A statistical tour of genomic data

Franck Picard

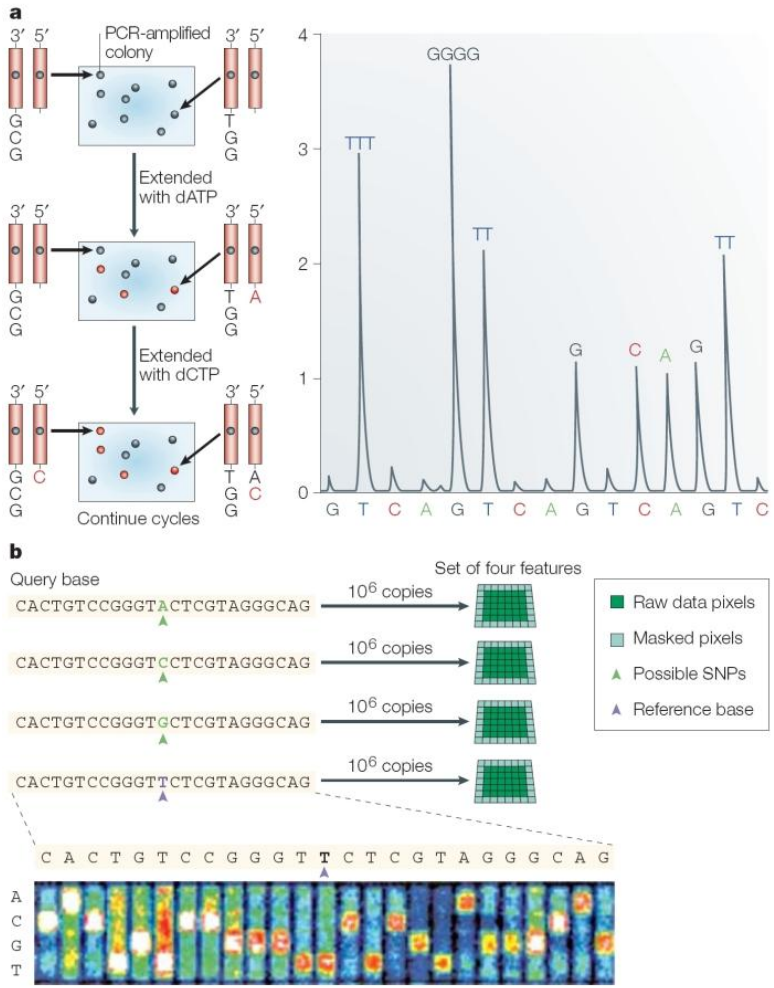
Laboratoire Biométrie et Biologie Evolutive, CNRS Univ. Lyon

Habilitation à diriger des recherches
December 4th 2014

Outline

Some definitions about high throughput genomics

- Based on automated techniques
- View molecular processes globally
- Gene expression and regulation, copy number variations
- Changed the scale of thinking of molecular biologists
- Complements the locus-specific or candidate approach
- -omics view of biological processes



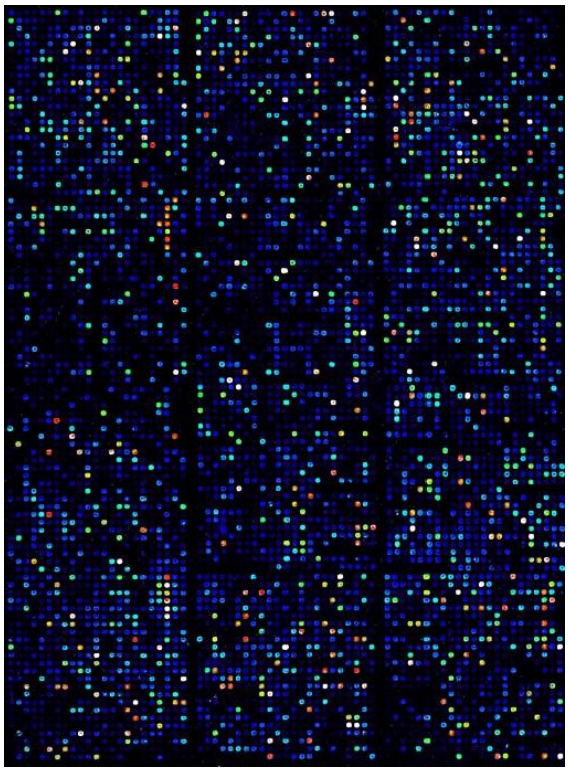
High throughput genomics and technological developments

- Deeply connected to the technological developments
- Two major technologies:

~1990 The microarray technology (DNA hybridization and fluorescence)

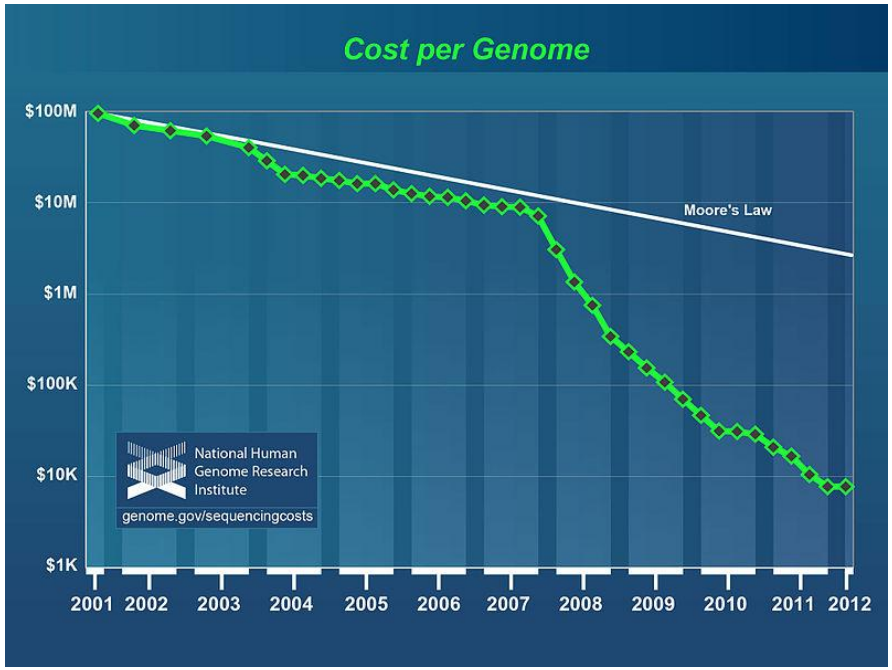
~2000 Next Generation Sequencing technologies (parallelization of short reads sequencing)

- Developments of storage capacities and computing power



Costs Drops and high availability of genomic data

- Cost drop of genome-scale data
- Access data:
  - with higher coverage
  - for more individuals
- Quantification of biological variability at the population genomics scale



Automated data production and automated analysis

- Statistics is only one piece of the Bioinformatics puzzle
- The purpose of statistical research applied to genomics:
  1. extract meaningful information by developping methods (if needed !)
  2. share statistical concepts (multiple testing, overfitting)
- Statistics is also an elvolving science
- The challenge lies in the sharing of modern statistical concepts for modern Genomics

Research Directions

- My personal direction has been to develop statistical models that were accounting for particular data structures
  - Functional models for data that are spatially organized (1D) along the genome
  - Random graph models for data that are in the form of networks
- What are the specific statistical questions that are raised by these structures ?

