Development of statistical methods for DNA copy number analysis in cancerology

Morgane Pierre-Jean

Supervisors : Catherine Matias and Pierre Neuvial

Laboratoire de Mathématique et de Modélisation d'Evry, LaMME

December 2nd, 2016

Outline

1 Introduction

2 Segmentation

3 Heterogeneity Model

4 Simulations

5 Application to real data sets

6 Conclusion

Outline

1. Introduction
   - Alterations in tumor cells
   - Notion of Heterogeneity
2. Segmentation
3. Heterogeneity Model
4. Simulations
5. Application to real data sets
6. Conclusion

Objectives

Alterations in tumor cells can be observed at several levels

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

Why study genetic alterations in cancers ?

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

Objectives

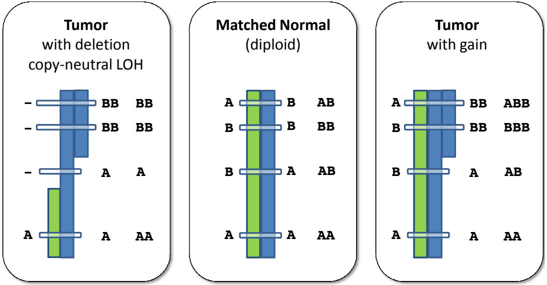
Alterations in tumor cells can be observed at several levels

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

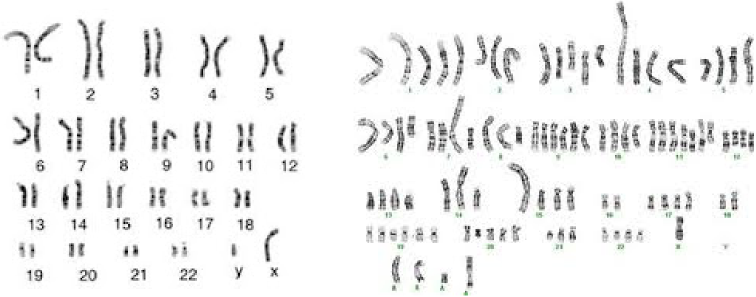
Why study genetic alterations in cancers ?

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

Illustration of alterations at level of DNA copy number



Human Karyotype



(a) Normal cell

(b) Tumor cell

How to measure DNA copy number more precisely ?

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

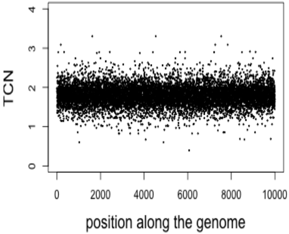
What kind of signals from SNPs arrays ?

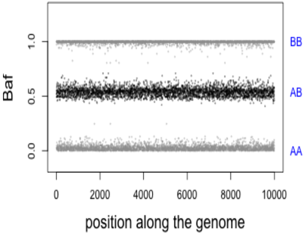
Total copy number

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

B allele fraction

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





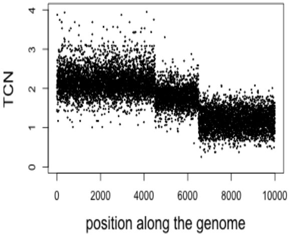
What kind of signals from SNPs arrays ?

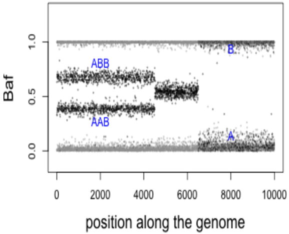
Total copy number

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

B allele fraction

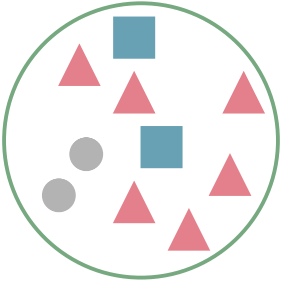
$$b_j = \frac{N_j^B}{c_j}$$



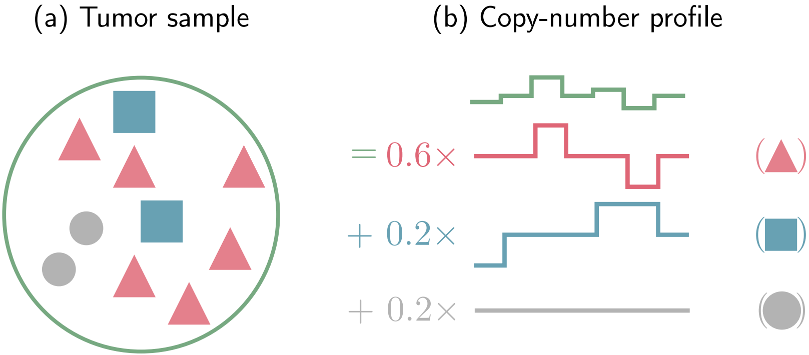


Notion of heterogeneity in cancers

- Differences between tumors of the same disease in different patients (inter-tumor heterogeneity)
- Differences between cancer cells within a single tumor of one patient (intra-tumor heterogeneity).



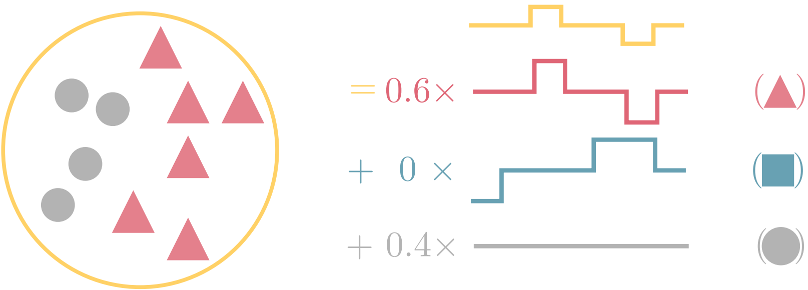
Heterogeneity illustration



(a) Tumor sample

(b) Copy-number profile

Heterogeneity illustration



(a) Tumor sample

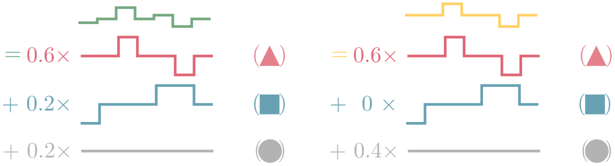
(b) Copy-number profile

Mathematical modelization

- $\mathrm{y}_{1\bullet} \in \mathbb{R}^J$ and $\mathrm{y}_{2\bullet} \in \mathbb{R}^J$ the observed DNA copy number profiles

$$\mathrm{y}_{1\bullet} = w_{11} z_{1\bullet} + w_{12} z_{2\bullet} + w_{13} z_{3\bullet}$$

$$\mathrm{y}_{2\bullet} = w_{21} z_{1\bullet} + w_{22} z_{2\bullet} + w_{23} z_{3\bullet}$$



