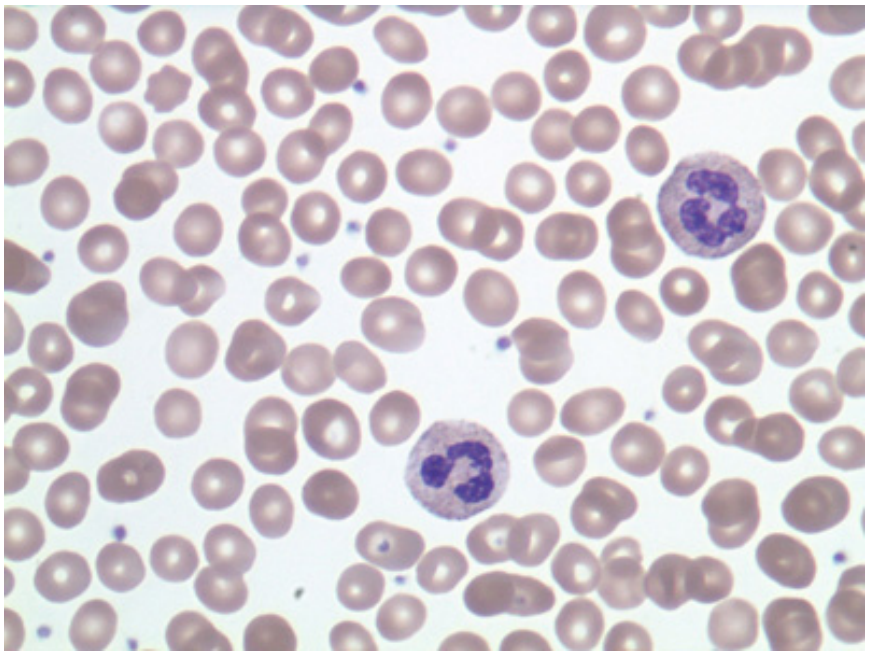
# Learning with structured sparsity in computational biology

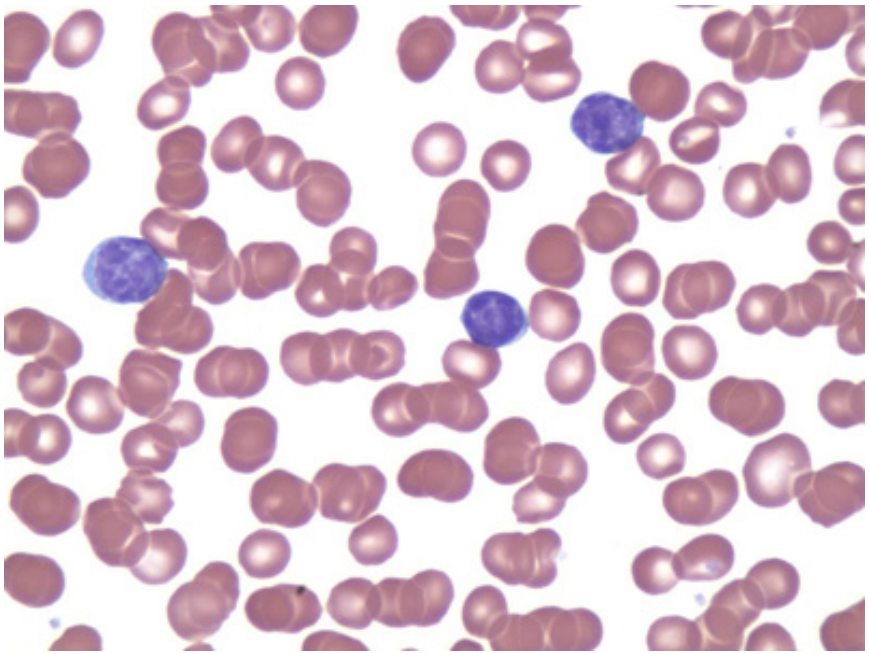
Jean-Philippe Vert



PRIB, Nice, France, June 18, 2013

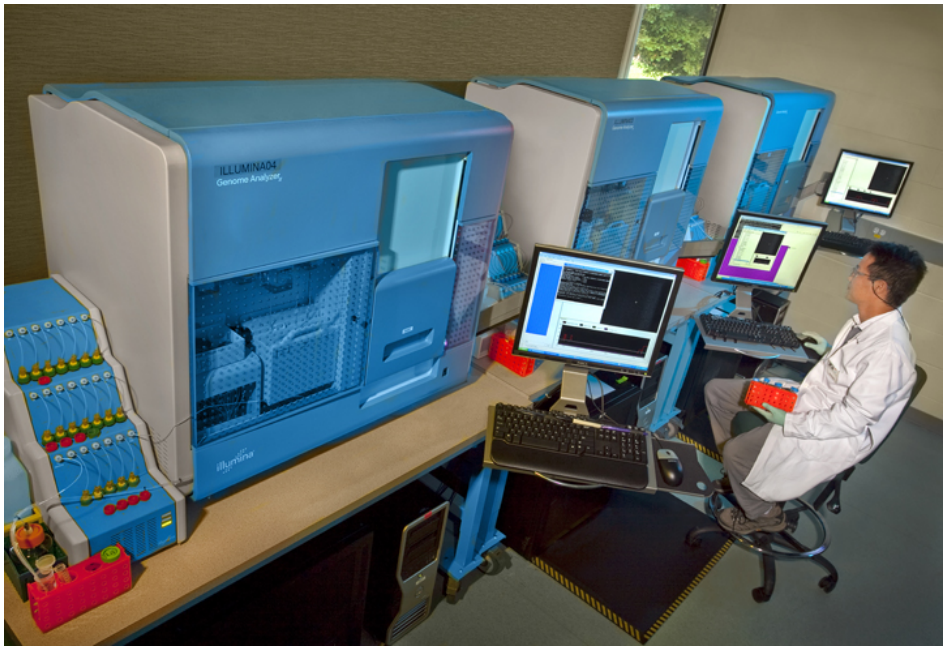
# Normal vs cancer cells



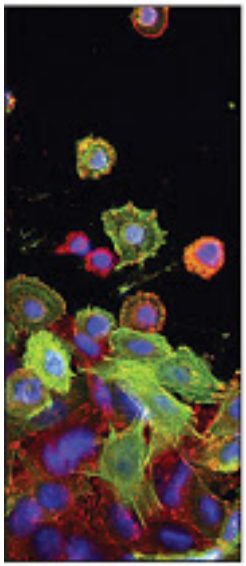


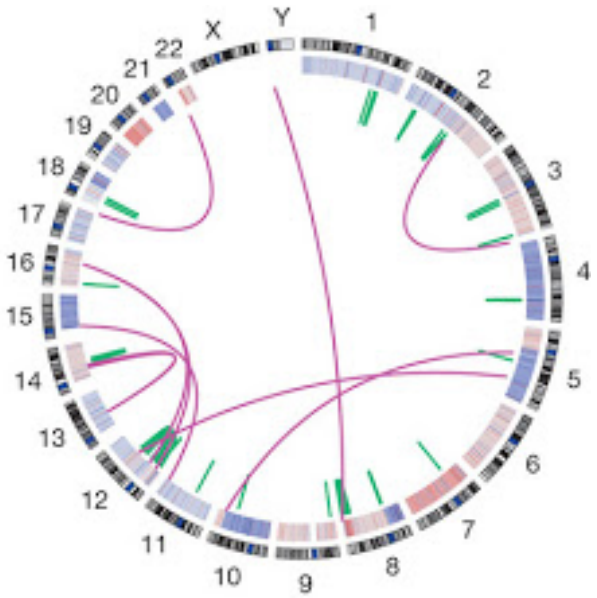
What goes wrong?
How to treat?

# Biology is now quantitative, "high-throughput"

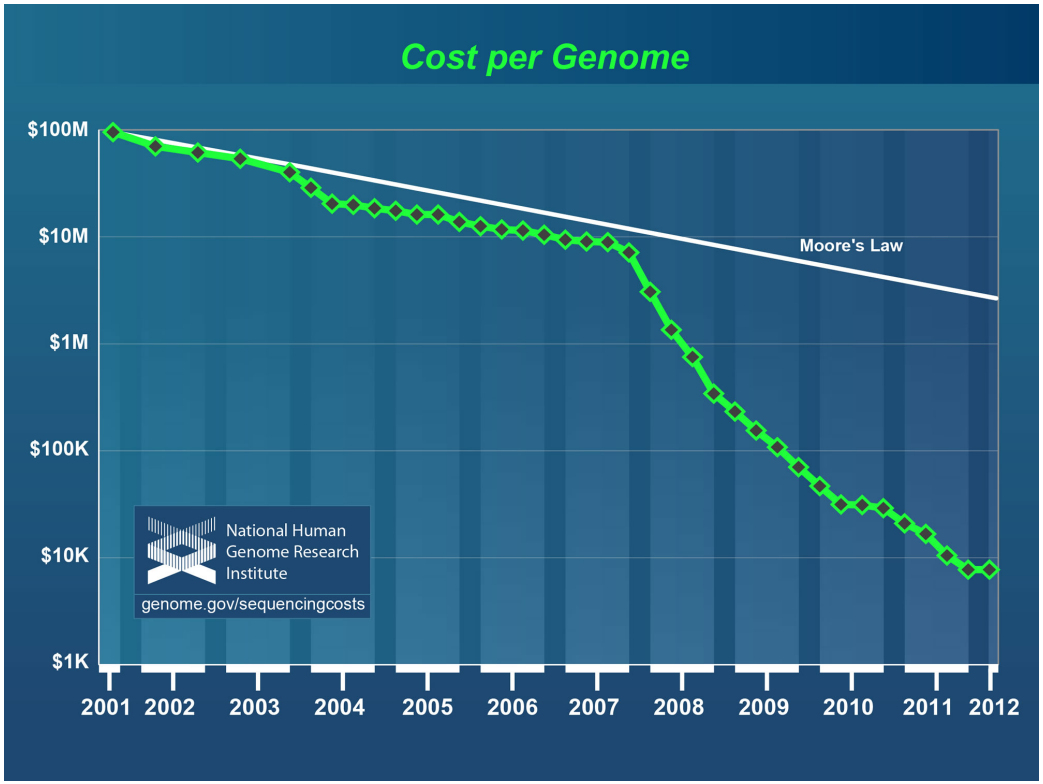



*DOE Joint Genome Institute*





# Big data in biology




- "The $1,000 genome, the $1 million interpretation" (B. Kopf)
- High-dimensional, heterogeneous, structured data. "Large $p$"
- `http://aws.amazon.com/1000genomes/`

# In this talk

$$\min_w R(w) + \lambda \Omega(w)$$

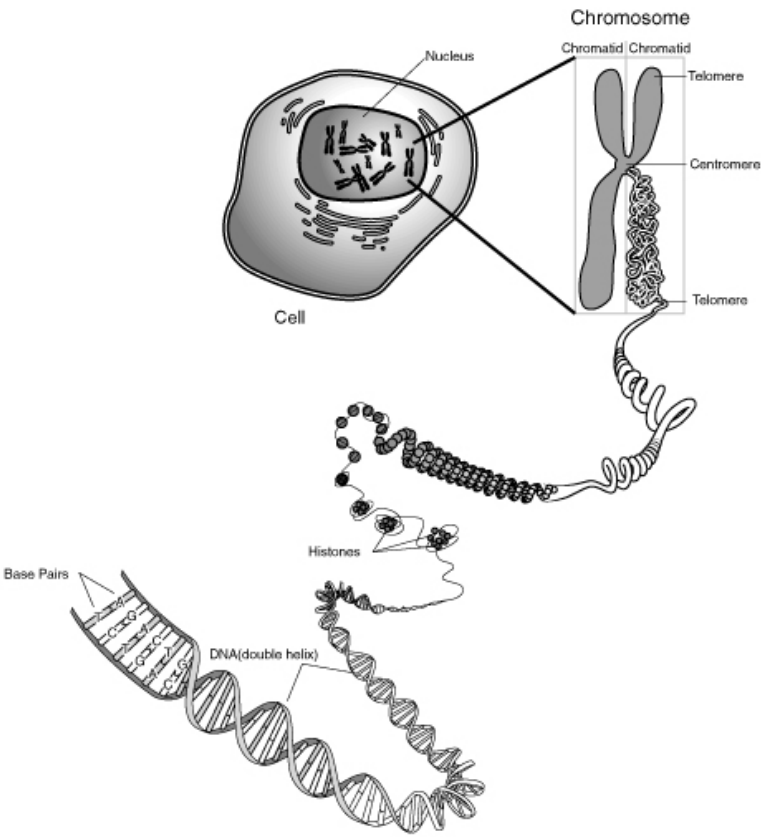
where:

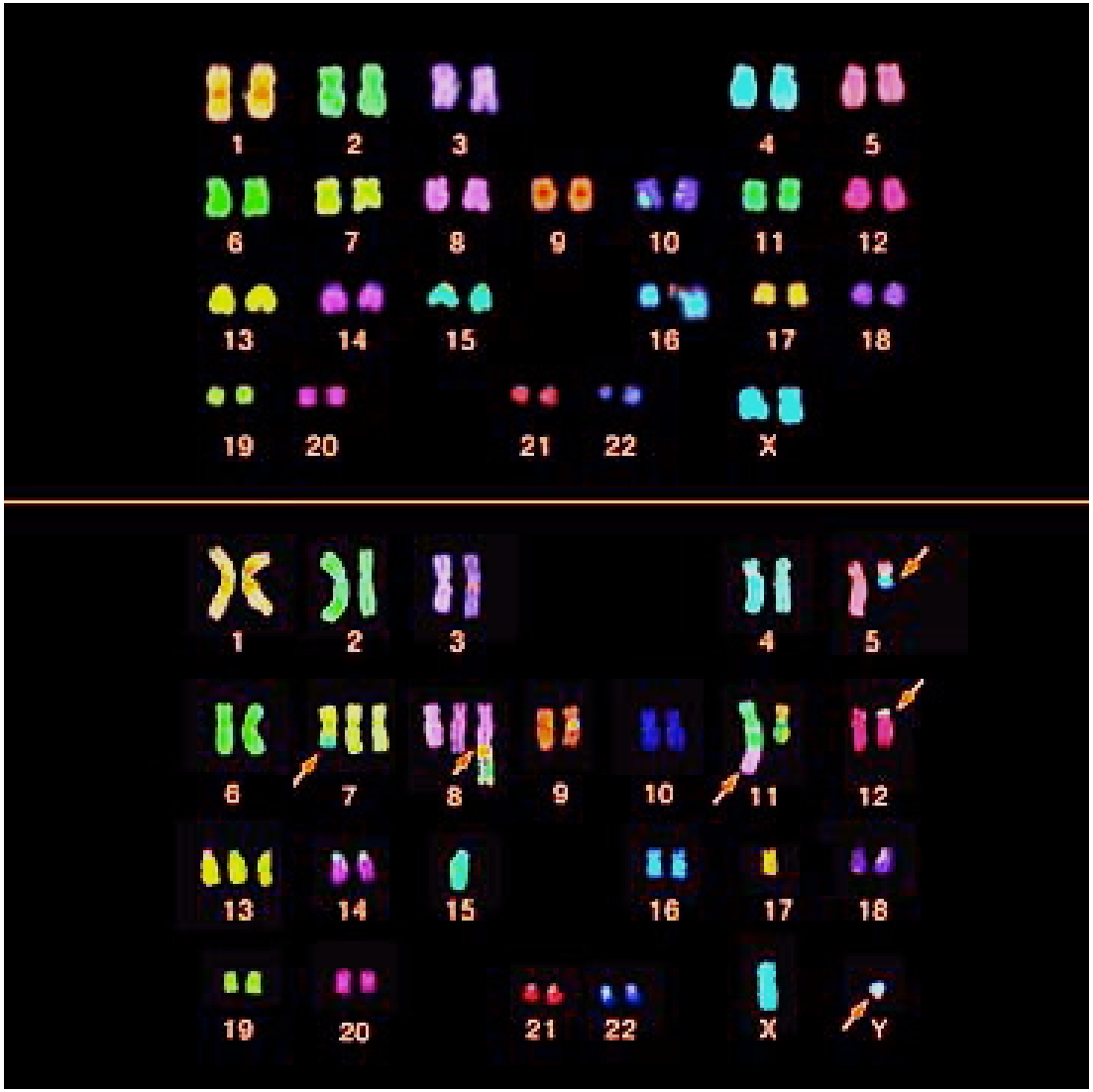
- $w$ is the hypothesis we want to infer from data
- $R(w)$ is a smooth convex "fitness" function
- $\Omega(w)$ is a non-smooth convex penalty, which favors particular solution

1. Mapping DNA breakpoints in cancer genomes
2. Isoform detection from RNA-seq data
3. Learning molecular classifiers with network information

# Outline

# Chromosomic aberrations in cancer

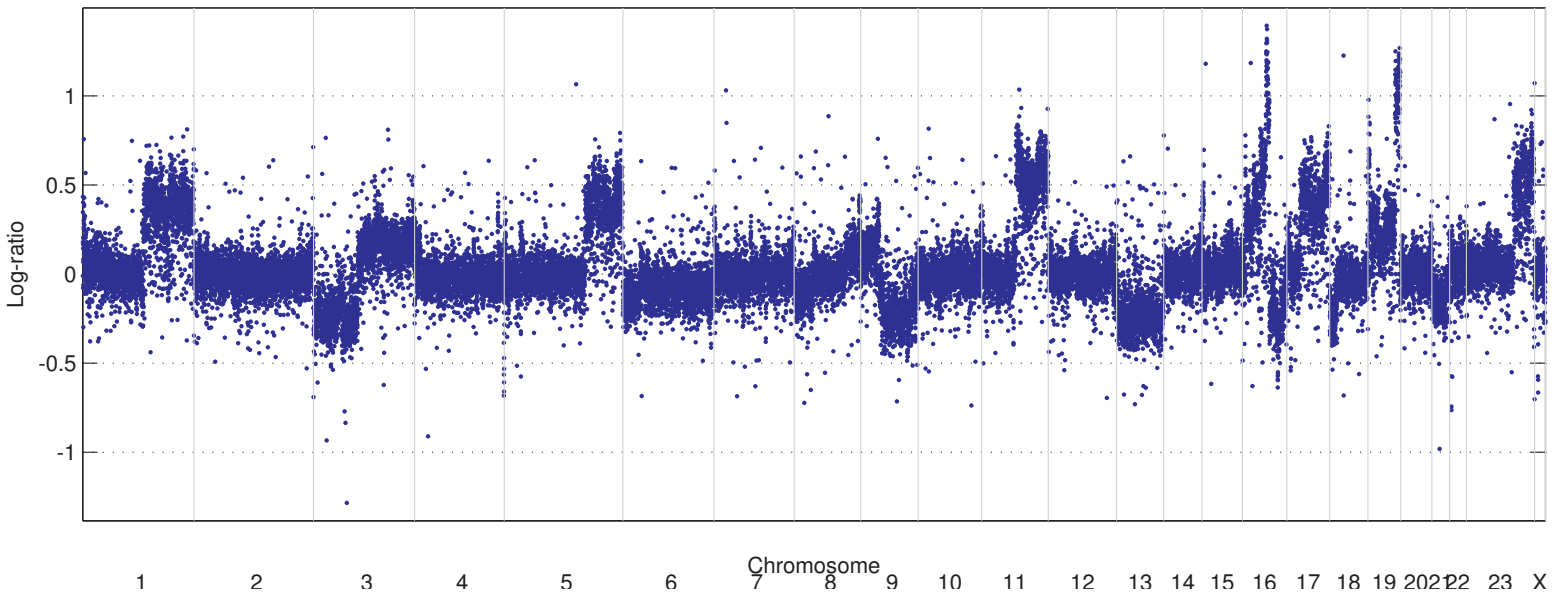


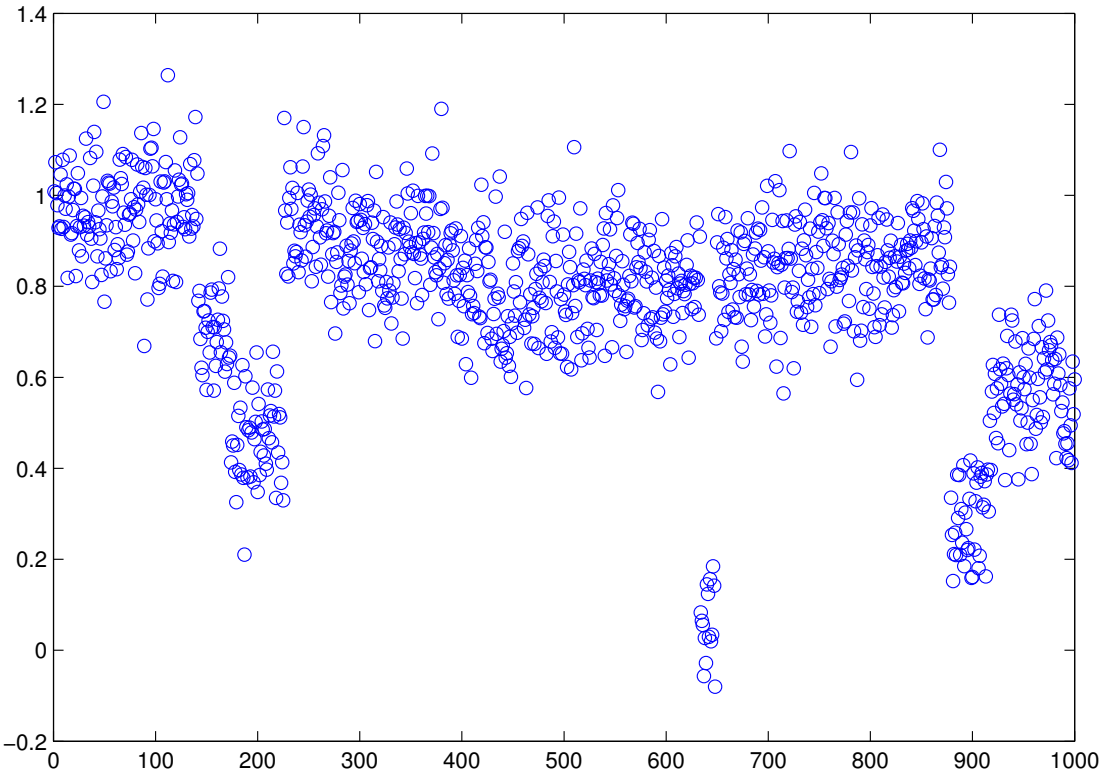
# Comparative Genomic Hybridization (CGH)

