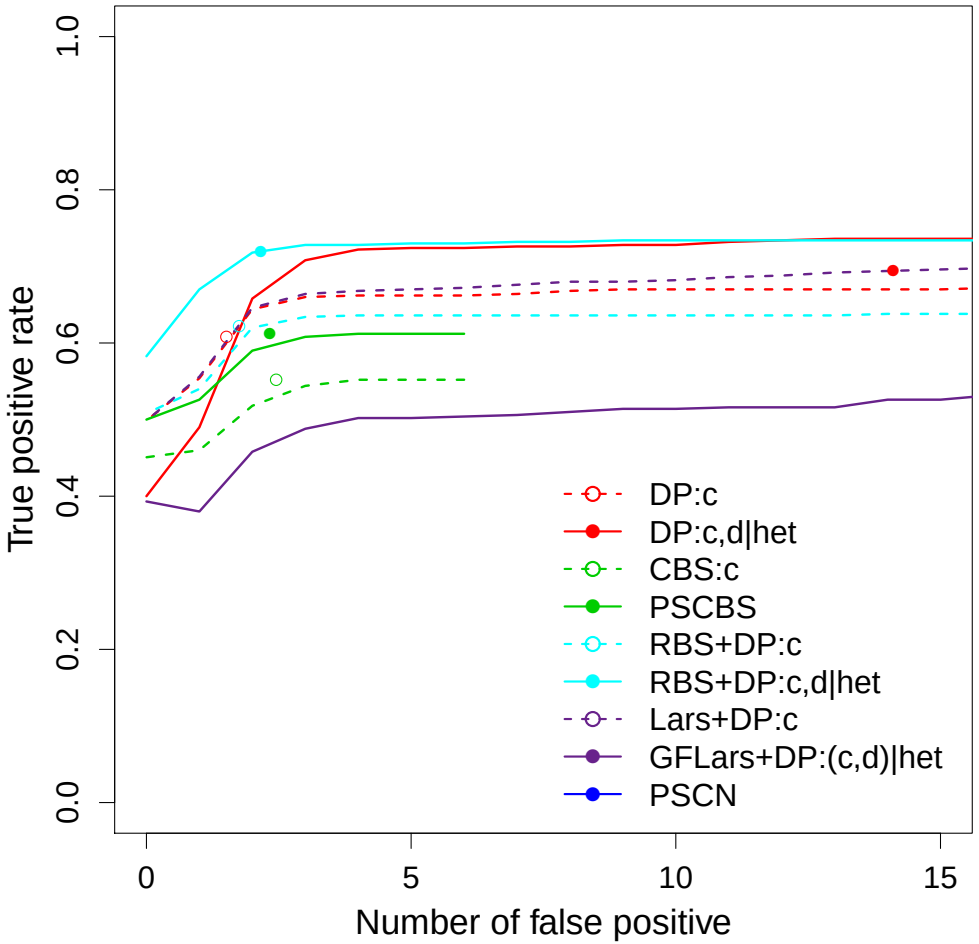


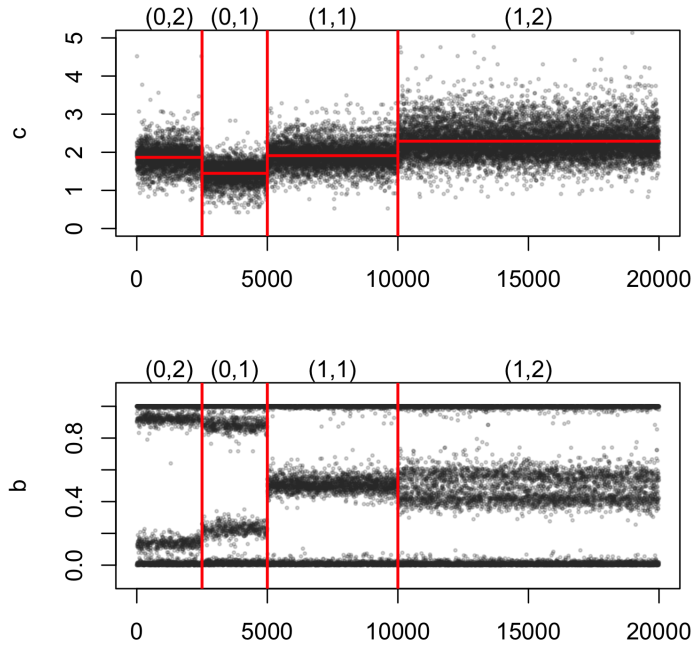


2d methds $\gg$ 1d methods

Taking both dimensions into account helps... or not

100 profiles, $n = 5000$, $K = 5$, purity $= 79\%$, precision $= 1$

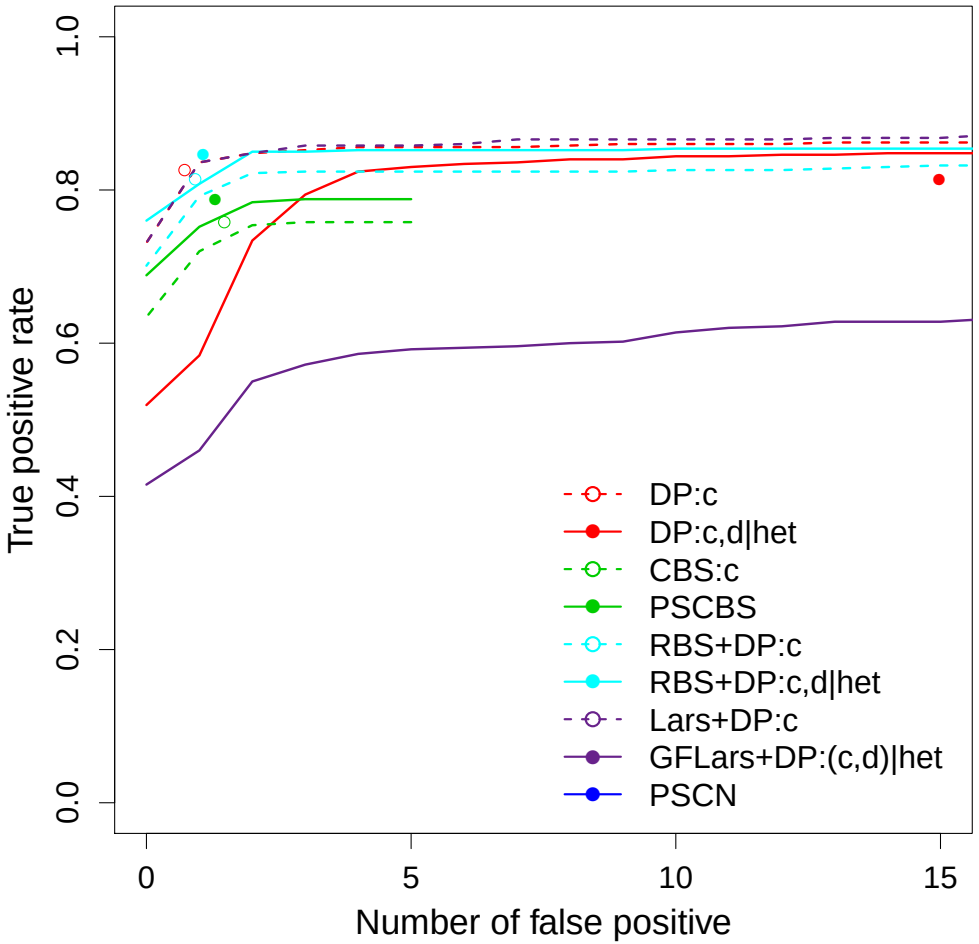


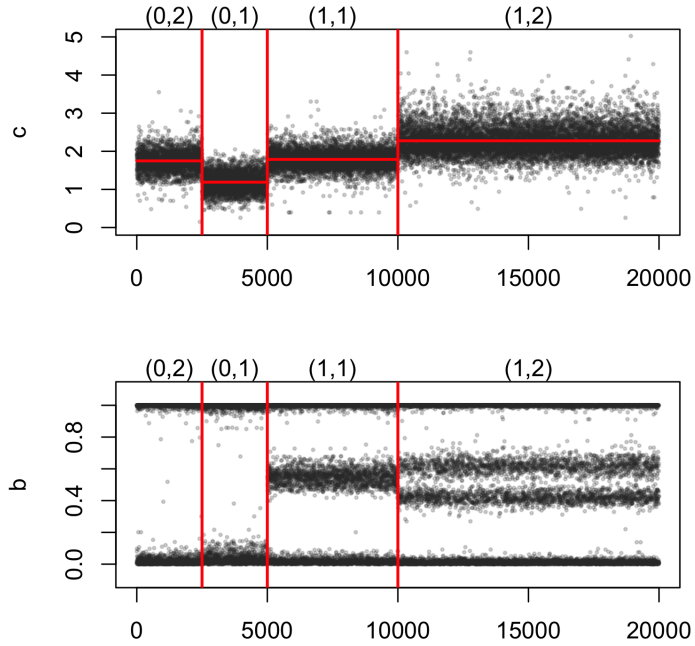


- 2d version of Fused Lars outperformed by 1d counterpart !
- reason : missing values in 'd' not handled

Influence of the proportion of normal cells

100 profiles, $n = 5000$, $K = 5$, purity $= 100\%$, precision $= 1$





Method ranking depends on % of normal cells