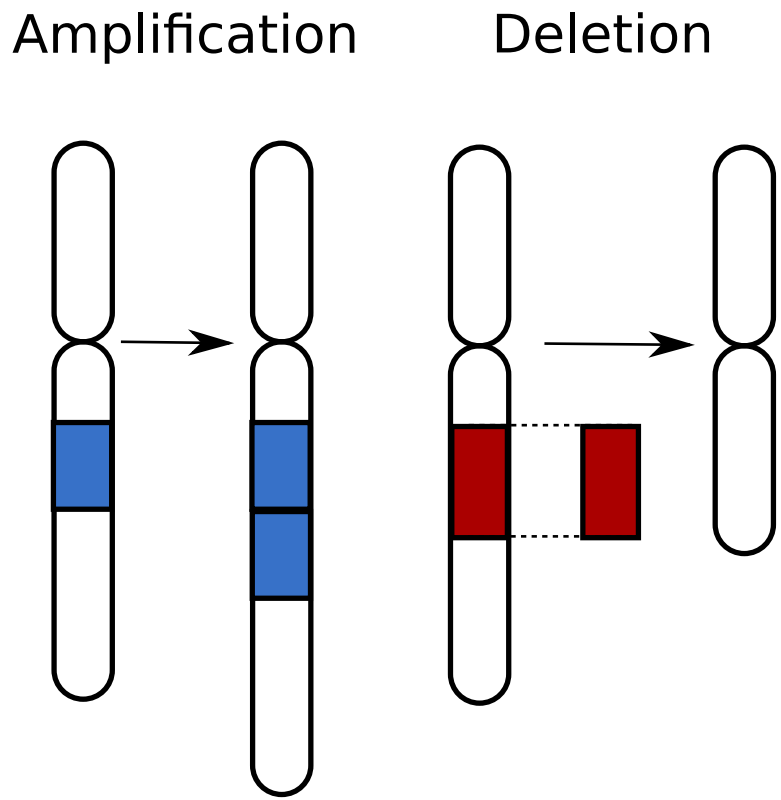
Organization of the presentation

1. Segmentation models for multisample copy number variations
   - computational issues
2. Functional Regression models in the Gaussian case
   - curve clustering
   - functional mixed models
   - dimension reduction
3. Functional Regression models in the Poisson case
   - dimension reduction and calibration

Outline

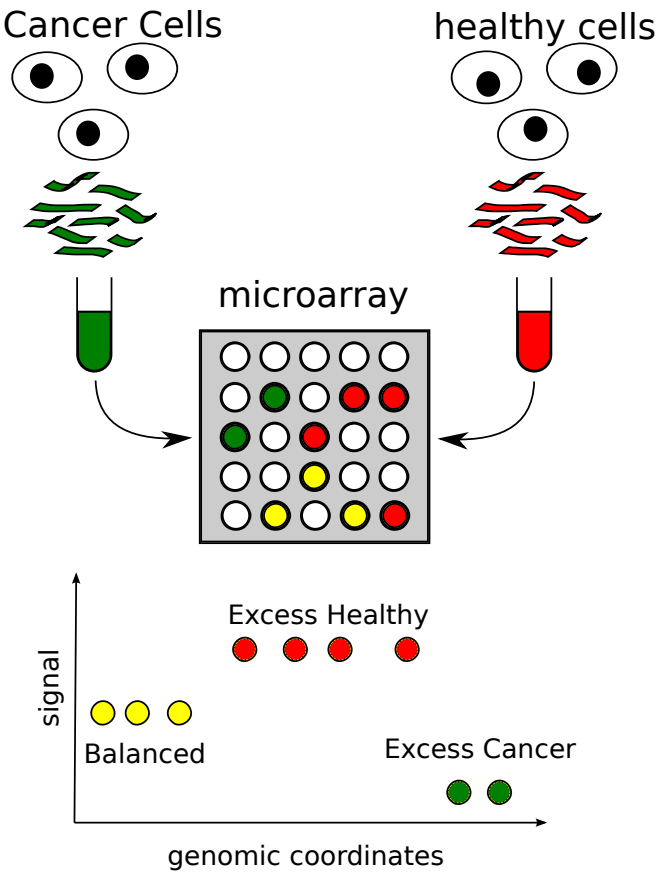
Copy Number variations and mapped genomic data

- Gene copy number is highly regulated (humans are diploid)
- Additional or deleted chromosomes cause syndromes
- The challenge : detect sub-chromosomal aberrations
- Change the resolution (from MegaBases to KiloBases)



Array-based Comparative Genomic Hybridization (CGH)

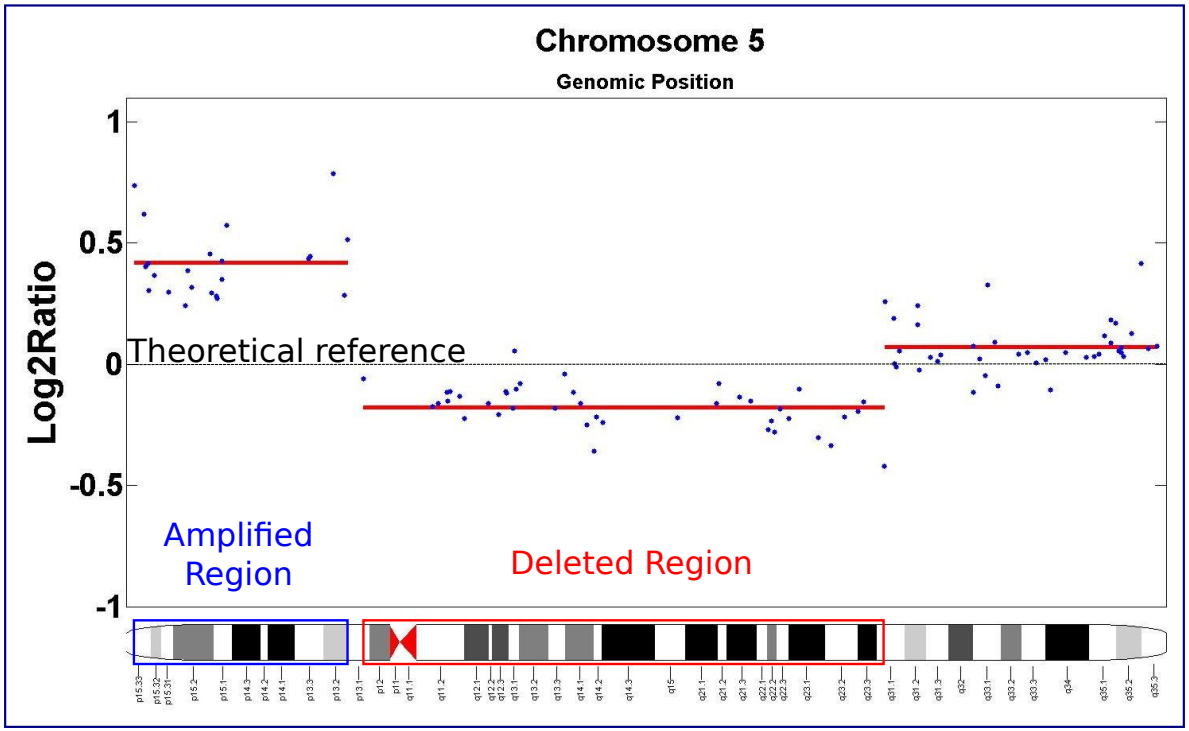
- Compare two genomes by hybridization (healthy / cancer)
- Use glass-fixed probes with mapped coordinates
- Measure relative DNA quantities at different loci on the genome
- genome-wide blind search for few kbs aberrations