## Motivation

- Comparative genomic hybridization (CGH) data measure the DNA copy number along the genome
- Very useful, in particular in cancer research to observe systematically variants in DNA content

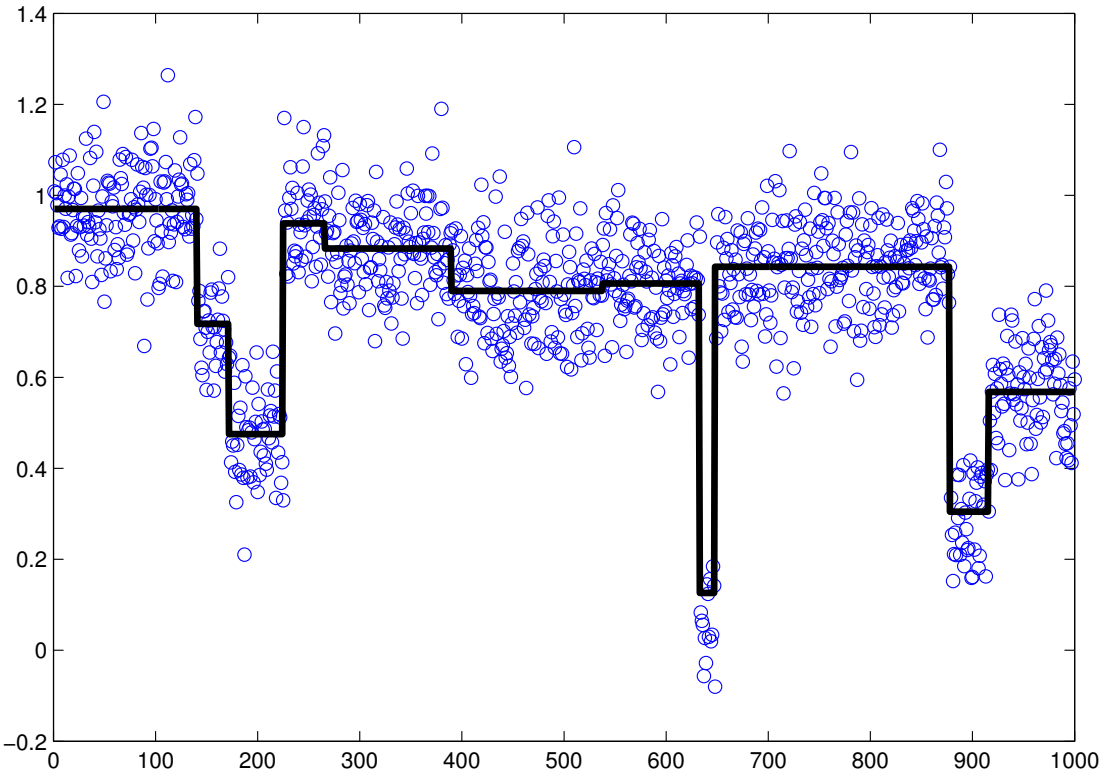



# Can we identify breakpoints and "smooth" each profile?



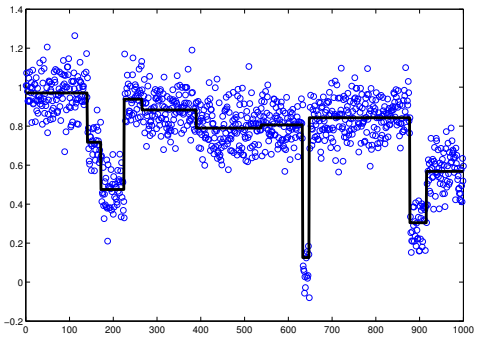

- A classical multiple change-point detection problem
- Should scale to lengths of order $10^6 \sim 10^9$

# Can we identify breakpoints and "smooth" each profile?




- A classical multiple change-point detection problem
- Should scale to lengths of order $10^6 \sim 10^9$

# An optimal solution

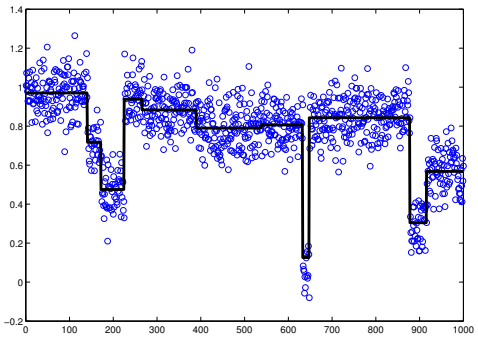



- For a signal $Y \in \mathbb{R}^p$, define an optimal approximation $\beta \in \mathbb{R}^p$ with $k$ breakpoints as the solution of

$$\min_{\beta \in \mathbb{R}^p} \| Y - \beta \|^2 \quad \text{such that} \quad \sum_{i=1}^{p-1} \mathbf{1}\left(U_{i+1} \neq U_i\right) \leq k$$

- This is an optimization problem over the $\binom{p}{k}$ partitions...
- Dynamic programming finds the solution in $O(p^2 k)$ in time and $O(p^2)$ in memory
- But: does not scale to $p = 10^6 \sim 10^9$...

# An optimal solution

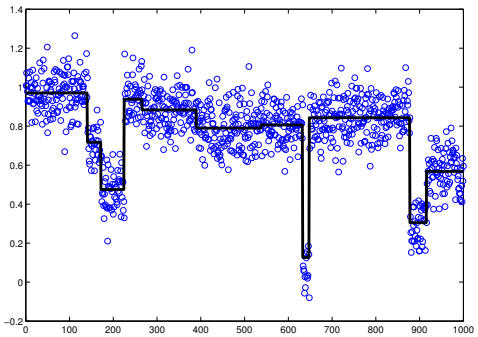



- For a signal $Y \in \mathbb{R}^p$, define an optimal approximation $\beta \in \mathbb{R}^p$ with $k$ breakpoints as the solution of

$$\min_{\beta \in \mathbb{R}^p} \| Y - \beta \|^2 \quad \text{such that} \quad \sum_{i=1}^{p-1} \mathbf{1}(U_{i+1} \neq U_i) \leq k$$

- This is an optimization problem over the $\binom{p}{k}$ partitions...
- Dynamic programming finds the solution in $O(p^2 k)$ in time and $O(p^2)$ in memory
- But: does not scale to $p = 10^6 \sim 10^9$...

# An optimal solution

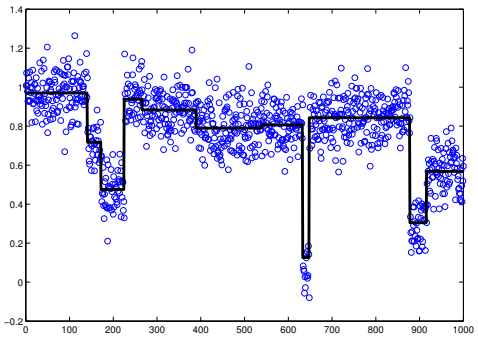



- For a signal $Y \in \mathbb{R}^p$, define an optimal approximation $\beta \in \mathbb{R}^p$ with $k$ breakpoints as the solution of

$$\min_{\beta \in \mathbb{R}^p} \| Y - \beta \|^2 \quad \text{such that} \quad \sum_{i=1}^{p-1} \mathbf{1}(U_{i+1} \neq U_i) \leq k$$

- This is an optimization problem over the $\binom{p}{k}$ partitions...
- Dynamic programming finds the solution in $O(p^2 k)$ in time and $O(p^2)$ in memory
- But: does not scale to $p = 10^6 \sim 10^9$...

# An optimal solution




- For a signal $Y \in \mathbb{R}^p$, define an optimal approximation $\beta \in \mathbb{R}^p$ with $k$ breakpoints as the solution of

$$\min_{\beta \in \mathbb{R}^p} \| Y - \beta \|^2 \quad \text{such that} \quad \sum_{i=1}^{p-1} \mathbf{1}(U_{i+1} \neq U_i) \leq k$$

- This is an optimization problem over the $\binom{p}{k}$ partitions...
- Dynamic programming finds the solution in $O(p^2 k)$ in time and $O(p^2)$ in memory
- But: does not scale to $p = 10^6 \sim 10^9$...

# Promoting sparsity with the $\ell_1$ penalty

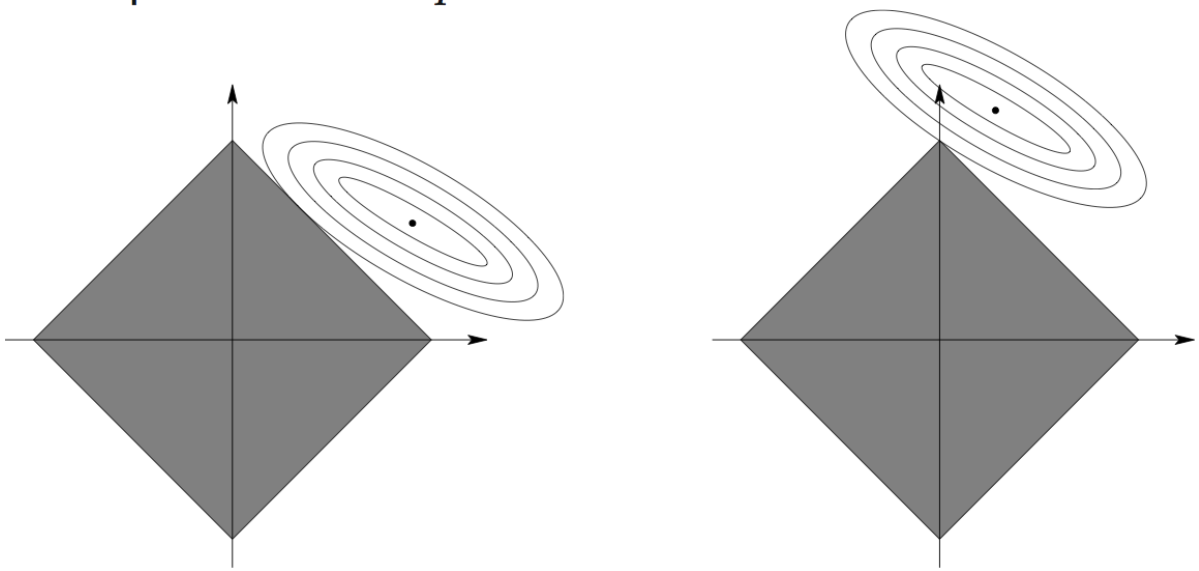
## The $\ell_1$ penalty (Tibshirani, 1996; Chen et al., 1998)

If $R(\beta)$ is convex and "smooth", the solution of

$$\min_{\beta \in \mathbb{R}^p} R(\beta) + \lambda \sum_{i=1}^{p} |\beta_i|$$

is usually sparse.

Geometric interpretation with $p = 2$



# Promoting piecewise constant profiles penalty

## The total variation / variable fusion penalty

If $R(\beta)$ is convex and "smooth", the solution of

$$\min_{\beta \in \mathbb{R}^p} R(\beta) + \lambda \sum_{i=1}^{p-1} |\beta_{i+1} - \beta_i|$$

is usually piecewise constant (Rudin et al., 1992; Land and Friedman, 1996).

Proof:

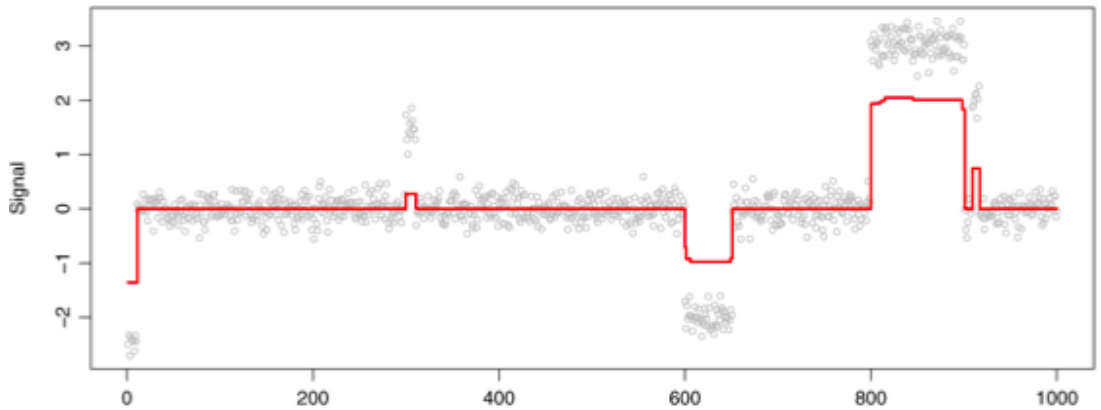
- Change of variable $u_i = \beta_{i+1} - \beta_i$, $u_0 = \beta_1$
- We obtain a Lasso problem in $u \in \mathbb{R}^{p-1}$
- $u$ sparse means $\beta$ piecewise constant

# TV signal approximator

$$\min_{\beta \in \mathbb{R}^p} \| Y - \beta \|^2 \quad \text{such that} \quad \sum_{i=1}^{p-1} | \beta_{i+1} - \beta_i | \leq \mu$$

Adding additional constraints does not change the change-points:

- $\sum_{i=1}^{p} | \beta_i | \leq \nu$ (Tibshirani et al., 2005; Tibshirani and Wang, 2008)
- $\sum_{i=1}^{p} \beta_i^2 \leq \nu$ (Mairal et al. 2010)



# Solving TV signal approximator

$$\min_{\beta \in \mathbb{R}^p} \| Y - \beta \|^2 \quad \text{such that} \quad \sum_{i=1}^{p-1} | \beta_{i+1} - \beta_i | \leq \mu$$

- QP with sparse linear constraints in $O(p^2)$ -> 135 min for $p = 10^5$ (Tibshirani and Wang, 2008)
- Coordinate descent-like method $O(p)$? -> 3s s for $p = 10^5$ (Friedman et al., 2007)
- For all $\mu$ with the LARS in $O(pK)$ (Harchaoui and Levy-Leduc, 2008)
- For all $\mu$ in $O(p \ln p)$ (Hoefling, 2009)
- For the first $K$ change-points in $O(p \ln K)$ (Bleakley and V., 2010)

# TV signal approximator as dichotomic segmentation

**Algorithm 1** Greedy dichotomic segmentation

**Require:** $k$ number of intervals, $\gamma(I)$ gain function to split an interval $I$ into $I_L(I), I_R(I)$

1: $I_0$ represents the interval $[1, n]$
2: $\mathcal{P} = \{I_0\}$
3: **for** $i = 1$ to $k$ **do**
4: $\quad I^* \leftarrow \arg\max_{I \in \mathcal{P}} \gamma(I^*)$
5: $\quad \mathcal{P} \leftarrow \mathcal{P} \backslash \{I^*\}$
6: $\quad \mathcal{P} \leftarrow \mathcal{P} \cup \{I_L(I^*), I_R(I^*)\}$
7: **end for**
8: **return** $\mathcal{P}$

**Theorem (V. and Bleakley, 2010; see also Hoefling, 2009)**

TV signal approximator performs "greedy" dichotomic segmentation

Apparently greedy algorithm finds the global optimum!

# TV signal approximator as dichotomic segmentation

**Algorithm 1** Greedy dichotomic segmentation

**Require:** $k$ number of intervals, $\gamma(I)$ gain function to split an interval $I$ into $I_L(I), I_R(I)$

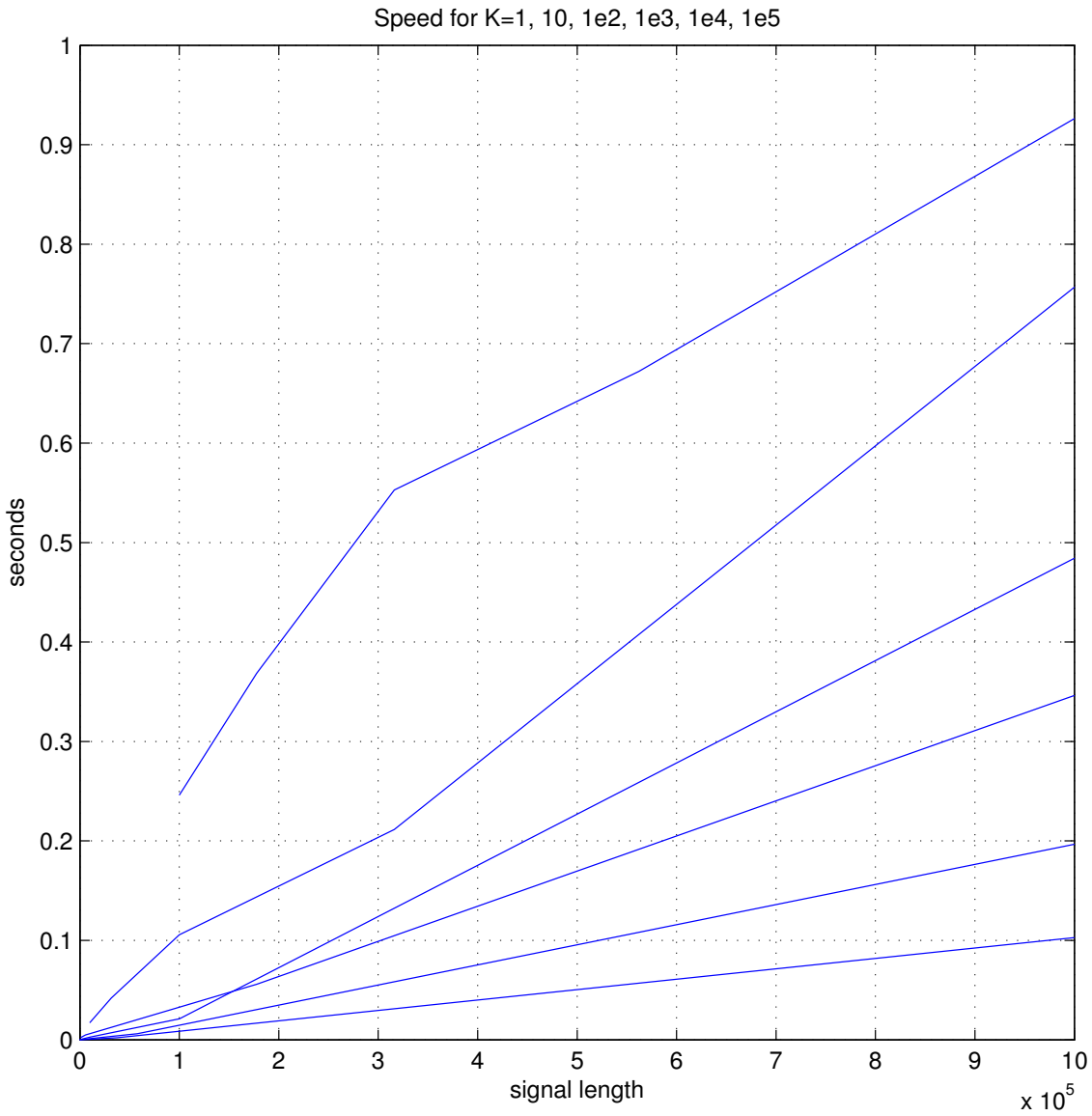
1: $I_0$ represents the interval $[1, n]$
2: $\mathcal{P} = \{I_0\}$
3: **for** $i = 1$ to $k$ **do**
4: $\quad I^* \leftarrow \arg\max_{I \in \mathcal{P}} \gamma(I^*)$
5: $\quad \mathcal{P} \leftarrow \mathcal{P} \backslash \{I^*\}$
6: $\quad \mathcal{P} \leftarrow \mathcal{P} \cup \{I_L(I^*), I_R(I^*)\}$
7: **end for**
8: **return** $\mathcal{P}$

