Statistical and computational challenges for population-based segmentation of copy-number profiles

Guillem Rigaill[1], Vincent Miele[2] and Franck Picard[2]

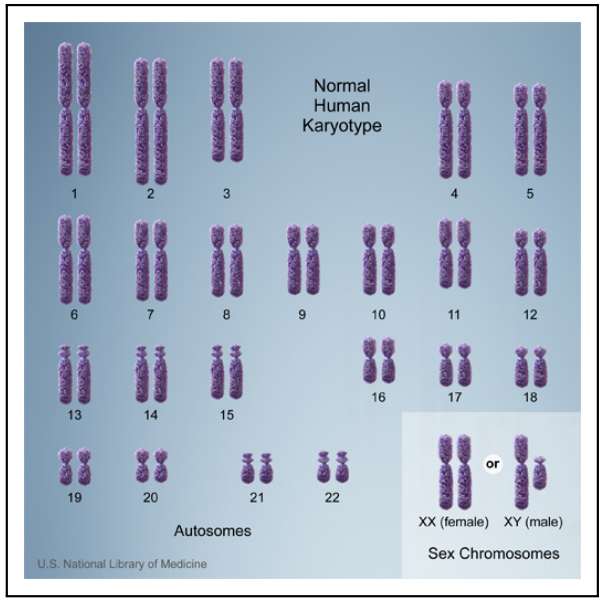
[1]Statistique et Génome, UMR CNRS 8071, USC INRA Université d'Evry, France
[2] Biométrie et Biologie Evolutive, UMR CNRS 5558 Université Lyon 1, France

Centre Blaise Pascal, Lyon November 2013

Outline

Karyotype and chromosome copy numbers

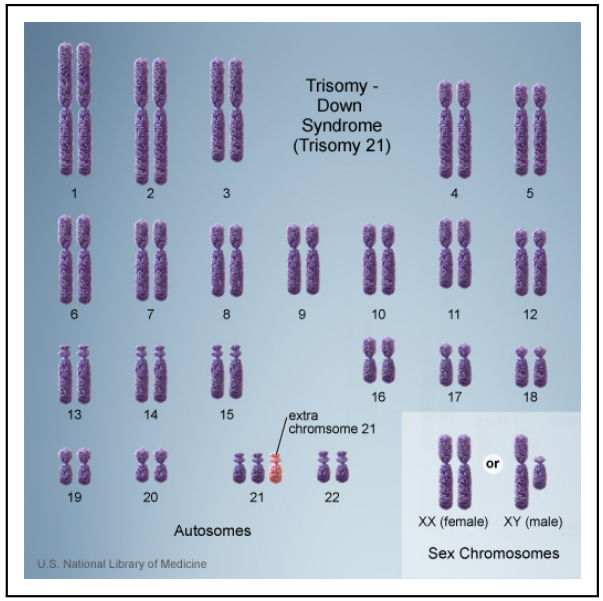
- Gene copy number is tightly regulated
- Humans: 22 pairs (autosomal)+ 1 pair sexual chromosomes
- At the chromosomal resolution, the karyotype is a visual tool to check for abnormalities
- Deviations from the reference number (2) result in massive disorders



Human karyotype (from NSF)

Karyotype and chromosome copy numbers

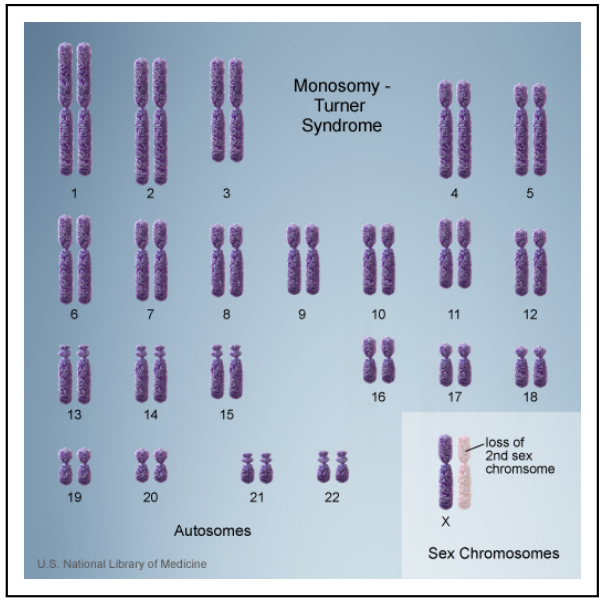
- Gene copy number is tightly regulated
- Humans: 22 pairs (autosomal)+ 1 pair sexual chromosomes
- At the chromosomal resolution, the karyotype is a visual tool to check for abnormalities
- Deviations from the reference number (2) result in massive disorders



Human karyotype (from NSF)

Karyotype and chromosome copy numbers

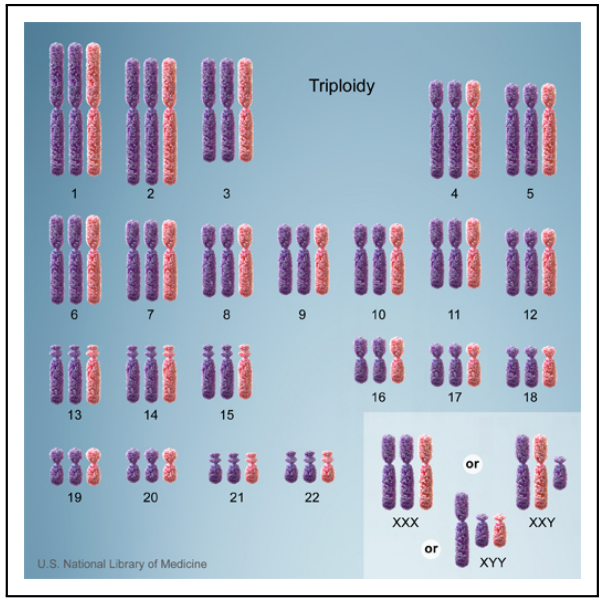
- Gene copy number is tightly regulated
- Humans: 22 pairs (autosomal)+ 1 pair sexual chromosomes
- At the chromosomal resolution, the karyotype is a visual tool to check for abnormalities
- Deviations from the reference number (2) result in massive disorders