$= 0.6\times$ (▲) $= 0.6\times$ (▲)

$+ 0.2\times$ (■) $+ \ 0 \ \times$ (■)

$+ 0.2\times$ (●) $+ 0.4\times$ (●)

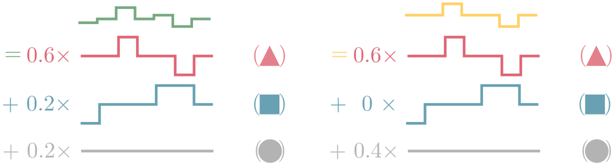
- Find w and z for the two profiles

Mathematical modelization

- $\mathrm{y}_{1\bullet} \in \mathbb{R}^J$ and $\mathrm{y}_{2\bullet} \in \mathbb{R}^J$ the observed DNA copy number profiles

$$\mathrm{y}_{1\bullet} = w_{11} z_{1\bullet} + w_{12} z_{2\bullet} + w_{13} z_{3\bullet}$$

$$\mathrm{y}_{2\bullet} = w_{21} z_{1\bullet} + w_{22} z_{2\bullet} + w_{23} z_{3\bullet}$$



- Find w and z for the two profiles

General mathematical modelization

- Let $\mathrm{y}_{i\bullet} \in \mathbb{R}^J$ the observed DNA copy number profiles

$$\mathrm{y}_{i\bullet} = \sum_{k=1}^{p} w_{ik} z_{k\bullet} + \epsilon$$

- Latent profiles assumed to be shared between the observed profiles
- Minimize $\sum_{i=1}^{n} \|\mathrm{y}_{i\bullet} - \sum_{k=1}^{p} w_{ik} z_{k\bullet}\|^2$ under some constraints.

Related works

- Matrix Factorization problem

$$\min_{W,Z} \|\mathrm{Y} - \mathrm{WZ}\|_F^2$$

- Penalized latent models to infer heterogeneity
  - Fused Lasso latent model `FLlat` (Nowak et al., 2011)
  - CGH analysis with Dictionary Learning `e-FLlat` (Masecchia et al., 2013)
  - Evolutionary history by next-generation sequencing `Canopy` (Jiang et al., 2016)

InCaSCN- Inferring Cancer Subclone using Copy Number

Features of method

- joint segmentation of all n profiles $\Rightarrow S-1$ breakpoints (Pierre-Jean et al., Briefings in Bionformatics, 2015)
- Integration of B allele fraction information by using transformations
- Biological interpretation of constraints on latent profiles of TCN and BAF and weight matrix W

Outline

1 Introduction

2 Segmentation
- Models
- Recursive Binary Segmentation for multiple samples

3 Heterogeneity Model

4 Simulations

5 Application to real data sets

6 Conclusion

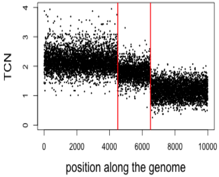
What is segmentation ?

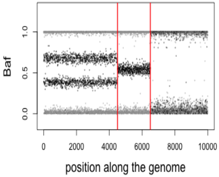
Total copy number

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

B allele fraction

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

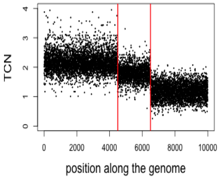




What is segmentation ?

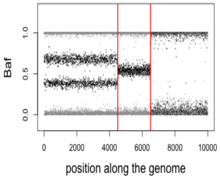
Total copy number

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



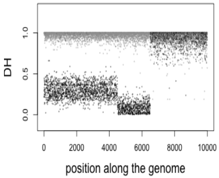
B allele fraction

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



Decrease of Heterozygosity

$$d_j = 2 \times |b_j - \tfrac{1}{2}|$$



What is segmentation ?

Total copy number

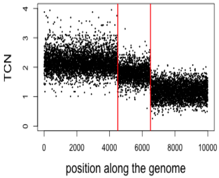
$$c_j = N_j^A + N_j^B$$

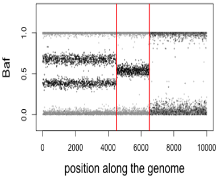
B allele fraction

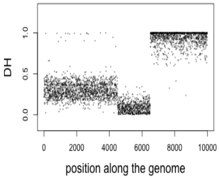
$$b_j = \frac{N_j^B}{c_j}$$

Decrease of Heterozygosity

$$d_j = 2 \times |b_j - \frac{1}{2}|$$







What is segmentation ?

Total copy number

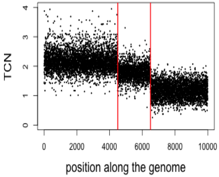
$$c_j = N_j^A + N_j^B$$

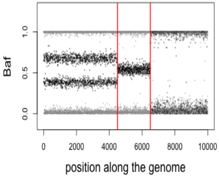
B allele fraction

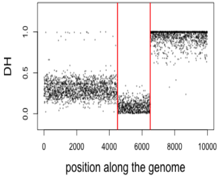
$$b_j = \frac{N_j^B}{c_j}$$

Decrease of Heterozygosity

$$d_j = 2 \times |b_j - \tfrac{1}{2}|$$







Segmentation methods

- Multiple change-point
- Recursive

- Total variation
- Hidden Markov Models
- Kernel methods

Segmentation methods

- Multiple change-point
- Recursive

- Total variation
- Hidden Markov Models
- Kernel methods

Segmentation methods

- Multiple change-point
- Recursive
    - Joint segmentation
- Total variation
- Hidden Markov Models
- Kernel methods
    - Change-point detection in whole distribution