**Theorem (V. and Bleakley, 2010; see also Hoefling, 2009)**

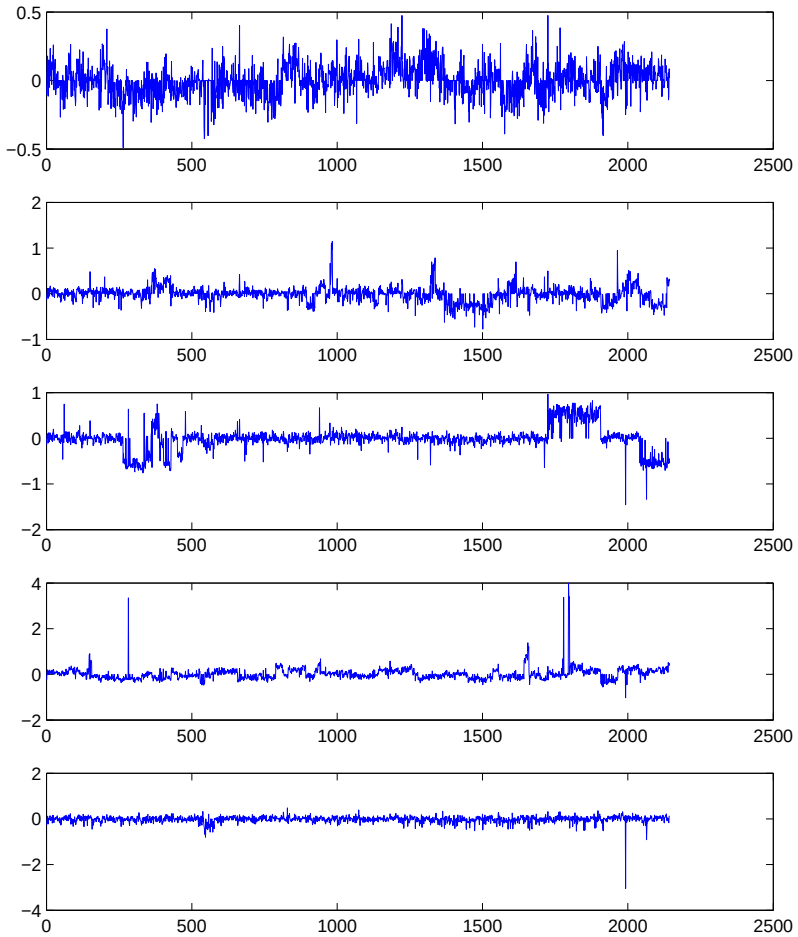
TV signal approximator performs "greedy" dichotomic segmentation

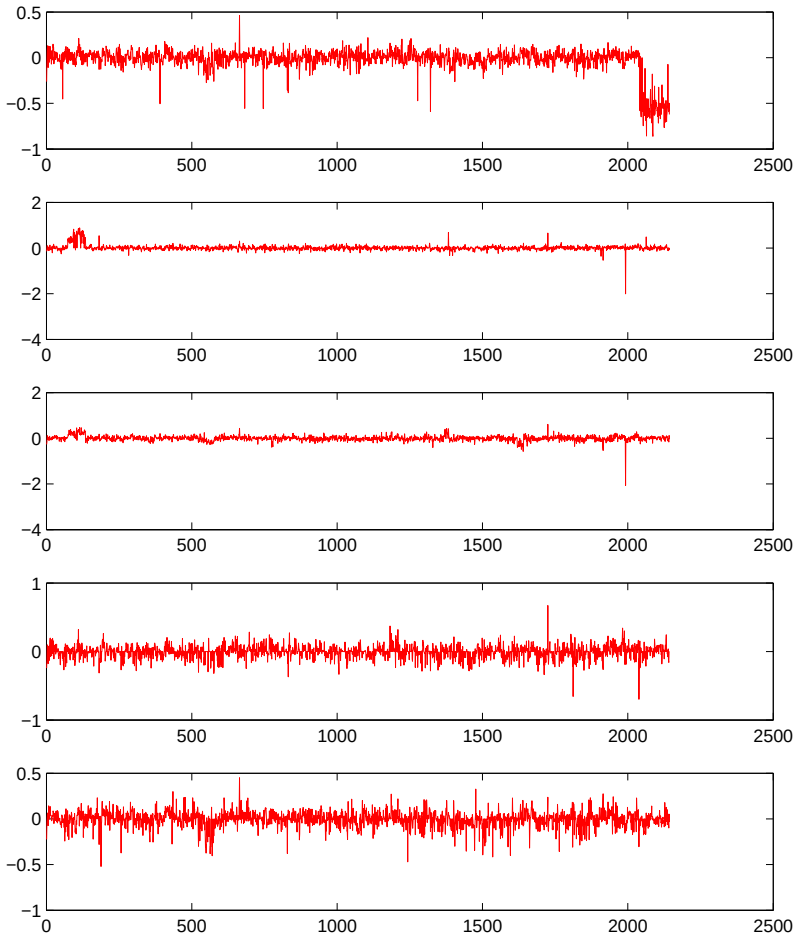
Apparently greedy algorithm finds the global optimum!

# Speed trial : 2 s. for $K = 100$, $p = 10^7$



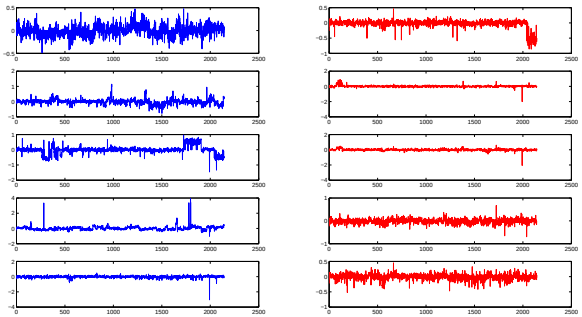
# Extension: cancer prognosis




*Aggressive (left) vs non-aggressive (right) melanoma*

# Fused lasso for supervised classification

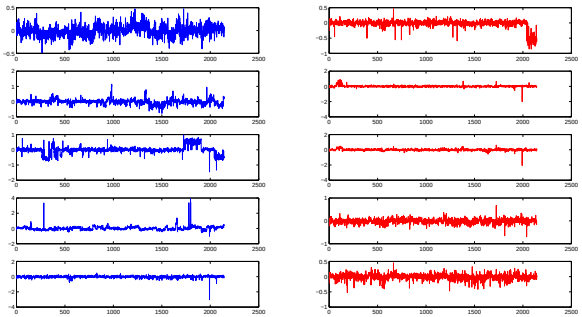


- Idea: find a linear predictor $f(Y) = \beta^\top Y$ that best discriminates the aggressive vs non-aggressive samples, subject to the constraints that it should be sparse and piecewise constant
- Mathematically:

$$\min_{\beta \in \mathbb{R}^p} R(\beta) + \lambda_1 \| \beta \|_1 + \lambda_2 \| \beta \|_{TV}$$

- Computationnally: this is convex optimization problem that can be solved very efficiently with proximal optimization methods (V. and Bleakley, 2012)

# Fused lasso for supervised classification

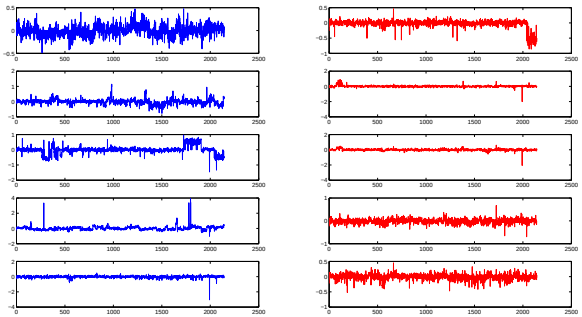


- Idea: find a linear predictor $f(Y) = \beta^\top Y$ that best discriminates the aggressive vs non-aggressive samples, subject to the constraints that it should be sparse and piecewise constant
- Mathematically:

$$\min_{\beta \in \mathbb{R}^p} R(\beta) + \lambda_1 \| \beta \|_1 + \lambda_2 \| \beta \|_{TV}$$

- Computationnally: this is convex optimization problem that can be solved very efficiently with proximal optimization methods (V. and Bleakley, 2012)

# Fused lasso for supervised classification

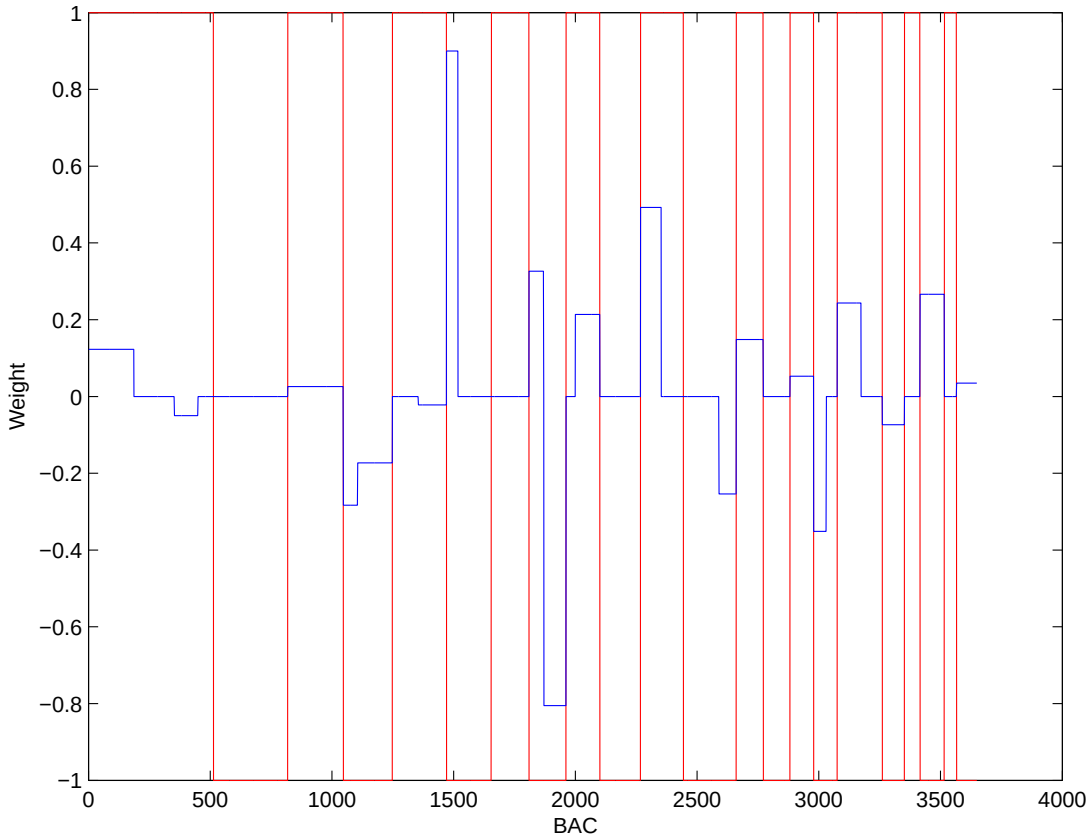


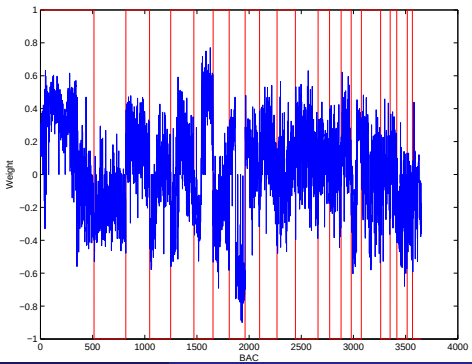
- Idea: find a linear predictor $f(Y) = \beta^\top Y$ that best discriminates the aggressive vs non-aggressive samples, subject to the constraints that it should be sparse and piecewise constant
- Mathematically:

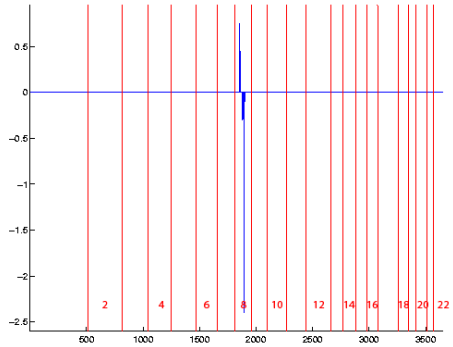
$$\min_{\beta \in \mathbb{R}^p} R(\beta) + \lambda_1 \| \beta \|_1 + \lambda_2 \| \beta \|_{TV}$$

- Computationnally: this is convex optimization problem that can be solved very efficiently with proximal optimization methods (V. and Bleakley, 2012)

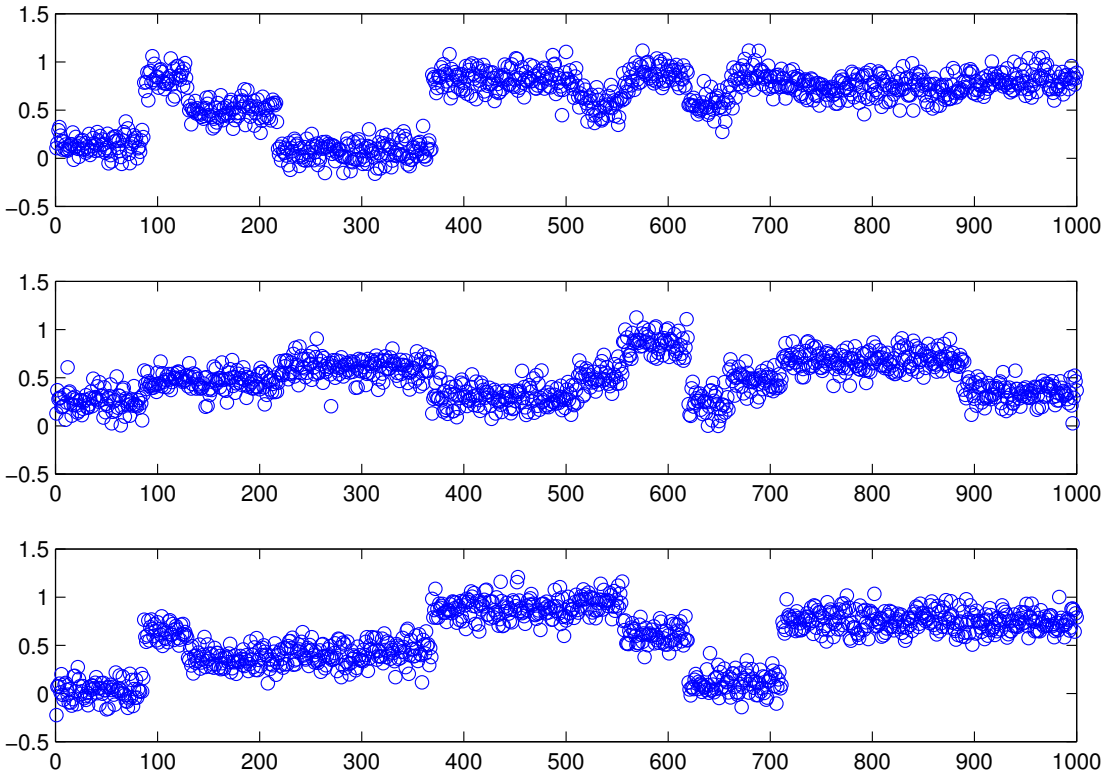
# Prognostic in melanoma (Rapaport et al., 2008)



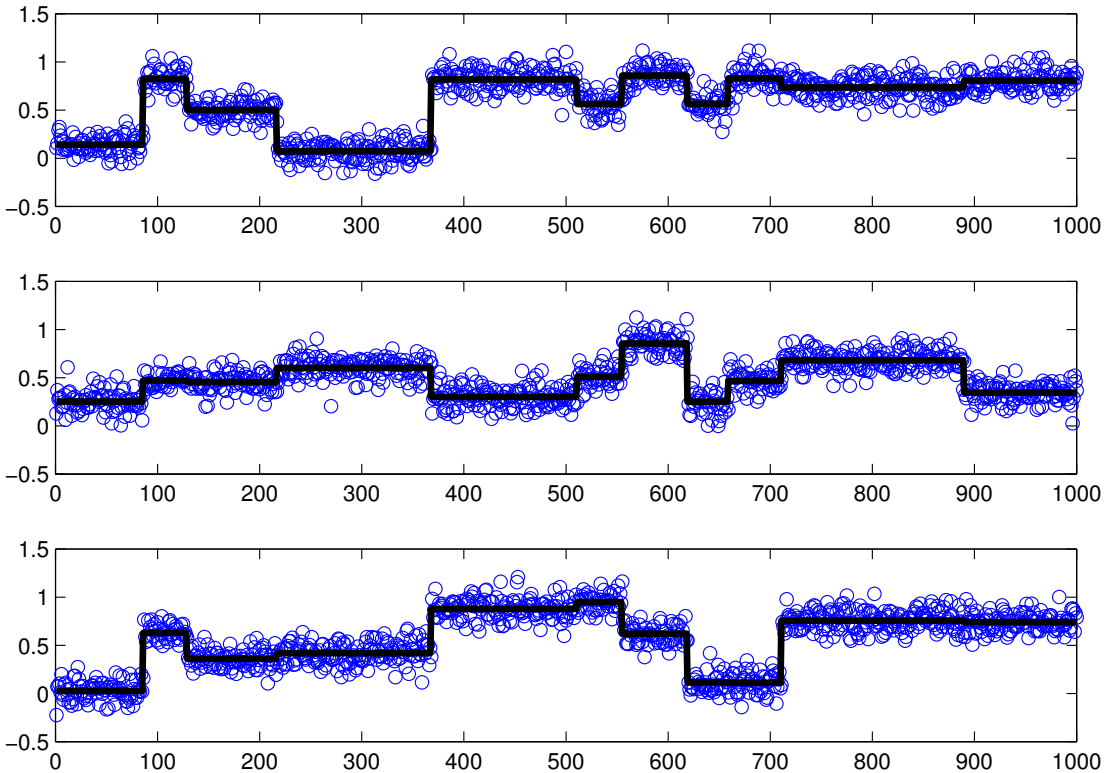





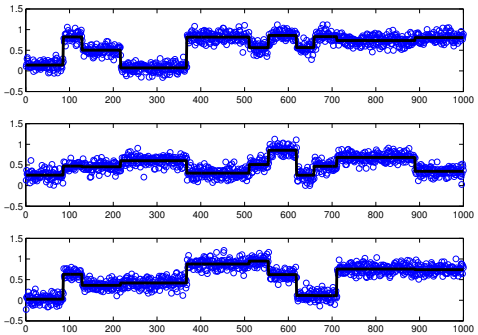
# Extension: finding multiple change points shared by several profiles



# Extension: finding multiple change points shared by several profiles



# "Optimal" segmentation by dynamic programming



- Define the "optimal" piecewise constant approximation $\hat{U} \in \mathbb{R}^{p \times n}$ of $Y$ as the solution of

$$\min_{U \in \mathbb{R}^{p \times n}} \| Y - U \|^2 \quad \text{such that} \quad \sum_{i=1}^{p-1} \mathbf{1}\left(U_{i+1,\bullet} \neq U_{i,\bullet}\right) \leq k$$

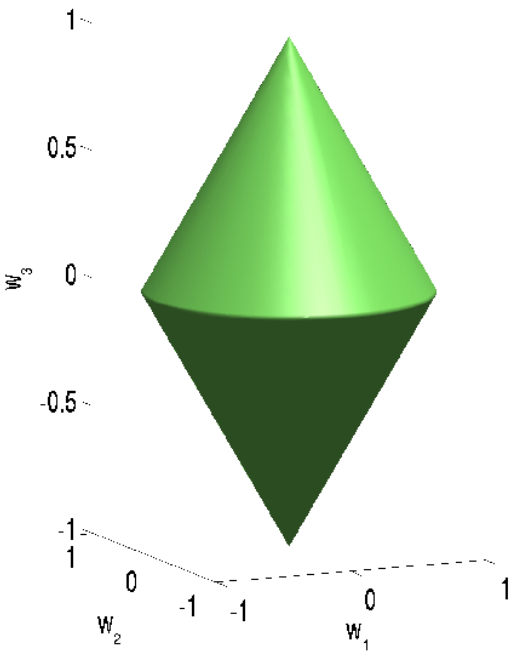
- DP finds the solution in $O(p^2 kn)$ in time and $O(p^2)$ in memory
- But: does not scale to $p = 10^6 \sim 10^9$...

# Selecting pre-defined groups of variables

## Group lasso (Yuan & Lin, 2006)

If groups of covariates are likely to be selected together, the $\ell_1/\ell_2$-norm induces sparse solutions *at the group level*:

$$\Omega_{group}(w) = \sum_g \|w_g\|_2$$




$$\begin{aligned}\Omega(w_1, w_2, w_3) &= \|(w_1, w_2)\|_2 + \|w_3\|_2 \\ &= \sqrt{w_1^2 + w_2^2} + \sqrt{w_3^2}\end{aligned}$$

# GFLseg (Bleakley and V., 2011)

Replace

$$\min_{U \in \mathbb{R}^{p \times n}} \| Y - U \|^2 \quad \text{such that} \quad \sum_{i=1}^{p-1} \mathbf{1}\left(U_{i+1,\bullet} \neq U_{i,\bullet}\right) \leq k$$

by

$$\min_{U \in \mathbb{R}^{p \times n}} \| Y - U \|^2 \quad \text{such that} \quad \sum_{i=1}^{p-1} w_i \| U_{i+1,\bullet} - U_{i,\bullet} \| \leq \mu$$

GFLseg = Group Fused Lasso segmentation

**Questions**

- Practice: can we solve it efficiently?
- Theory: does it recover the correct segmentation?