Human karyotype (from NSF)

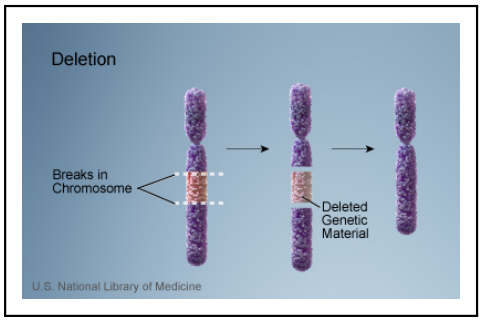
Karyotype and chromosome copy numbers

- Gene copy number is tightly regulated
- Humans: 22 pairs (autosomal)+ 1 pair sexual chromosomes
- At the chromosomal resolution, the karyotype is a visual tool to check for abnormalities
- Deviations from the reference number (3) result in massive disorders

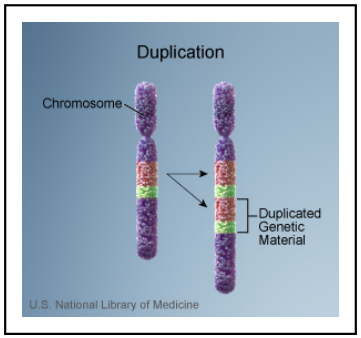


Human karyotype (from NSF)

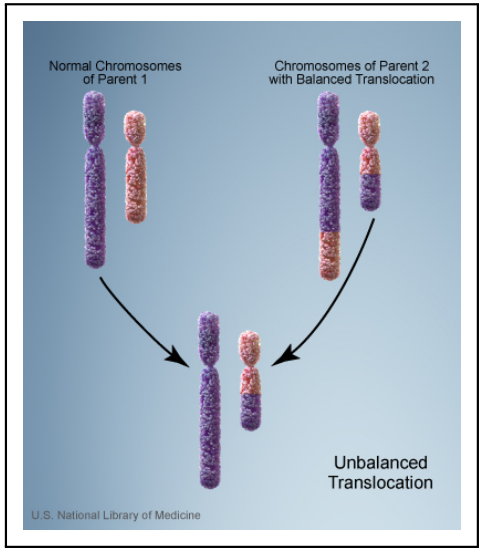
Sub-Chromosomal Aberrations



deletion



duplication

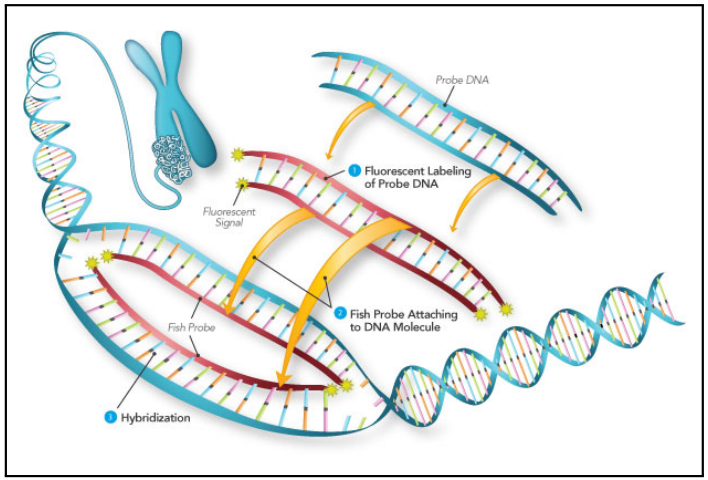


translocation

Mapping aberrations at low resolution has been a technical challenge in cytogenetics

Mapping using Fluorescent In Situ Hybridization

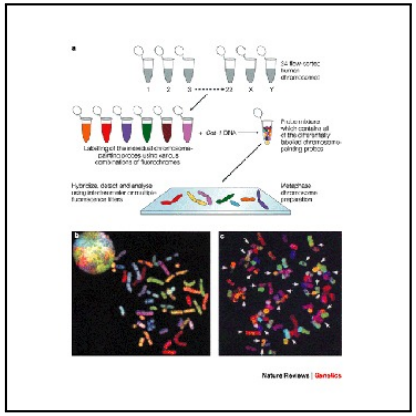
- Consider a known sequence of ~ 1 Mb and link it with a fluorochrome
- Mix it in presence of denatured chromosomes
- Check if the probe hybridizes somewhere
- If the probe comes from another chromosome, map the aberration

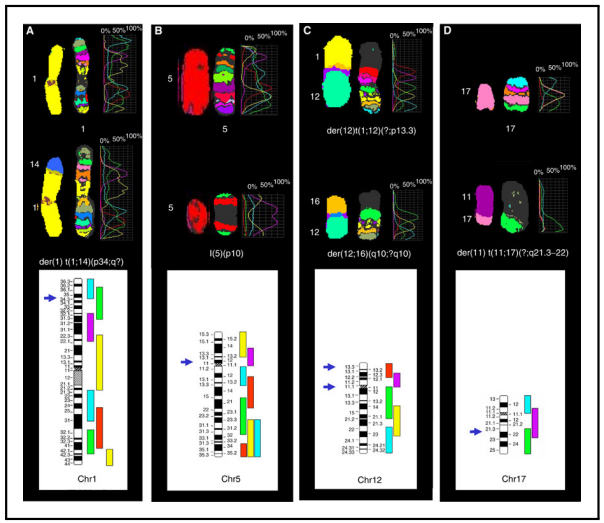


Going FISHing ? (from NSF)