A change-point model

- Biological assumption : DNA copy number signal is piecewise constant in the mean
- Statistical model for $S-1$ change points at $(t_1, ...t_{S-1})$:

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

where $\forall s \in \{1, \ldots, S\}, \forall j \in [t_{S-1}, t_S[\qquad \gamma_j = \Gamma_s$

A change-point model

- Biological assumption : DNA copy number signal is piecewise constant in the mean
- Statistical model for $S-1$ change points at $(t_1, ...t_{S-1})$:

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

where $\forall s \in \{1, \ldots, S\}, \forall j \in [t_{S-1}, t_S[\qquad \gamma_j = \Gamma_s$

Complexity

- Challenges : S and $(t_1, ...t_{S-1})$ are unknown
- For a fixed S, the number of possible partitions : $C_{J-1}^{S-1} = \mathcal{O}(J^{S-2})$

Two-step approaches for joint segmentation

Gey and Lebarbier (2008) and Vert and Bleakley (2010)

First step :

- Running a fast but approximate segmentation method (RBS)

Second step

- Pruning the final set of breakpoints using dynamic programming that is slower but exact

Versatility of RBS

- Possibility to have different scales
- TCN-DoH segmentation
- Several TCN signals
- Several TCN-DoH signals

Two-step approaches for joint segmentation

Gey and Lebarbier (2008) and Vert and Bleakley (2010)

First step :

- Running a fast but approximate segmentation method (RBS)

Second step

- Pruning the final set of breakpoints using dynamic programming that is slower but exact

Versatility of RBS

- Possibility to have different scales
- TCN-DoH segmentation
- Several TCN signals
- Several TCN-DoH signals

Binary Segmentation

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

$$Z_j = \frac{\left(\frac{S_j}{j} - \frac{S_J - S_j}{J - j}\right)}{\sqrt{\frac{1}{j} + \frac{1}{J-j}}}, \tag{1}$$

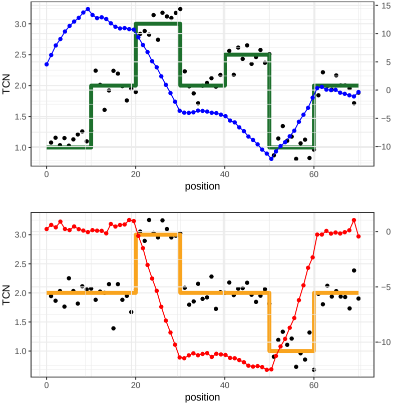
And $S_j = \sum_{1 \leq t \leq j} c_j$

If ($d > 1$) : the likelihood ratio statistic becomes $\max_{1 \leq j \leq J} \|Z_j\|_2^2$

First step : Recursive Binary Segmentation (RBS)

Complexity : $O(dJlog(S))$

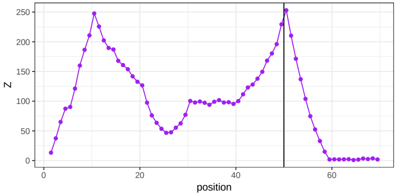
- First breakpoint
- For each j : we compute Z_j : $t_1 = \arg\max_{1 \leq j \leq J} \|Z_j\|_2^2$



First step : Recursive Binary Segmentation (RBS)

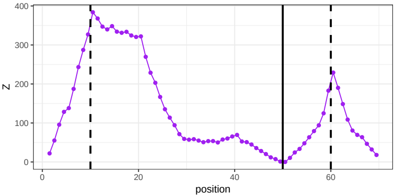
Complexity : $O(dJlog(S))$

- First breakpoint
- For each j : we compute Z_j : $t_1 = \arg\max_{1 \leq j \leq J} \|Z_j\|_2^2$



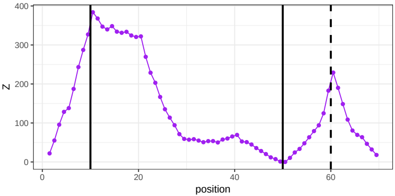
First step : Recursive Binary Segmentation (RBS)

- Second breakpoint :
  - $\max_{1 \leq j \leq t_1} \|Z_j\|_2^2$
  - $\max_{t_1 < j \leq J} \|Z_j\|_2^2$
- Compute RSE for each segment.
- Keep the RSE that yield the maximum gain
- Add the breakpoint to the active set



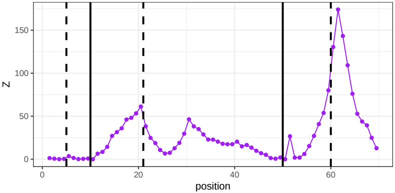
First step : Recursive Binary Segmentation (RBS)

- Second breakpoint :
  - $\max_{1 \leq j \leq t_1} \|Z_j\|_2^2$
  - $\max_{t_1 < j \leq J} \|Z_j\|_2^2$
- Compute RSE for each segment.
- Keep the RSE that yield the maximum gain
- Add the breakpoint to the active set



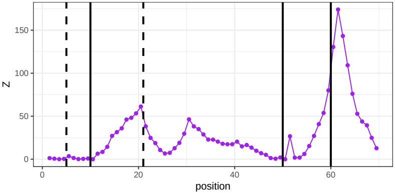
First step : Recursive Binary Segmentation (RBS)

- Third breakpoint :
  - $\max_{1 \leq j \leq t_1} \|Z_j\|_2^2$
  - $\max_{t_1 < j \leq t_2} \|Z_j\|_2^2$
  - $\max_{t_2 < j \leq J} \|Z_j\|_2^2$
- Compute RSE for each segment.
- Keep the RSE that yield the maximum gain
- Add the breakpoint to the active set



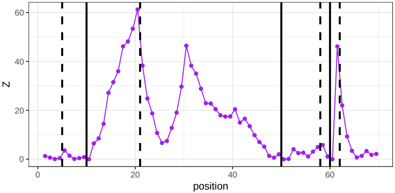
First step : Recursive Binary Segmentation (RBS)