Genome wide investigation of copy number variations in populations

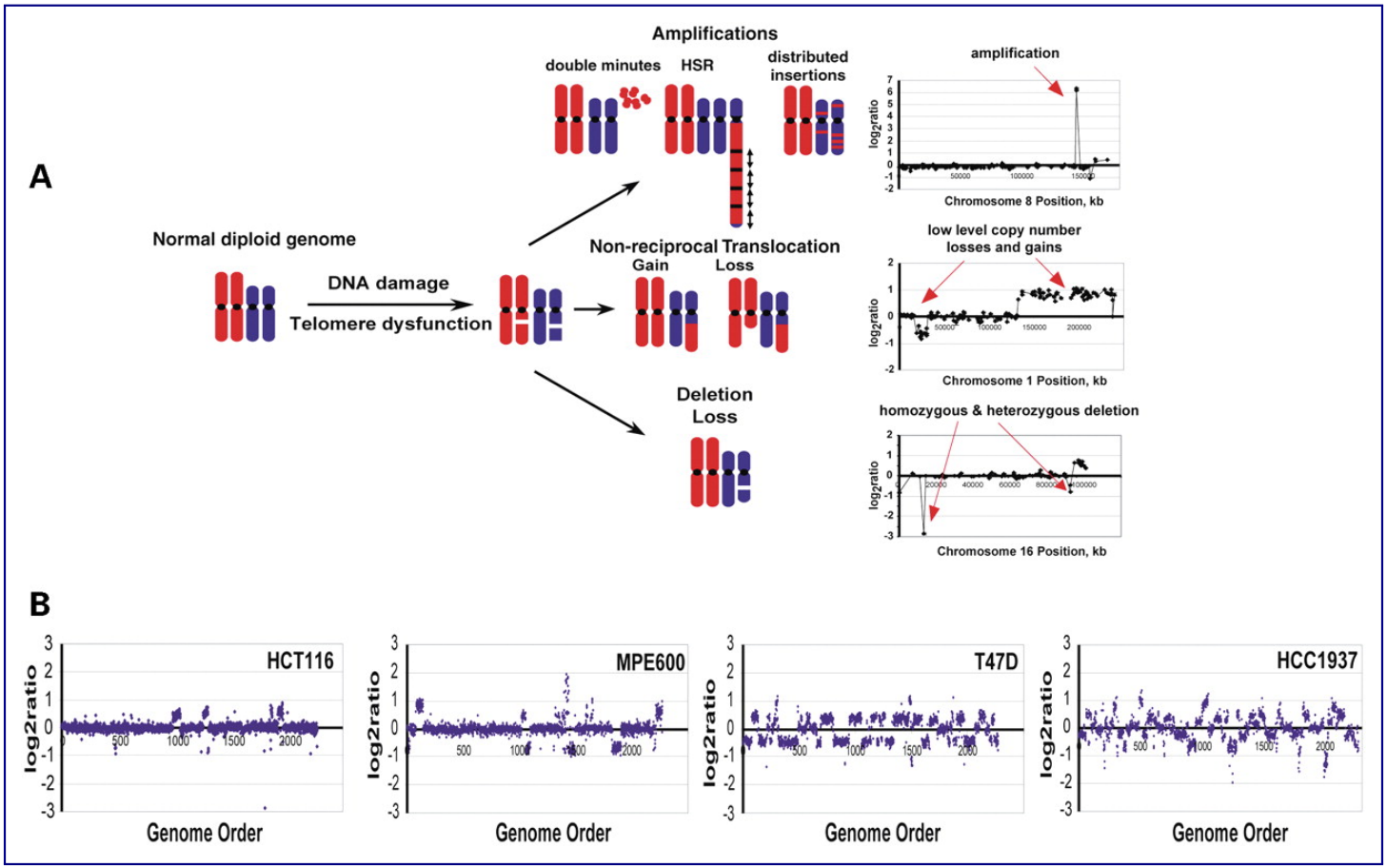
How to read CGH profiles

- Data are mapped on their genomic coordinates
- The signal Y_t is a $\log_2$ ratio of fluorescence mapped onto the genome



The signal is expected to be piece-wise constant

Tracking Genomic Aberrations in Cancer Genomes



Segmentation models: definitions and notations

- We consider the regression model:

$$Y_t = \mu(t) + E_t$$

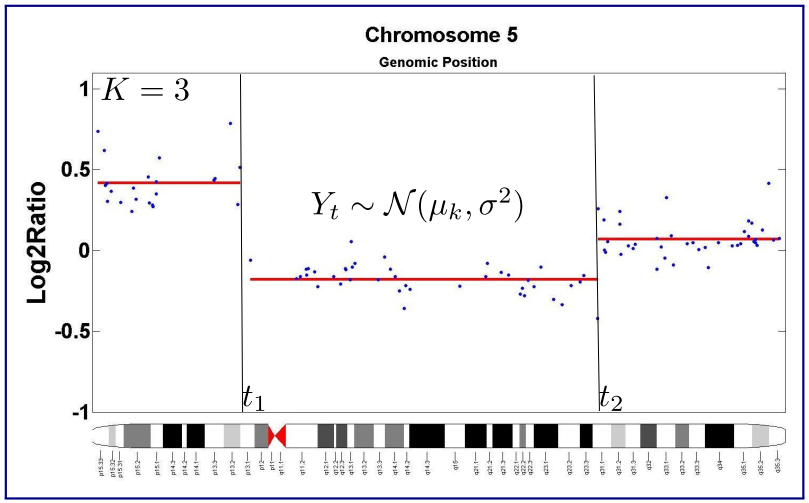
- Suppose there exists $K+1$ change-points $t_0 < \ldots < t_K$ such that μ is constant between two changes and different from a change to another.
- $\mathcal{I}_k =]t_{k-1}, t_k]$: interval of stationarity, μ_k the mean of the signal between two changes:

$$\forall t \in \mathcal{I}_k, \quad Y_t = \mu_k + E_t, \quad E_t \sim \mathcal{N}(0, \sigma^2).$$

Estimation framework

- The parameters:
  $\mathbf{T} = \{t_0, \ldots, t_K\}$,
  $\boldsymbol{\mu} = \{\mu_1, \ldots, \mu_K\}$ and σ^2.
- K is fixed in a first step
- When K and $\mathbf{T}$ are known the MS estimator of $\boldsymbol{\mu}$ is:

$$\widehat{\mu}_k = \frac{1}{t_k - t_{k-1}} \sum_{t=t_{k-1}+1}^{t_k} Y_t,$$



Main challenge : estimation of $\mathbf{T}$

$$\widehat{\mathbf{T}} = \arg\min_{\mathbf{T}} \{\mathrm{RSS}_K(\mathbf{T}, \boldsymbol{\mu})\}.$$

$$\mathrm{RSS}_K(\mathbf{T}, \boldsymbol{\mu}) = \sum_{k=1}^{K} \sum_{t=t_{k-1}+1}^{t_k} (Y_t - \widehat{\mu}_k)^2$$

A computational challenge to find the breaks

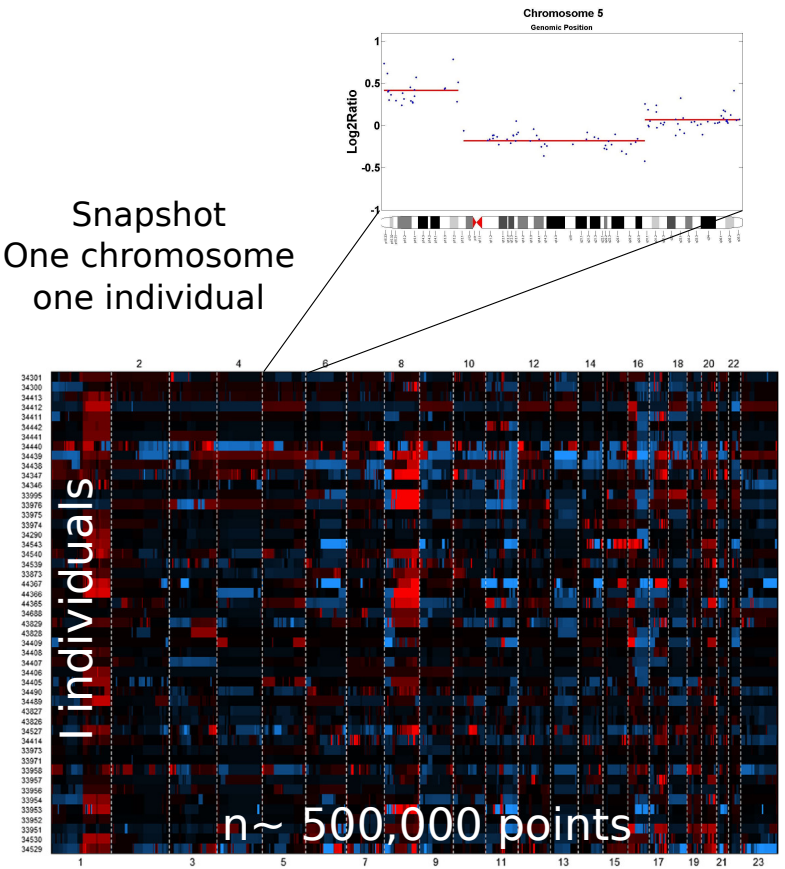
- Partition n data points into K segments: complexity $\mathcal{O}(n^K)$.
- Dynamic Programming reduces the complexity to $\mathcal{O}(Kn^2)$
- Application of the shortest path algorithm
- $\text{RSS}_k(i,j)$ cost of the path connecting i to j in k segments:

$$\begin{aligned} \forall 0 \leq i < j \leq n, \;\; \text{RSS}_1(i,j) &= \sum_{t=i+1}^{j} (y_t - \bar{y}_{ij})^2, \\ \forall 1 \leq k \leq K-1, \;\; \text{RSS}_{k+1}(1,j) &= \min_{1 \leq h \leq j} \left\{ \text{RSS}_k(1,h) + \text{RSS}_1(h+1,j) \right\}. \end{aligned}$$

- Provides the optimal position of breakpoints $\widehat{\mathbf{T}}$

MultiSample Segmentation

- Copy number variations are now studied at the population level
- The challenge is to account for genomic order and for the size of the dataset.
- The resolution as increased and signals are $n \sim 500,000$ long
- Joint segmentation allowing patient-specific breaks while sharing common noise and biases.