Multicolor Fish and Comparative Genomic Hybridization

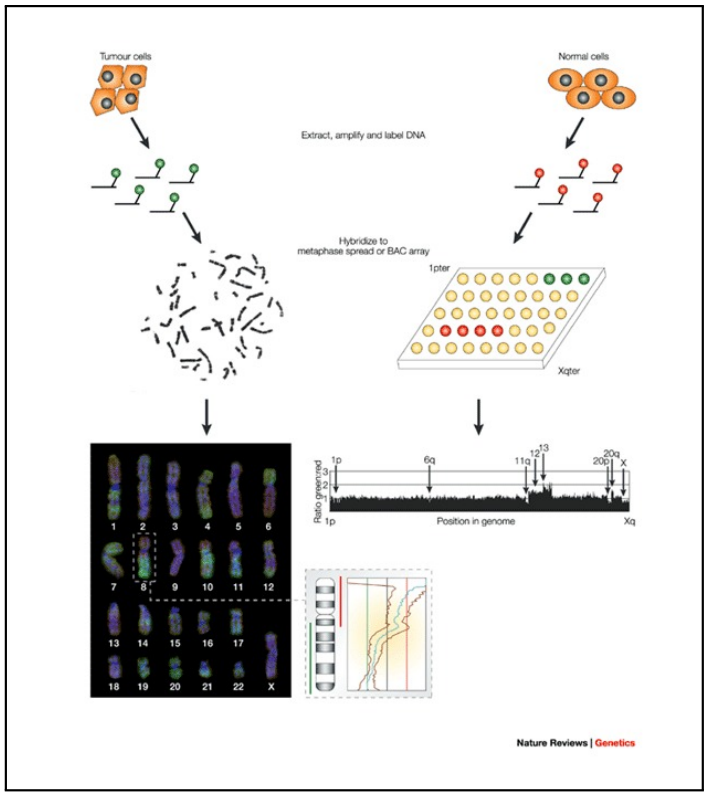
- Consider a set of reference sequences of size $\sim x$Mb
- Link them with different fluorochromes
- Mix in presence of denatured chromosomes
- Check if the probes hybridize somewhere
- If the probe comes from another chromosome, map the aberrations





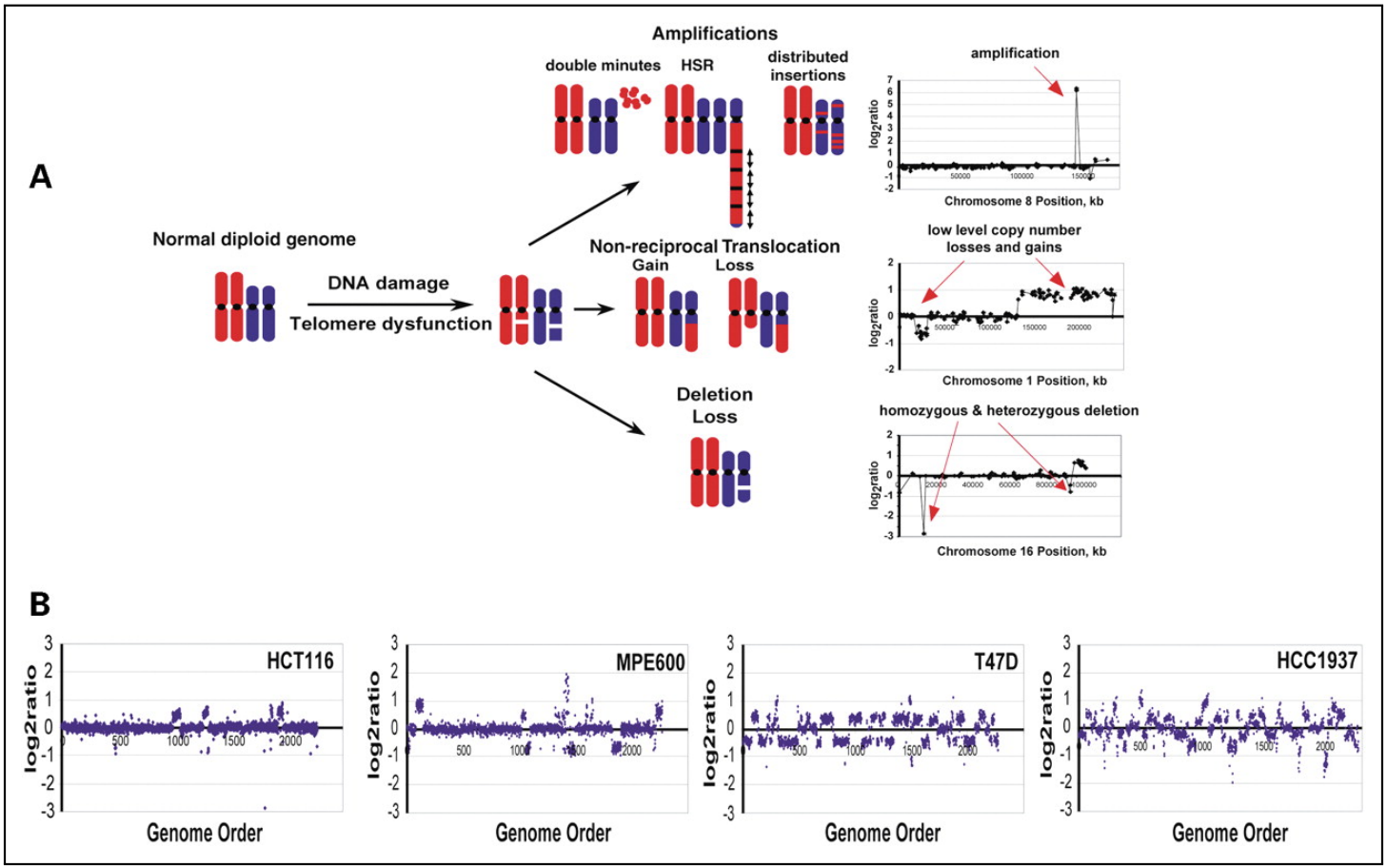
Application of the microarray technology to CGH