Conclusion

A flexible framework for generating realistic copy-number data

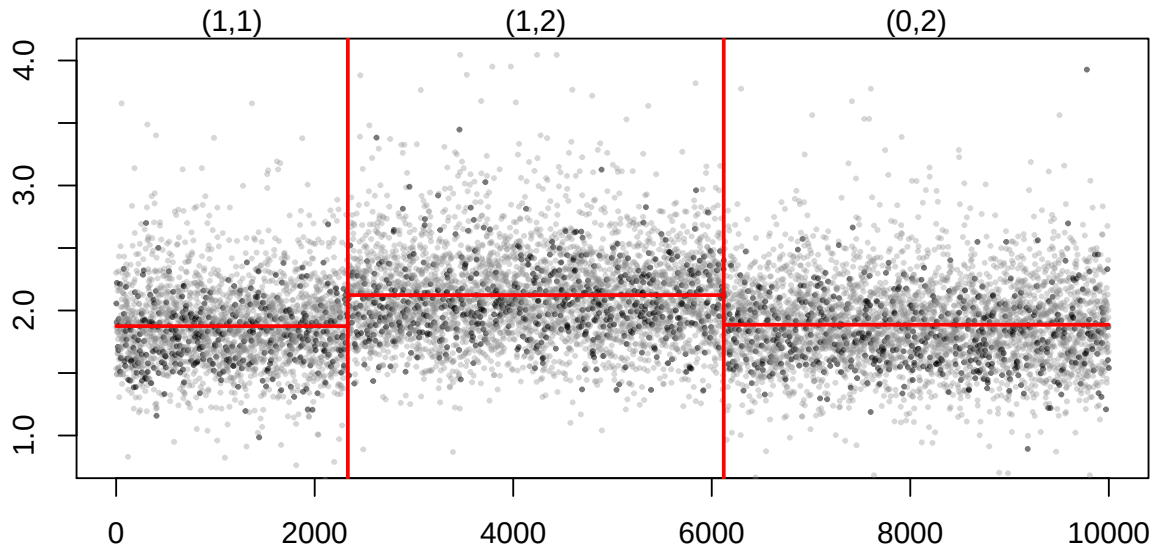
- based on real copy-number data
- SNR governed by biological parameters
- allows for synthetic data generation

Application to joint segmentation of SNP-array data

- No method is uniformly better
- Key biological parameters :
  - % informative values in each dimension
  - % normal cells in the biological sample

Outline

Segment-level copy numbers are not integers



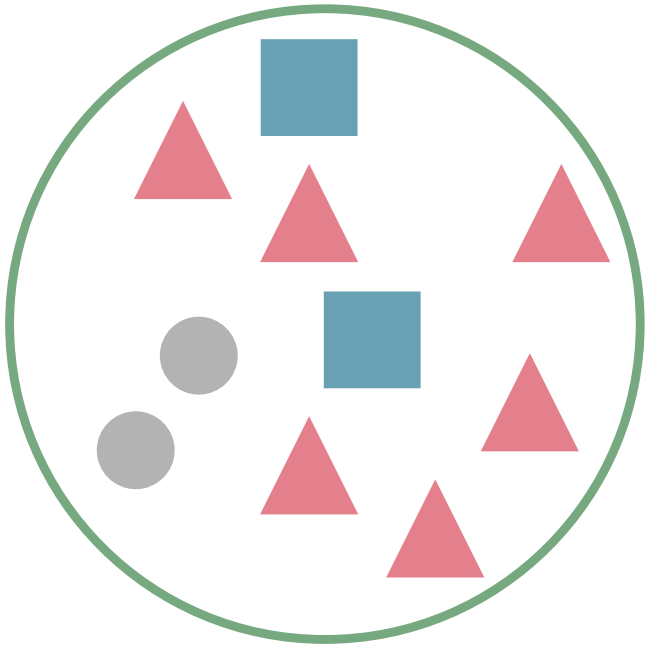
Possible reasons :

