Joint segmentation of aCGH profiles using fast group LARS

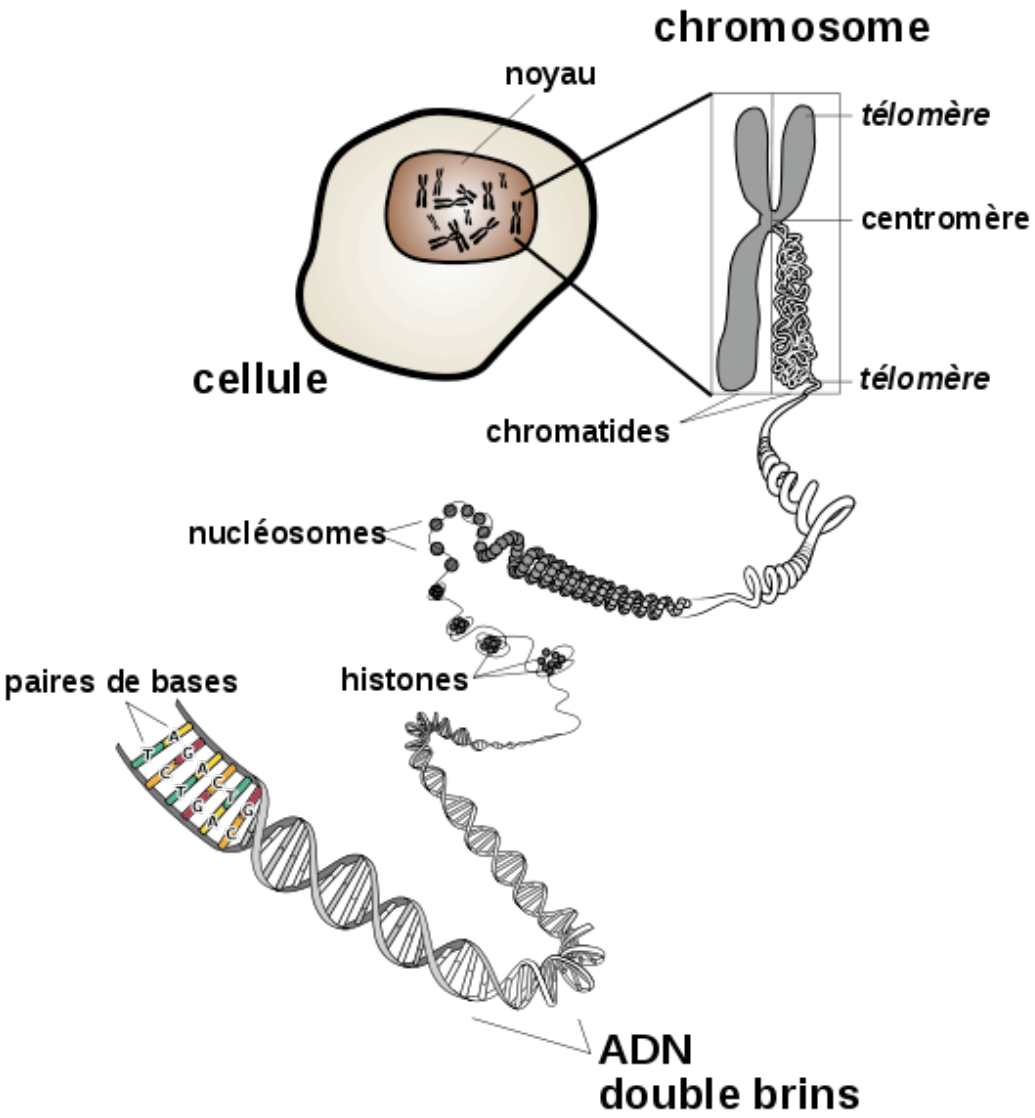
K. Bleakley and J.-P. Vert

Cambridge, UK, September 2010

Outline

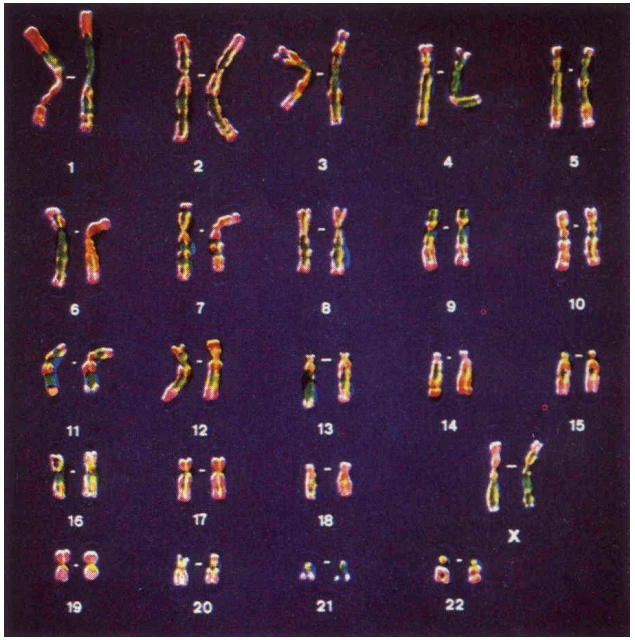
- Biology
- Mathematics
- Simulations and pretty pictures
- Real data