- The microarray technology was mainly developed for expression data. Application to CGH in 2003 (array-CGH)
- Probes are $\sim$ kbs long and fixed on a glass support
- Two genomes are compared by measuring the relative quantity of DNA at different loci



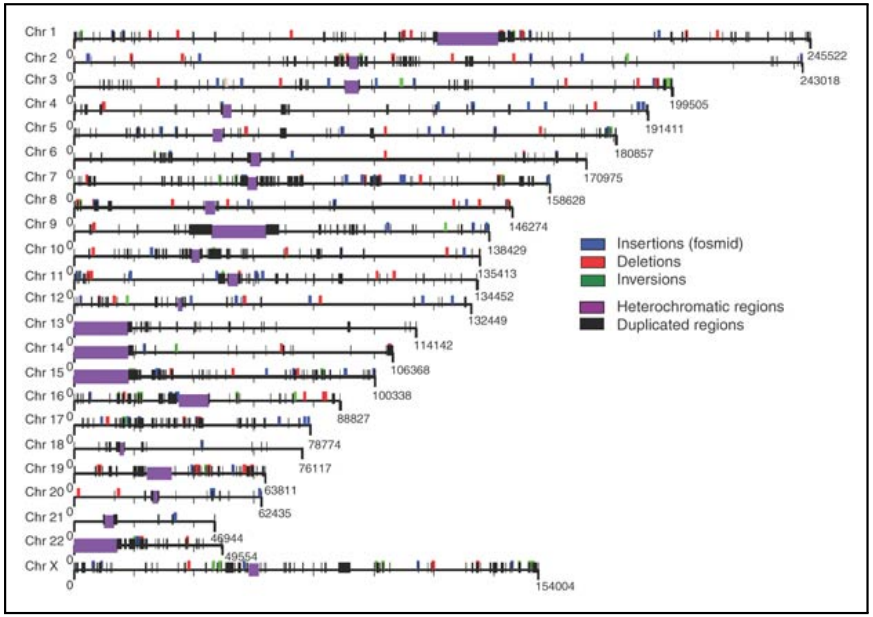
Array CGH allows a genome-wide blind search for $\sim$kbs aberrations

Tracking Genomic Aberrations in Cancer Genomes



Tracking Genomic Variation in Healthy Genomes

- In 2005 a study published the map of Copy Number Variations in healthy individuals
- Most initial studies of genetic variation concentrated on individual nucleotide sequences (SNPs)
- CNVs have become new genetic markers to study human diseases and evolution

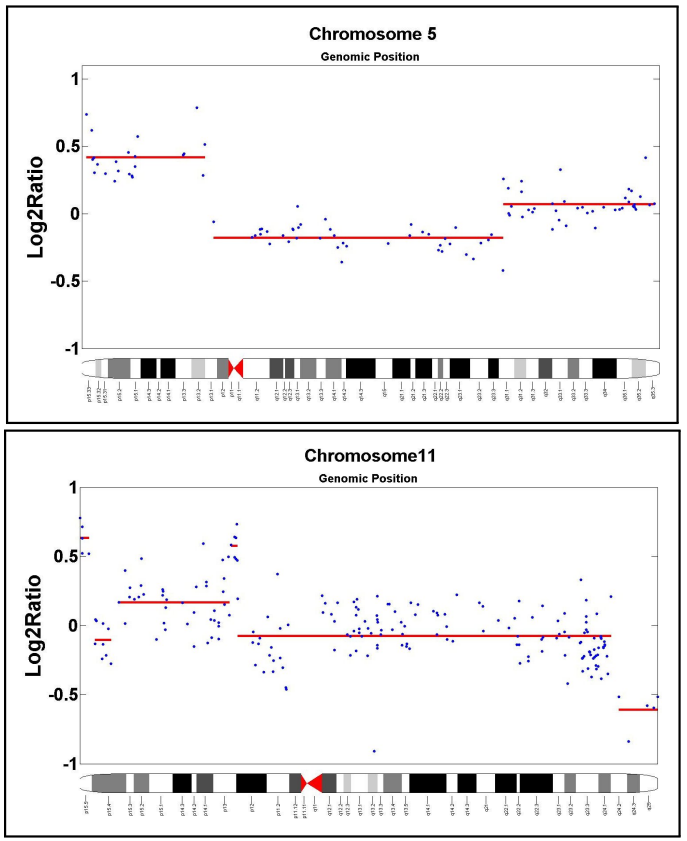


from `http://www.nature.com/`

Outline

Nature of array CGH data

- The signal Y_t is a $\log_2$ ratio of fluoresence organized along the genome (t)
- When $Y_t \sim 0$ the region has no imbalance between test-reference
- When $Y_t > 0$ (resp < 0) the test genome shows gains (resp. deletions)
- How many segments ? where ? status ?