- normal contamination
- tumor heterogeneity
- overall ploidy

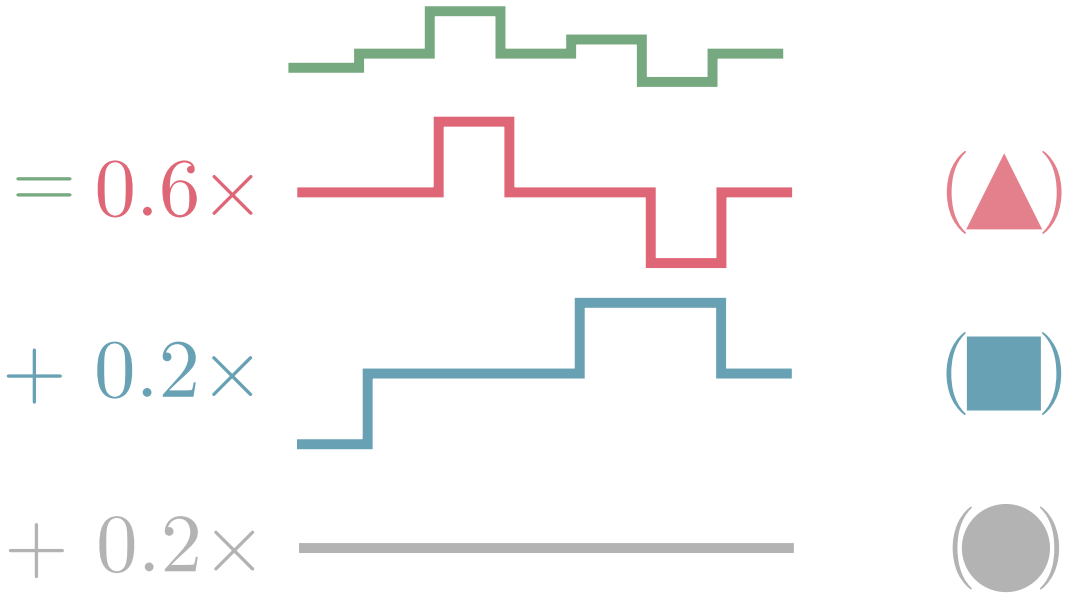
Heterogeneity of a tumor sample

A statistician's view

(a) Tumor sample



(b) Copy-number profile



Heterogeneity of two tumor samples

A statistician's view

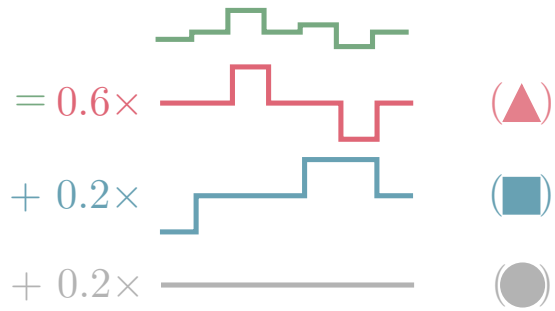
Sample 1

Sample 2

Natural assumption : the latent features are **shared across samples**

Basic model

$$\mathbf{Y}_i = \sum_{k=1}^{p} w_{ik}\mathbf{Z}_k + \mathbf{E}_i$$



- $\mathbf{Y}_i \in \mathbb{R}^L$: copy-number profile of sample i
- $\mathbf{Z}_k \in \mathbb{R}^J$: copy-number profile of the k-th latent profile
- w_{ik} : weight of latent profile k in sample i
- $\mathbf{E}_i \in \mathbb{R}^J$: reconstruction errors for sample i.

Goal

Given $(Y_i)_{1\leq i\leq n}$, estimate $\mathbf{Z}_k$ and w_{ik} for all $i = 1 \ldots n$ and $k = 1, \ldots p$.

NB : **$\mathbf{Z}_k$ does not depend on the sample index i**

Multi-sample latent feature model

$$\mathbf{Y} = \mathbf{WZ} + \mathbf{E}$$

- $\mathbf{Y}$ is the $n \times J$ matrix of copy-number signals for each sample,
- $\mathbf{W}$ is the $n \times p$ matrix of weights for each archetype,
- $\mathbf{Z}$ is the $p \times J$ matrix of copy-number signals for each archetype,

Parameter estimation

- **identifiability** issues
- many approaches from different literatures : NMF, artchetypal analysis, dictionary learning

State of the art

Nowak et al, 2011

Constraints :

- latent profiles Z_k are piecewise constant
- ℓ^2 constraint on the weights for identifiability

FFLAT

$$\min_{\mathbf{W}\in\mathbb{R}^{np},\mathbf{Z}\in\mathbb{R}^{Jp}} \left\{ \|\mathbf{Y}-\mathbf{W}\mathbf{Z}\|^2 + \mu\|\mathbf{Z}\|_1 + \lambda\left\|\mathbf{D}\mathbf{Z}^\top\right\|_1 \right\}$$
$$\text{s.t.} \quad \mathbf{W}_i\mathbf{W}_i^\top \leq 1 \quad \forall i = 1,\dots n, \quad (1)$$

where $\mathbf{D} = \begin{pmatrix} -1 & 1 & & & \\ & -1 & 1 & & \\ & & \ddots & \ddots & \\ & & & -1 & 1 \end{pmatrix}$

State of the art

Masecchia et al, (2013, 2015)

Additional constraints :

- weights are non-negative
- location-dependent weights (chromosome boundaries)

e-FFLAT

$$\min_{\mathbf{W}\in\mathbb{R}^{np},\mathbf{Z}\in\mathbb{R}^{Jp}} \left\{ \|\mathbf{Y}-\mathbf{W}\mathbf{Z}\|^2 + \mu \|\mathbf{Z}\|_1 + \lambda \left\| \theta \mathbf{D}\mathbf{Z}^\top \right\|_1 \right\}$$
$$\text{s.t.} \quad \mathbf{W}_i\mathbf{W}_i^\top \leq 1, \quad \mathbf{W}_i \succeq 0 \quad \forall i = 1,\dots n, \quad (2)$$

where $\theta \in \mathbb{R}^{L-1}$ encode user-given weights

Contributions

- remove the Lasso constraint
- constrain $\sum_k w_{ik} = 1$
- work with two-dimensional copy number signals
- work on segment-level data (after joint segmentation)

Optimization problem considered

$$\min_{\mathbf{W}\in\mathbb{R}^{np},\mathbf{Z}_m\in\mathbb{R}^{Jp}} \left\{ \sum_{m=1}^{2} \|\mathbf{Y}_m - \mathbf{W}\mathbf{Z}_m\|^2 + \lambda_m \left\|\mathbf{D}\mathbf{Z}_m^\top\right\|_1 \right\}$$
$$\text{s.t.} \quad \mathbb{1}_p^\top \mathbf{W}_i = 1, \quad \mathbf{W}_i \succeq 0 \quad \forall i = 1, \ldots n, \quad (3)$$

Parameter estimation

This optimization problem is not jointly convex in $(\mathbf{W}, \mathbf{Z}_1, \mathbf{Z}_2)$!