Section 1: Biology



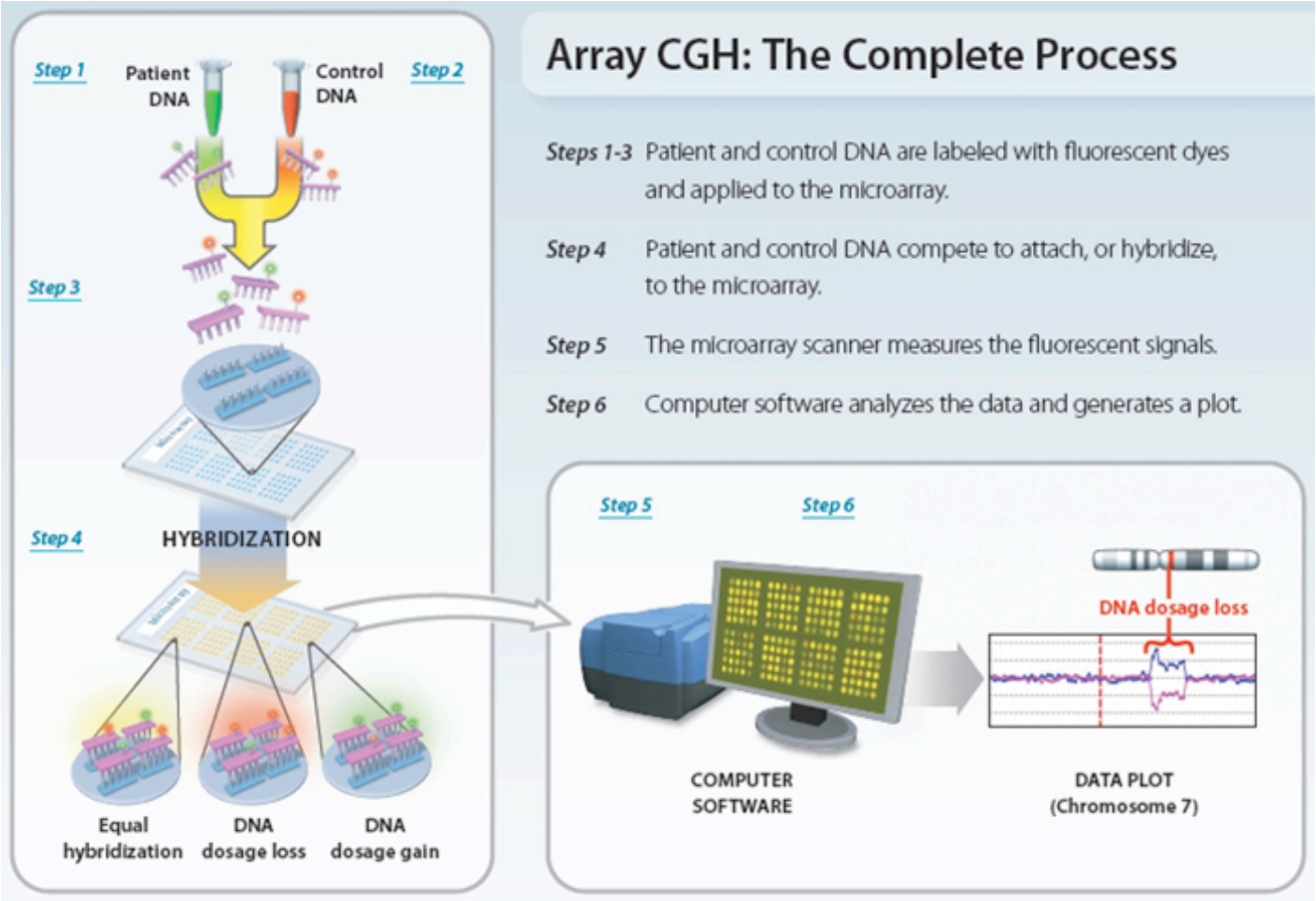
Copy number variation

- 2 copies of each chromosome (except X in men)