Modeling & Computing Strategies

- Hidden Markov models [2, 7, 6]
  - Introduce a hidden Markovian sequence to model copy number
  - Recover the hidden sequence by Forward-Backward Algorithm
- Segmentation Models [10, 8]
  - Suppose that there exist abrupt changes in the signal
  - Detect jumps using a partitioning algorithm

Many comparative studies have shown the efficiency of segmentation methods on those data [12]

We focus on computational aspects of segmentations

Segmentation models: definitions and notations

- We observe a Gaussian process $\mathbf{Y} = \{Y_1, \ldots, Y_n\}$ with

$$Y_t \sim \mathcal{N}(\mu_t, \sigma^2).$$

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

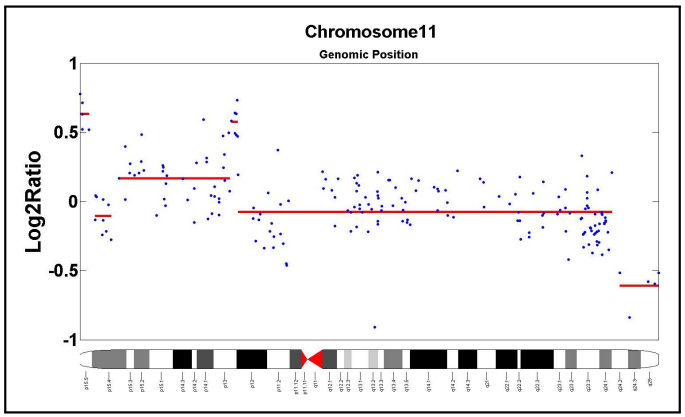
$$\forall t \in I_k, \quad Y_t = \mu_k + E_t, \quad E_t \sim \mathcal{N}(0, \sigma^2).$$

Calling Segments status by segmentation/clustering

- Segments can be in different `states` (Deleted, Normal, Amplified) which impacts the level of segments
- The idea is to introduce a hidden indicator variable Z_{kp} such that

$$\forall t \in I_k, \text{if } Z_{kp} = 1, \quad Y_t = m_p + E_t, \quad E_t \sim \mathcal{N}(0, \sigma^2).$$

- Segment levels are shared across the genome (m_{deleted} is the same for all deleted segments for instance).



Parameters and estimation strategy

- The parameters: $\mathbf{T} = \{t_0, \ldots, t_K\}$, $\boldsymbol{\mu} = \{\mu_1, \ldots, \mu_K\}$ and σ^2.
- The estimation is done for a given K which is estimated afterwards.
- The log-likelihood of the model is:

$$\log \mathcal{L}_K(\mathbf{Y}; \mathbf{T}, \boldsymbol{\mu}, \sigma^2) = \sum_{k=1}^{K} \sum_{t=t_{k-1}+1}^{t_k} f(y_t; \mu_k, \sigma^2).$$

- When K and $\mathbf{T}$ are known, how to estimate $\boldsymbol{\mu}$?
- When K is known, how to estimate $\mathbf{T}$?
- How to choose K ?

Parameter estimation

- When K and $\mathbf{T}$ are known the estimation of $\boldsymbol{\mu}$ is straightforward:

$$\begin{aligned}
\widehat{\mu}_k &= \frac{1}{\widehat{t}_k - \widehat{t}_{k-1}} \sum_{t=\widehat{t}_{k-1}+1}^{\widehat{t}_k} y_t, \\
\widehat{\sigma}^2 &= \frac{1}{n} \sum_{k=1}^{K} \sum_{t=\widehat{t}_{k-1}+1}^{\widehat{t}_k} (y_t - \widehat{\mu}_k)^2.
\end{aligned}$$

- Find $\widehat{\mathbf{T}}$ such that:

$$\widehat{\mathbf{T}} = \arg\max_{\mathbf{T}} \left\{ \log \mathcal{L}_K(\mathbf{Y}; \mathbf{T}, \boldsymbol{\mu}, \sigma^2) \right\}.$$

Dynamic Programming to optimize the log-likelihood

- Partition n data points into K segments: complexity $\mathcal{O}(n^K)$.
- DP reduces the complexity to $\mathcal{O}(n^2)$ when K is fixed.
- Shortest path problem: "subpaths of optimal paths are themselves optimal".
- $\text{RSS}_k(i,j)$ cost of the path connecting i to j in k segments:

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

Model selection for segmentation

- The number of segments K should be estimated:

$$\widehat{K} = \arg\max_K \left\{ \log \mathcal{L}_K(\mathbf{Y}; \widehat{\mathbf{T}}, \widehat{\boldsymbol{\mu}}, \widehat{\sigma}^2) - \beta \text{pen}(K) \right\}.$$

- Difficulty: C_{n-1}^{K-1} possible partitions for a model with K segments.
- Non-asymptotic theory provides a general form for $pen(K)$ [5]:

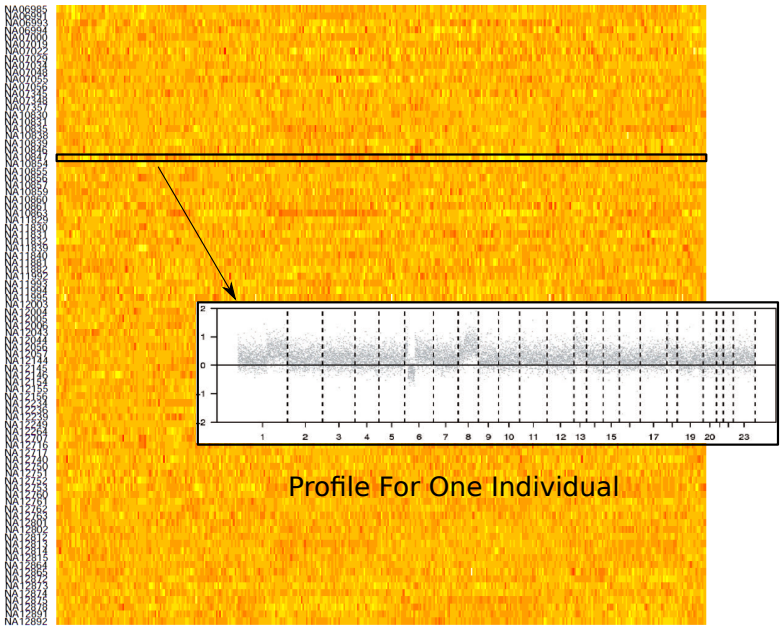
$$\beta pen(K) = \frac{K}{n}\sigma^2 \times \left(c_1 + c_2 \log \frac{n}{K} \right).$$