Algorithm

- Initialization : clustering
- for $t \leftarrow 1, \ldots T$,
  1. $\mathbf{W}^{(t)} \leftarrow \underset{\mathbf{W} \in \mathbb{R}^{np}}{\arg\min} \sum_{m=1}^{2} \left\| \mathbf{Y}_m - \mathbf{W}\mathbf{Z}_m^{(t-1)} \right\|^2 \quad \text{s.t.} \quad \mathbb{1}_p \mathbf{W}_i = 1, \quad \mathbf{W}_i \succeq 0,$
  2. $\mathbf{Z}_1^{(t)} \leftarrow \underset{\mathbf{Z}_1 \in \mathbb{R}^{Sp}}{\arg\min} \|\mathbf{Y}_1 - \mathbf{W}^{(t)}\mathbf{Z}_1\|^2 + \lambda_1 \|\mathbf{D}\mathbf{Z}_1^\top\|_1$
  3. $\mathbf{Z}_2^{(t)} \leftarrow \underset{\mathbf{Z}_2 \in \mathbb{R}^{Sp}}{\arg\min} \|\mathbf{Y}_2 - \mathbf{W}^{(t)}\mathbf{Z}_2\|^2 + \lambda_2 \|\mathbf{D}\mathbf{Z}_2^\top\|_1$

This can be done using standard optimization tools :

- Step 1 : linear inverse problem
- Steps 2 and 3 : lasso problems

Parameter calibration

Adapted from Nowak et al, 2011

3 tuning parameters : λ_1, λ_2, p

1. for each p, calibrate λ_1 and λ_2 using a BIC criterion

$$(nS) \times \log\left(\frac{\|\mathbf{Y} - \widehat{\mathbf{W}}\widehat{\mathbf{Z}}\|^2}{nS}\right) + k(\widehat{\mathbf{Z}}) \log(nS)$$

2. use the percentage of variance explained (PVE, aka R^2) to estimate p

$$\text{PVE}(p) = 1 - \frac{\|\mathbf{Y} - \widehat{\mathbf{W}}\widehat{\mathbf{Z}}\|^2}{\|\mathbf{Y} - \overline{\mathbf{Y}}\|^2},$$

Performance evaluation methods

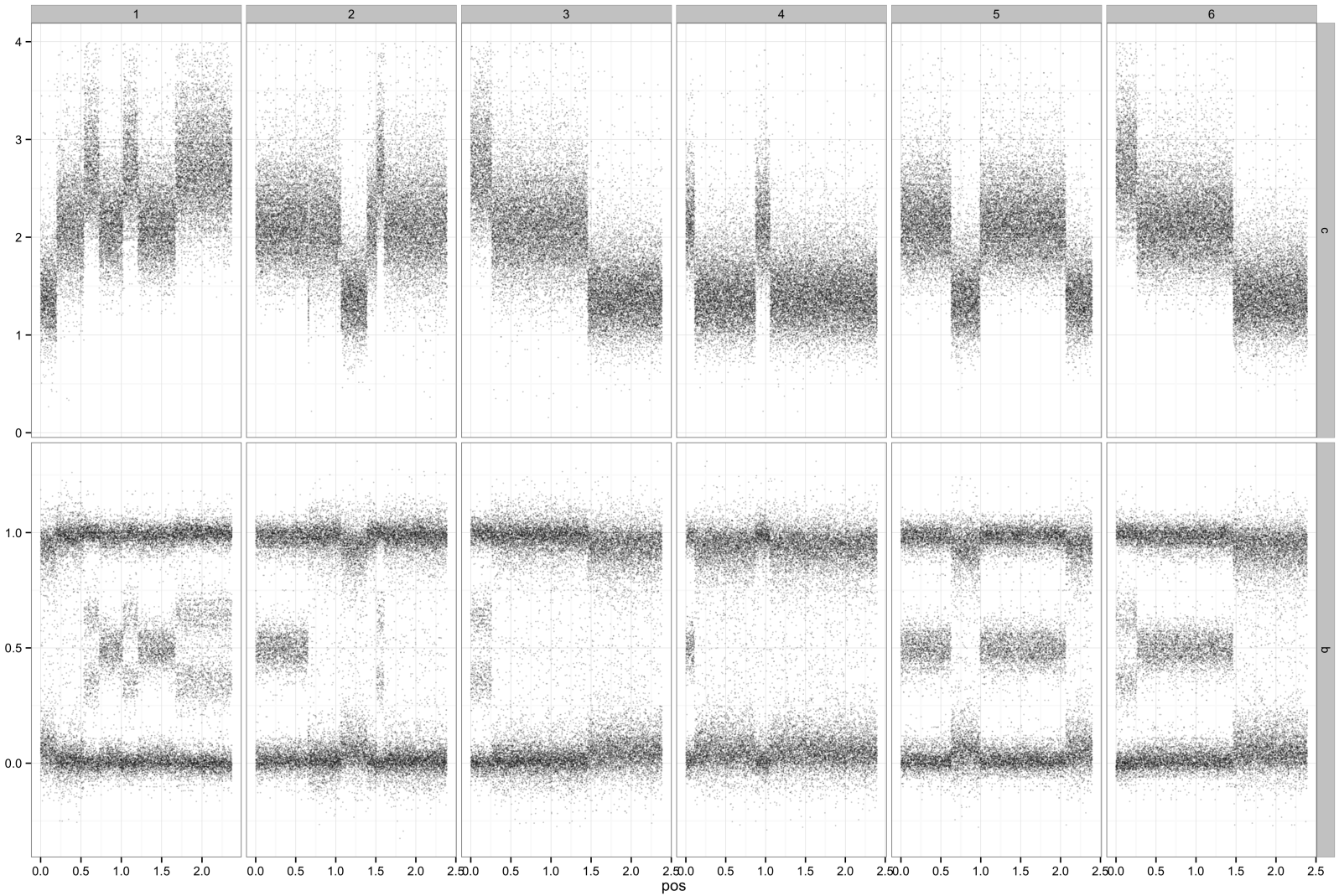
Criteria

- ability to recover the correct number of latent profiles
- quality of the reconstruction of weights and latent profiles
- ability to recover the true copy number alterations