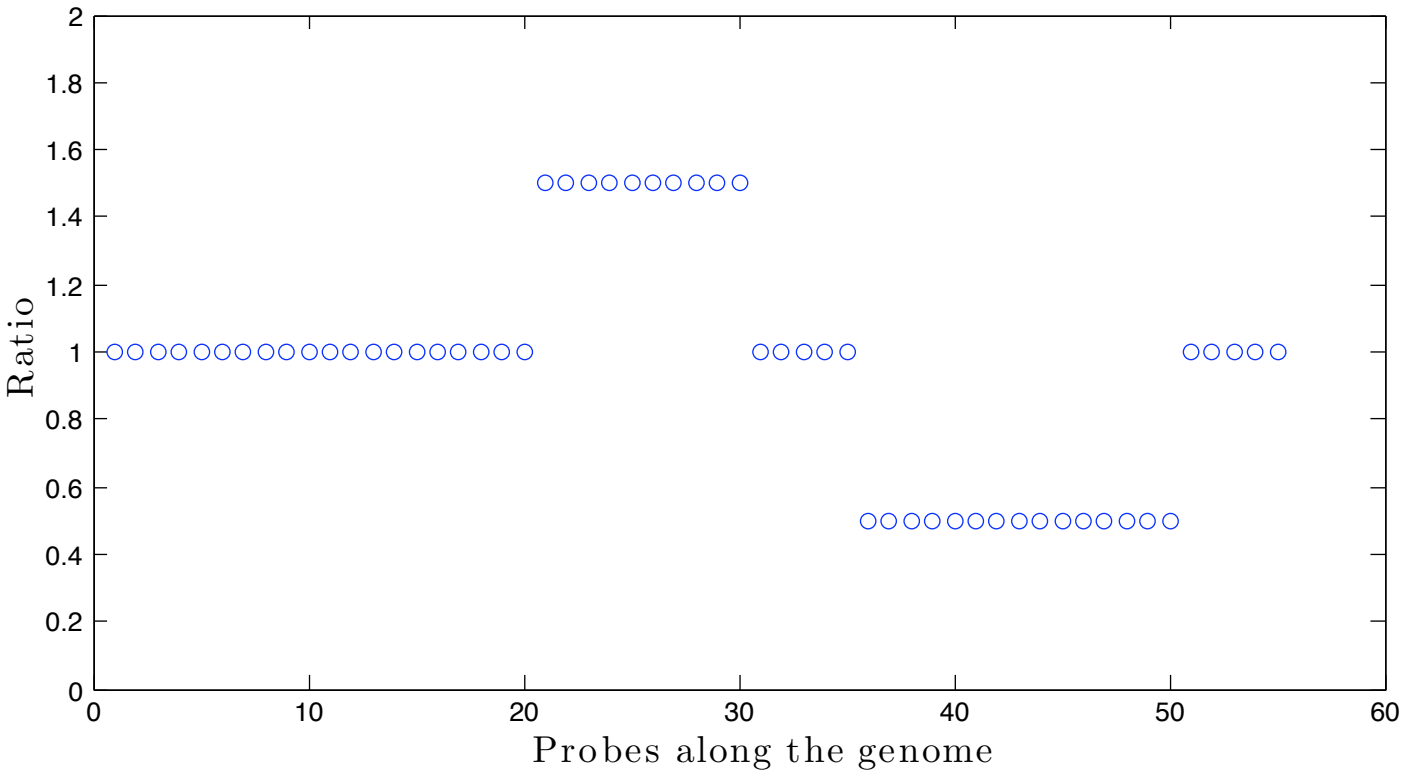


- cancer cells: copy number variation
- oncogenes: amplification
- tumour suppressor genes: deletion

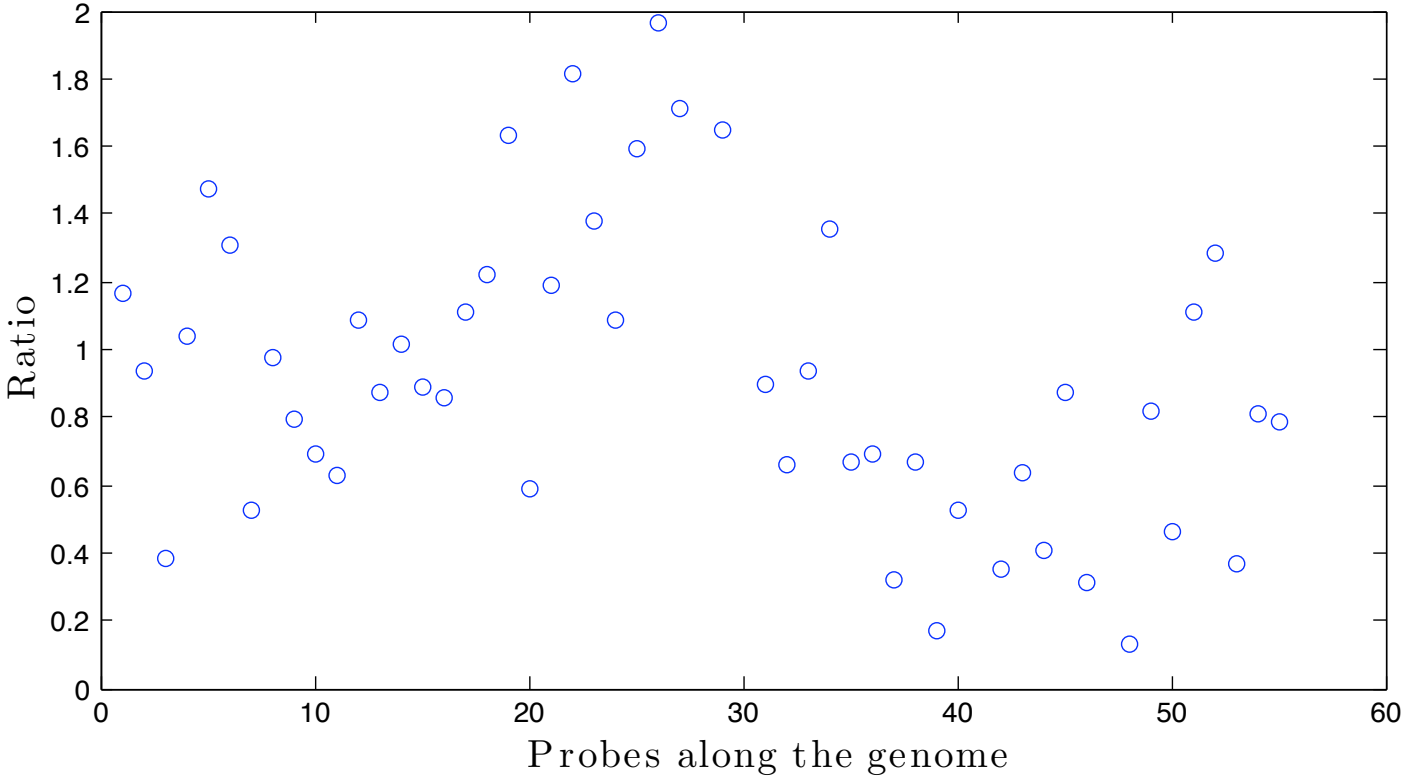
Measuring the copy number



In theory


Ratio
Probes along the genome
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
0
10
20
30
40
50
60