- Other methods are based on an adaptive estimation of K [4, 10] or on a modification of the BIC [13]

Outline

Using Multiple Arrays to assess CNA/CNV

- Population-based analysis for cancer and human genetics
- Multiple Arrays Analysis
  - Find breaks using all samples
  - Find *reccurrent* breaks
- What is specific/common ?
  - Shared biases
  - Specific CN



Use multiple samples to increase the power of detection

Modelling individual-specific breakpoints [10]

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

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

- $\boldsymbol{\mu}_i$ specific levels of segments
- $\mathbf{T}_i$ specific incidence matrix of the breaks

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

- Signal levels associated to CN status are shared across arrays:

$$\{Z_{kp}^i = 1\},\ \forall t \in \mathcal{I}_k^i,\ Y_i(t) = m_p + \varepsilon_i(t),\ \varepsilon_i(t) \sim \mathcal{N}(0, \sigma^2).$$

$$\mathbf{Y}_i = \mathbf{T}_i \mathbf{Z}_i \mathbf{m} + \mathbf{E}_i$$

Segmentation of Multiple Arrays [9]

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

$$RSS_K(\boldsymbol{\mu}, \mathbf{T}) = \|\mathbf{Y} - \mathbf{T}\boldsymbol{\mu}\|^2 \quad = \quad \sum_{i=1}^{I} \sum_{k=1}^{k_i} RSS_k^i(\boldsymbol{\mu}_i, \mathbf{T}_i)$$

- Global DP would lead to a $\mathcal{O}(n^2 I^2)$ complexity.
- But there is a constraint : $\sum_i k_i = K$, (K unknown) thus:

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

A two-stage Dynamic Programming procedure - 1

- Find all optimal breaks for each profile using a "classical DP"
- $\widehat{\mathbf{T}}^i(k_i)$ the set of optimal breaks with k_i segments for profile i.
- Find $\widehat{\mathbf{T}}^i(k_i),\ \forall k_i = 1, \ldots, k_{\max}$ segments by minimizing $RSS^i_{K_i}(\mathbf{T}_i, \boldsymbol{\mu}_i)$ for each series.

$$\forall i \in [1, I]\ \ \{\widehat{\mathbf{T}}_i, \widehat{\boldsymbol{\mu}}_i\} = \underset{\mathbf{T}_i, \boldsymbol{\mu}_i}{\arg\min} \left\{ RSS^i_{k_i}(\mathbf{T}_i, \boldsymbol{\mu}_i) \right\}$$

A two-stage Dynamic Programming procedure - 2

- Optimal allocation of segments to series

$$\forall i \in [1:I],$$
$$\{\widehat{k}_1, \ldots, \widehat{k}_i\} = \underset{k_1+\ldots+k_i=K}{\arg\min} RSS_K\left(\widehat{\mathbf{T}}^1(k_1), \ldots, \widehat{\mathbf{T}}^i(k_i)\right)$$

$$\widehat{\mathbf{T}}(K) = \left\{\widehat{\mathbf{T}}^1(\widehat{k}_1), \ldots, \widehat{\mathbf{T}}^I(\widehat{k}_I)\right\}.$$

- This procedure is optimal with a complexity $\mathcal{O}(In^2 k_{max} + k_{\max}^2 I^3)$.

Enrich the model to account for common genomic biases ?

- There exist common genomic biases that are shared by all profiles. How to correct them ?
- The simplest way to model this trend is to introduce a common background function $b(t)$ such that:

$$\forall t \in]t^i_{k-1}, t^i_k], \quad Y_i(t) = \mu_{ik} + b(t) + E_i(t).$$

- This produces a new model that mixes piece-wise constant functions and other undetermined functions

Regularization of the trend using splines

- Control the second derivative of $\mathbf{b}$ using a penalty:

$$\min_{\mathbf{T},\boldsymbol{\mu},\boldsymbol{\theta}} \left\{ \frac{1}{2} \|\mathbf{Y} - \mathbf{T}\boldsymbol{\mu} - \mathbf{X}\mathbf{b}\|_2^2 + \lambda l \int \left[b''(t)\right]^2 \mathrm{d}t \right\}.$$

- $\{\mathbf{W}\}_{jk} = W_j(t_k)$ a n-dim. set of natural spline functions: $\mathbf{b} = \mathbf{W}\boldsymbol{\theta}$

$$\min_{\mathbf{T},\boldsymbol{\mu},\boldsymbol{\theta}} \left\{ \frac{1}{2} \|\mathbf{Y} - \mathbf{T}\boldsymbol{\mu} - \mathbf{X}\mathbf{W}\boldsymbol{\theta}\|_2^2 + \lambda l \boldsymbol{\theta}^T \boldsymbol{\Omega} \boldsymbol{\theta} \right\},$$

- The solution is given by:

$$\widehat{\boldsymbol{\theta}} \quad = \quad \left\{ \mathbf{W}^T\mathbf{W} + \lambda\boldsymbol{\Omega} \right\}^{-1} \mathbf{W}^T \left(\mathbf{X}^T \left[\mathbf{Y} - \mathbf{T}\boldsymbol{\mu}^{[h]} \right] / l \right).$$