Data

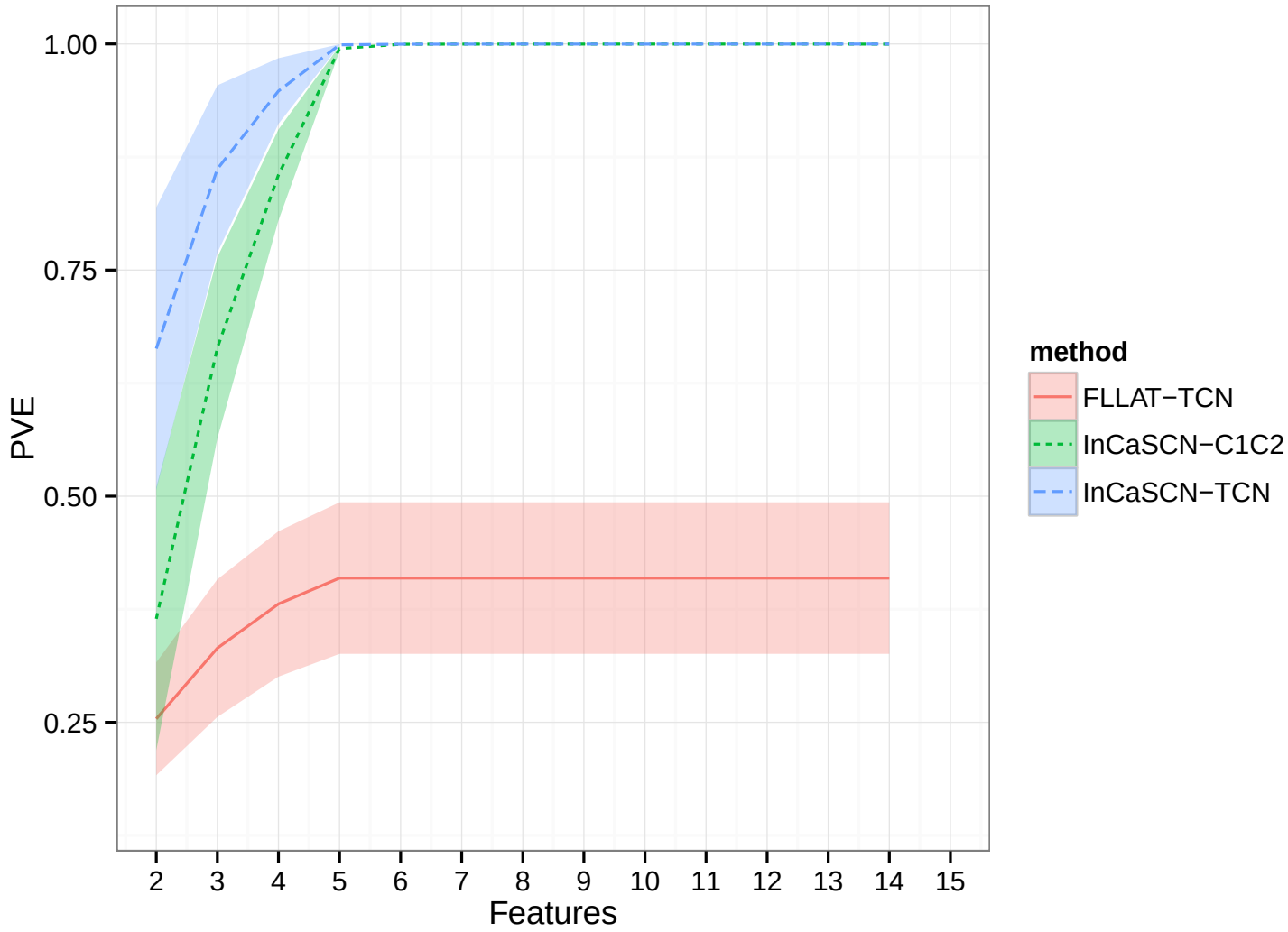
resampling of real, annotated data sets using the **acnr** and **jointseg** packages

Example of simulated latent profiles



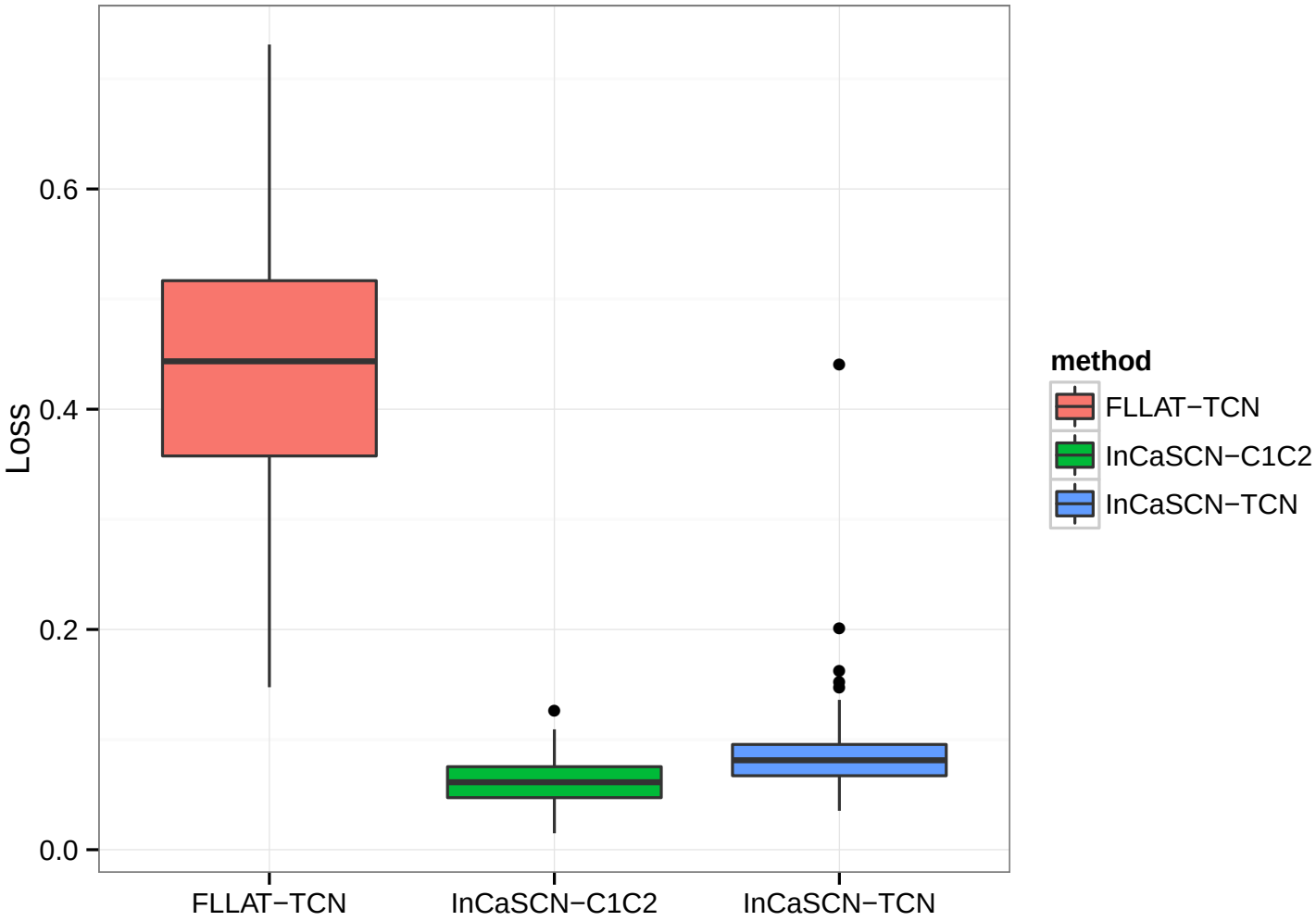
Estimation of the number of latent profiles