- Third breakpoint :
  - $\max_{1 \leq j \leq t_1} \|Z_j\|_2^2$
  - $\max_{t_1 < j \leq t_2} \|Z_j\|_2^2$
  - $\max_{t_2 < j \leq J} \|Z_j\|_2^2$
- Compute RSE for each segment.
- Keep the RSE that yield the maximum gain
- Add the breakpoint to the active set



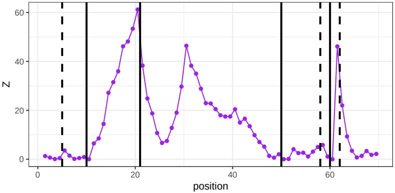
First step : Recursive Binary Segmentation (RBS)

- Third breakpoint :
  - $\max_{1 \leq j \leq t_1} \|Z_j\|_2^2$
  - $\max_{t_1 < j \leq t_2} \|Z_j\|_2^2$
  - $\max_{t_2 < j \leq J} \|Z_j\|_2^2$
- Compute RSE for each segment.
- Keep the RSE that yield the maximum gain
- Add the breakpoint to the active set



First step : Recursive Binary Segmentation (RBS)

- Third breakpoint :
  - $\max_{1 \leq j \leq t_1} \|Z_j\|_2^2$
  - $\max_{t_1 < j \leq t_2} \|Z_j\|_2^2$
  - $\max_{t_2 < j \leq J} \|Z_j\|_2^2$
- Compute RSE for each segment.
- Keep the RSE that yield the maximum gain
- Add the breakpoint to the active set



Summary

Contributions to segmentation methods

- Implementation of a fast joint segmentation followed by a pruning. (jointseg package)
- Kernel methods (preprint submitted to CSDA)
- Evaluation of performance (Pierre-Jean et al., Briefings in Bionformatics, 2015)

Outline

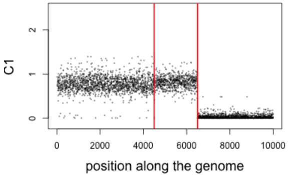
1. Introduction
2. Segmentation
3. Heterogeneity Model
   - BAF integration
   - Model
   - Algorithm
   - Model selection
4. Simulations
5. Application to real data sets
6. Conclusion

Integrating BAF through Parental copy numbers

What is parental copy number ?
$d_j = 2|b_j - 1/2|$ for AB SNPs

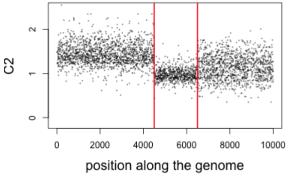
Minor copy number

$$c_j^1 = c_j(1 - d_j)/2$$



Major copy number

$$c_j^2 = c_j(1 + d_j)/2$$



Model on parental copy number

$$\min_{W,Z^1,Z^2} \|Y^1 - WZ^1\|_F^2 + \lambda_1 \sum_{k=1}^{p} \sum_{s=1}^{S-1} |z^1_{k,s+1} - z^1_{k,s}| \qquad (2)$$

$$\|Y^2 - WZ^2\|_F^2 + \lambda_2 \sum_{k=1}^{p} \sum_{s=1}^{S-1} |z^2_{k,s+1} - z^2_{k,s}|$$

s. t $w_{i\bullet} \in \Delta_p$ where
$\Delta_p = \left\{ w \in \mathbb{R}^p \quad s.t. \quad w \geq 0 \quad and \quad \sum_{k=1}^{p} w_k = 1 \right\}$

Final algorithm

Algorithm 1 Find weights and latent profiles

```
**Parameters** : $\lambda_1, \lambda_2$ and $p$
**INIT** : Matrices $Y \in \mathbb{R}^{n \times S}$, $Y^1 \in \mathbb{R}^{n \times S}$ and $Y^2 \in \mathbb{R}^{n \times S}$ and matrix $Z_0^1$ and $Z_0^2 \in \mathbb{R}^{p \times S}$, and
**for** $l = 0, 1, 2, \dots$ **do**
    Minimize in $W$ with $Z_l^1$ and $Z_l^2$ fixed
    Minimize in $Z^1$ with $W_l$ fixed
    Minimize in $Z^2$ with $W_l$ fixed
    $W_l$, $Z_l^1$ and $Z_l^2$ are updated
    Check if $\|W_{l-1} - W_l\|_2^2 < \epsilon$ or $max_{it}$ is reached
**end for**
```

Final algorithm

Algorithm 2 Find weights and latent profiles

```
**Parameters :** $\lambda_1, \lambda_2$ and $p$
**INIT :** Matrices $Y \in \mathbb{R}^{n \times S}$, $Y^1 \in \mathbb{R}^{n \times S}$ and $Y^2 \in \mathbb{R}^{n \times S}$ and matrix $Z_0^1$ and $Z_0^2 \in \mathbb{R}^{p \times S}$, and
**for** $l = 0, 1, 2, \dots$ **do**
    Minimize in $W$ with $Z_l^1$ and $Z_l^2$ fixed
    Minimize in $Z^1$ with $W_l$ fixed
    Minimize in $Z^2$ with $W_l$ fixed
    $W_l$, $Z_l^1$ and $Z_l^2$ are updated
    Check if $\|W_{l-1} - W_l\|_2^2 < \epsilon$ or $max_{it}$ is reached
**end for**
```

Solving 4 : Inference of W

- Weights of each patient can be treated independently
- Solve n least-squares problems with equality constraint plus inequality constraints for the non-negativity of the coefficient
- linear inverse problem that can be solved in R with the package **limSolve**.

Solving 5 and 6 : Inference of latent profiles

- for a fixed W cut into two independent LASSO problems in (Z_1, Z_2)
- Use matrix algebra and properties of the vectorization operator
- Obtain LASSO problem that can be solved in R with the package **glmnet**.

Choice of λ_1 and λ_2 values when p is fixed

- Use a BIC criterion
- We search to minimize :

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

where $k(Z^T)$ is the number of breakpoints.

- This criterion helps to strike a balance between over-fit and under-fit models.