# GFLseg (Bleakley and V., 2011)

Replace

$$\min_{U \in \mathbb{R}^{p \times n}} \| Y - U \|^2 \quad \text{such that} \quad \sum_{i=1}^{p-1} \mathbf{1} \left( U_{i+1,\bullet} \neq U_{i,\bullet} \right) \leq k$$

by

$$\min_{U \in \mathbb{R}^{p \times n}} \| Y - U \|^2 \quad \text{such that} \quad \sum_{i=1}^{p-1} w_i \| U_{i+1,\bullet} - U_{i,\bullet} \| \leq \mu$$

GFLseg = Group Fused Lasso segmentation

## Questions

- Practice: can we solve it efficiently?
- Theory: does it recover the correct segmentation?

# GFLseg as a group Lasso problem

- Make the change of variables:

$$\begin{aligned} \gamma &= U_{1,\bullet} \,, \\ \beta_{i,\bullet} &= w_i \left( U_{i+1,\bullet} - U_{i,\bullet} \right) \quad \text{for } i = 1, \ldots, p-1 \,. \end{aligned}$$

- TV approximator is then equivalent to the following group Lasso problem (Yuan and Lin, 2006):

$$\min_{\beta \in \mathbb{R}^{(p-1) \times n}} \| \bar{Y} - \bar{X} \beta \|^2 + \lambda \sum_{i=1}^{p-1} \| \beta_{i,\bullet} \| \,,$$

where $\bar{Y}$ is the centered signal matrix and $\bar{X}$ is a particular $(p-1) \times (p-1)$ design matrix.

# TV approximator implementation

$$\min_{\beta \in \mathbb{R}^{(p-1)\times n}} \| \bar{Y} - \bar{X}\beta \|^2 + \lambda \sum_{i=1}^{p-1} \| \beta_{i,\bullet} \| ,$$

## Theorem

The TV approximator can be solved efficiently:

- approximately with the group LARS in $O(npk)$ in time and $O(np)$ in memory
- exactly with a block coordinate descent + active set method in $O(np)$ in memory

# Proof: computational tricks...

Although $\bar{X}$ is $(p-1) \times (p-1)$:

- For any $R \in \mathbb{R}^{p \times n}$, we can compute $C = \bar{X}^\top R$ in $O(np)$ operations and memory
- For any two subset of indices $A = (a_1, \ldots, a_{|A|})$ and $B = (b_1, \ldots, b_{|B|})$ in $[1, p-1]$, we can compute $\bar{X}^\top_{\bullet,A} \bar{X}_{\bullet,B}$ in $O(|A||B|)$ in time and memory
- For any $A = (a_1, \ldots, a_{|A|})$, set of distinct indices with $1 \leq a_1 < \ldots < a_{|A|} \leq p-1$, and for any $|A| \times n$ matrix $R$, we can compute $C = \left(\bar{X}^\top_{\bullet,A} \bar{X}_{\bullet,A}\right)^{-1} R$ in $O(|A|n)$ in time and memory

# Speed trial

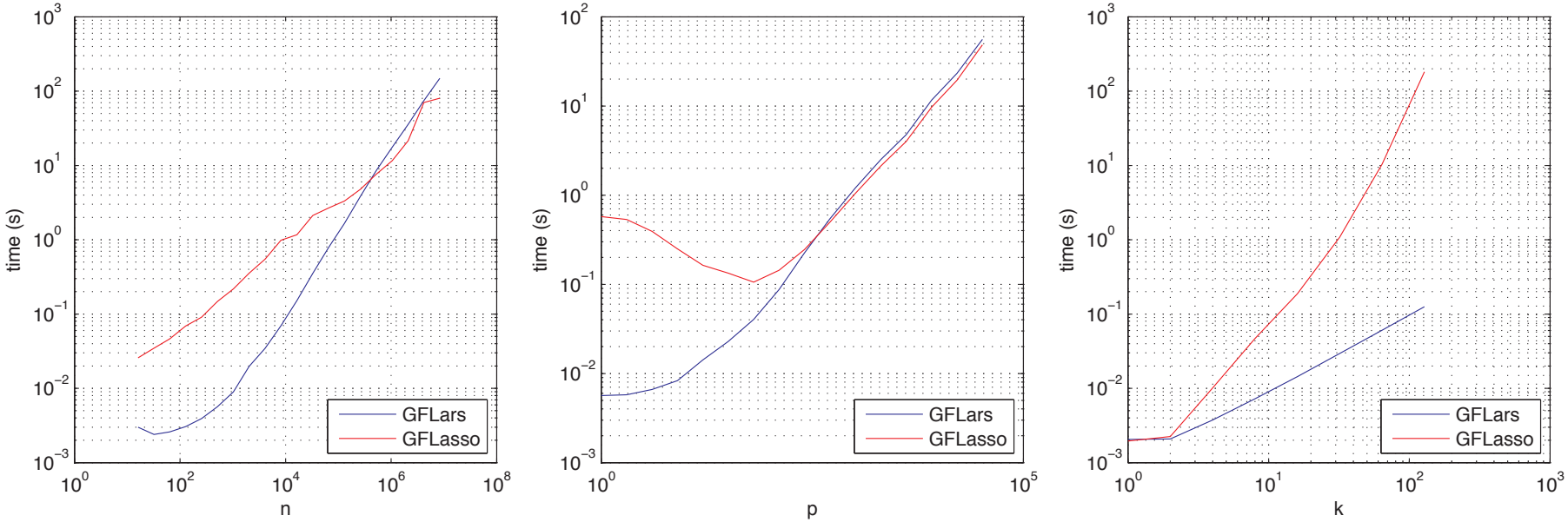
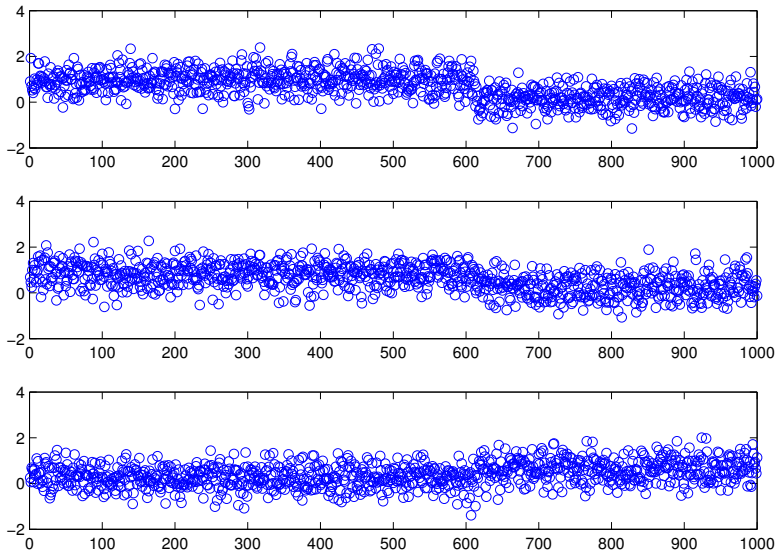



Figure 2: **Speed trials for group fused LARS (top row) and Lasso (bottom row).** *Left column:* varying $n$, with fixed $p = 10$ and $k = 10$; *center column:* varying $p$, with fixed $n = 1000$ and $k = 10$; *right column:* varying $k$, with fixed $n = 1000$ and $p = 10$. Figure axes are log-log. Results are averaged over 100 trials.

# Consistency

Suppose a single change-point:

- at position $u = \alpha p$
- with increments $(\beta_i)_{i=1,\ldots,n}$ s.t. $\bar{\beta}^2 = \lim_{k\to\infty} \frac{1}{n} \sum_{i=1}^{n} \beta_i^2$
- corrupted by i.i.d. Gaussian noise of variance $\sigma^2$




Does the TV approximator correctly estimate the first change-point as $p$ increases?

# Consistency of the unweighted TV approximator

$$\min_{U \in \mathbb{R}^{p \times n}} \| Y - U \|^2 \quad \text{such that} \quad \sum_{i=1}^{p-1} \| U_{i+1,\bullet} - U_{i,\bullet} \| \leq \mu$$

## Theorem

*The unweighted TV approximator finds the correct change-point with probability tending to* 1 *(resp. 0) as* $n \to +\infty$ *if* $\sigma^2 < \tilde{\sigma}_\alpha^2$ *(resp.* $\sigma^2 > \tilde{\sigma}_\alpha^2$*), where*

$$\tilde{\sigma}_\alpha^2 = p\bar{\beta}^2 \frac{(1-\alpha)^2(\alpha - \frac{1}{2p})}{\alpha - \frac{1}{2} - \frac{1}{2p}} .$$

- correct estimation on $[p\epsilon, p(1-\epsilon)]$ with $\epsilon = \sqrt{\frac{\sigma^2}{2p\bar{\beta}^2}} + o(p^{-1/2})$ .
- wrong estimation near the boundaries

# Consistency of the weighted TV approximator

$$\min_{U \in \mathbb{R}^{p \times n}} \| Y - U \|^2 \quad \text{such that} \quad \sum_{i=1}^{p-1} w_i \| U_{i+1,\bullet} - U_{i,\bullet} \| \leq \mu$$

## Theorem

*The weighted TV approximator with weights*

$$\forall i \in [1, p-1] \,, \quad w_i = \sqrt{\frac{i(p-i)}{p}}$$

*correctly finds the first change-point with probability tending to* 1 *as* $n \to +\infty$.

- we see the benefit of increasing $n$
- we see the benefit of adding weights to the TV penalty

# Proof sketch

- The first change-point $\hat{i}$ found by TV approximator maximizes $F_i = \| \hat{c}_{i,\bullet} \|^2$, where

$$\hat{c} = \bar{X}^\top \bar{Y} = \bar{X}^\top \bar{X} \beta^* + \bar{X}^\top W \,.$$

- $\hat{c}$ is Gaussian, and $F_i$ is follows a non-central $\chi^2$ distribution with

$$G_i = \frac{EF_i}{p} = \frac{i(p-i)}{pw_i^2}\sigma^2 + \frac{\bar{\beta}^2}{w_i^2 w_u^2 p^2} \times \begin{cases} i^2 (p-u)^2 & \text{if } i \leq u \,, \\ u^2 (p-i)^2 & \text{otherwise.} \end{cases}$$

- We then just check when $G_u = \max_i G_i$

# Consistency for a single change-point

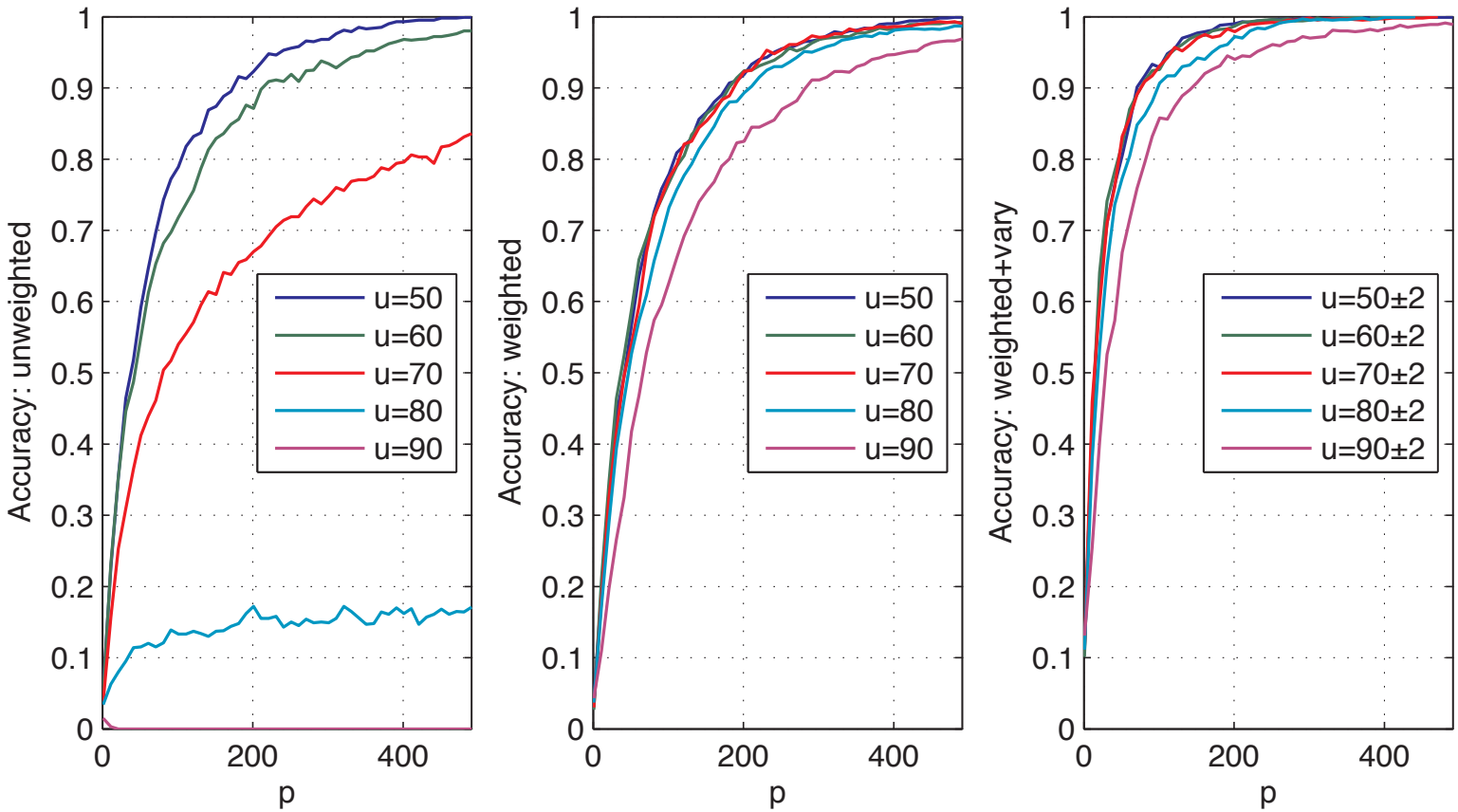



Figure 3: **Single change-point accuracy for the group fused Lasso.** Accuracy as a function of the number of profiles $p$ when the change-point is placed in a variety of positions $u = 50$ to $u = 90$ (left and centre plots, resp. unweighted and weighted group fused Lasso), or: $u = 50 \pm 2$ to $u = 90 \pm 2$ (right plot, weighted with varying change-point location), for a signal of length 100.

# Estimation of several change-points

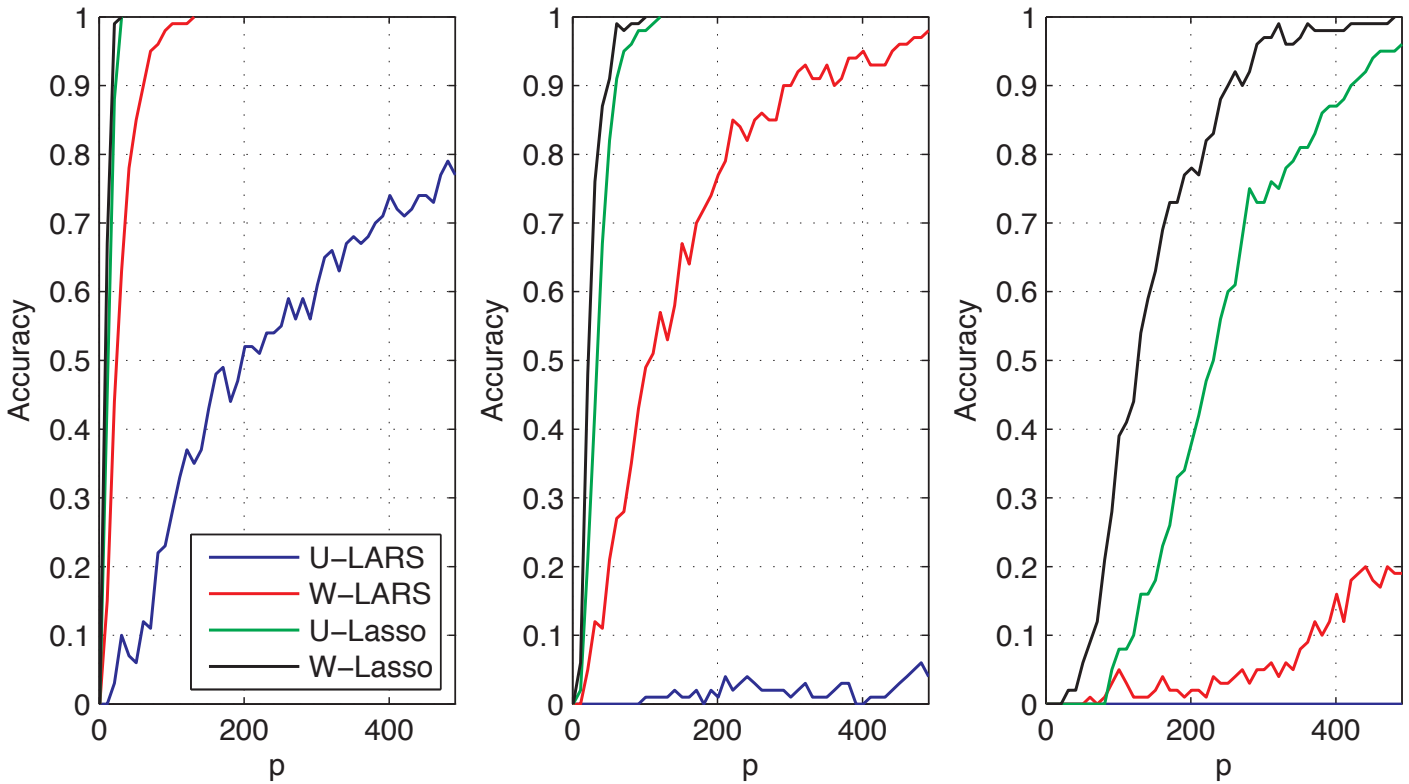
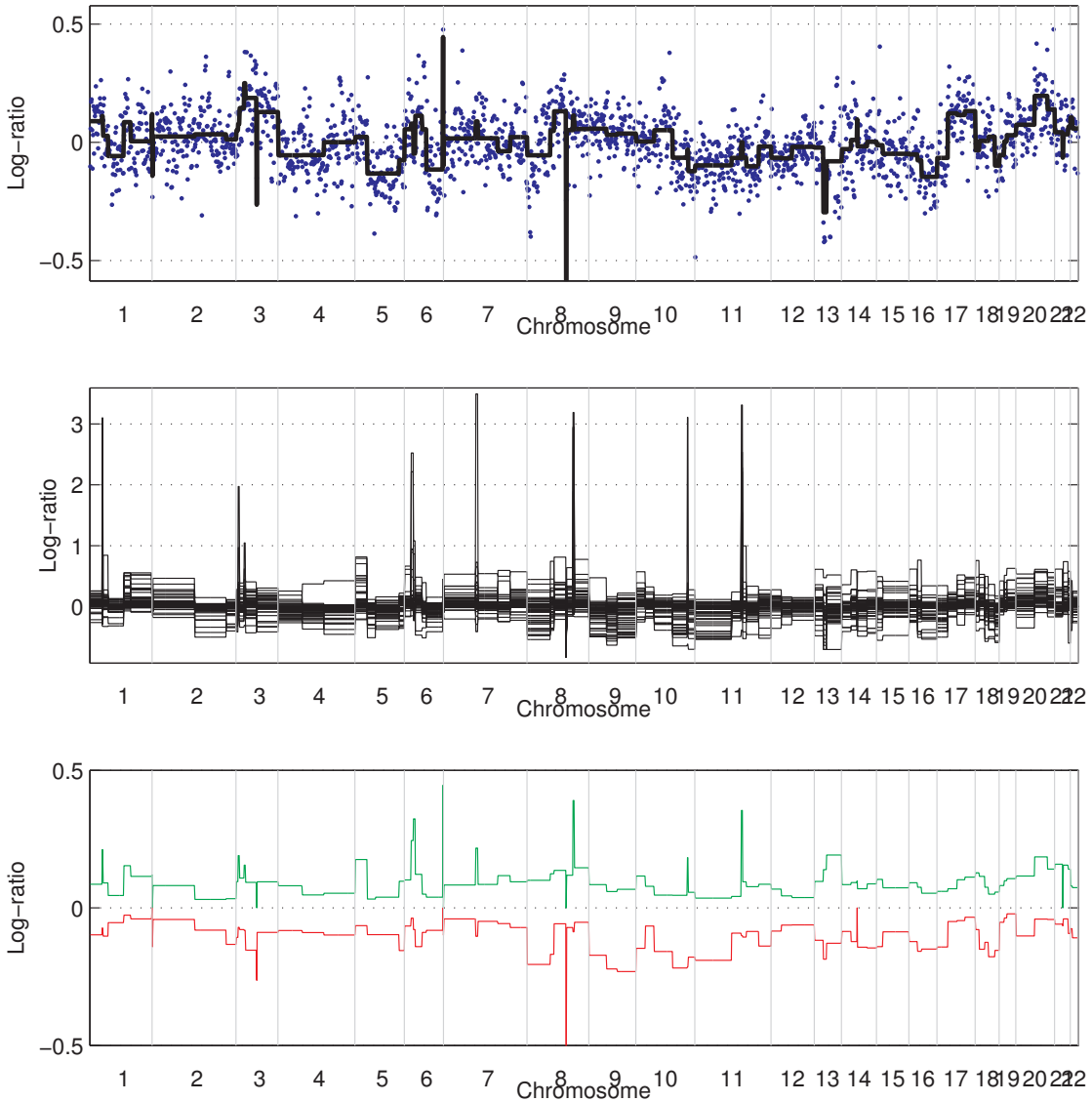



Figure 4: **Multiple change-point accuracy.** Accuracy as a function of the number of profiles $p$ when change-points are placed at the nine positions $\{10, 20, \ldots, 90\}$ and the variance $\sigma^2$ of the centered Gaussian noise is either $0.05$ (left), $0.2$ (center) and $1$ (right). The profile length is $100$.