Truth= 6 latent profiles



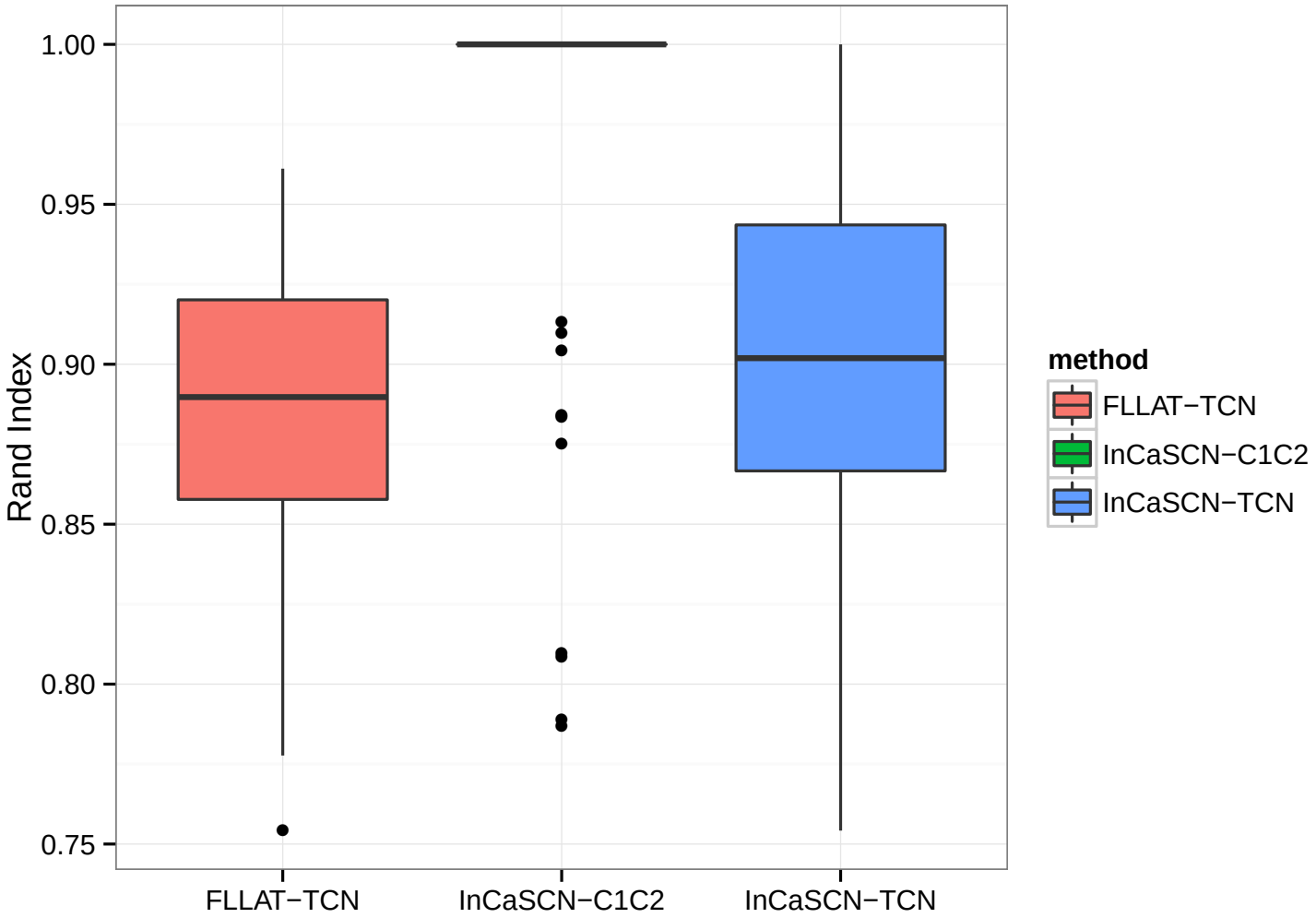
Quality of weights reconstruction

ℓ^2 loss of the weight matrix $\mathbb{E}(\|W - \hat{W}\|^2)$



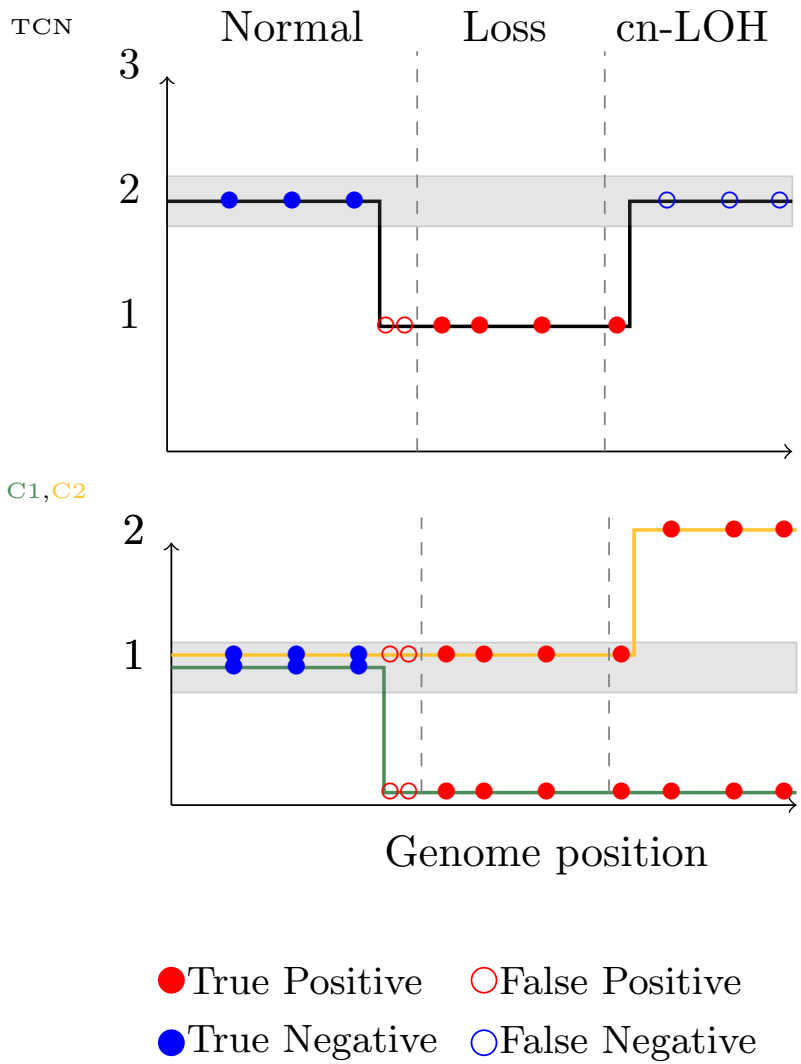
Quality of weights reconstruction