In reality


Ratio
Probes along the genome
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
0
10
20
30
40
50
60

Section 2: Mathematics

- reconstruct the underlying piecewise constant signal
- identify gains and losses

- Hidden Markov Models (J. Fridlyand, ...)
- dynamic programming (F. Picard, S. Robin, ...)
- sparsity (R. Tibshirani, T. Hastie, ...)

Dynamic programming

- p probes and K breakpoints
- minimise some criteria over the $\binom{p}{K}$ partitions
  e.g., mean square error
- dynamic programming $O(p^K) \rightarrow O(p^2)$

- today, $p \simeq$ 1 million
- tomorrow (next generation sequencing), $p \simeq$ 1 billion

The Lasso (Tibshirani, 1996)

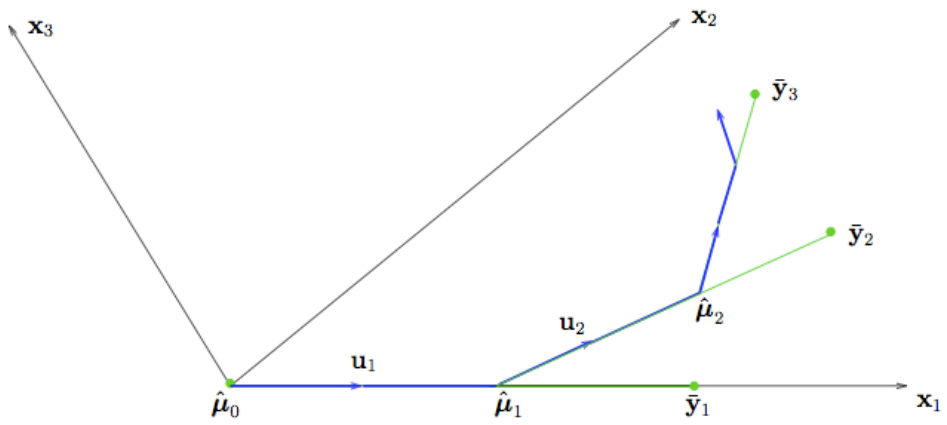
- $\min \|Y - X\beta\|^2$ subject to $\sum |\beta_i| \leq \mu$.
- for small μ, most β_i are equally zero $\rightarrow$ sparse variable selection.

LARS (Efron et al. 2004)

- LARS = Least Angle Regression.
- Y and X and β as before.

- 1) start with all coefficients β_i as 0.
- 2) find variable x_i (column of X) most correlated with Y.
- 3) increase β_i until get an x_j such that
  $Y - x_i\beta_i$ equally correlated with x_i and x_j.
- 4) proceed in equiangular direction to x_i and x_j
  until $\ldots x_k$, etc.

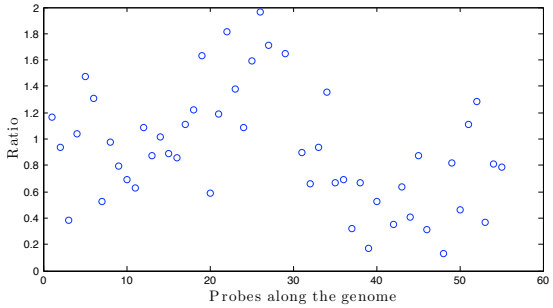
LARS (Efron et al. 2004)



- simple formula calculates when next variable enters (and $\beta_1, \ldots, \beta_p$).
- only one variable enters at a time if data in general position.
- whole solution path in time it takes to do a least-squares fit!
- tiny modification gives the Lasso!

The Fused Lasso

- Let Y be the vector of length p of values along the genome.



- Let X be the identity matrix of size p.

- The Lasso:
  $\min \|Y - X\beta\|^2 \quad \text{subject to} \quad \sum |\beta_i| \leq \mu.$

- i.e., $\min \sum (y_i - \beta_i)^2 \quad \text{subject to} \quad \sum |\beta_i| \leq \mu.$
- the Lasso is no good for copy number variation/CGH stuff.

The Fused Lasso

- $\min \|Y - X\beta\|^2$ subject to $\sum |\beta_i| \leq \mu$ and $\sum_{i=2}^{p} |\beta_i - \beta_{i-1}| \leq \nu$.
- keeps sparsity of coefficients.
- induces sparsity in difference between consecutive coefficients.
- i.e. induces piecewise-constant profiles!

- but loses link with superfast Lars algorithm.
- requires convex optimization $\rightarrow$ slow.

The Fused Lasso but without the Lasso

- Harchaoui and Lévy-Leduc (2008).
- forget about the constraint $\sum |\beta_i| \leq \mu$.
- just solve:
  $\min \|Y - X\beta\|^2$ subject to $\sum_{i=1}^{p} |\beta_i - \beta_{i-1}| \leq \nu$.
- by convention, $\beta_0 = 0$ (and remember $X = I_p$).