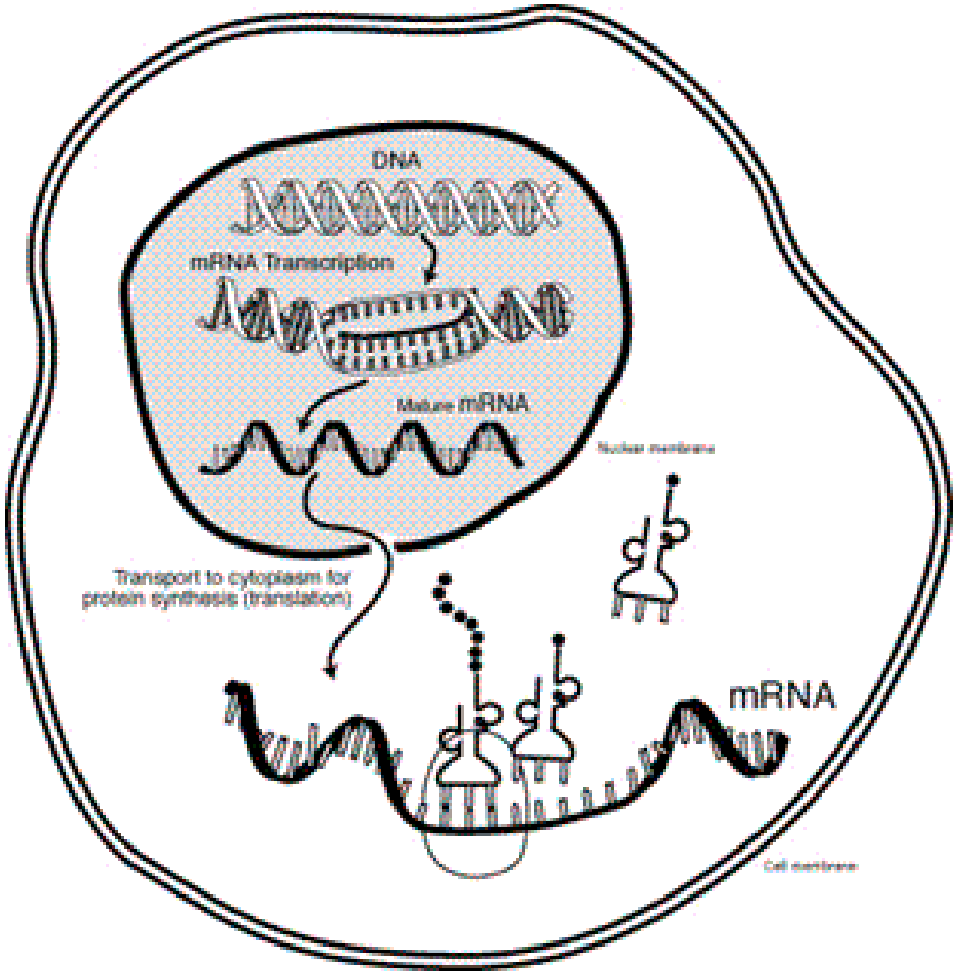
# Application: detection of frequent abnormalities

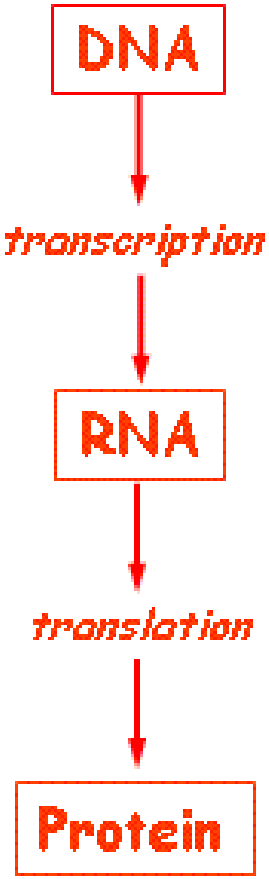


# Outline

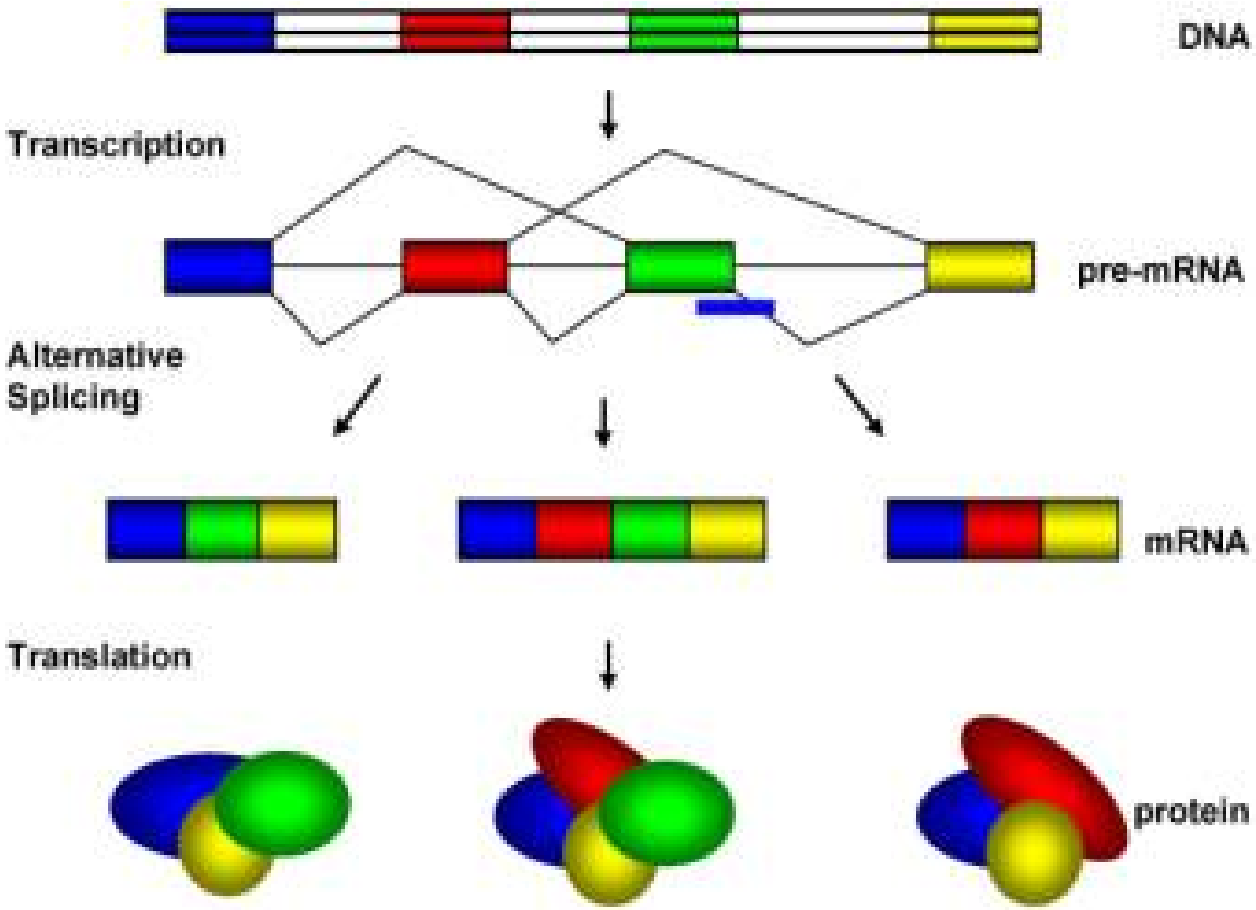
1. Mapping DNA breakpoints in cancer genomes
2. Isoform detection from RNA-seq data
3. Learning molecular classifiers with network information

# Central dogma

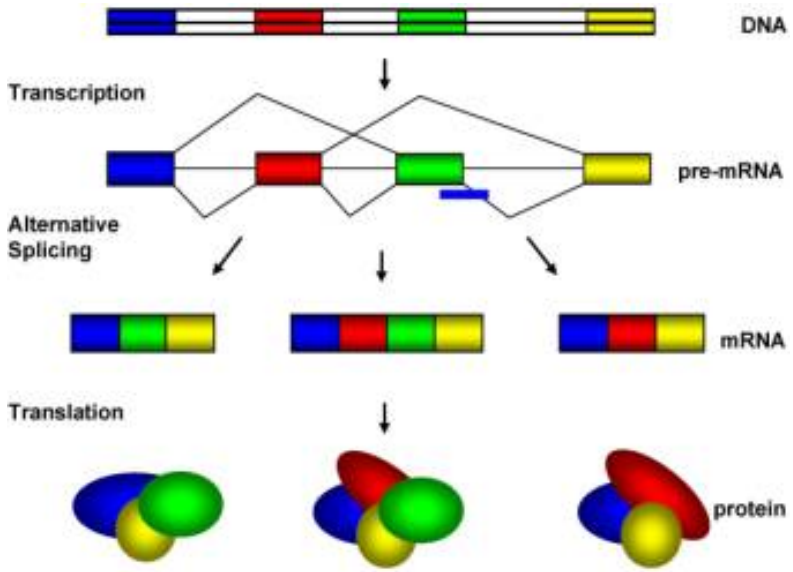


# Alternative splicing: 1 gene = many proteins



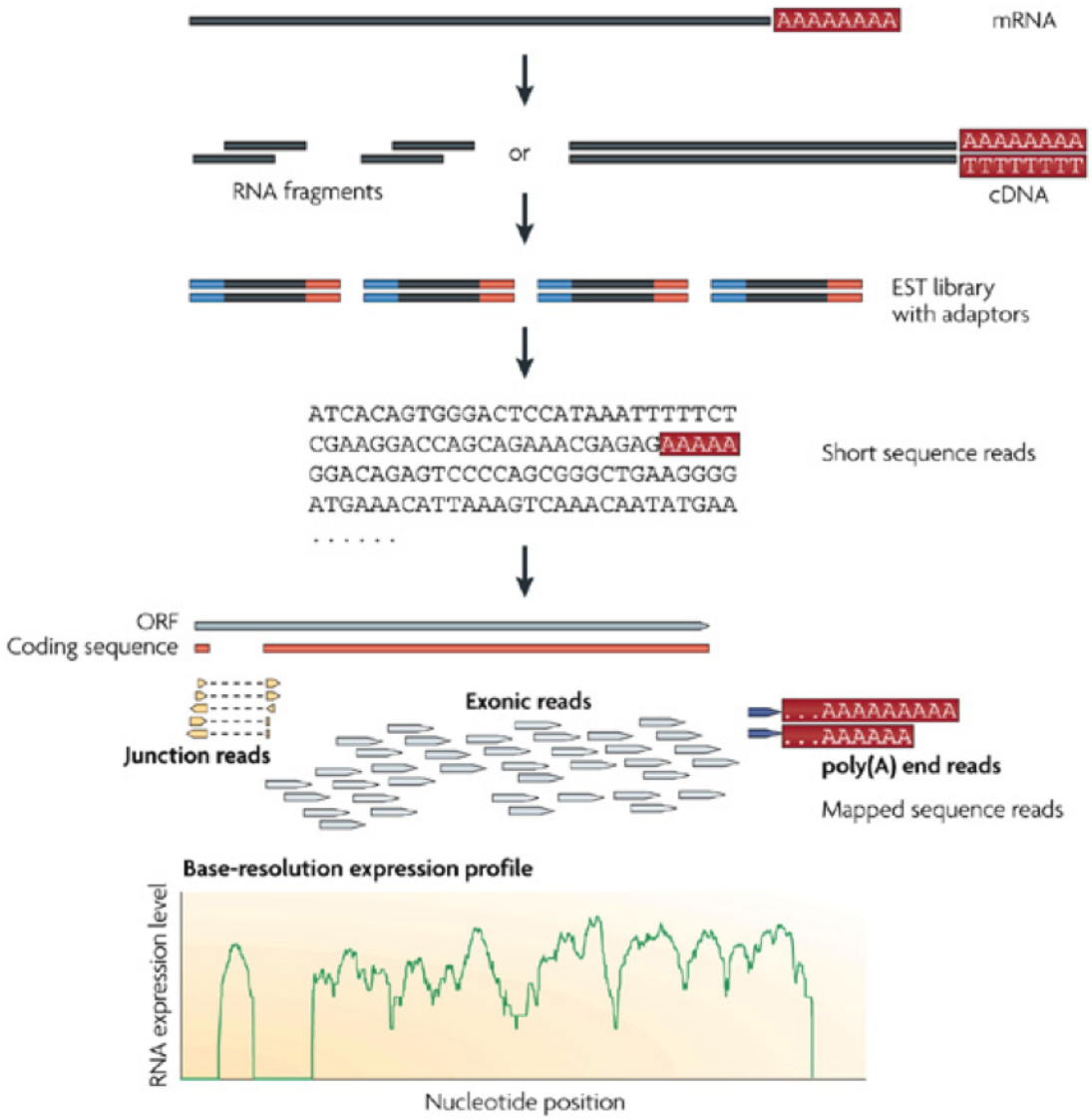
# The isoform identification and quantification problem



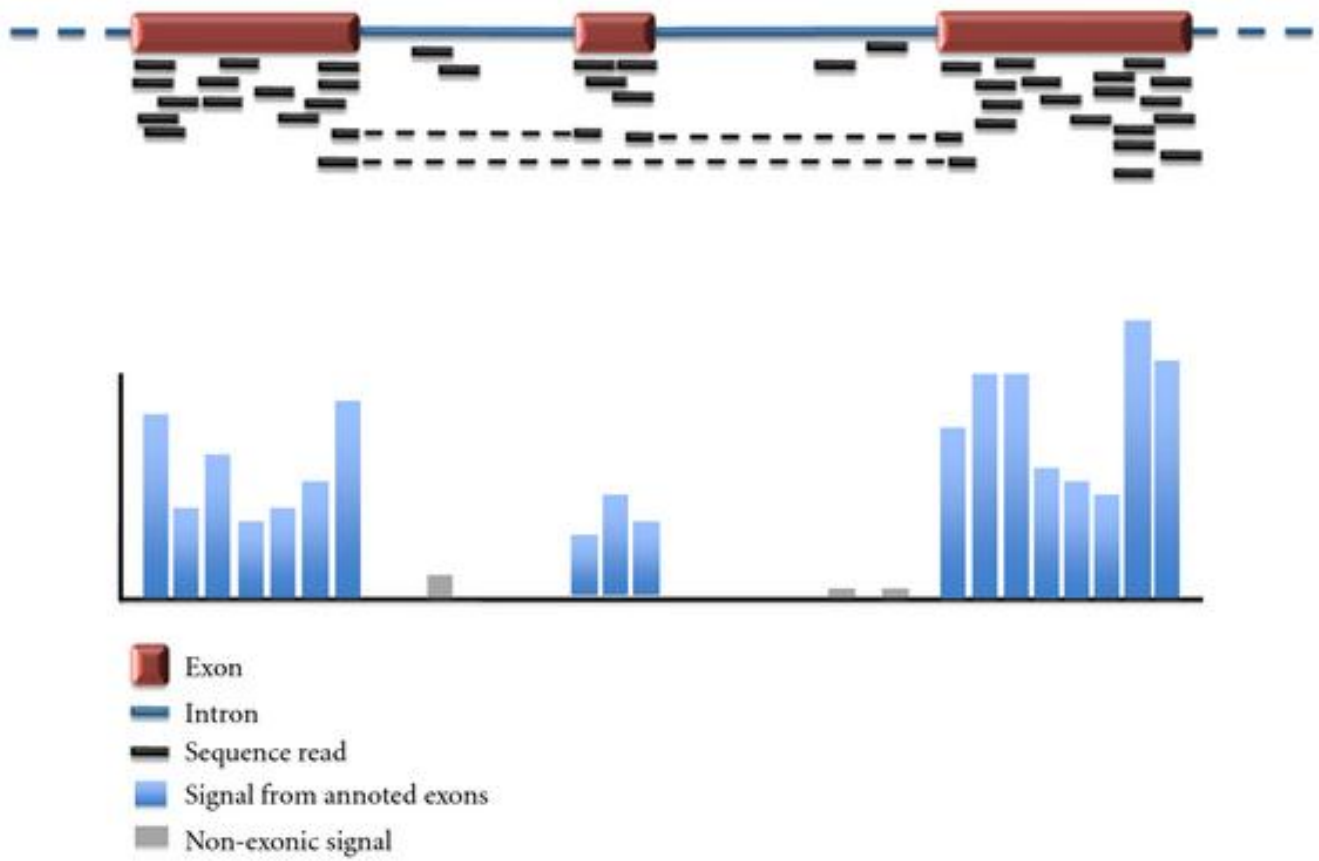

Given a biological sample (e.g., cancer tissue), can we:

1. identify the isoform(s) of each gene present in the sample?
2. quantify their abundance?

# RNA-seq measures mRNA abundance by sequencing short fragments

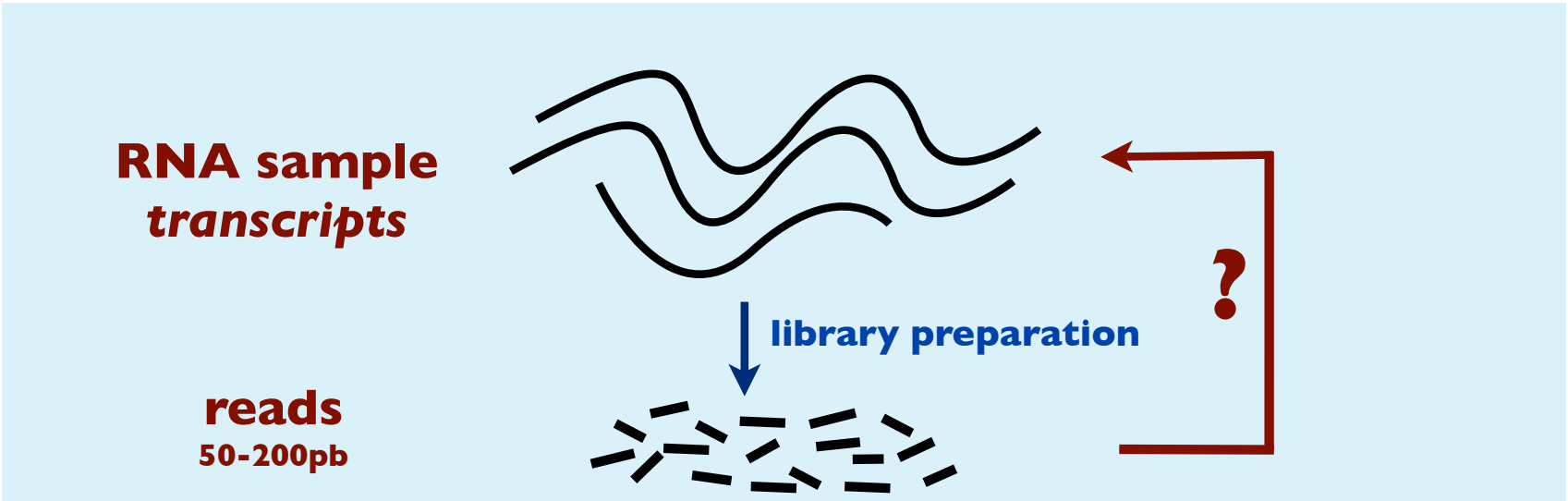


# RNA-seq and alternative splicing




(Costa et al., 2011)

# From RNA-seq to isoforms




**Transcripts Quantification using annotations**

- RQuant (Bohnert et al. 2009)
- FluxCapacitor (Montgomery et al. 2010)
- IsoEM (Nicolae et al. 2011)
- eXpress (Roberts et al. 2013)