Choice of p

- Use the percentage of variation explained (PVE) for each p, where the PVE is defined as :

$$PVE_P \quad = \quad 1 - \frac{\sum_{i=1}^{n}\sum_{j=1}^{S}\left(y_{ij} - \sum_{k=1}^{p}\hat{w}_{ik}\hat{z}_{kj}\right)^2}{\sum_{i=1}^{n}\sum_{j=1}^{S}\left(y_{ij} - \bar{y}_i\right)^2}$$

where $\bar{y}_i = \frac{\sum_{j=1}^{S} y_{ij}}{S}$.

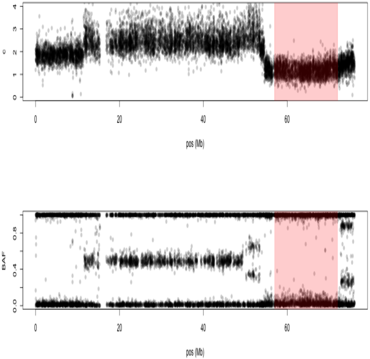
Outline

1. Introduction
2. Segmentation
3. Heterogeneity Model
4. Simulations
   - Generating data with known truth
   - Framework
5. Application to real data sets
6. Conclusion

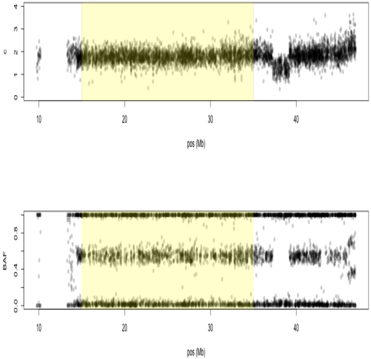
Proposed approach

Step 1- Annotate a real data set

Loss of one copy (Chr18)



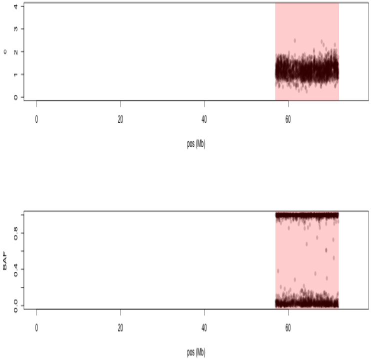
Normal region (Chr21)



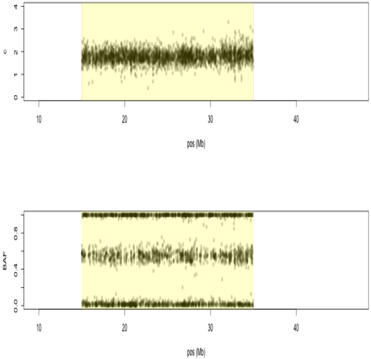
Proposed approach

Step 1- Annotate a real data set

Loss of one copy (Chr18)

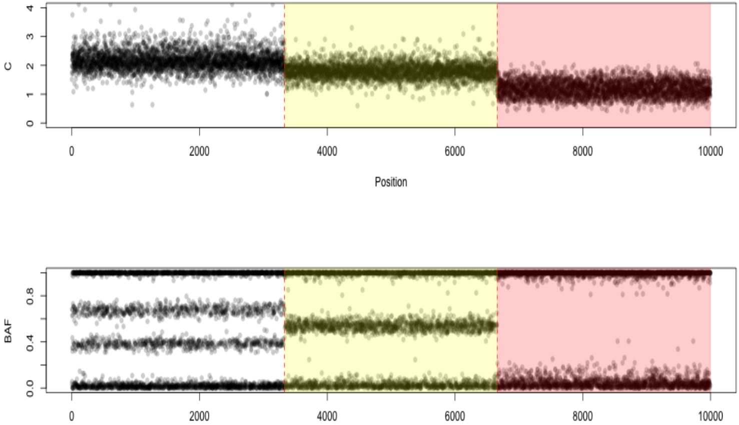


Normal region (Chr21)



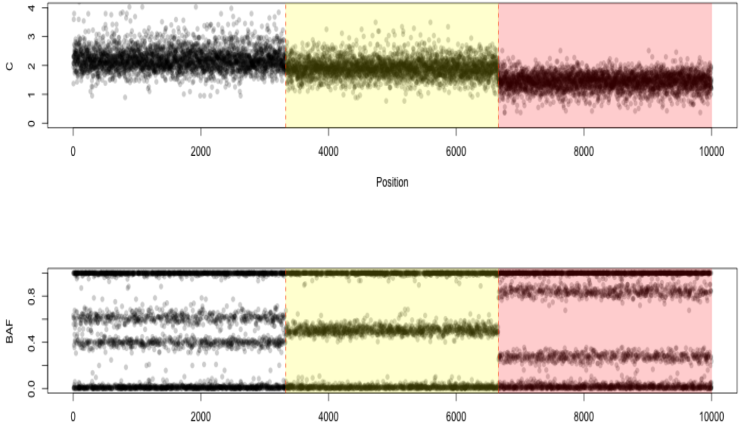
Proposed approach

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



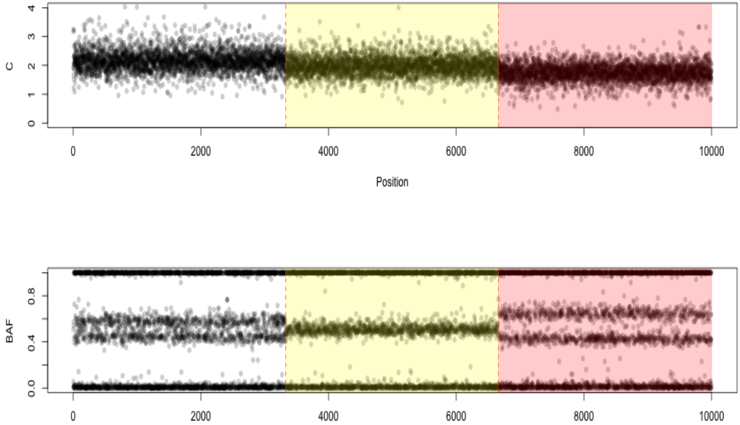
Proposed approach

Step 2 - Synthetic data generation by resampling 79% tumor cells



Proposed approach

Step 2 - Synthetic data generation by resampling 50% tumor cells



Summary

Advantages

1. More realistic noise Hocking et al. (2013)
2. SNR is controlled with the proportion of tumor cells Staaf et al. (2008); Rasmussen et al. (2011)
3. Variety of simulated profiles Willenbrock and Fridlyand (2005)
4. True and false positive evaluation Hocking et al. (2013)