Joint segmentation model for multiple samples*

- $Y_i(t)$: the signal for individual $i = 1, ...I$ with segments $\{\mathcal{I}_k^i\}$

$$\forall t \in \mathcal{I}_k^i,\ Y_i(t) = \mu_i(t) + \varepsilon_i(t),\ \varepsilon_i(t) \sim \mathcal{N}(0, \sigma^2).$$

- $\mu_i(t)$ is also piece-wise constant with k_i segments ($\sum_i k_i = K$)
- $\boldsymbol{\mu}_i = [\mu_{1i}, \ldots, \mu_{k_i i}]$ is the mean signal for patient i.
- $\mathbf{T}_i$ specific (unknown) incidence matrix of the breaks

$$\mathbf{Y}_i = \mathbf{T}_i \boldsymbol{\mu}_i + \mathbf{E}_i$$

- If common biases are shared:

$$\mathbf{Y}_i = \mathbf{T}_i \boldsymbol{\mu}_i + \mathbf{X}_i \boldsymbol{\beta} + \mathbf{E}_i$$

*Picard et al. (2011a, 2011b), Biostatistics, CSDA

Segmentation of multiple arrays

- The RSS is additive wrt the individuals and to the number of segments.

$$\mathrm{RSS}_K(\mathbf{T}, \boldsymbol{\mu}) = \sum_{i=1}^{I} \sum_{k=1}^{k_i} \mathrm{RSS}_k^i(\mathbf{T}_i, \boldsymbol{\mu}_i)$$

- Global Dyn. Prog. $\mathcal{O}(Kn^2I^2)$ complexity.
- But there is a constraint : $\sum_i k_i = K$, (K fixed) thus:

$$\min_{\{\mathbf{T},\boldsymbol{\mu}\}} \mathrm{RSS}_K(\mathbf{T}, \boldsymbol{\mu}) \quad = \quad \min_{k_1+\ldots+k_I=K} \left\{ \sum_{i=1}^{I} \min_{\mathbf{T}_i,\boldsymbol{\mu}_i} \mathrm{RSS}_{k_i}^i(\mathbf{T}_i, \boldsymbol{\mu}_i) \right\}.$$

Optimal joint segmentation*

- We propose a two-step Dynamic Programming for joint segmentation
- Optimal segmentation for each individual and best segments allocation
- Provide `cghseg` R package

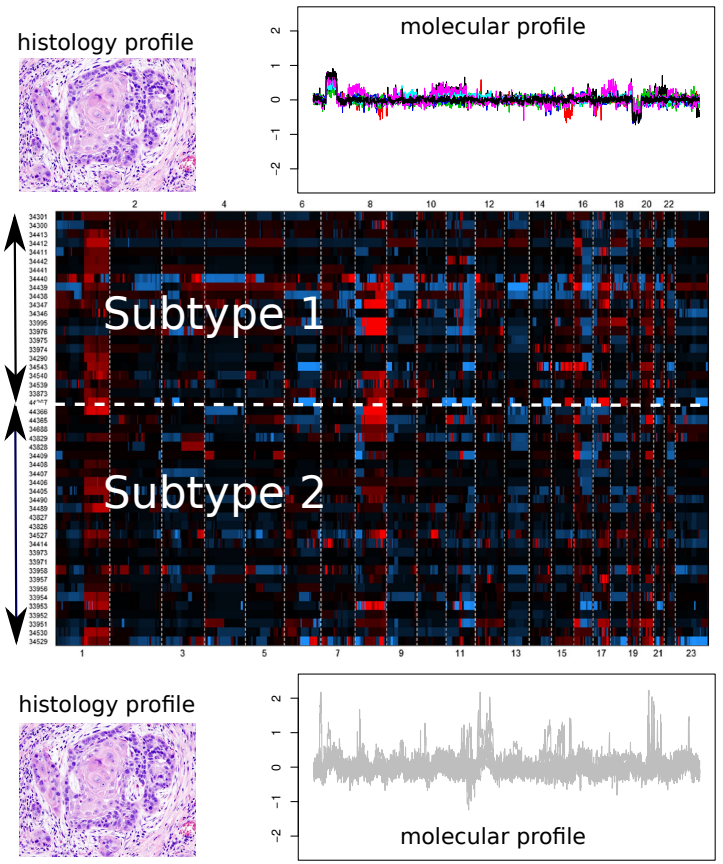
n (observations/profile)	$20,000$			$100,000$		
I (number of profiles)	256	512	1024	256	512	1024
Average CPU time (min)	6	15	54	31	70	253
Memory usage (Gb)	0.4	0.8	1.8	1.7	3.7	7.9

*Rigaill et al. (2014) Comp. Int. Meth. for Bioinfo. and Biostat.

Outline

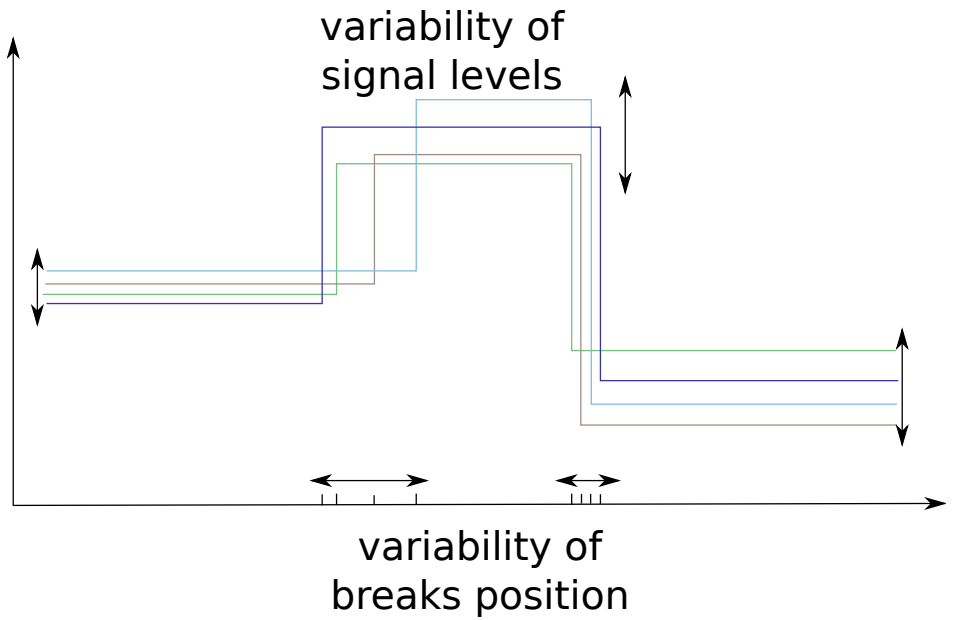
Molecular classification of diseases

- Link molecular features to patient outcome
- Task: clustering (subgroups discovery)
- Integrate the genomic organization of the data



Accounting for inter-individual variability

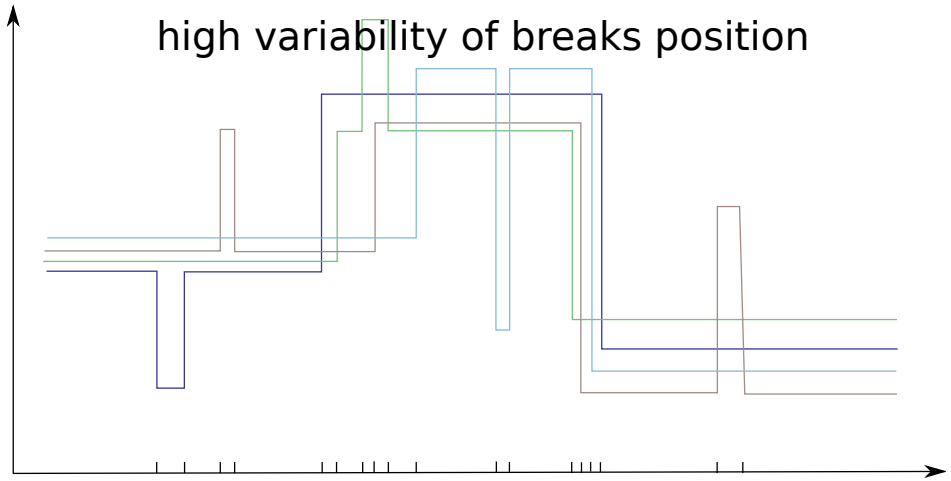
- Variability of sampled individuals
- Signal levels and breaks position are (highly) variable
- Focus on the shape of a shared profile + variations