Outline

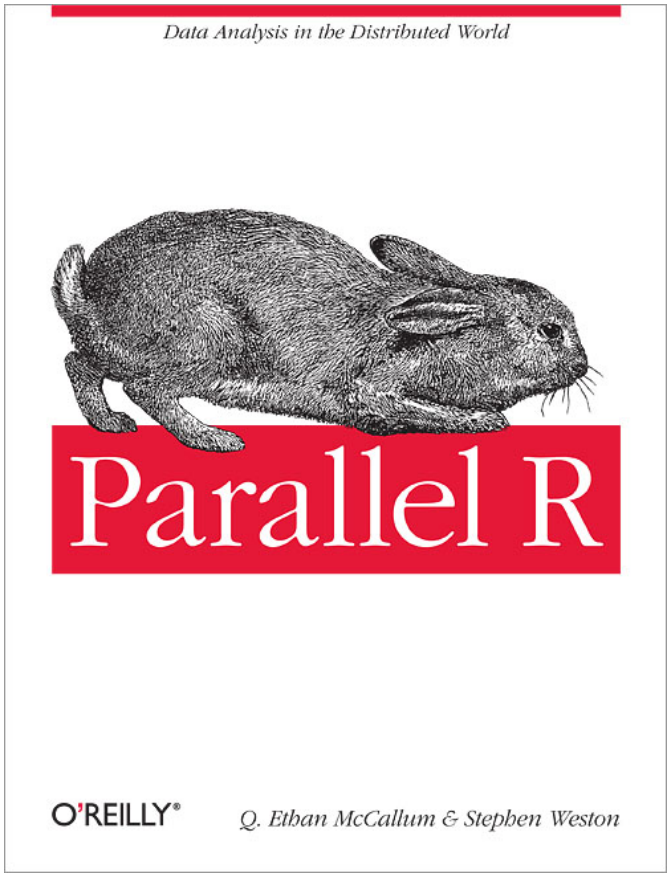
Pruning Strategy

- Pruning Strategies reduce the computational burden of Dynamic Programming ([11, 3])
- The idea is to prune the set of candidates while computing potential segmentations
- The complexity is reduced from $\mathcal{O}(Kn^2)$ to $\mathcal{O}(n)$ or $\mathcal{O}(n\log(n))$
- This linearization allows segmentation to be used on very long signals (microarrays, sequencing)

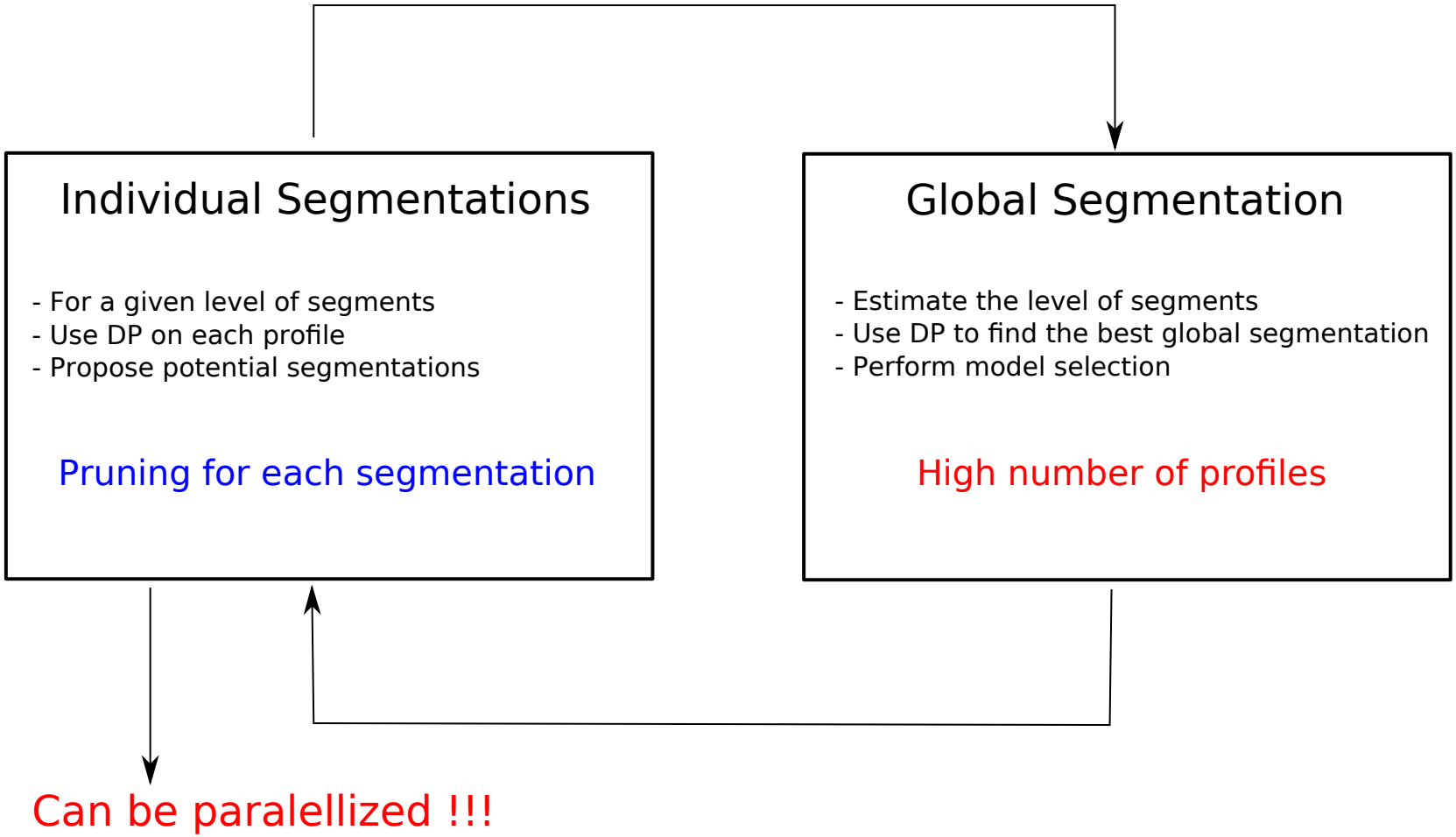
Parallelization in cghseg

- cghseg is a R-package dedicated to segmentation
- Most computers have multi-core architectures (from laptops to many-core servers)
- It has become essential to adapt software to computer architectures