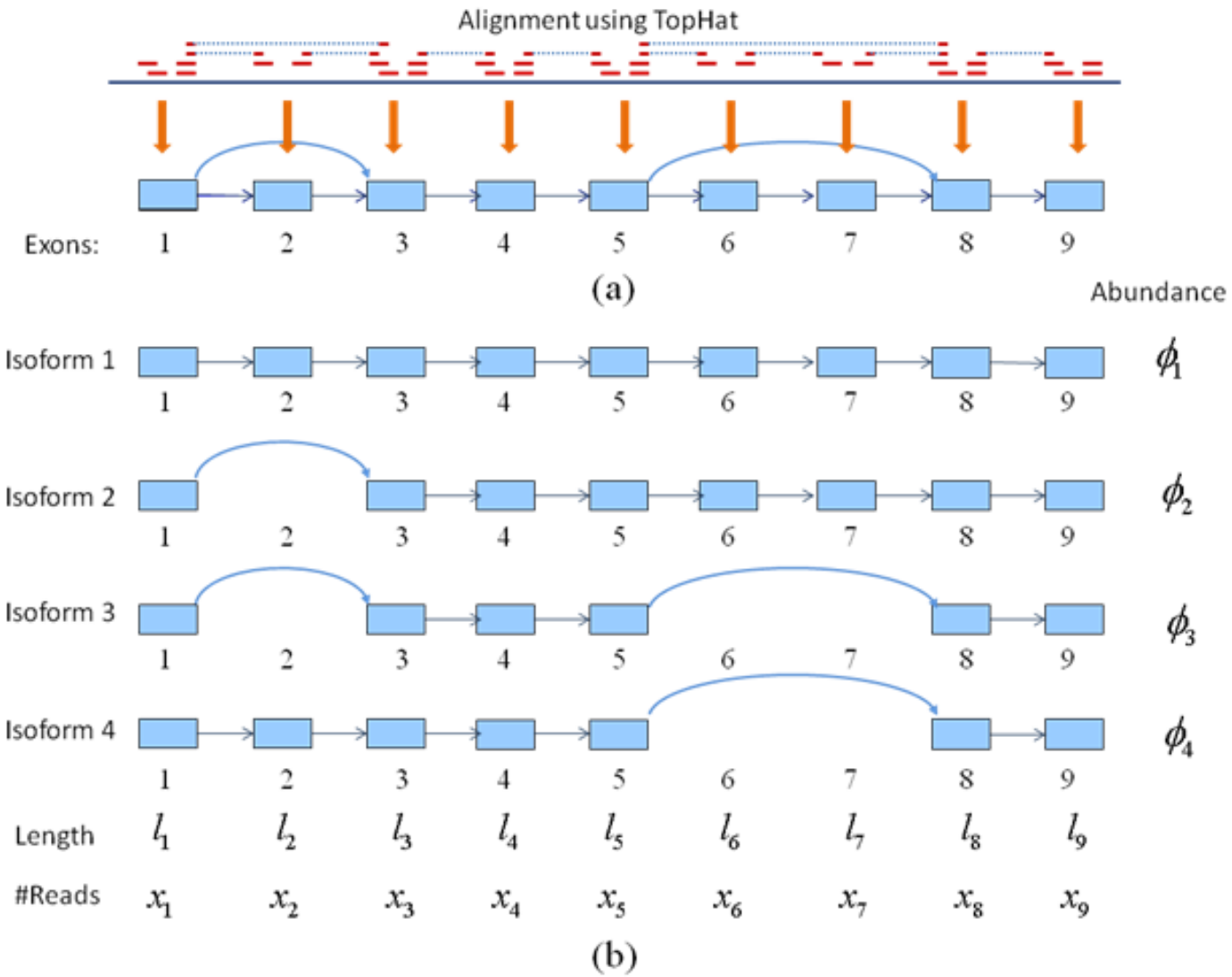
**De Novo approaches**

- OASES (Schultz et al. 2012)
- Trinity (Grabherr et al. 2011)
- Kissplice (Sacomoto et al. 2012)

**Genome-based Transcripts Reconstruction**

- Scripture (Guttman et al. 2010)
- Cufflinks (Trapnell et al. 2010)
- IsoLasso (Li et al. 2011a)
- NSMAP (Xia et al. 2011)
- SLIDE (Li et al. 2011b)
- iReckon (Mezlini et al. 2012)
- **FlipFlop**

# The isoform deconvolution problem




(Xia et al., 2011)

# More formally

$e$ exons
$c$ candidate isoforms (up to $2^e - 1$)
$\phi \in \mathbb{R}_+^c$ the vector of abundance of isoforms (unknown!)
$U$ binary matrix:

$$
\begin{array}{c}
 \\
\mathit{isoform}_1 \\
\mathit{isoform}_2 \\
\vdots \\
\mathit{isoform}_c
\end{array}
\begin{array}{c}
\begin{array}{ccccccc}
\mathit{exon}_1 & \cdots & \mathit{exon}_e & \mathit{junction}_{1,2} & \cdots & \mathit{junction}_{e_1,e}
\end{array} \\
\left(\begin{array}{ccccccc}
1 & \cdots & 1 & 1 & \cdots & 1 \\
1 & \cdots & 0 & 1 & \cdots & 0 \\
 & \cdots & & & \cdots & \\
0 & \cdots & 1 & 0 & \cdots & 0
\end{array}\right)
\end{array}
$$

$U^\top \phi$ the abundance of each exon/junction.

Goal: estimate $\phi$ from the observed reads on each exon/junction

# Isoform deconvolution with the Lasso

Estimate $\phi$ sparse by solving:

$$\min_{\phi \in \mathbb{R}^c_+} R(U^\top \phi) + \lambda \| \phi \|_1$$

- IsoLasso (Li et al., 2011)
- NSMAP (Xia et al., 2011)
- SLIDE (Li et al., 2011)

Works well BUT computationally challenging to enumerate all candidate isoforms (up to $2^e$) for large genes!

# Fast isoform deconvolution with the Lasso

**Theorem (Bernard, Mairal, Jacob and V., 2012)**

The isoform deconvolution problem

$$\min_{\phi \in \mathbb{R}_+^c} R(U^\top \phi) + \lambda \| \phi \|_1$$

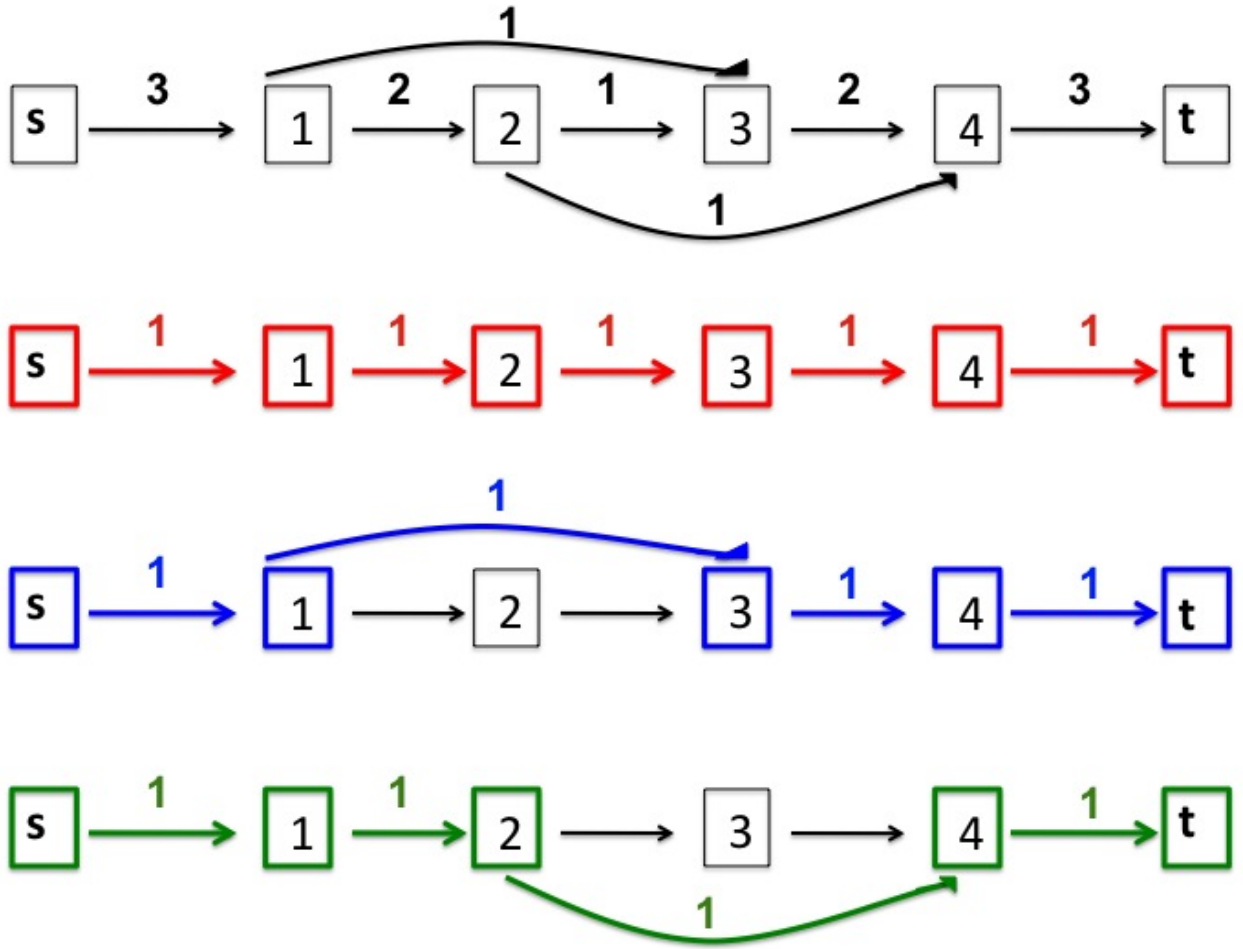
can be solved in polynomial time in the number of exon.

Key ideas

1. $U^\top \phi$ corresponds to a flow on the graph
2. Reformulation as a convex cost flow problem (Mairal and Yu, 2012)
3. Recover isoforms by flow decomposition algorithm

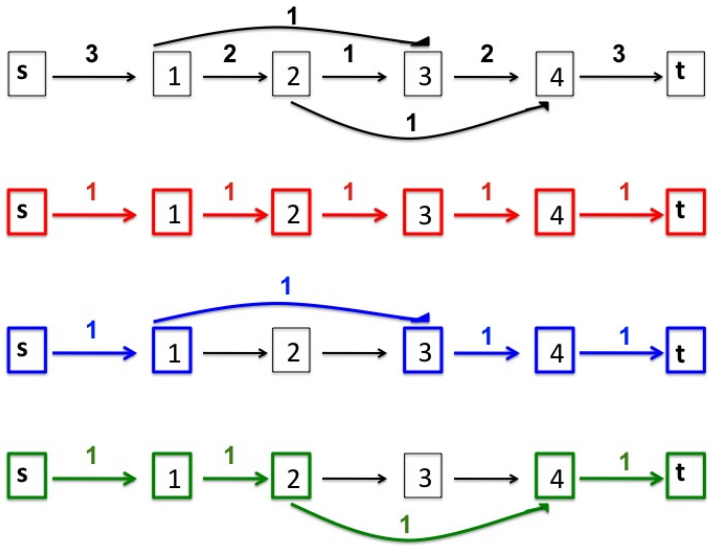
"Feature selection on an exponential number of features in polynomial time"

# From isoforms to flows




- Isoforms are paths
- Linear combinations of isoforms are flows

# Isoform deconvolution as convex cost flow problem




$$\min_{\phi \in \mathbb{R}_+^c} R(U^\top \phi) + \lambda \| \phi \|_1$$

is equivalent to

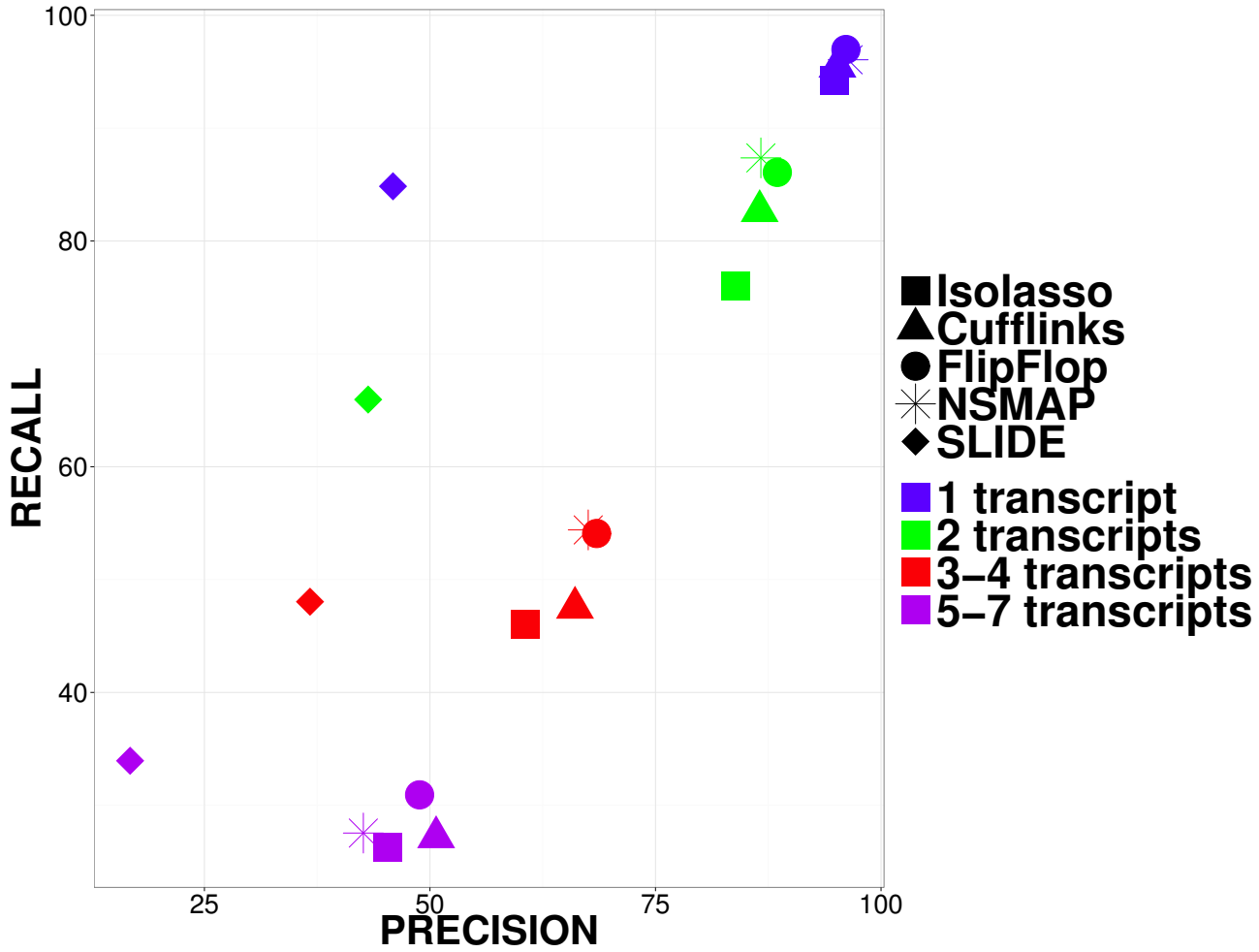
$$\min_{f \text{ flow}} R(f) + \lambda f_t$$

# FlipFlop (Bernard, Jacob, Mairal and V., 2012)

$$\min_{\phi \in \mathbb{R}^c_+} R(U^\top \phi) + \lambda \| \phi \|_1$$

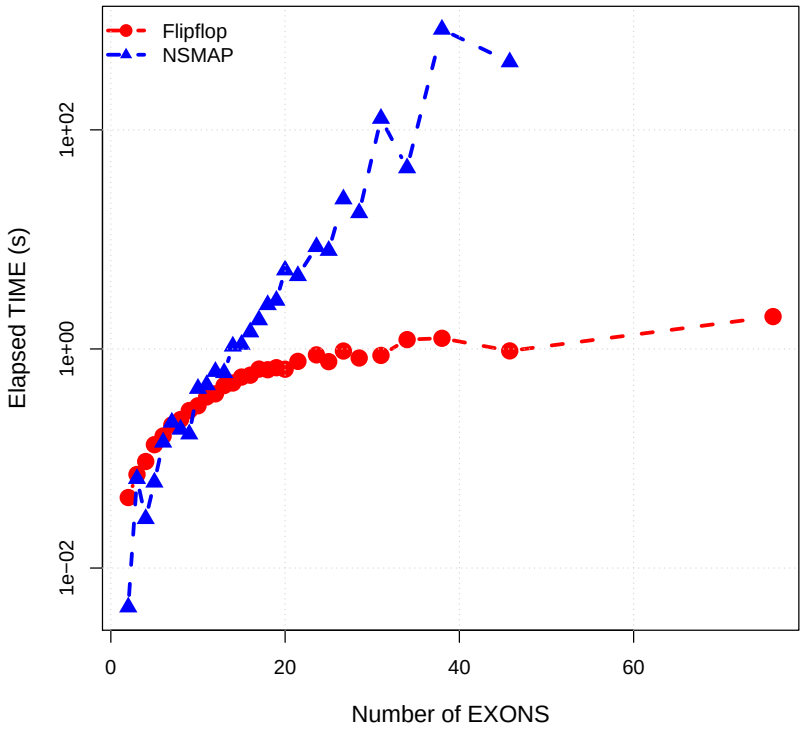
- Cufflink : *a priori* selection of isoforms (minimum graph cover)
- IsoLasso : pre-filtering of candidate isoforms using various heuristics
- NSMAP, SLIDE : limit the maximum number of exons
- FlipFlop : exact optimization without pre-filtering in polynomial time

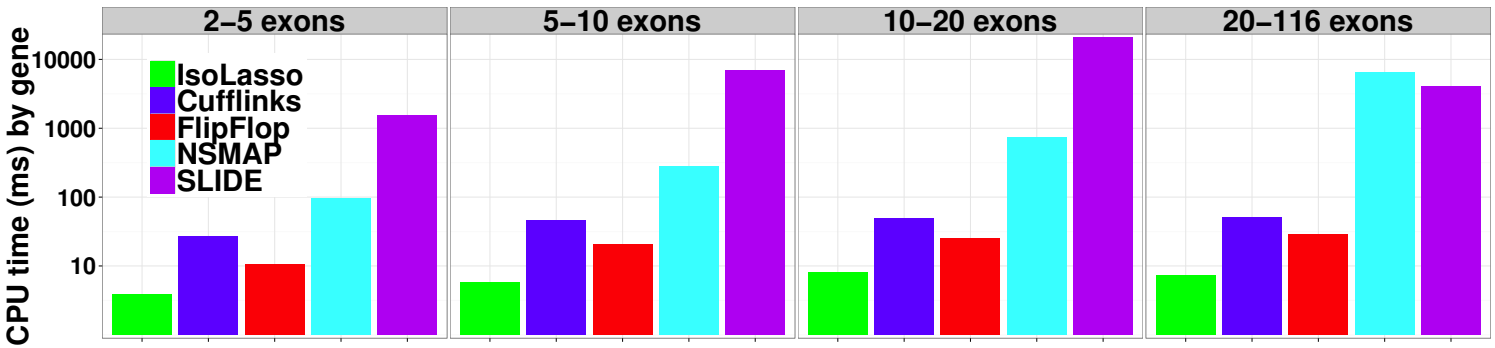
# Performance in isoform identification




Simulated data (hg19, 1137 genes on chr1, 1million 75 bp single-end reads by transcript levels).