Functional-based clustering models

Accounting for inter-individual variability

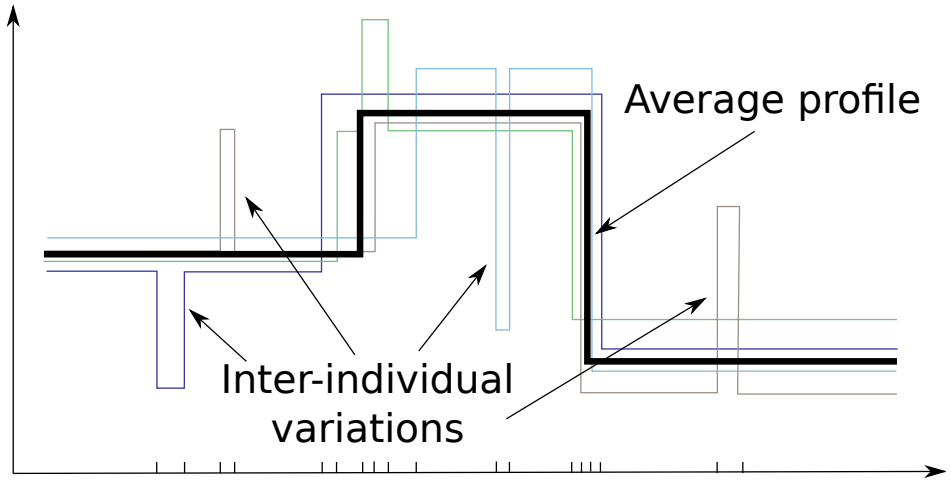
- Variability of sampled individuals
- Signal levels and breaks position are (highly) variable
- Focus on the shape of a shared profile + variations



Functional-based clustering models

Accounting for inter-individual variability

- Variability of sampled individuals
- Signal levels and breaks position are (highly) variable
- Focus on the shape of a shared profile + variations



Functional-based clustering models

Functional clustering model

- Cluster individuals on L unknown clusters based on functional observations
- Suppose there exists hidden label variables

$$\zeta_{i\ell} \sim \mathcal{M}(1, \pi_1, \ldots, \pi_L)$$

- The mixture model becomes (given $\{\zeta_{i\ell} = 1\}$):

$$Y_i(t) \quad = \quad \mu_\ell(t) + E_i(t)$$

- μ_ℓ are approximated by wavelets
  - Modelling curves with irregularities
  - Computationaly efficiency
  - Dimension Reduction

Definition of wavelets and wavelet coefficients

- Wavelets provide an orthonormal basis of $L^2_{[0,1]}$ with a scaling function ϕ and a mother wavelet ψ such that:

$$\left\{\phi_{j_0 k}(t), k = 0, \ldots, 2^{j_0} - 1; \psi_{jk}(t), j \geq j_0, k = 0, \ldots, 2^j - 1\right\}$$

- Any function $Y \in L^2_{[0,1]}$ is then expressed in the form:

$$Y_i(t) = \sum_{k=0}^{2^{j_0}-1} c^*_{i,j_0 k} \phi_{j_0 k}(t) + \sum_{j \geq j_0} \sum_{k=0}^{2^j - 1} d^*_{i,jk} \psi_{jk}(t)$$

where $c^*_{i,j_0 k} = \langle Y_i, \phi_{j_0 k} \rangle$ and $d^*_{i,jk} = \langle Y_i, \psi_{jk} \rangle$ are the theorical scaling and wavelet coefficients.

The DWT and empirical wavelet coefficients

- Denote by $\mathbf{W}$ an orthogonal matrix of filters (wavelet specific),
- The Discrete Wavelet Transform is given by

$$\underset{[M\times M]}{\mathbf{W}} \underset{[M\times 1]}{\mathbf{Y}_i(\mathbf{t})} = \begin{bmatrix} \mathbf{c}_i \\ \mathbf{d}_i \end{bmatrix}$$

- $(\mathbf{c}_i, \mathbf{d}_i)$ are empirical scaling and wavelet coefficients
- Linear model in the coefficients domain (given $\{\zeta_{i\ell} = 1\}$):

$$\begin{aligned} \mathbf{W}\mathbf{Y}_i(\mathbf{t}) &= \mathbf{W}\boldsymbol{\mu}_\ell(\mathbf{t}) + \mathbf{W}\mathbf{E}_i(\mathbf{t}) \\ \begin{bmatrix} \mathbf{c}_i \\ \mathbf{d}_i \end{bmatrix} &= \begin{bmatrix} \boldsymbol{\alpha}_\ell \\ \boldsymbol{\beta}_\ell \end{bmatrix} + \boldsymbol{\varepsilon}_i,\ \boldsymbol{\varepsilon}_i \sim \mathcal{N}(\mathbf{0}_M, \sigma_\varepsilon^2 \mathbf{I}_M) \end{aligned}$$

Functional Clustering Mixed Models*

- Introduce inter-individual functional variability (given $\{\zeta_{i\ell} = 1\}$):

$$Y_i(t) = \mu_\ell(t) + U_i(t) + E_i(t),\ U_i(t) \perp E_i(t)$$

- U_i is a Gaussian stochastic process with kernel $K_\ell(\bullet, t)$, that models individual-specific changes
- In the coefficients domain, and given $\{\zeta_{i\ell} = 1\}$:

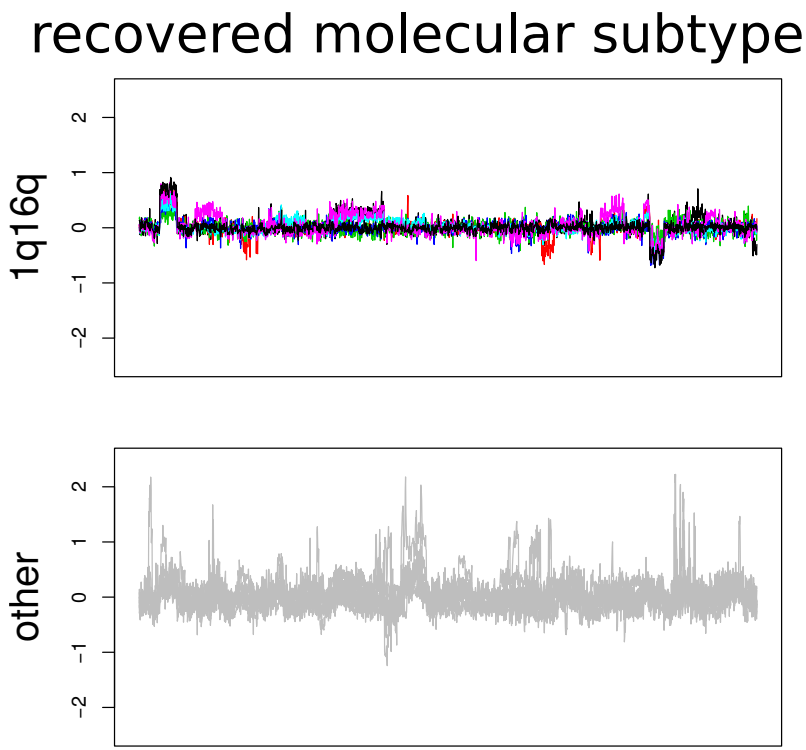
$$\begin{bmatrix} \mathbf{c}_i \\ \mathbf{d}_i \end{bmatrix} = \begin{bmatrix} \boldsymbol{\alpha}_\ell \\ \boldsymbol{\beta}_\ell \end{bmatrix} + \begin{bmatrix} \boldsymbol{\nu}_i \\ \boldsymbol{\theta}_i \end{bmatrix} + \boldsymbol{\varepsilon}_i,$$

$$\begin{bmatrix} \boldsymbol{\nu}_i \\ \boldsymbol{\theta}_i \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}_M, \begin{bmatrix} \mathbf{G}_\nu & 0 \\ 0 & \mathbf{G}_\theta \end{bmatrix}\right)$$

*Giacofci et al. (2013) Biometrics

Application to array CGH data

- We applied this model on breast tumors data
- We retrieve biologically meaningful clusters
- First estimations of the inter-individual variability
- We provide a `R` package for curve-clustering
- PhD of M. Giacofci (co-Adv. S. Lambert-Lacroix)