Rand index between clustering of samples on W and on $\hat{W}$



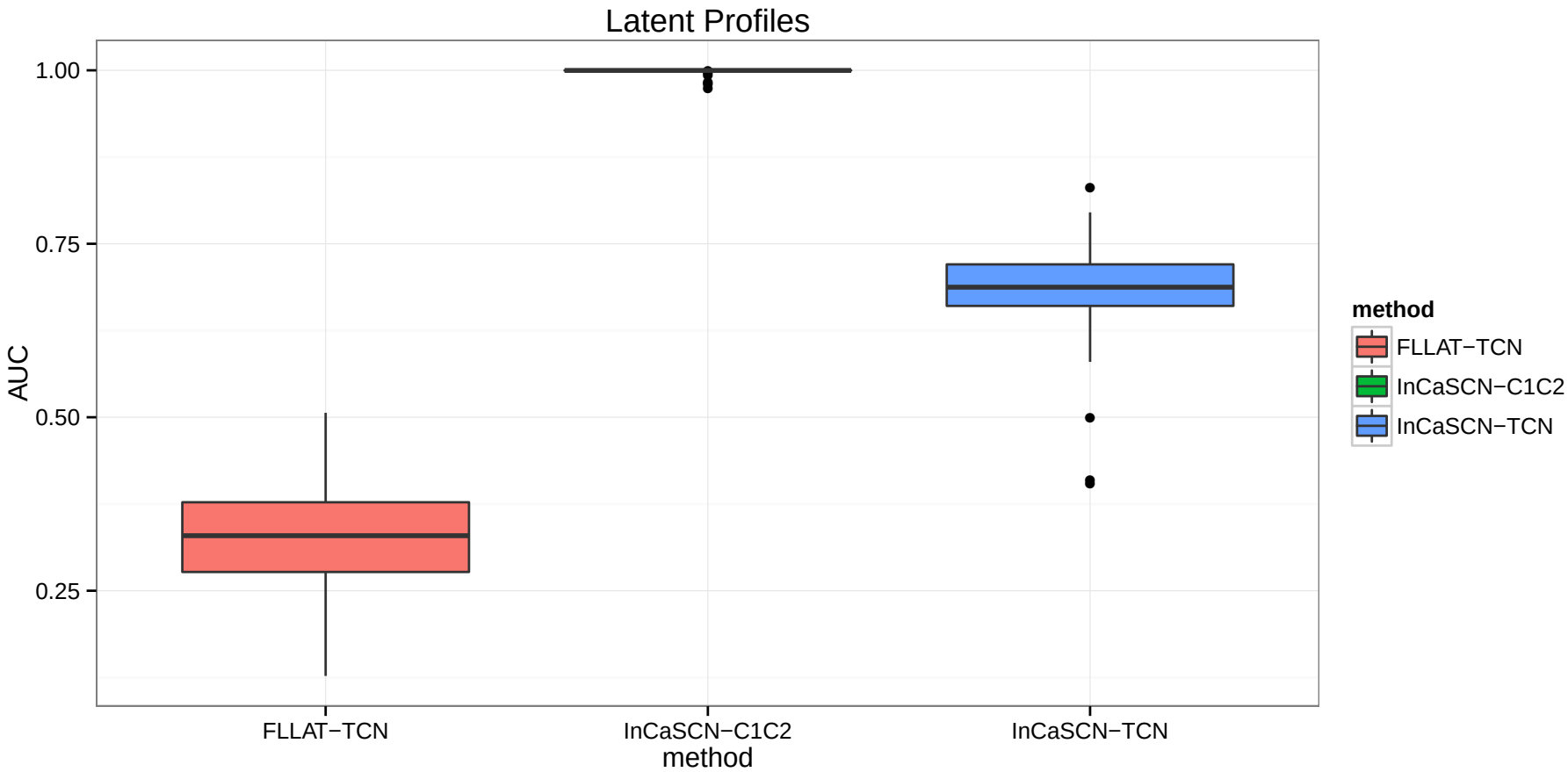
Ability to recover the true copy number alterations

Definition of true and false positives



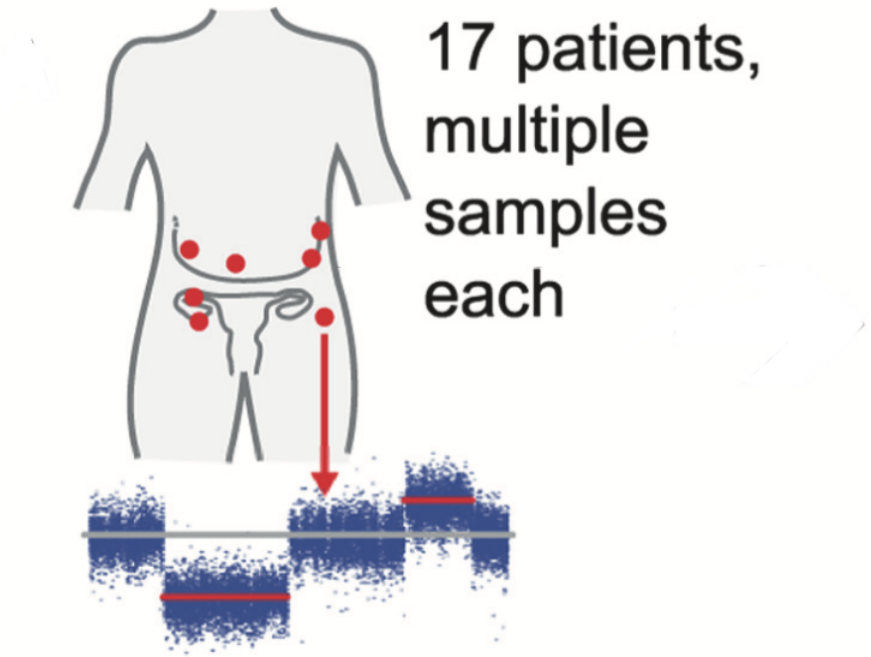
Ability to recover the true copy number alterations