- change of variable: $u_i = \beta_i - \beta_{i-1}$.
- result:
  $\min \|Y - Xu\|^2$ subject to $\sum_{i=1}^{p} |u_i| \leq \nu$,
  with:

$$X = \begin{pmatrix} 1 & 0 & 0 & 0 & \dots \\ 1 & 1 & 0 & 0 & \dots \\ 1 & 1 & 1 & 0 & \dots \\ \vdots & & & & \end{pmatrix}$$

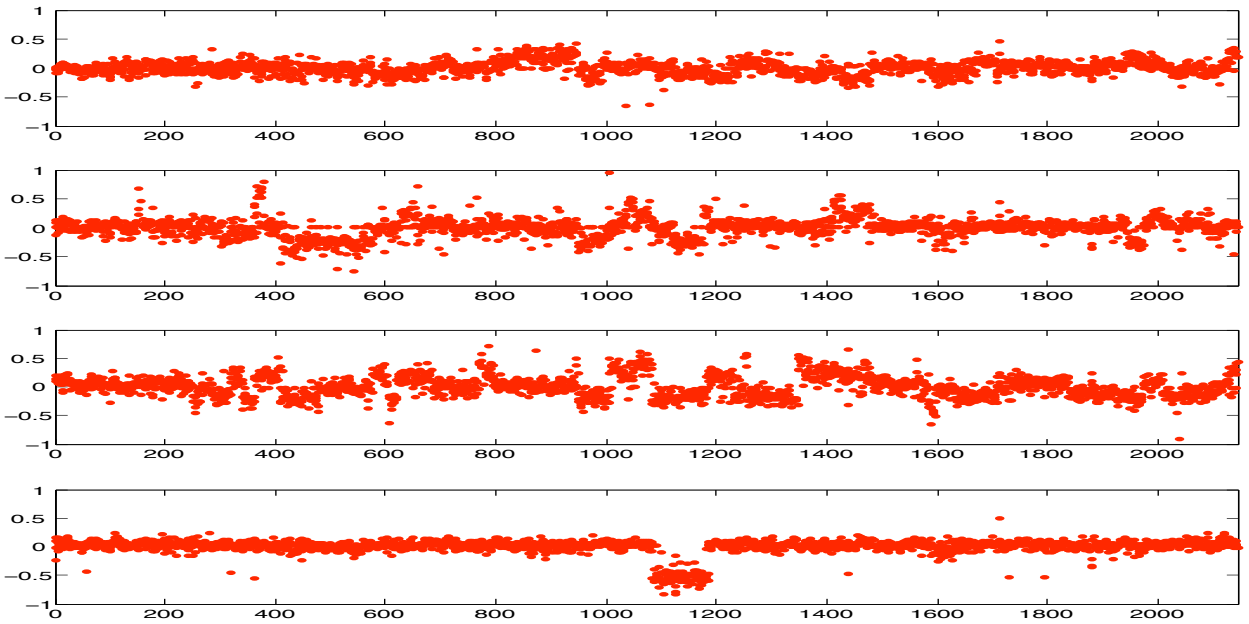
A fast implementation

- the LARS package in R has to store X, a $p \times p$ matrix.
- if $p \simeq 1$ million, your computer will not be happy.

- Harchaoui and Lévy-Leduc used the special structure of X to avoid storing it.
- e.g. to find the first non-zero u_i, find column of X that is most correlated with Y.
- first non-zero coefficient (breakpoint!) $= u_{p-i+1}$, with i the index of the maximum of $|\text{cumsum}(y_p, y_{p-1}, \ldots, y_1)|$.
- this and other tools $\rightarrow$ find k breakpoints in $O(pk)$.