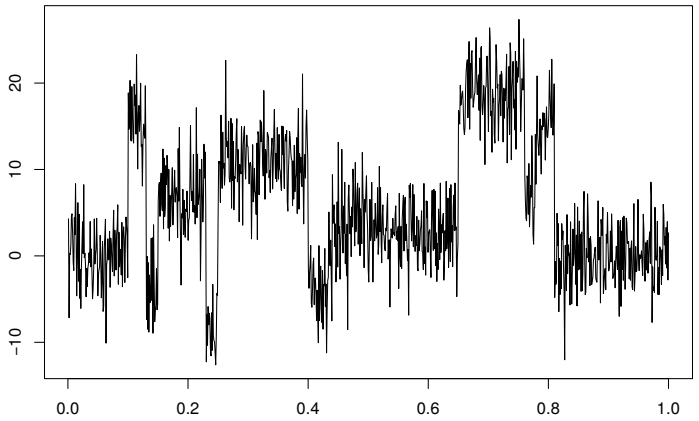
Outline

Dimension Reduction

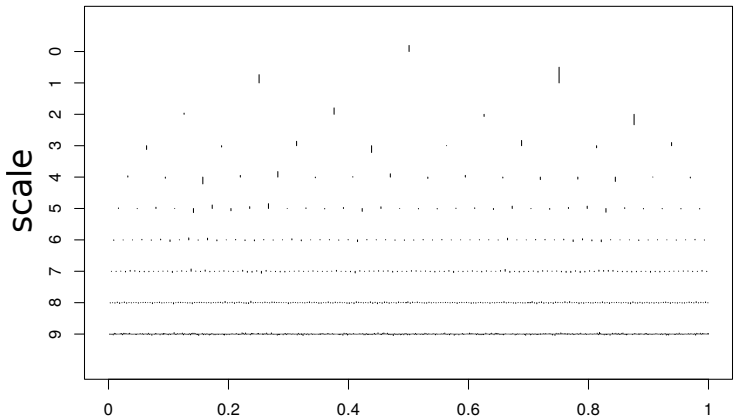
- One strength of wavelets is their compression property
- Few wavelet coefficients are necessary to summarize the signal (spatial adaptivity)
- The model is sparse in the coefficients domain

How to perform simultaneous clustering and thresholding (with random effects)?

$$Y(t) = \mu(t) + E(t)$$



$$\beta_{jk}$$

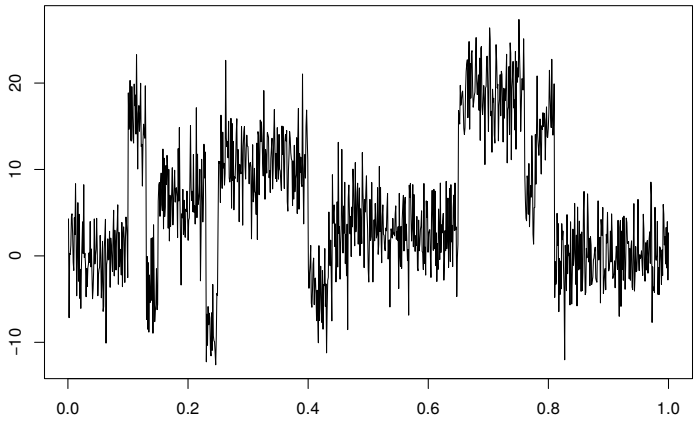


Dimension Reduction

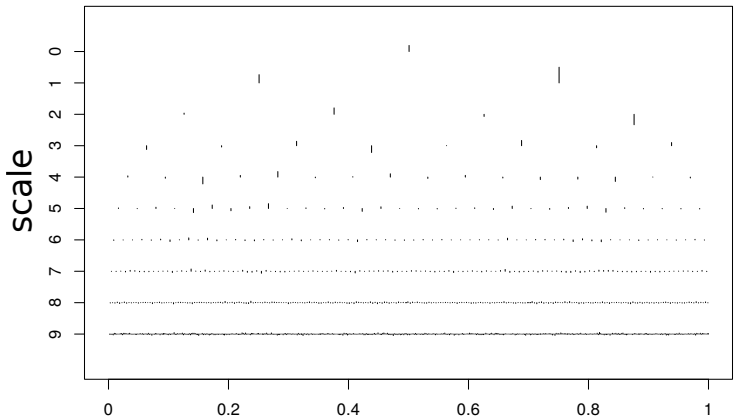
- One strength of wavelets is their compression property
- Few wavelet coefficients are necessary to summarize the signal (spatial adaptivity)
- The model is sparse in the coefficients domain

How to perform simultaneous clustering and thresholding (with random effects)?

$$Y(t) = \mu(t) + E(t)$$



$$\beta_{jk}$$



Dimension Reduction and for curve clustering

- Linear model in the coefficients domain (identity design matrix)

$$\mathbf{d} = \boldsymbol{\beta}_\ell + \boldsymbol{\varepsilon}$$

- Restate the problem as a variable selection problem
- Equivalence between the lasso and soft thresholding (orthogonal design):

$$\widehat{\boldsymbol{\beta}}_\ell(\lambda) \in \arg\min_{\boldsymbol{\beta}_\ell} \left\{ \sum_i \zeta_{i\ell} \left\| \mathbf{d}_i - \boldsymbol{\beta}_\ell \right\|^2 + \lambda \|\boldsymbol{\beta}_\ell\|_1 \right\}$$

- Estimated coefficients are of the form:

$$\widehat{\beta}^\ell_{jk}(\lambda) = \text{sign}(d^\ell_{\bullet,jk}) \left(|d^\ell_{\bullet,jk}| - \lambda \right)_+$$

Dimension Reduction for functional mixed models

- Focus on the functional mixed model

$$Y_i(t) = \mu(t) + U_i(t) + E_i(t), \;\; U_i(t) \sim \mathcal{N}\left(0, K(\bullet, t)\right)$$

- In the coefficients domain:

$$\mathbf{d}_i = \boldsymbol{\beta} + \boldsymbol{\theta}_i + \boldsymbol{\varepsilon}_i, \;\; \boldsymbol{\theta}_i \sim \mathcal{N}(0, \mathbf{G}_\theta)$$

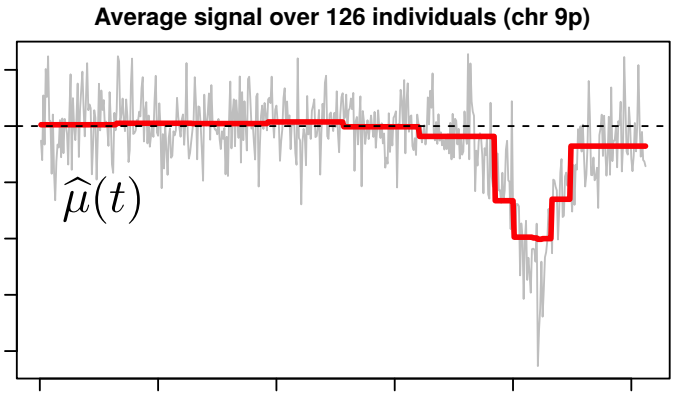
- First theoretical results on the reconstruction properties of $\widehat{\mu}$ *
- What about the characterization of $U_i(t)$?

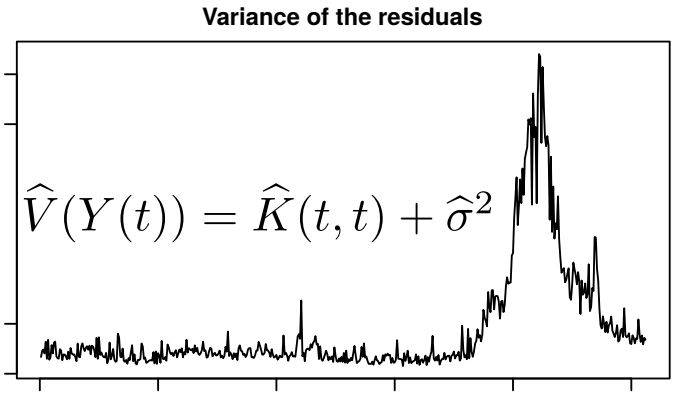
*Giacofci et al. *in prep*

Structure of the random effects variance

- U_i is characterized by $K(\bullet, t)$
- The inter-individual variance shows spatial heterogeneity
- Sparsity assumption for $\mathbf{G}_\theta$
- Assumption that $\mathbf{G}_\theta$ is diagonal:

$$\mathbf{G}_\theta = \underset{jk}{\text{diag}}\left(2^{-j\eta}\gamma_{jk}^2\right)$$





Joint selection of fixed and variance coefficients

- Mixed Linear model in high dimension:

$$\left(\widehat{\boldsymbol{\beta}}(\lambda_\beta), \widehat{\boldsymbol{\gamma}}(\lambda_\gamma)\right) \in \arg\min_{\boldsymbol{\beta},\boldsymbol{\gamma}} \left\{ -\log \mathcal{L}(\mathbf{d}; \boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma^2) + \lambda_\beta \|\boldsymbol{\beta}\|_1 + \lambda_\gamma \|\boldsymbol{\gamma}\|_1 \right\}$$