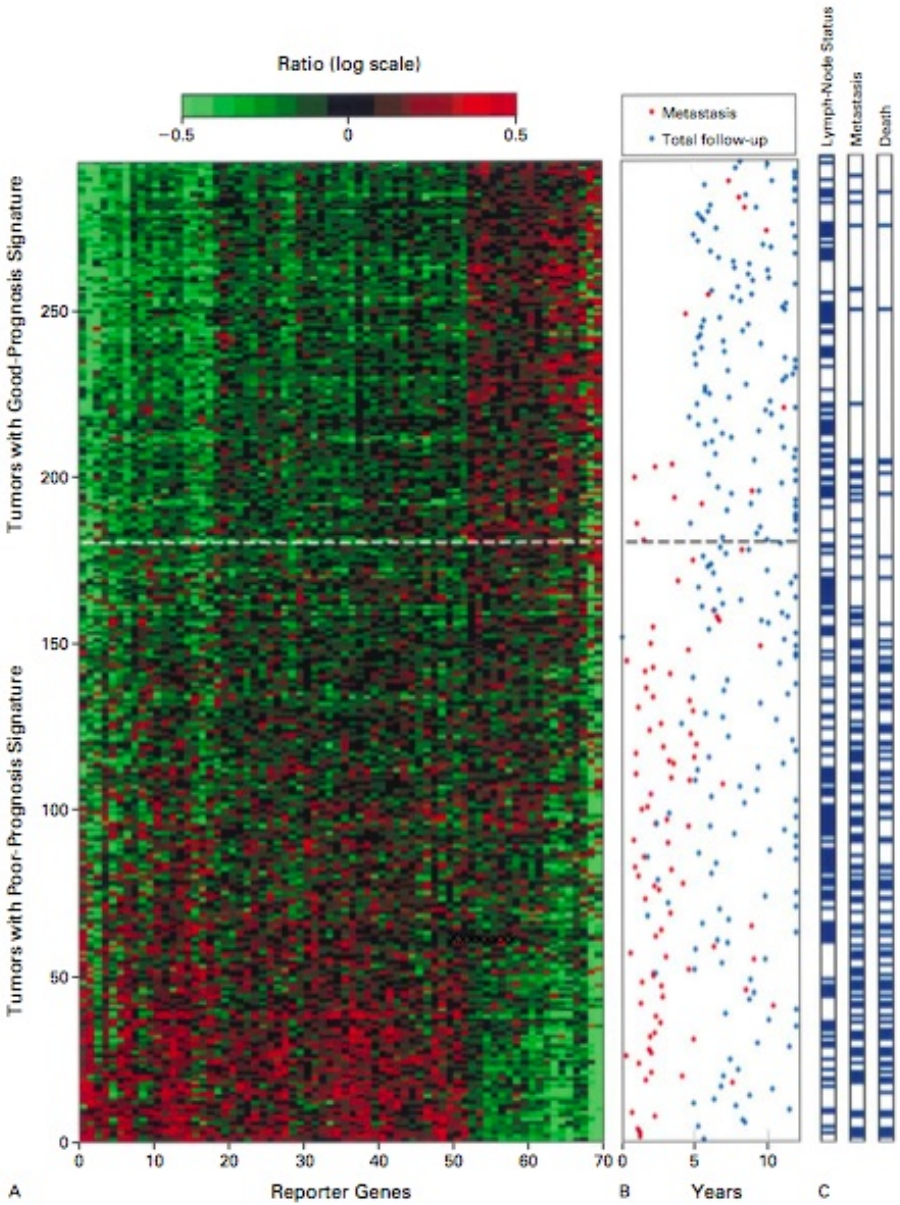
# Speed trial

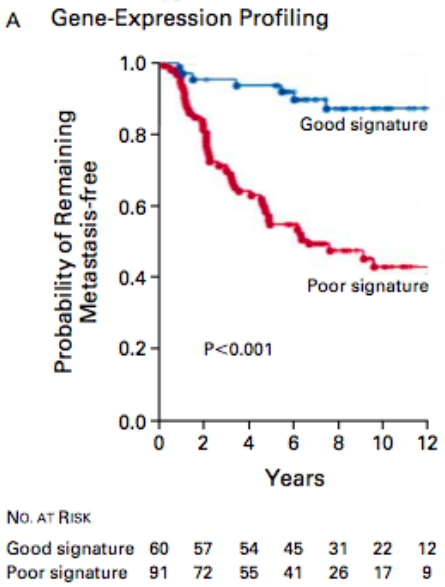

Flipflop
NSMAP
Elapsed TIME (s)
1e−02
1e+00
1e+02
0
20
40
60
Number of EXONS

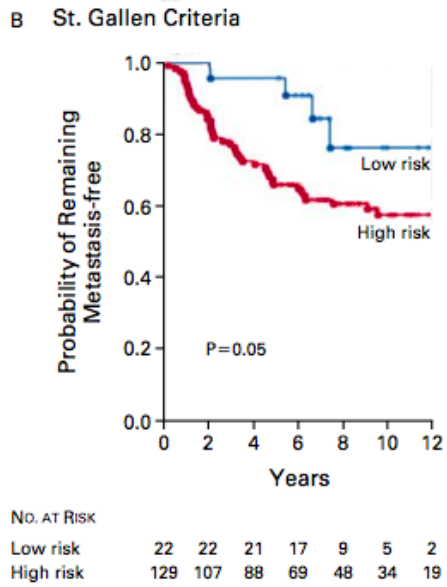

2–5 exons
5–10 exons
10–20 exons
20–116 exons
CPU time (ms) by gene
10000
1000
100
10
IsoLasso
Cufflinks
FlipFlop
NSMAP
SLIDE

# Outline

# Breast cancer prognosis



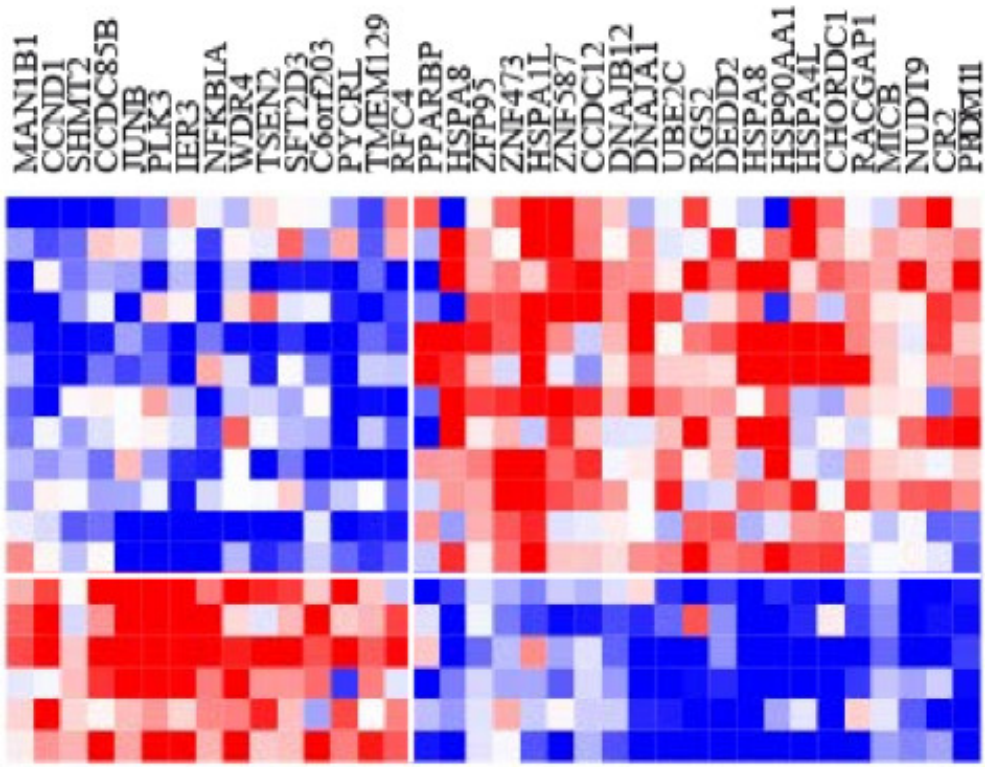



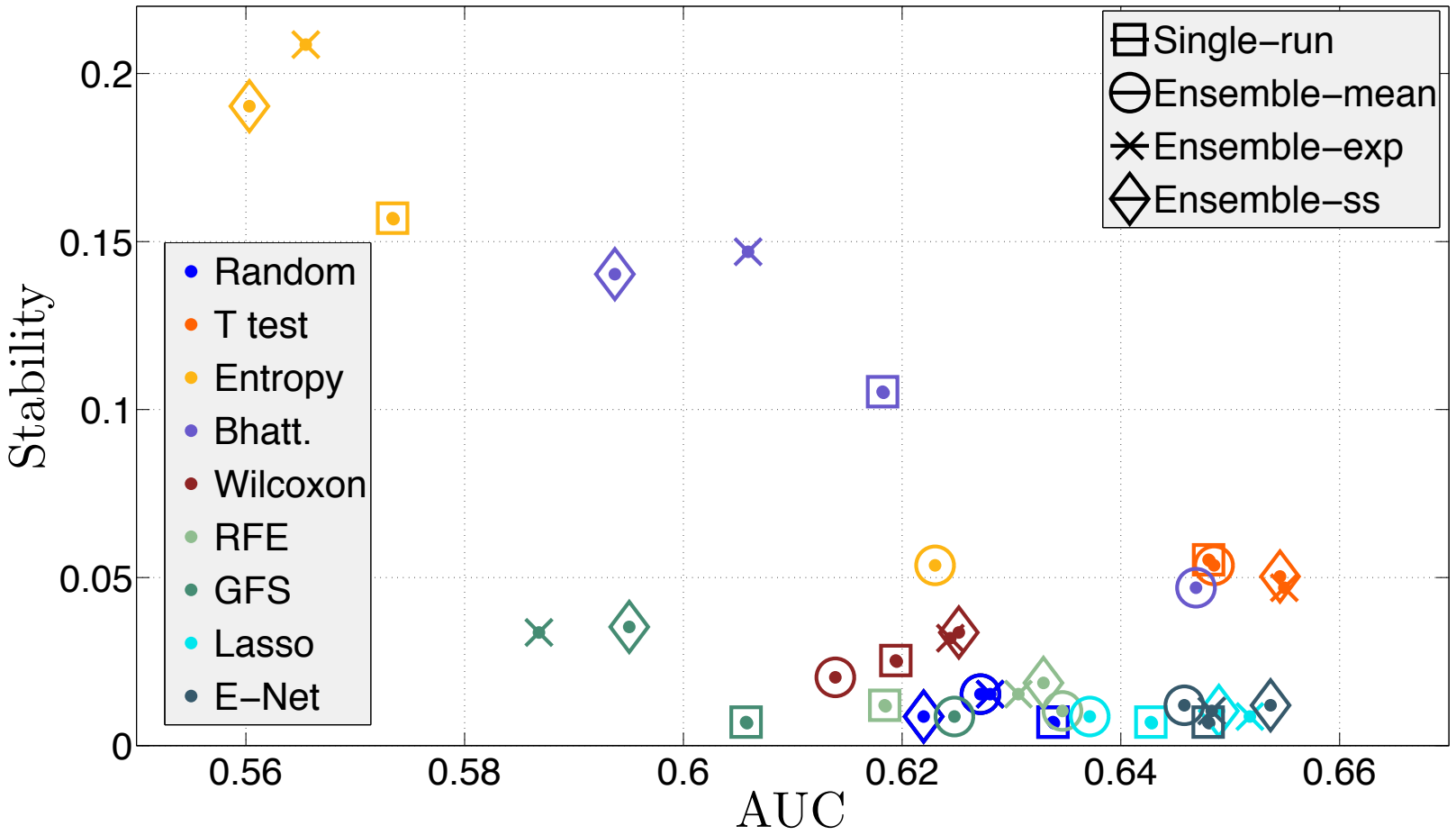
# Gene selection, molecular signature

## The idea

- We look for a limited set of genes that are sufficient for prediction.
- Selected genes should inform us about the underlying biology

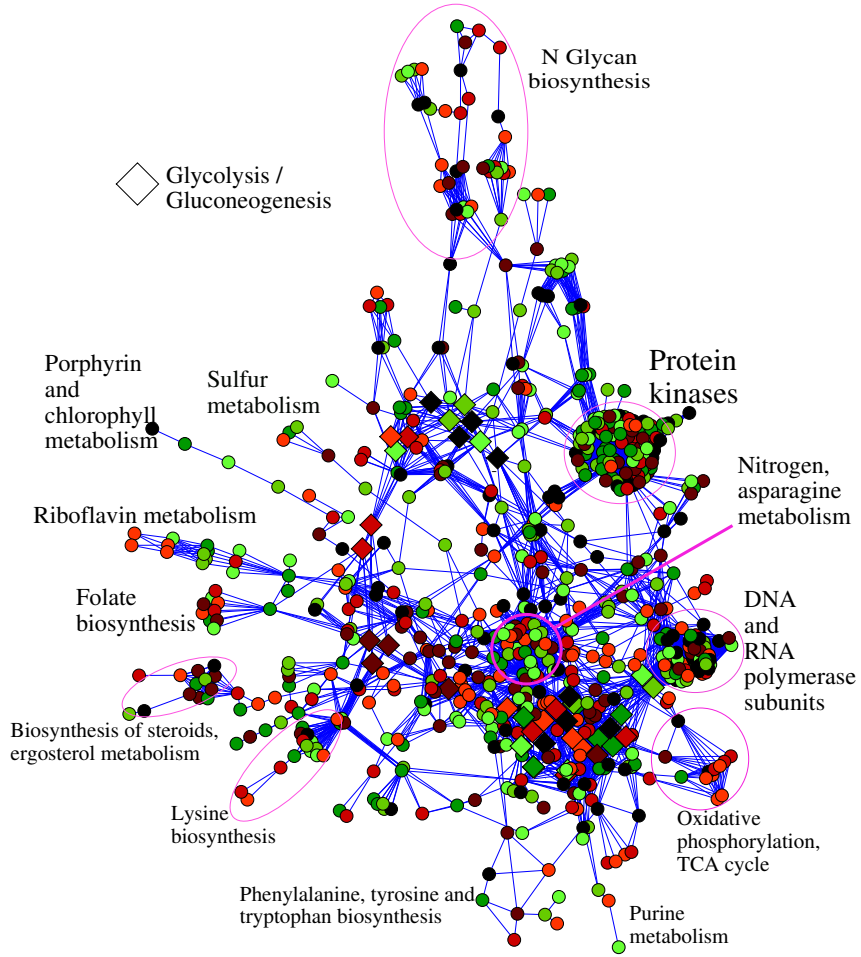


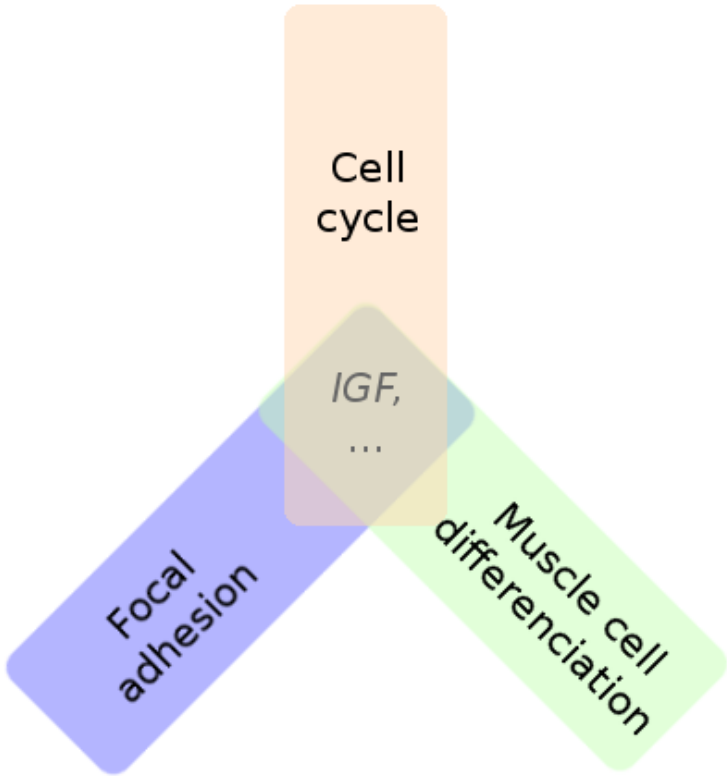
# Lack of stability of signatures




*Haury et al. (2011)*

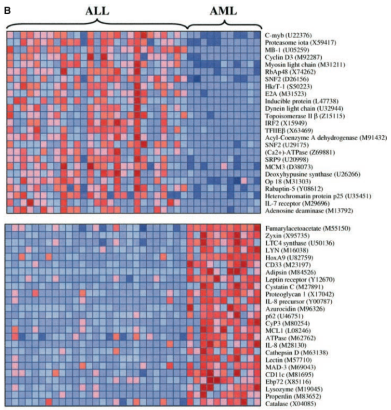
# Gene networks, gene groups

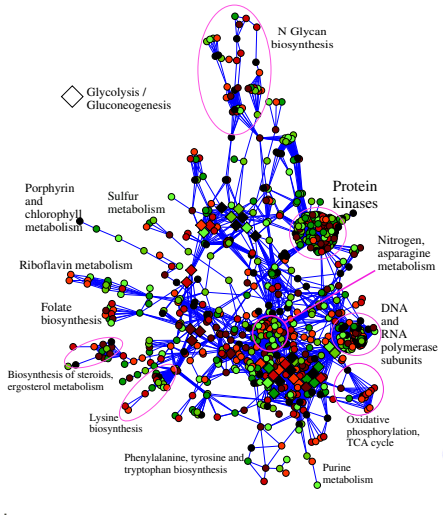


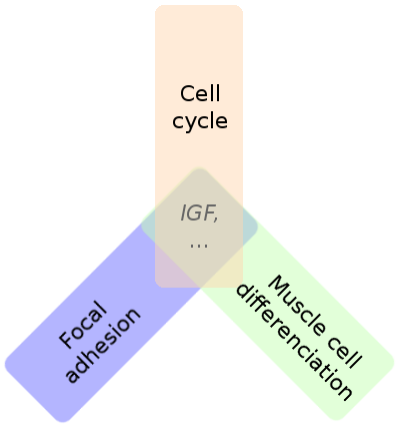



# Structured feature selection

- Basic biological functions usually involve the coordinated action of several proteins:
  - Formation of protein complexes
  - Activation of metabolic, signalling or regulatory pathways
- How to perform structured feature selection, such that selected genes
  - belong to only a few groups?
  - form a small number of connected components on the graph?



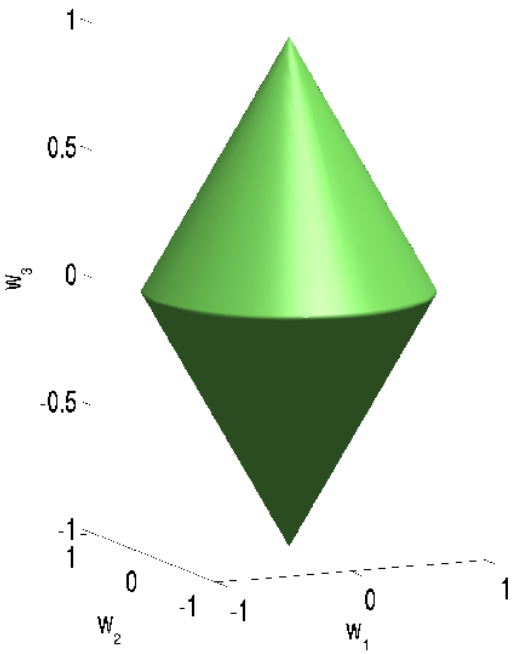



# Selecting pre-defined groups of variables

## Group lasso (Yuan & Lin, 2006)

If groups of covariates are likely to be selected together, the $\ell_1/\ell_2$-norm induces sparse solutions *at the group level*:

$$\Omega_{group}(w) = \sum_g \|w_g\|_2$$

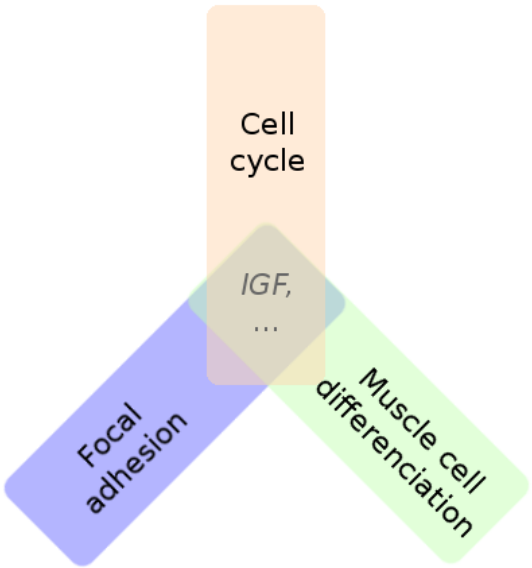



$$\Omega(w_1, w_2, w_3) = \|(w_1, w_2)\|_2 + \|w_3\|_2$$

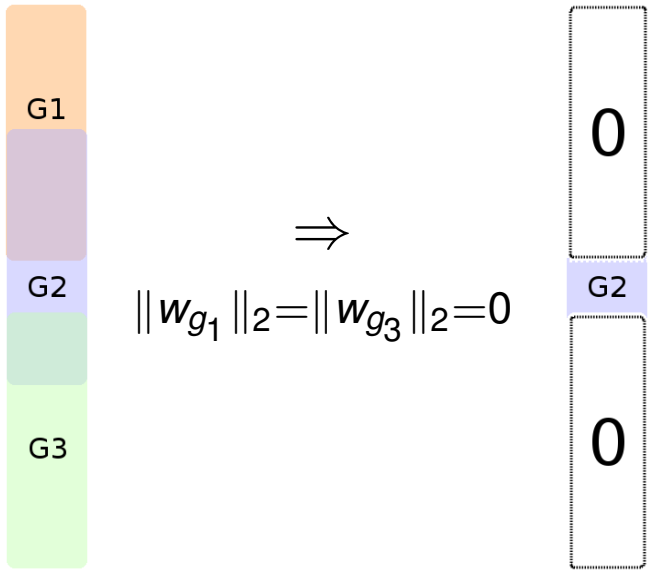
# Group lasso with overlapping groups

## Idea 1: shrink groups to zero (Jenatton et al., 2009)

- $\Omega_{group}(w) = \sum_g \|w_g\|_2$ sets groups to 0.
- One variable is selected $\Leftrightarrow$ all the groups to which it belongs are selected.