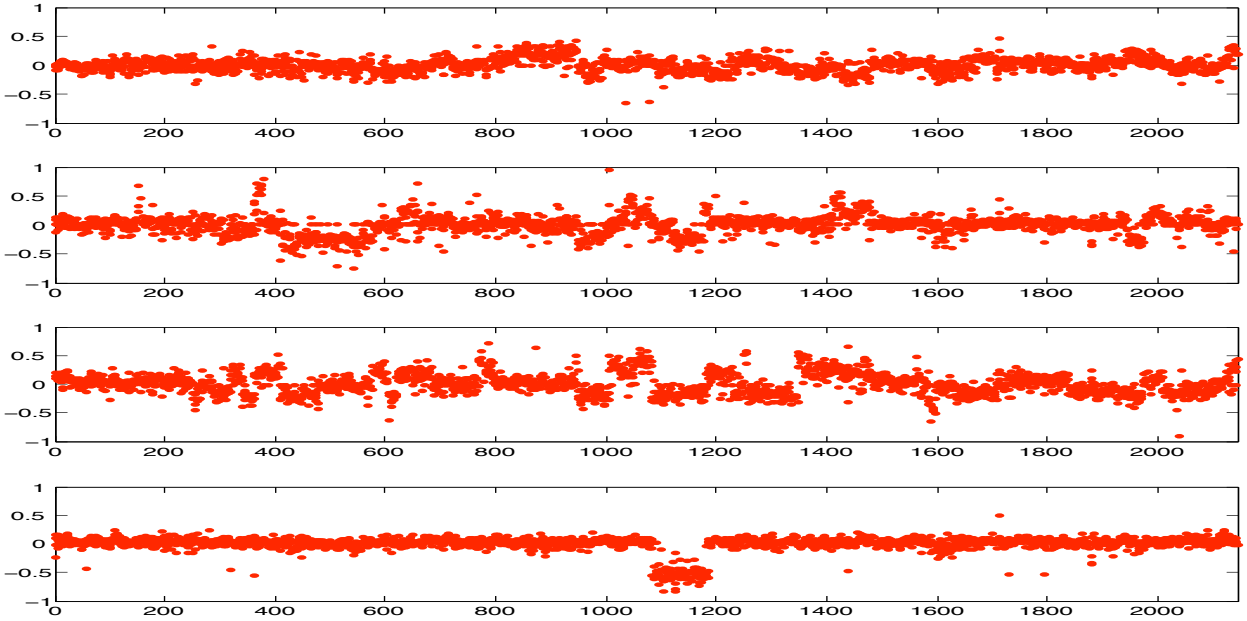
From one profile to many

- many individuals with the same cancer
- do they share copy number variations?
- actually, yes.
- but not always.
- and rarely all the same ones. e.g. bladder cancer:



From one profile to many

- either segment separately → look for shared,
- or segment together and force shared.



Formulation of the joint segmentation problem

- Y now an $n \times p$ matrix (n individuals, p probes).
- Let β be an $n \times p$ and β_i its ith column.

- $\min \|Y - \beta\|^2$ subject to $\sum_{i=1}^{p} \|\beta_i - \beta_{i-1}\|_2 \leq \nu$
- forces a given breakpoint on all profiles at the same time!

Group Lasso and Group Lars

- change of variable: $u_i = \beta_i - \beta_{i-1}$.
- problem becomes:
  $\min_{u \in \mathbb{R}^{np}} \left\| Y - \sum_{i=1}^{p} X_i u_i \right\|^2$ subject to $\sum_{i=1}^{p} \|u_i\| < \nu$,
  with Y as an np-dimensional vector, and X_i an $np \times n$ matrix.
- this is exactly a group Lasso.
- variables are selected in groups.
- here, selecting a group = selecting a shared breakpoint.

Group Lasso and Group Lars

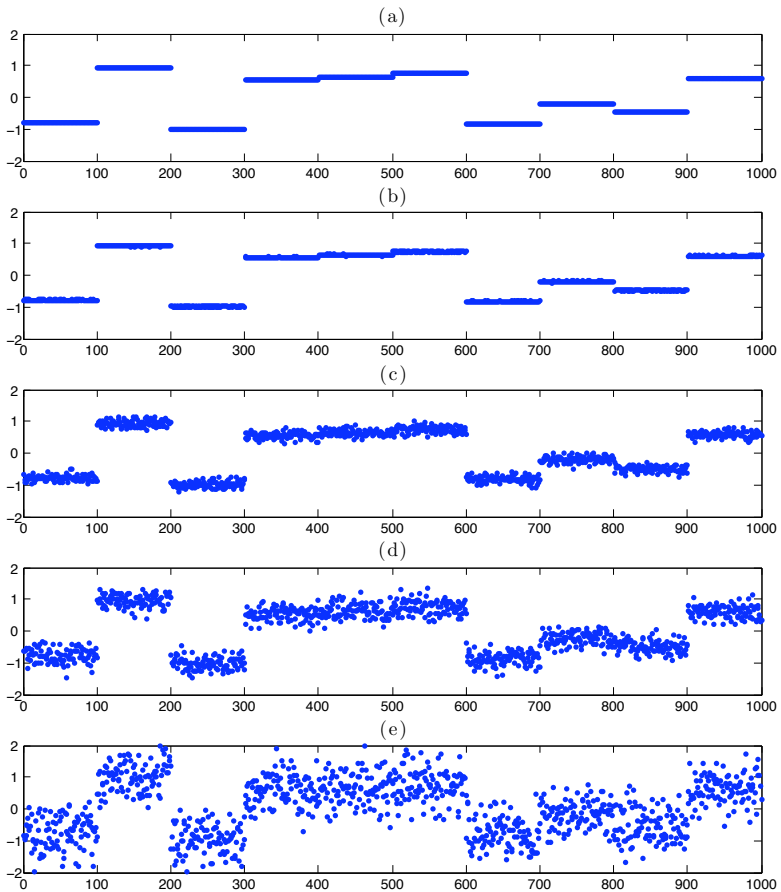
- group Lasso and group Lars introduced by Yuan and Lin (2006).
- group Lasso loses the piecewise constant property
  $\longrightarrow$ slower algorithm to get the whole solution path.

- group Lars = same geometric interpretation as Lars.
- it retains the piecewise constant property
  $\longrightarrow$ can create fast path algorithm in $O(knp)$.