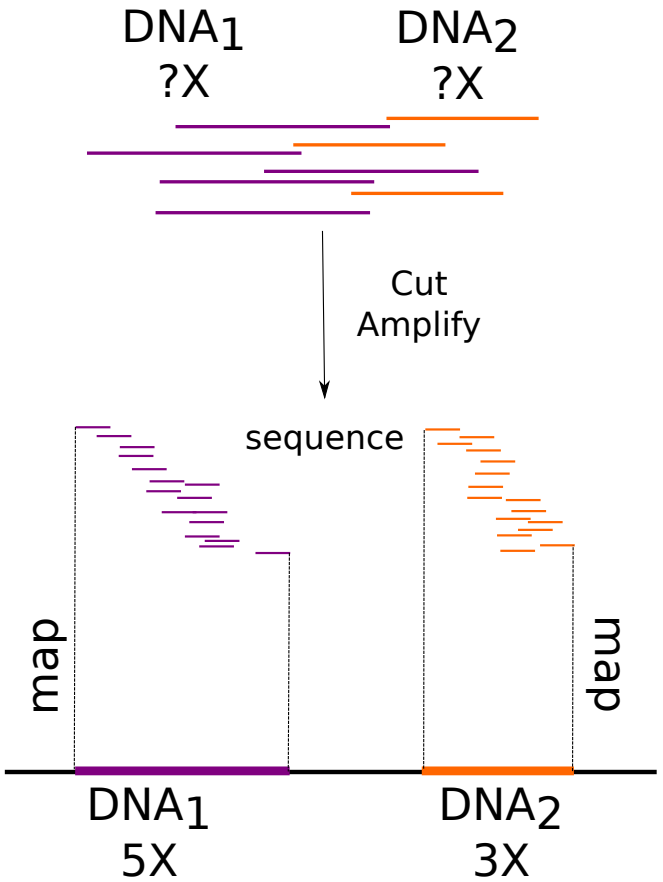
- Calibration issues (λ_β, λ_γ, regularity of U_i)
- First oracle properties for $\widehat{\boldsymbol{\beta}}$, and $\widehat{\boldsymbol{\gamma}}$
- What are the reconstruction properties of the predicted random effects ?

$$\widehat{U}_i(t) = \mathbb{E}\left(U_i(t) | Y_i(t)\right)$$

Outline

Next Generation Sequencing Data

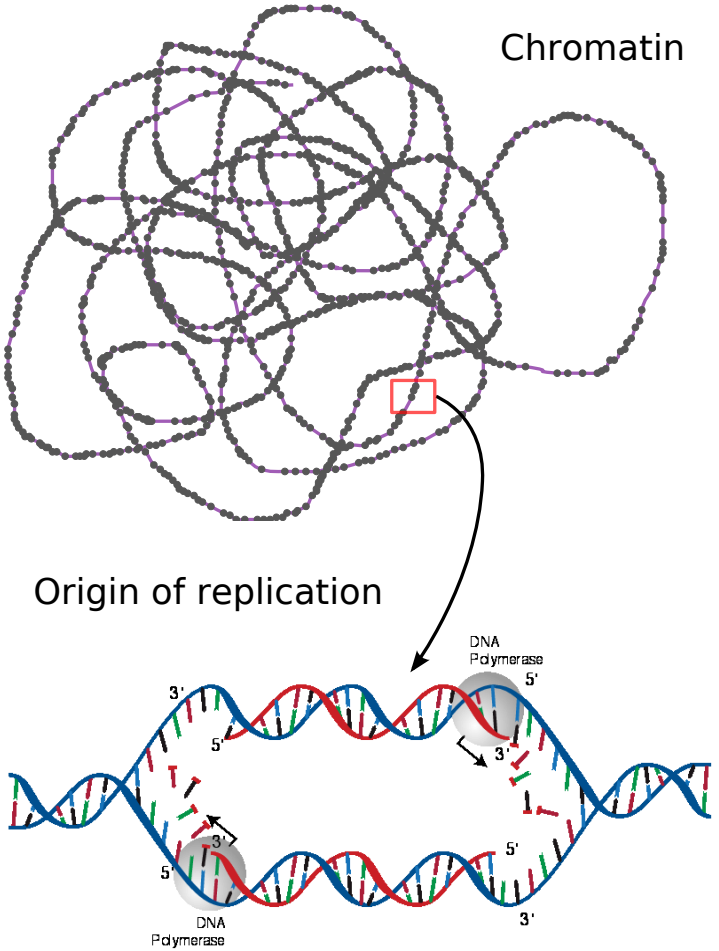
- Massive parallel sequencing of DNA molecules
- Can be used to quantify DNA in a sample
- Expression, copy numbers, DNA-prot. interactions
- Focus on mapped data



Outline of the OriSeq project

- DNA replication: duplication of 1 molecule into 2 daughter molecules
- The exact duplication of mammalian genomes is strongly controlled
- Spatial control (loci choice)
- Temporal control (firing timing)

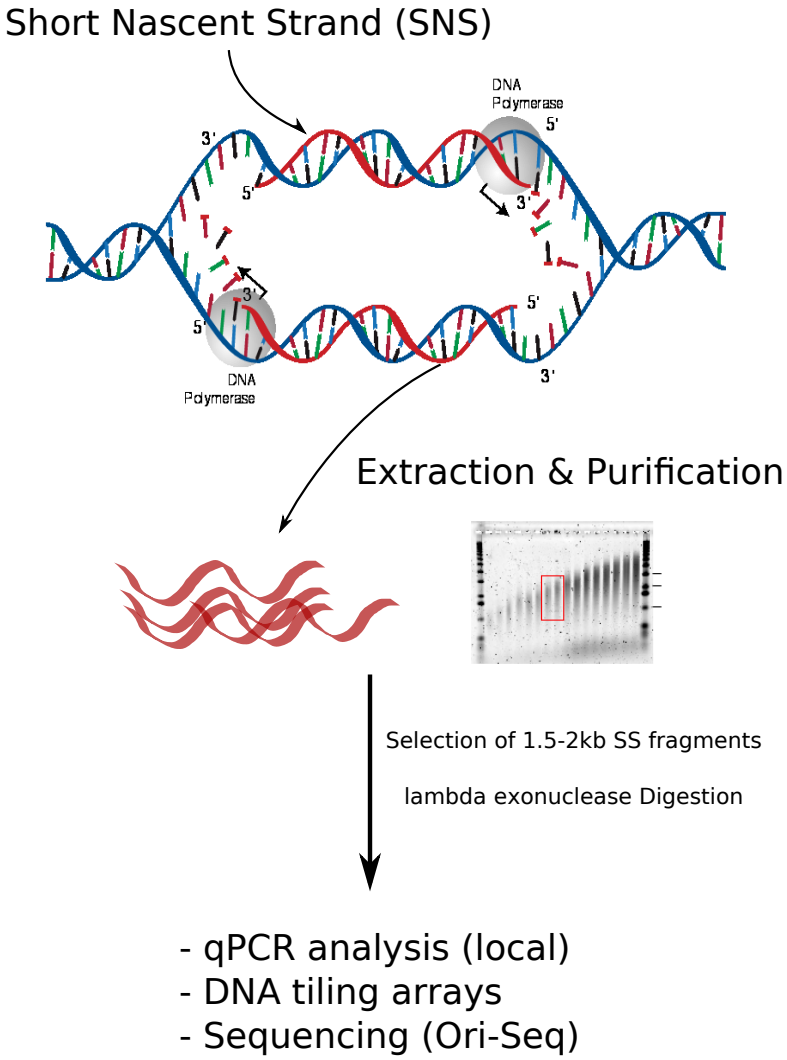
$\Rightarrow$ What are the (epi)genetic determinants of these controls ?