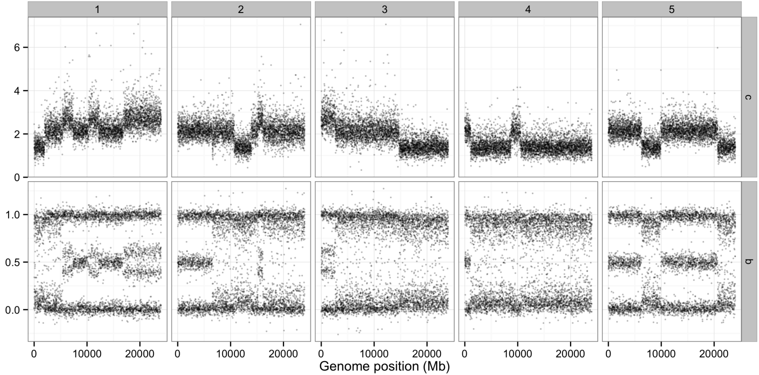
Application

1. Performance of segmentation methods
2. Evaluation of heterogeneity model

Characteristics

- 100 data sets simulated
- 30 tumor samples and 5 latent profiles based on realistic simulation framework
- Each matrix W is different for the 100 data sets

Simulated latent profiles

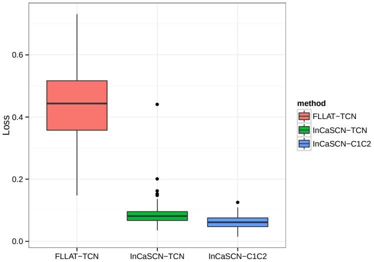


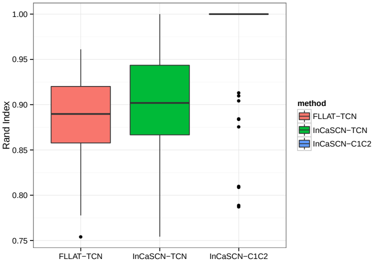
Performance evaluation

We compared performance of three methods :

- InCaSCN on parental copy number profiles
- InCaSCN on total copy number profiles
- FLLAT on total copy number profiles (Nowak et al., 2011)

Better estimation and interpretation of weights by using InCaSCN

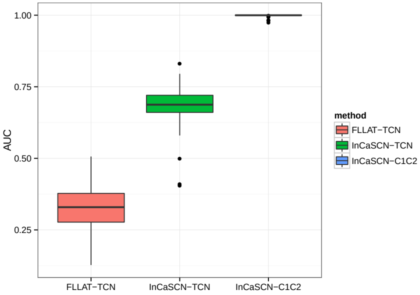




Inferred latent profiles from InCaSCN recover the true alterations

Evaluation

- Characterize each region as normal or altered for latent profiles
- AUC close to 1 : altered regions have been recovered with a few number of mistakes



Conclusion

- InCaSCN enables to recover both :
  - simulated latent profiles
  - weights with a small error
- Results on simulation are very promising for the application to real data sets.

Outline

1 Introduction

2 Segmentation

3 Heterogeneity Model

4 Simulations

5 Application to real data sets
- Inter-tumoral heterogeneity application
- Intra-tumoral heterogeneity application

6 Conclusion

Collaboration with Institut Curie

- Fabien Reyal's team (RT^2 : Residual Tumor and Response to Treatment)
- Triple-negative breast cancer (TNBC)
  - 16 patients
  - Micro-biopsy of the Primary Tumor at diagnosis
  - Neo-adjuvant chemotherapy before surgery
  - Primary Tumor size reduced but incomplete –> Residual
    - 10 patients with Primary Tumor and Residual samples
    - 6 patients with an additional metastasis Lymph Node sample
- Whole exome sequencing data
- RNAseq data

Collaboration with Institut Curie

- Fabien Reyal's team (RT^2 : Residual Tumor and Response to Treatment)
- Triple-negative breast cancer (TNBC)
  - 16 patients
  - Micro-biopsy of the Primary Tumor at diagnosis
  - Neo-adjuvant chemotherapy before surgery
  - Primary Tumor size reduced but incomplete –> Residual
    - 10 patients with Primary Tumor and Residual samples
    - 6 patients with an additional metastasis Lymph Node sample
- Whole exome sequencing data
- RNAseq data

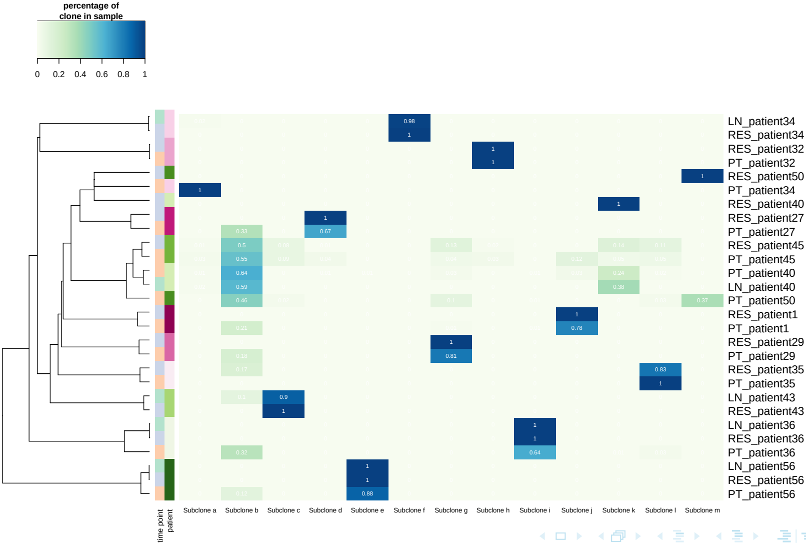
Collaboration with Institut Curie

- Fabien Reyal's team (RT^2 : Residual Tumor and Response to Treatment)
- Triple-negative breast cancer (TNBC)
    - 16 patients
    - Micro-biopsy of the Primary Tumor at diagnosis
    - Neo-adjuvant chemotherapy before surgery
    - Primary Tumor size reduced but incomplete –> Residual
        - 10 patients with Primary Tumor and Residual samples
        - 6 patients with an additional metastasis Lymph Node sample
- Whole exome sequencing data
- RNAseq data