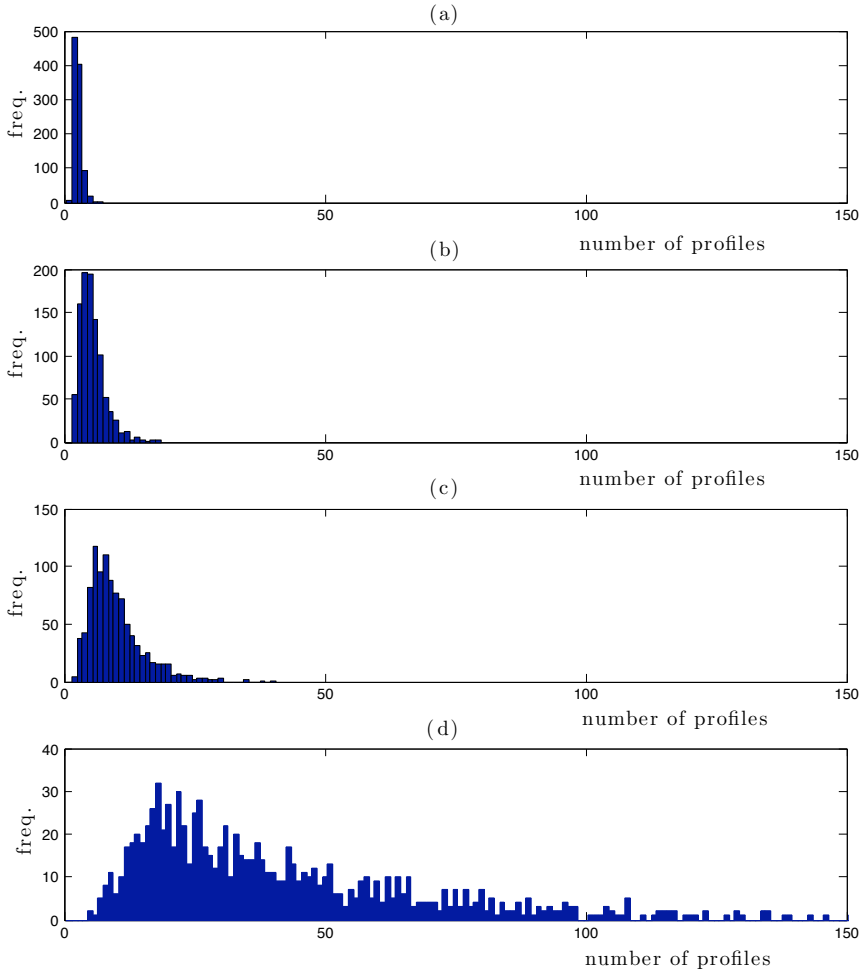
- e.g., 20 individuals, 1 million probes:
  takes 6.1 seconds to find each new shared breakpoint!

Simulation

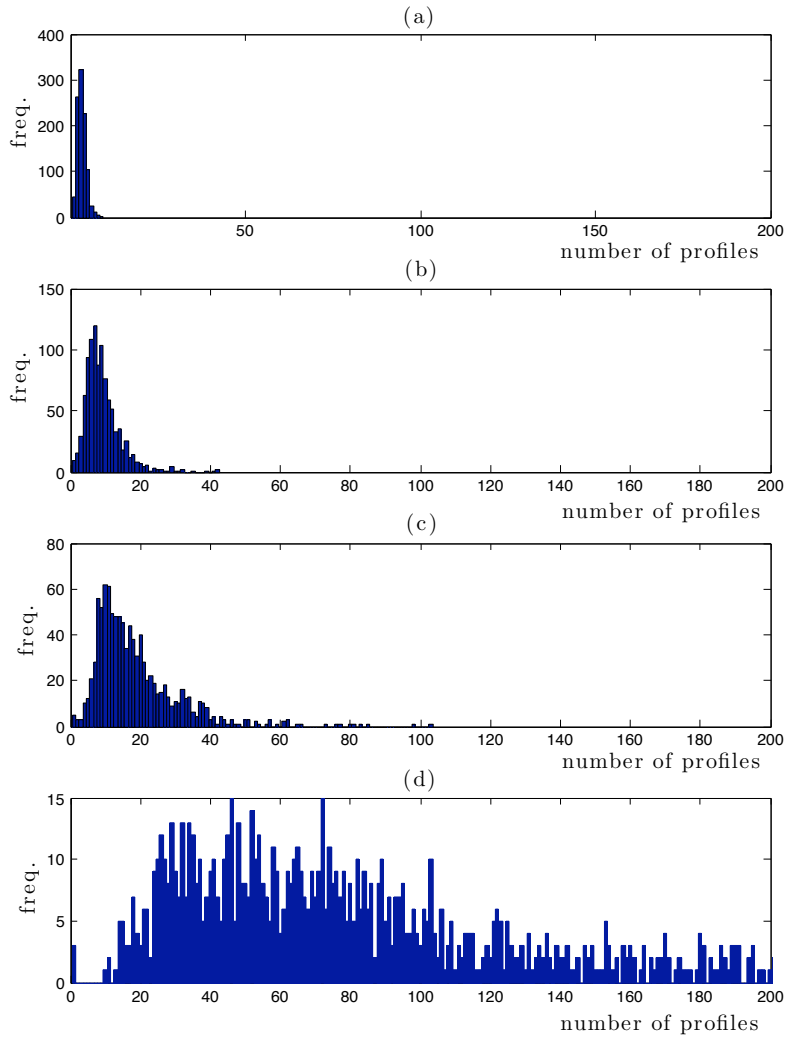
- add differing amounts of noise to piecewise-constant signal:



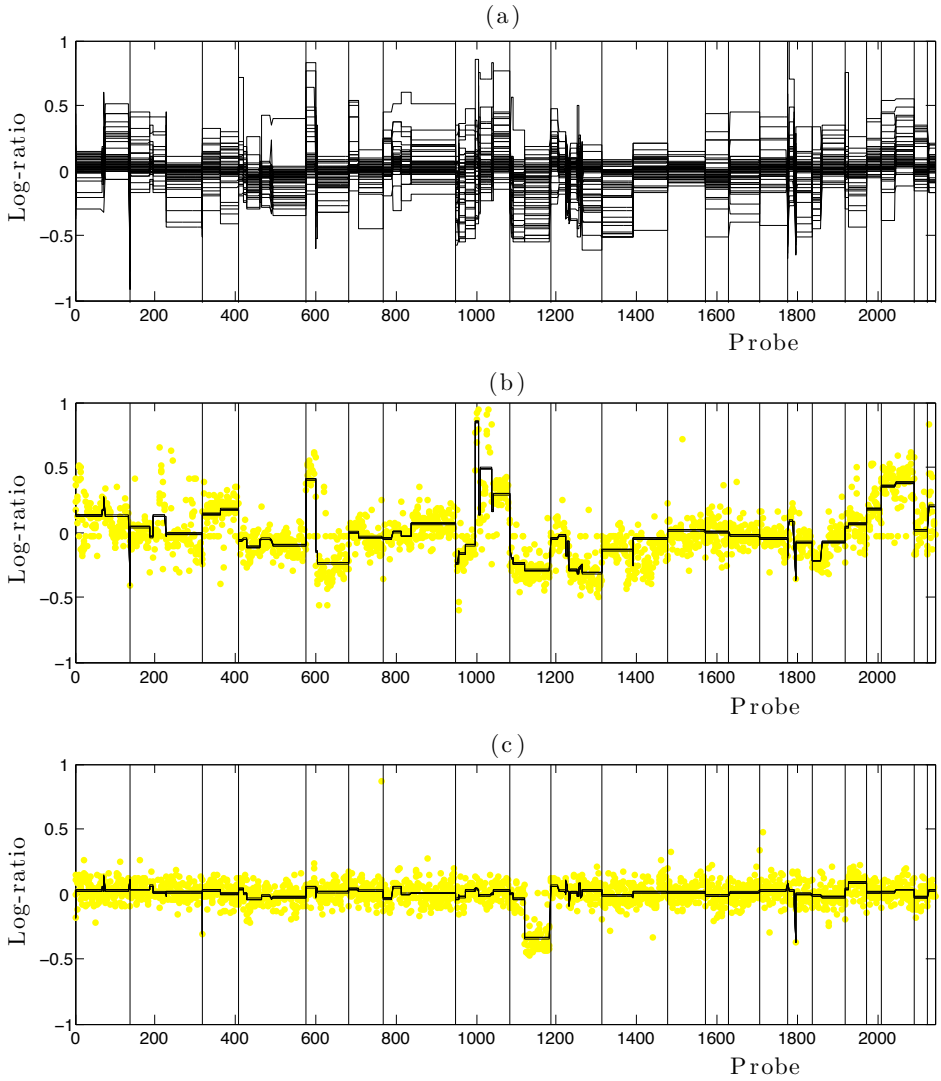
Distribution of number of required individuals: all shared

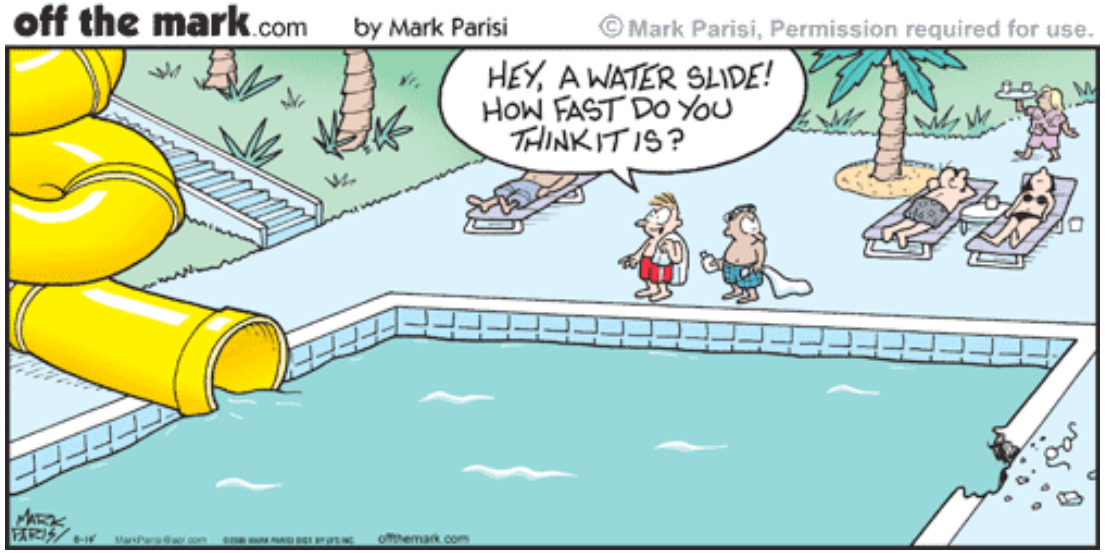


Distribution of number of required individuals: some shared



Bladder tumour profiles



Conclusions and future work

- fast
- algorithm does what it should
- theoretical results
- next generation sequencing
- fast $\longrightarrow$ faster!