IGF selection $\Rightarrow$ selection of unwanted groups

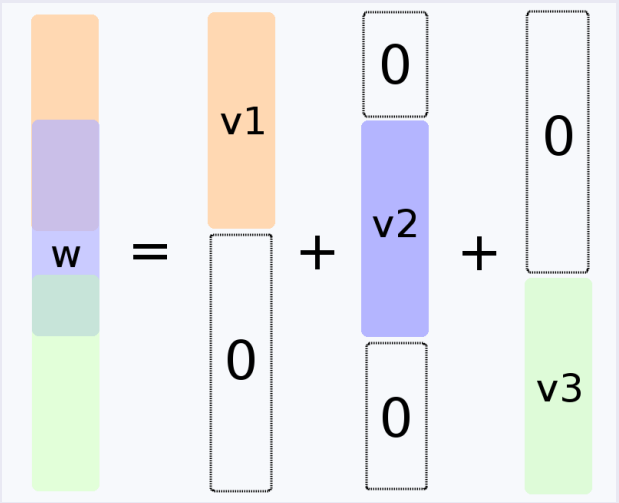



Removal of *any* group containing a gene $\Rightarrow$ the weight of the gene is 0.

# Group lasso with overlapping groups

## Idea 2: latent group Lasso (Jacob et al., 2009)

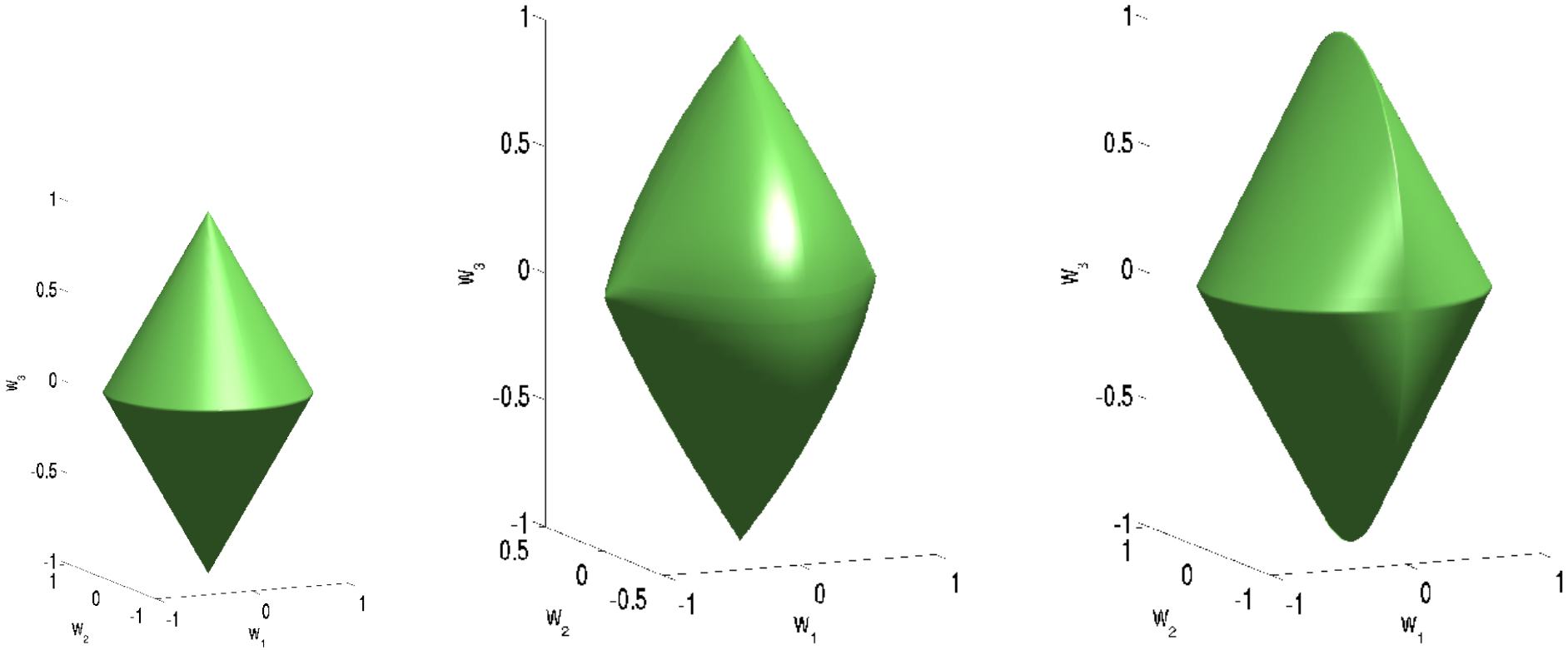
$$\Omega^{\mathcal{G}}_{\text{latent}}(w) \triangleq \begin{cases} \min_v \sum_{g \in \mathcal{G}} \|v_g\|_2 \\ w = \sum_{g \in \mathcal{G}} v_g \\ \text{supp}(v_g) \subseteq g. \end{cases}$$




## Properties

- Resulting support is a *union* of groups in $\mathcal{G}$.
- Possible to select one variable without selecting all the groups containing it.
- Equivalent to group lasso when there is no overlap

# Overlap and group unity balls




Balls for $\Omega^{\mathcal{G}}_{\text{group}}(\cdot)$ (middle) and $\Omega^{\mathcal{G}}_{\text{latent}}(\cdot)$ (right) for the groups $\mathcal{G} = \{\{1,2\},\{2,3\}\}$ where $w_2$ is represented as the vertical coordinate. Left: group-lasso ($\mathcal{G} = \{\{1,2\},\{3\}\}$), for comparison.

# Theoretical results

## Consistency in group support (Jacob et al., 2009)

- Let $\bar{w}$ be the true parameter vector.
- Assume that there exists a unique decomposition $\bar{v}_g$ such that $\bar{w} = \sum_g \bar{v}_g$ and $\Omega^{\mathcal{G}}_{\text{latent}}(\bar{w}) = \sum \|\bar{v}_g\|_2$.
- Consider the regularized empirical risk minimization problem $L(w) + \lambda \Omega^{\mathcal{G}}_{\text{latent}}(w)$.

Then

- under appropriate mutual incoherence conditions on $X$,
- as $n \to \infty$,
- with very high probability,

the optimal solution $\hat{w}$ admits a unique decomposition $(\hat{v}_g)_{g \in \mathcal{G}}$ such that

$$\left\{ g \in \mathcal{G} | \hat{v}_g \neq 0 \right\} = \left\{ g \in \mathcal{G} | \bar{v}_g \neq 0 \right\}.$$

# Theoretical results

## Consistency in group support (Jacob et al., 2009)

- Let $\bar{w}$ be the true parameter vector.
- Assume that there exists a unique decomposition $\bar{v}_g$ such that $\bar{w} = \sum_g \bar{v}_g$ and $\Omega^{\mathcal{G}}_{\text{latent}}(\bar{w}) = \sum \|\bar{v}_g\|_2$.
- Consider the regularized empirical risk minimization problem $L(w) + \lambda \Omega^{\mathcal{G}}_{\text{latent}}(w)$.

Then

- under appropriate mutual incoherence conditions on $X$,
- as $n \to \infty$,
- with very high probability,

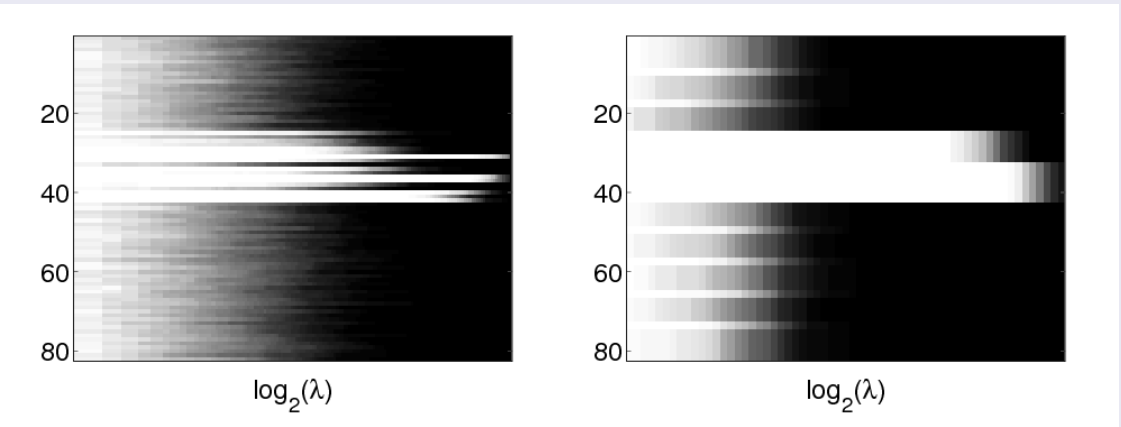
the optimal solution $\hat{w}$ admits a unique decomposition $(\hat{v}_g)_{g \in \mathcal{G}}$ such that

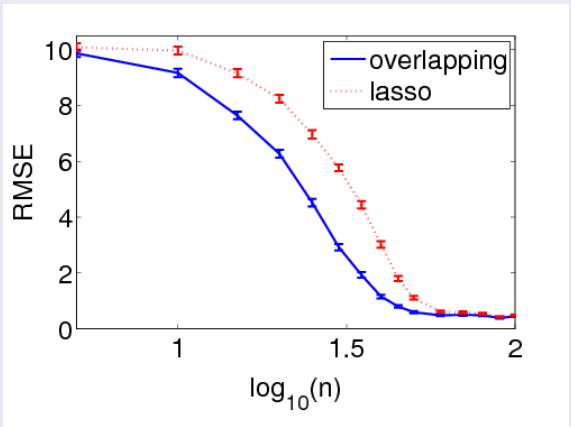
$$\left\{ g \in \mathcal{G} | \hat{v}_g \neq 0 \right\} = \left\{ g \in \mathcal{G} | \bar{v}_g \neq 0 \right\}.$$

# Experiments

## Synthetic data: overlapping groups

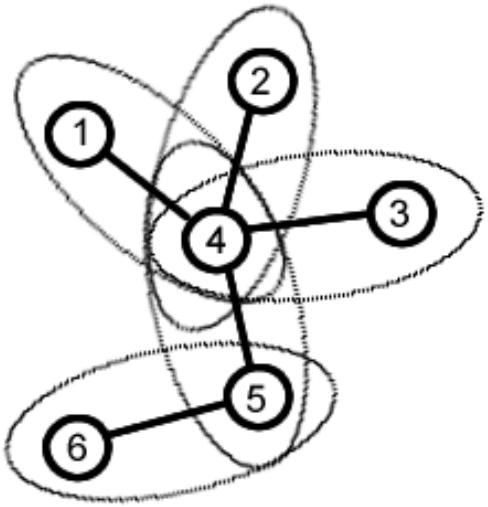
- 10 groups of 10 variables with 2 variables of overlap between two successive groups :$\{1, \ldots, 10\}, \{9, \ldots, 18\}, \ldots, \{73, \ldots, 82\}$.
- Support: union of 4*th* and 5*th* groups.
- Learn from 100 training points.

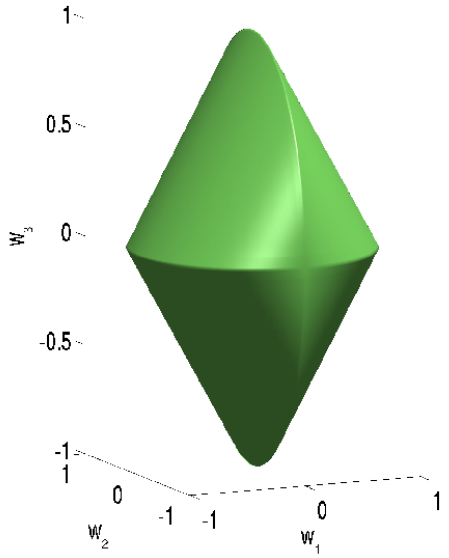



Frequency of selection of each variable with the lasso (left) and $\Omega^{\mathcal{G}}_{\text{latent}}(.)$ (middle), comparison of the RMSE of both methods (right).

# Graph lasso




## Two solutions

$$\Omega^{\mathcal{G}}_{\text{group}}(\beta) = \sum_{i \sim j} \sqrt{\beta_i^2 + \beta_j^2}\,,$$

$$\Omega^{\mathcal{G}}_{\text{latent}}(\beta) = \sup_{\alpha \in \mathbb{R}^p : \forall i \sim j, \|\alpha_i^2 + \alpha_j^2\| \leq 1} \alpha^\top \beta\,.$$

# Preliminary results

## Breast cancer data

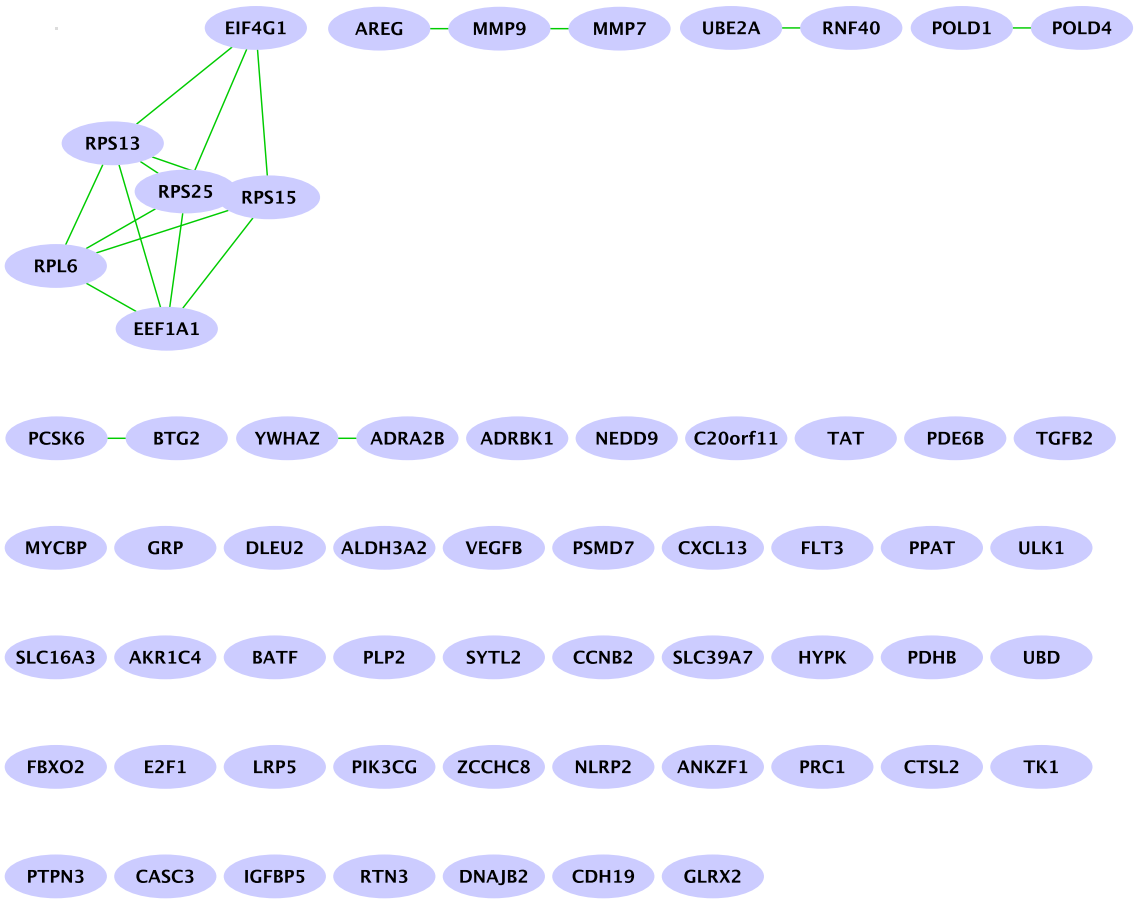
- Gene expression data for 8,141 genes in 295 breast cancer tumors.
- Canonical pathways from MSigDB containing 639 groups of genes, 637 of which involve genes from our study.

| METHOD | $\ell_1$ | $\Omega^{\mathcal{G}}_{\text{LATENT}}(.)$ |
|---|---|---|
| ERROR | $0.38 \pm 0.04$ | $0.36 \pm 0.03$ |
| MEAN $\sharp$ PATH. | 130 | 30 |

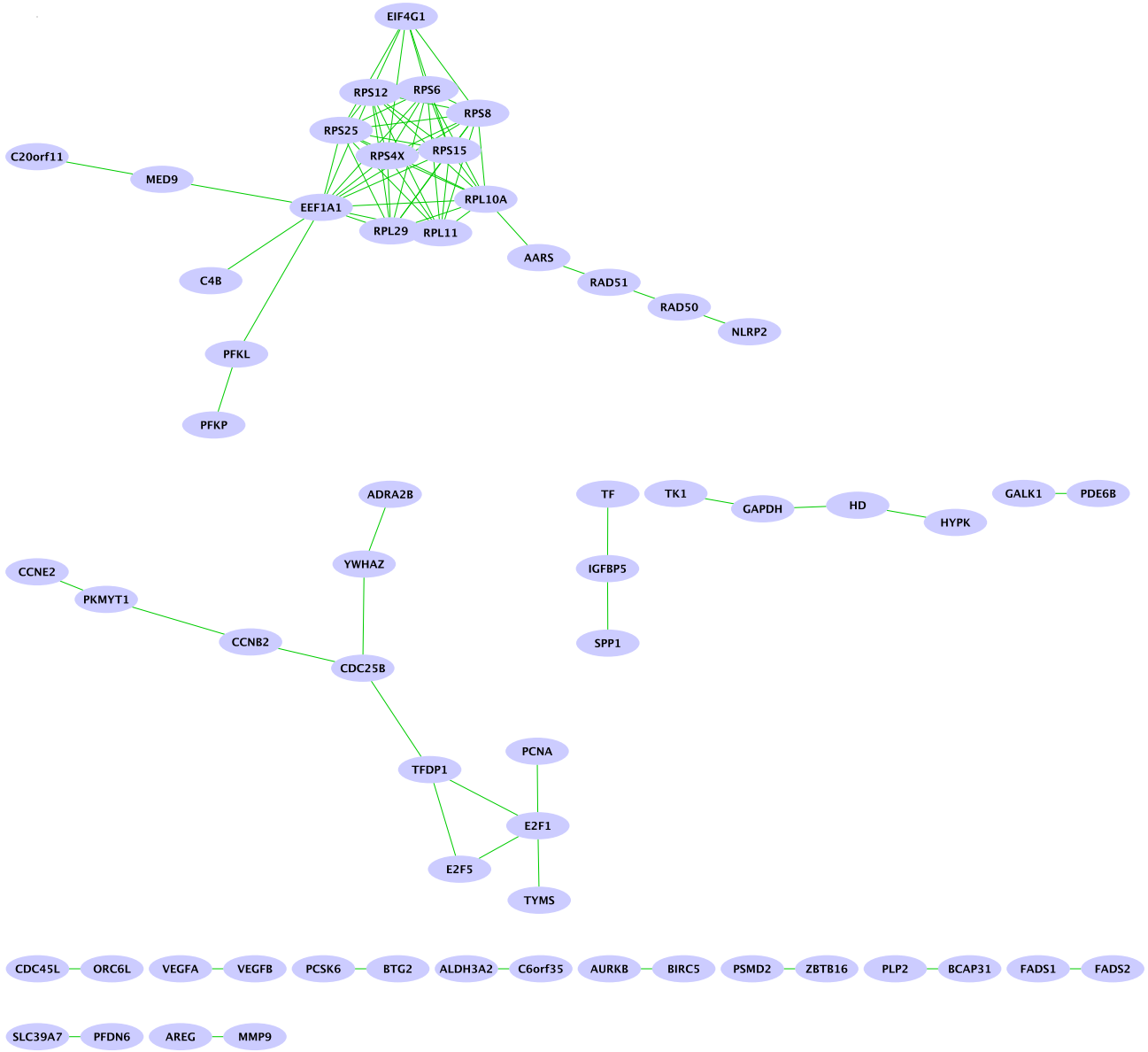
- Graph on the genes.

| METHOD | $\ell_1$ | $\Omega_{graph}(.)$ |
|---|---|---|
| ERROR | $0.39 \pm 0.04$ | $0.36 \pm 0.01$ |
| AV. SIZE C.C. | 1.03 | 1.30 |

# Lasso signature



# Graph Lasso signature



# Conclusions

- Convex sparsity-inducing penalties as a way to incorporate prior knowledge
- Specific implementations for specific problems:
  - greedy dichotomic segmentation for fused lasso
  - fast group Lasso for joint segmentation
  - network flow optimization of lasso over the paths of a graph
  - efficient proximity operator computation of latent group lasso
- Often, feature selection is consistent (although we pay a price when features are very correlated), stability selection may help
- Numerous applications in bioinformatics and beyond!

# Acknowledgements

Kevin Bleakley (INRIA), Laurent Jacob (UC Berkeley) Guillaume Obozinski (INRIA), Anne-Claire Haury (ParisTech), Julien Mairal (UC Berkeley/INRIA), Elsa Bernard (ParisTech), Fantine Mordelet (Duke), Paola Vera-Licona (Institut Curie)

*Post-docs available in Paris!*

AGENCE NATIONALE DE LA RECHERCHE ANR

FRANCE-BERKELEY FUND
FONDS FRANCE-BERKELEY
UNIVERSITY OF CALIFORNIA, BERKELEY | LE MINISTERE DES AFFAIRS ETRANGERES

European Research Council
erc