Areas under the ROC curve

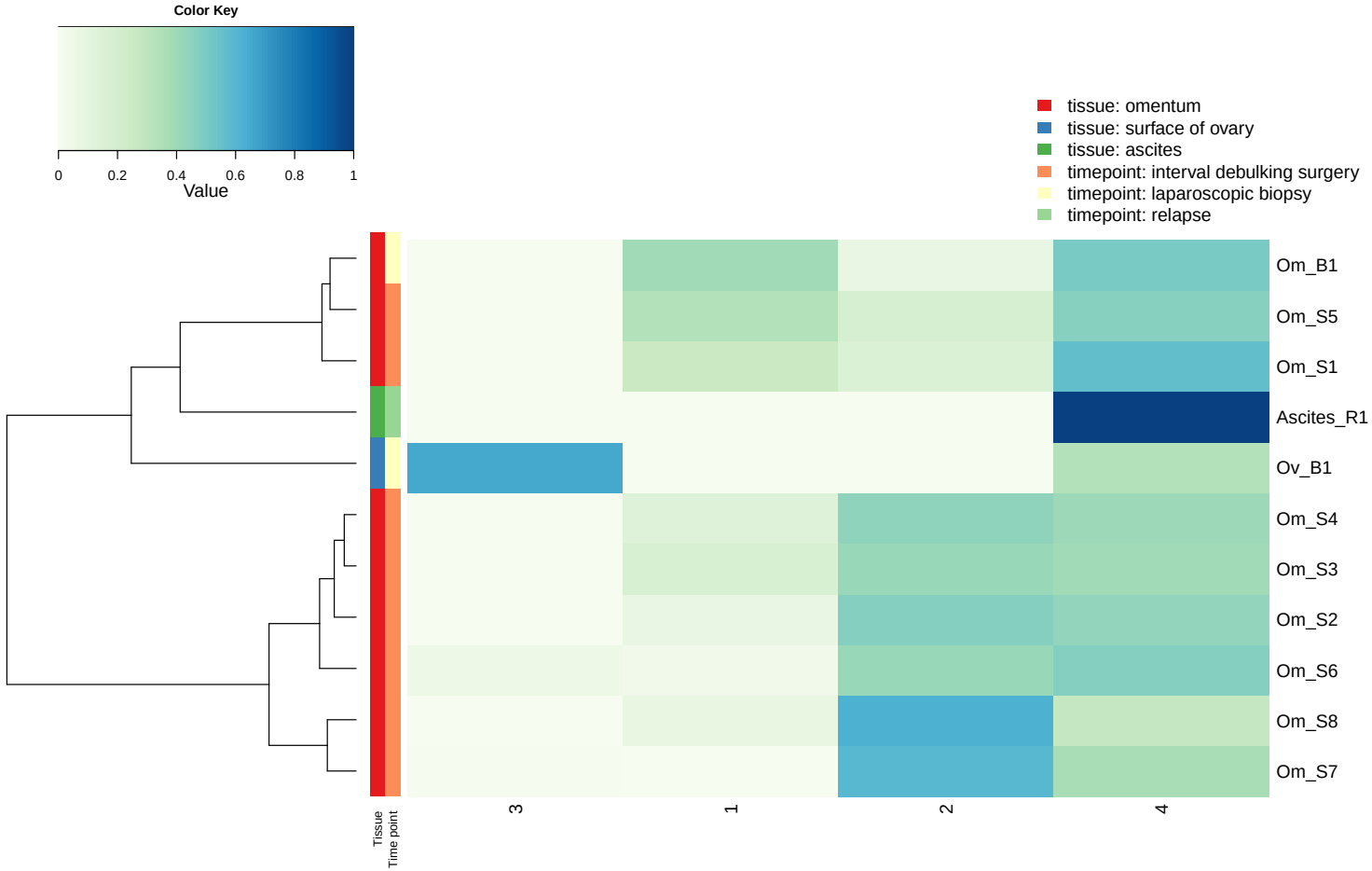


Spatial and temporal heterogeneity of ovarian cancer

Schwarz et al, PLoS Medicine, 2015



Results on patient 8



Acknowledgements

Laboratoire de Mathématiques et Modélisation d'Évry

- **Morgane Pierre-Jean**
- Guillem Rigaill
- Franck Samson

AgroParisTech/INRA MIA Paris

- Julien Chiquet

UCSF Epidemiology and biostatistics

- Henrik Bengtsson

References : evaluation methods

J Staaf, D Lindgren, J Vallon-Christersson, A Isaksson, and et al.
Segmentation-based detection of allelic imbalance and loss-of-heterozygosity in cancer cells using whole genome SNP arrays.
Genome Biol, 9(9) :R136, October 2008.

Hanni Willenbrock and Jane Fridlyand.
A comparison study : applying segmentation to array-CGH data for downstream analyses.
Bioinformatics, 21(22) :4084–91, Nov 2005.

Toby Hocking, Gudrun Schleiermacher, Isabelle Janoueix-Lerosey, Valentina Boeva, Julie Cappo, Olivier Delattre, Francis Bach, and Jean-Philippe Vert.
Learning smoothing models of copy number profiles using breakpoint annotations.
BMC Bioinformatics, 14(1) :164, 2013.

David Mosén-Ansorena, Ana Aransay, and Naiara Rodríguez-Ezpeleta.
Comparison of methods to detect copy number alterations in cancer using simulated and real genotyping data.
BMC bioinformatics, 13(1) :192, 2012.

References : segmentation methods

K. Bleakley and J.-P. Vert.
The group fused lasso for multiple change-point detection.
Technical report, Mines ParisTech, 2011.

Olshen AB et al.
Parent-specific copy number in paired tumor-normal studies using circular binary segmentation
Bioinformatics, (2011).

S. Gey and E. Lebarbier.
Using CART to Detect Multiple Change Points in the Mean for Large Sample
Technical report, *Statistics for Systems Biology research group*, 2008.

F. Picard and E. Lebarbier and M. Hoebeke and G. Rigaill and B. Thiam and S. Robin.
Joint segmenation, calling and normalization of multiple CGH profiles.
Biostatistics, 2011.

Chen, H., Xing, H. and Zhang, N.R.
Estimation of parent specific DNA copy number in tumors using high-density genotyping arrays.
PLoS Comput Biol,2011.

References : more segmentation methods


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


Olshen AB, Venkatraman ES, Lucito R, Wigler M.
Circular binary segmentation for the analysis of array-based DNA copy number data.
Biostatistics, (2004).


Zhang, Nancy R. and Siegmund, David O. and Ji, Hanlee and Li, Jun Z.
Detecting simultaneous changepoints in multiple sequences.
Biometrika, (2010)


Lai, Tze Leung and Xing, Haipeng and Zhang, Nancy
Stochastic segmentation models for array-based comparative genomic hybridization data analysis.
Biostatistics, (2008)


Z. Harchaoui and C. Lévy-Leduc.
Catching change-points with lasso.
Advances in Neural Information Processing Systems, 2008.

Many more informative probes for total copy numbers

Chip type : Affymetrix GenomeWideSNP_6

	All units	CN units	SNP units
Frequency	1,856,069	946,705	909,364
Proportion	100%	51%	49%

Unit types

	All units	AA	AB	BB
Frequency	1,856,069	326,500	251,446	331,418
Proportion	100%	18%	**14%**	18%

SNPs by genotype call for sample TCGA-23-1027