Collaboration with Institut Curie

- Fabien Reyal's team (RT^2 : Residual Tumor and Response to Treatment)
- Triple-negative breast cancer (TNBC)
    - 16 patients
    - Micro-biopsy of the Primary Tumor at diagnosis
    - Neo-adjuvant chemotherapy before surgery
    - Primary Tumor size reduced but incomplete –> Residual
        - 10 patients with Primary Tumor and Residual samples
        - 6 patients with an additional metastasis Lymph Node sample
- Whole exome sequencing data
- RNAseq data

Results


percentage of
clone in sample
0
0.2
0.4
0.6
0.8
1
LN_patient34
RES_patient34
RES_patient32
PT_patient32
RES_patient50
PT_patient34
RES_patient40
RES_patient27
PT_patient27
RES_patient45
PT_patient45
PT_patient40
LN_patient40
PT_patient50
RES_patient1
PT_patient1
RES_patient29
PT_patient29
RES_patient35
PT_patient35
LN_patient43
RES_patient43
LN_patient36
RES_patient36
PT_patient36
LN_patient56
RES_patient56
PT_patient56
0.98
1
1
1
1
1
1
1
0.33
0.67
0.5
0.13
0.14
0.11
0.55
0.12
0.64
0.24
0.59
0.38
0.46
0.1
0.37
1
0.21
0.78
1
0.18
0.81
0.17
0.83
1
0.1
0.9
1
1
1
0.32
0.64
1
1
0.12
0.88
Subclone a
Subclone b
Subclone c
Subclone d
Subclone e
Subclone f
Subclone g
Subclone h
Subclone i
Subclone j
Subclone k
Subclone l
Subclone m
time point
patient

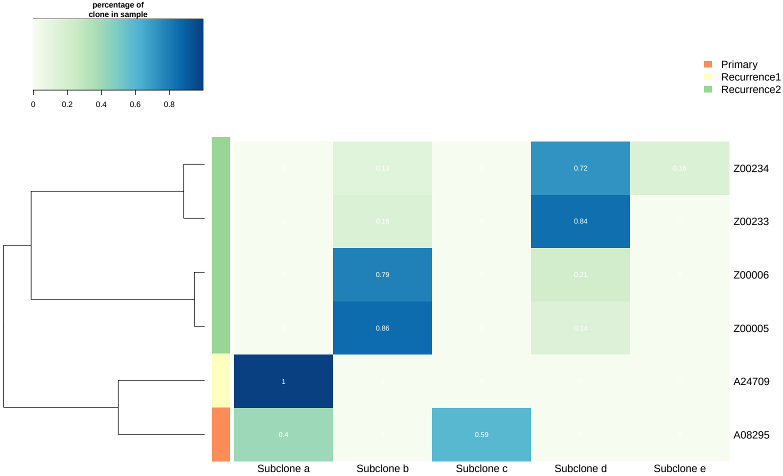
Conclusion on the application

- Only one latent profile (subclone B) common across the patients
- Patients are mainly grouped together
- For two patients (40 and 50), it seems that the resistant clone is already present in PT and becomes largely predominant in RES
- Same results from RNAseq analysis (B. Sadacca)

Collaboration with UCSF

- Henrik Bengtsson and Joe Costello
- Glioblastoma
  - 96 patients
  - Primary Tumor samples
  - Recurrence 1 with several samples
  - Sometimes Recurrence 2 with several samples
- Whole exome sequencing data
- Preprocessing with sequenza

Results


percentage of
clone in sample
0
0.2
0.4
0.6
0.8
Primary
Recurrence1
Recurrence2
0
0.13
0
0.72
0.16
Z00234
0
0.16
0
0.84
0
Z00233
0
0.79
0
0.21
0
Z00006
0
0.86
0
0.14
0
Z00005
1
0
0
0
0
A24709
0.4
0
0.59
0
0
A08295
Subclone a
Subclone b
Subclone c
Subclone d
Subclone e

Conclusions

Conclusions

- One resistant subclone already present in PT
- New cancer in Recurrence 2

Conclusions on the model

- Fast and efficient algorithm
- Application to other data sets
- Similar results than the model that uses mutations

Outline

1. Introduction
2. Segmentation
3. Heterogeneity Model
4. Simulations
5. Application to real data sets
6. Conclusion

Contributions

- Segmentation Methods
- Realistic simulation framework
- Performance of segmentation methods
- Heterogeneity
- Bioinformatic pipelines under several R packages
  - jointseg
  - acnr
  - InCaSCN

Contributions

- Segmentation Methods
- Realistic simulation framework
- Performance of segmentation methods
- Heterogeneity
- Bioinformatic pipelines under several R packages
  - jointseg
  - acnr
  - InCaSCN

Perspectives

- Exploring DNA copy number latent profiles
- Link to clinical outcomes
- Discover biomarkers
- Collaboration with UCSF

Thank you for your attention

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.

T. Hocking, G. Schleiermacher, I. Janoueix-Lerosey, V. Boeva, J. Cappo, O. Delattre, F. Bach, and J.-P. Vert. Learning smoothing models of copy number profiles using breakpoint annotations. *BMC Bioinformatics*, 14(1) :164, 2013.

Y. Jiang, Y. Qiu, A. J. Minn, and N. R. Zhang. Assessing intratumor heterogeneity and tracking longitudinal and spatial clonal evolutionary history by next-generation sequencing. *Proceedings of the National Academy of Sciences*, 113(37) :E5528–E5537, 2016. doi : 10.1073/pnas.1522203113. URL http://www.pnas.org/content/113/37/E5528.abstract.