Use straightforward parallelization to perform segmentation on large-numerous signals

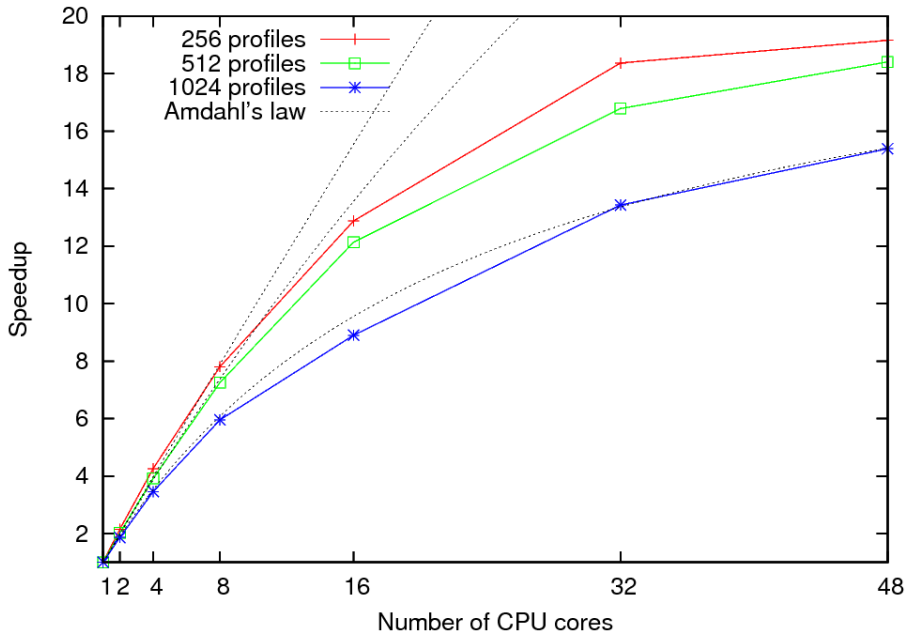


Computing scheme



Parallelization in cghseg

- Compare the observed speedup to theoretical speedup (Amdahl's law [1])
- The speedup of cghseg follows the Amdahl's law when the number of profiles is high
- The gain decreases with the number of profiles due to overheads associated with the used of the parallel R-package



total time $\simeq$ time(sequential) + time(parallel)/nb of cores

Next Gen. Computing/Next Gen. experiments

- Considering multiple Arrays allows the joint assessment of Copy Number Aberrations/Variations at the cohort level
- We solve the computational issue of joint segmentation using a X2 Stage DP with linearization
- The method is implemented in the `cghseg` 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

Conclusions

- The analysis of copy number profiles has been very challenging from a statistical and computational point of view
- When providing methods, check the scalability (500K probes)
- Developments had to be done using mathematical+computing skills
- Many question arise from these data, in particular the impact of the inter-individual variability
- Project: shift towards functional mixed models for genomics

References

[1] Gene M. Amdahl. Validity of the single processor approach to achieving large scale computing capabilities. In *Proceedings of the April 18-20, 1967, spring joint computer conference*, AFIPS '67 (Spring), pages 483–485. ACM, 1967.

[2] J. Fridlyand, A. Snijders, D. Pinkel, D.G. Albertson, and A.N. Jain. Hidden markov models approach to the analysis of array CGH data. *Journal of Multivariate Analysis*, 90(1):132–1533, 2004.

[3] R. Killick, P. Fearnhead, and I. A. Eckley. Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107(500):1590–1598, 2012.

[4] M. Lavielle. Using penalized contrasts for the change-point problem. *Signal Processing*, 85(8):1501–1510, 2005.

[5] E. Lebarbier. Detecting multiple change-points in the mean of Gaussian process by model selection. *Signal Processing*, 85:717–736, 2005.

[6] J.C. Marioni, N.P. Thorne, and S. Tavare. BioHMM: a heterogeneous hidden Markov model for segmenting array CGH data. *Bioinformatics*, 22(9):1144–1146, 2006.

[7] P. Nicolas, L. Bize, F. Muri, M. Hoebeke, F. Rodolphe, S. D. Ehrlich, B. Prum, and P. Bessieres. Mining Bacillus subtilis chromosome heterogeneities using hidden Markov models. *Nucleic Acids Res.*, 30(6):1418–1426, Mar 2002.

[8] AB. Olshen, ES. Venkatraman, R. Lucito, and M. Wigler. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, 5(4):557–572, 2004.

[9] F. Picard, Lebarbier E., Hoebeke M., Rigaill G., Thiam B., and Robin S. Joint segmentation, calling and normalization of multiple array CGH profiles. *Biostatistics*, 12(3):413–428, 2011.

[10] F. Picard, S. Robin, M. Lavielle, C. Vaisse, and J-J. Daudin. A statistical approach for CGH microarray data analysis. *BMC Bioinformatics*, 6:27, 2005.

[11] G. Rigaill. Pruned dynamic programming for optimal multiple change-point detection. *Arxiv:1004.0887*, 1:1–9, April 2010.

[12] H. Willenbrock and J. Fridlyand. A comparison study: applying segmentation to array CGH data for downstream analyses. *Bioinformatics*, 21(22):4084–4091, 2005.

[13] N. R. Zhang and D. O. Siegmund. A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data. *Biometrics*, 63(1):22–32, 2007.