

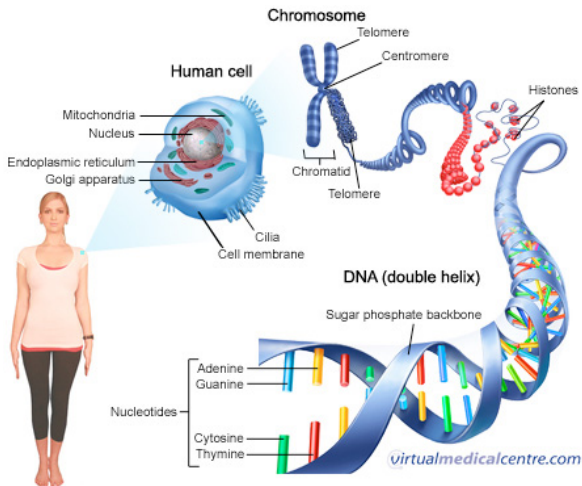
# Machine Learning for Personalized Medicine

Jean-Philippe Vert



LIG, Grenoble, November 6, 2014

# What's in your body



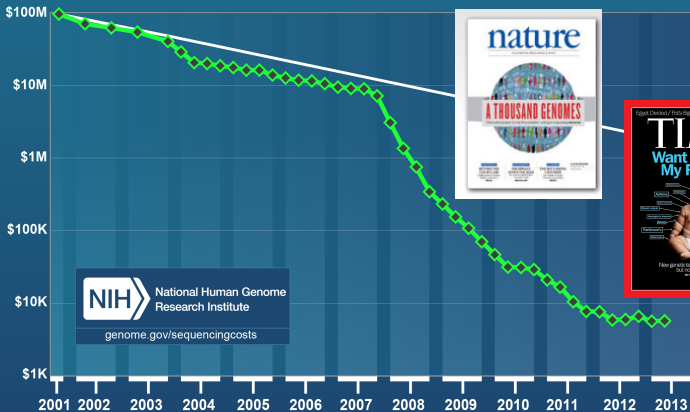
1 body =  $10^{14}$  human cells (and 100x more non-human cells)

1 cell =  $6 \times 10^9$  ACGT coding for 20,000 genes

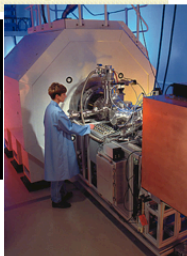
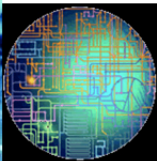
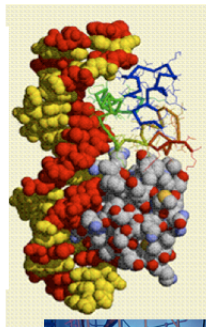
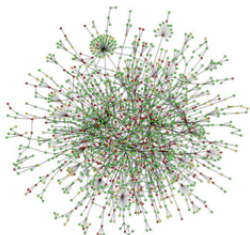
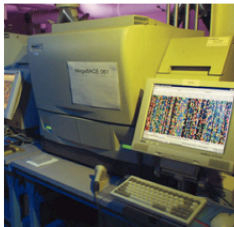
# Sequencing revolution