S. Masecchia, S. Salzo, A. Barla, and A. Verri. A dictionary learning based method for acgh segmentation. In *Proceedings of the European Symposium on Artificial Neural Networks*, 2013.

G. Nowak, T. Hastie, J. R. Pollack, and R. Tibshirani. A fused lasso latent feature model for analyzing multi-sample acgh data. *Biostatistics*, page kxr012, 2011.

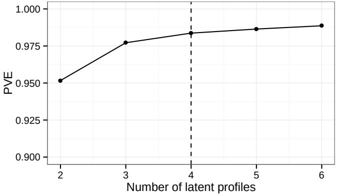
M. Rasmussen, M. Sundström, H. Göransson Kultima, J. Botling, and et al. Allele-specific copy number analysis of tumor samples with aneuploidy and tumor heterogeneity. *Genome Biol*, 12(10) :R108, Oct. 2011.

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, Oct. 2008.

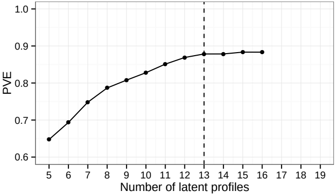
J.-P. Vert and K. Bleakley. Fast detection of multiple change-points shared by many signals using group LARS. *Advances in Neural Information Processing Systems*, 23 :2343–2351, 2010.

H. Willenbrock and J. Fridlyand. A comparison study : applying segmentation to array-CGH data for downstream analyses. *Bioinformatics*, 21(22) :4084–91, Nov 2005. doi : 10.1093/bioinformatics/bti677.

Selection of number of latent profiles



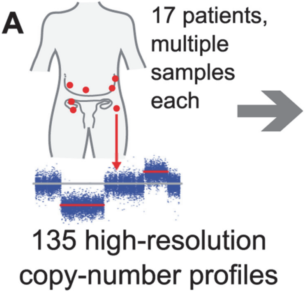
HSGOC



TNBC

Intra-tumoral heterogeneity

- Public data set
- High serious grade ovarian cancer (HSGOC)

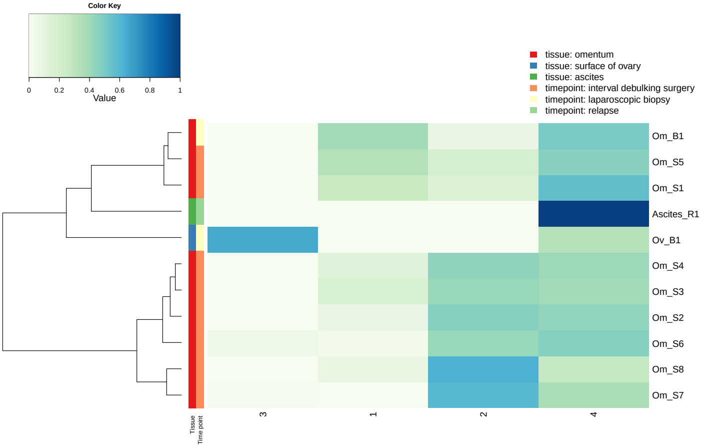

A
17 patients, multiple samples each
135 high-resolution copy-number profiles

- Quantify heterogeneity
- Reconstruct tumor evolution

Results

- We focused on one patient with 11 samples
  - Ovary (Biopsy)
  - Omentum
  - Ascites (relapse)
- We select a model with 4 latent profiles

Results : Weight matrix



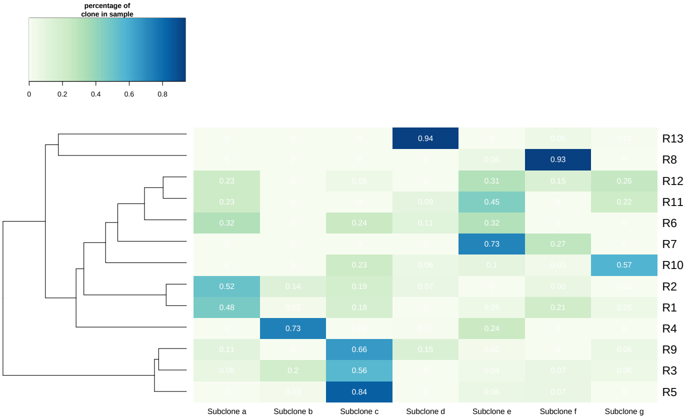
Conclusions an Perspectives

- One clone seems to be not resistant to the drug (latent profile 3)
- There may exist only one resistant clone to the drugs that led to a relapse (latent profile 4)
- exploring if there are not known genes that can be responsible for the resistance

Spatial Intra-tumoral heterogeneity

- Public data set
- Kidney cancer
- Several patients with several samples at various location.

Kidney cancer application



	Subclone a	Subclone b	Subclone c	Subclone d	Subclone e	Subclone f	Subclone g
R13	0	0	0	0.94	0	0.05	0.01
R8	0	0	0	0	0.06	0.93	0
R12	0.23	0	0.05	0	0.31	0.15	0.26
R11	0.23	0	0	0.09	0.45	0	0.22
R6	0.32	0	0.24	0.11	0.32	0	0
R7	0	0	0	0	0.73	0.27	0
R10	0	0	0.23	0.06	0.1	0.03	0.57
R2	0.52	0.14	0.19	0.07	0	0.08	0.01
R1	0.48	0.02	0.18	0	0.05	0.21	0.05
R4	0	0.73	0.01	0.02	0.24	0	0
R9	0.11	0	0.66	0.15	0.02	0	0.06
R3	0.08	0.2	0.56	0	0.04	0.07	0.06
R5	0	0.03	0.84	0	0.06	0.07	0

Sequencing information

- Illumina Hi-Seq 2500 pair-end aligned on hg19
- Depth : WEG : 100x
- bwa for alignement (soft clapping remove head and tail and map on the middle)
- reads sizes reads : 100 bases

Random Features

For a signal of length J.

Method	computation	Storage
Kernel	$\mathcal{O}(SJ^2)$	$\mathcal{O}(SJ)$
Approximation	$\mathcal{O}(p^2 J)$	$\mathcal{O}(SJ)$
Random Feature	$\mathcal{O}(SMJ)$	$\mathcal{O}(MJ)$