Mapping human replication origins: a technical challenge

- "bubbles" are small and instable (last only minutes by cycle)
- no clear consensus sequence (like in S. Cerevisiae)
- their specification is associated with both DNA sequence and chromatin structure

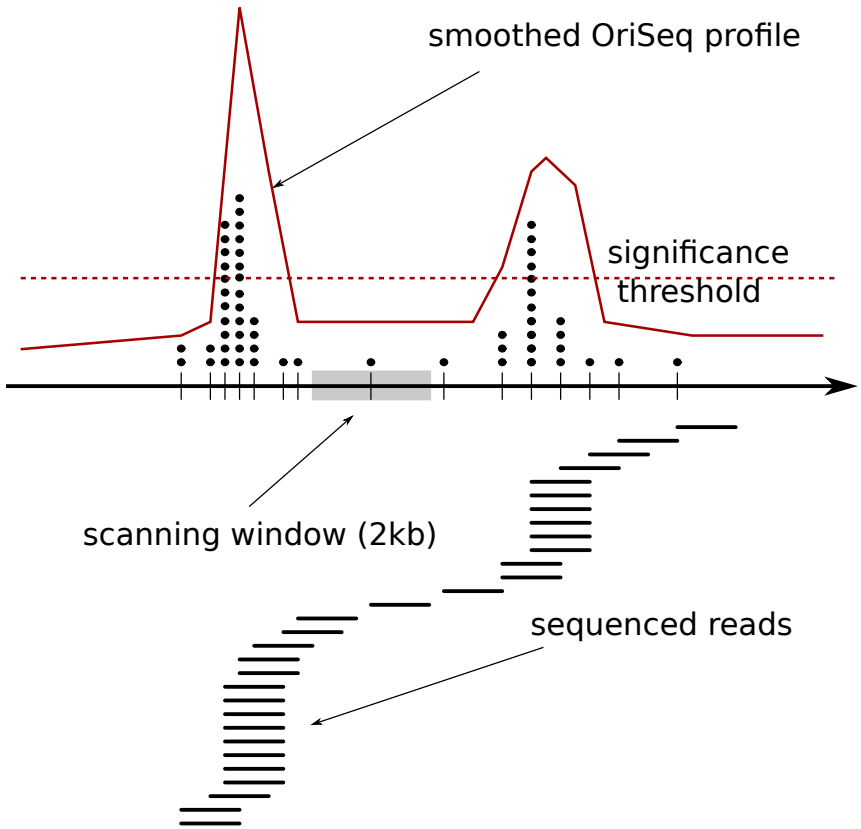
$\Rightarrow$ *Origin*-Omics: SNS Sequencing



Sliding windows to detect Replication origins*

- Model reads occurrences and accumulation by Compound Poisson model
- Detect significant enrichment by sliding windows
- Calibrate the threshold on regions using a FWER control :

$$\mathbb{P}\left(\max_t \{S_h(t)\} \geq x\right) \leq \alpha$$

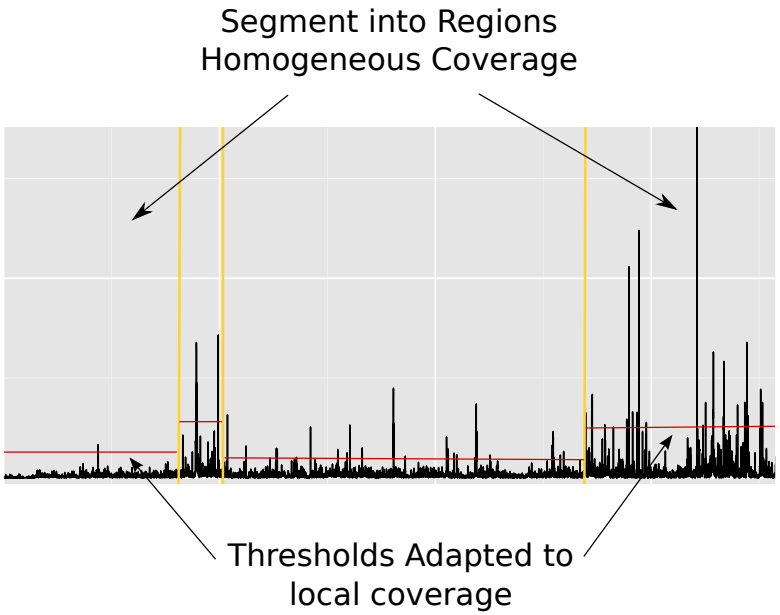


*Picard et al. (2014) PloS Genet.

Sliding windows to detect Replication origins*

- Model reads occurrences and accumulation by Compound Poisson model
- Detect significant enrichment by sliding windows
- Calibrate the threshold on regions using a FWER control :

$$\mathbb{P}\left(\max_t \{S_h(t)\} \geq x\right) \leq \alpha$$



*Picard et al. (2014) PloS Genet.

Back to functional models

- Model sequencing data by functional Poisson regression

$$Y_t | X_t \sim \mathcal{P}\left(f(X_t)\right),$$

- Consider a functional dictionary with p elements $\{\varphi_1, \ldots, \varphi_p\}$:

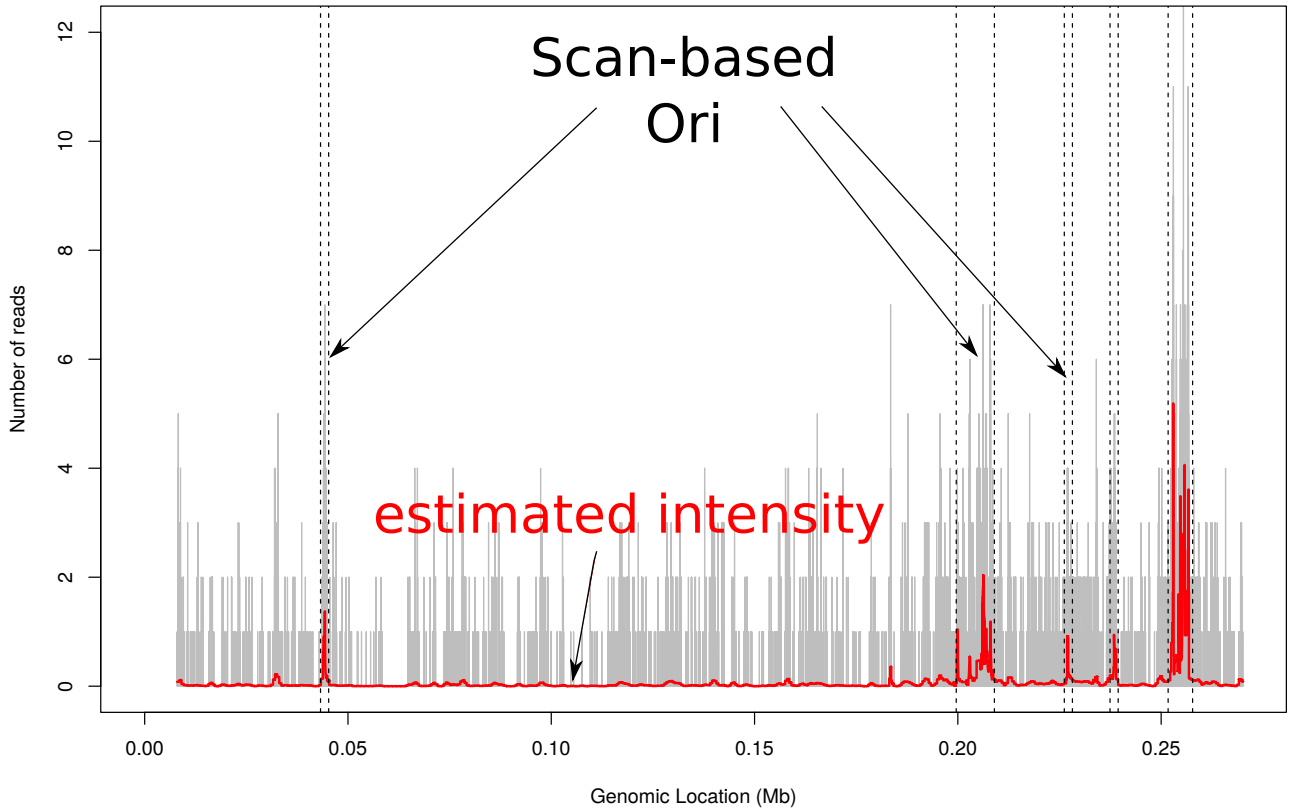
$$\log f(x) = \sum_{j=1}^{p} \beta_j \varphi_j(x)$$

- Selection can be performed by the lasso such that:

$$\widehat{\boldsymbol{\beta}} \in \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\arg\min} \left\{ -\log \mathcal{L}(\boldsymbol{\beta}) + \sum_{j=1}^{p} \lambda_j |\beta_j| \right\}.$$

Promising results on OriSeq*

- $(\lambda_j)_j$ calibrated theoretically (concentration)
- PhD student S. Ivanoff (co-adv. V. Rivoirard)



*Ivanoff et al. in prep

Outline

Functional models for genomics

- Flexible framework for modeling, computational efficiency, dimension reduction
- Perspectives for the mixed functional model (array CGH data ? other applications ?)
- Perspectives for the analysis of NGS data (variable coefficients models)
- Perspectives for the analysis of 3D data

Acknowledgements

LBBE	Collaborators	Students	SGDG & friends
C. Gautier	L. Duret	I. Bardet	L. Jacob
M. Gouy	A.-L. Fougères	G. Durif	V. Miele
D. Mouchiroud	S. Lambert-Lacroix	M. Giacofci	P. Veber
M.-F. Sagot	G. Marais	S. Ivanoff	V. Viallon
	M.-N. Prioleau	F. Mifsud	
	P. Reynaud-Bouret	L. Modolo	
	V. Rivoirard	A. Muyle	
	E. Roquain		