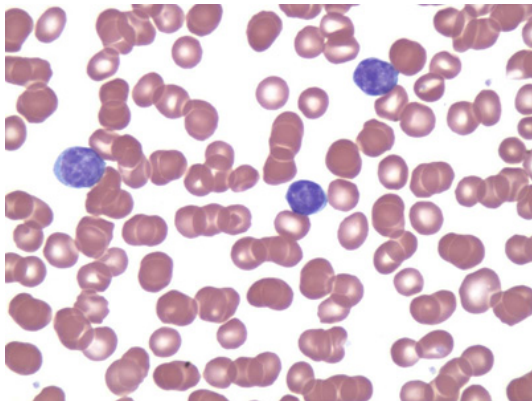
## Cost per Genome



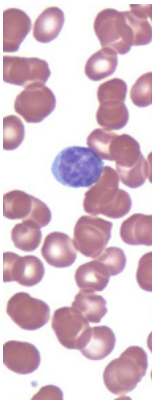
# Many various data



# A cancer cell

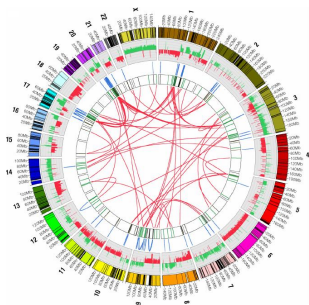


# A cancer cell





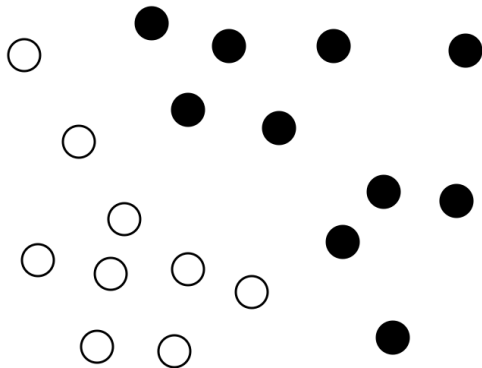
# Opportunities



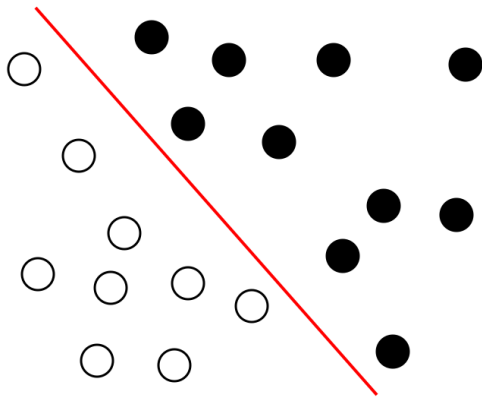
- What is your risk of developing a cancer? (*prevention*)
- After diagnosis and treatment, what is the risk of relapse? (*prognosis*)
- What specific treatment will cure your cancer? (*personalized medicine*)



# Machine learning formulation

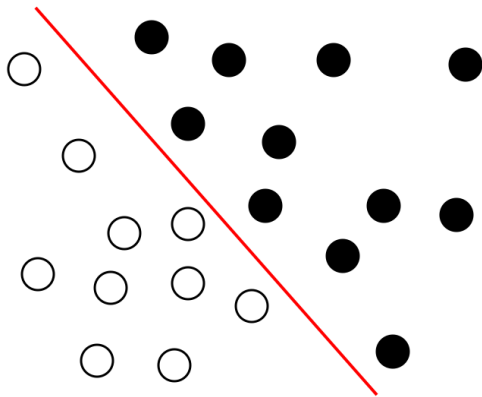


# Machine learning formulation

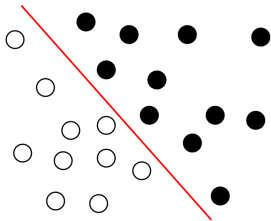




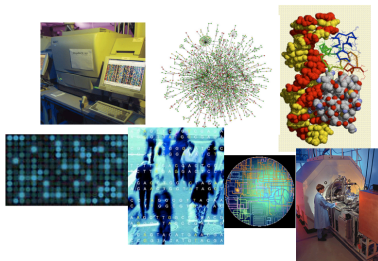
# Machine learning formulation



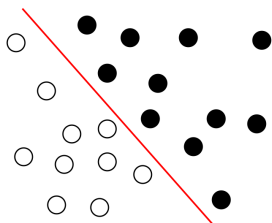
# Challenges



- High dimension
- Few samples
- Structured data
- Heterogeneous data
- Prior knowledge
- Fast and scalable implementations
- Interpretable models



# Learning with regularization



Learn

$$f_{\beta}(\mathbf{x}) = \beta^{\top} \mathbf{x}$$

by solving

$$\min_{\beta \in \mathbb{R}^p} R(f_{\beta}) + \lambda \Omega(\beta)$$

- $R(f_{\beta})$  empirical risk
- $\Omega(\beta)$  penalty

# Outline

- 1 FlipFlop: fast isoform prediction from RNA-seq data
- 2 Learning molecular classifiers with network information
- 3 Kernel bilinear regression for toxicogenomics

# Outline

- 1 FlipFlop: fast isoform prediction from RNA-seq data
- 2 Learning molecular classifiers with network information
- 3 Kernel bilinear regression for toxicogenomics

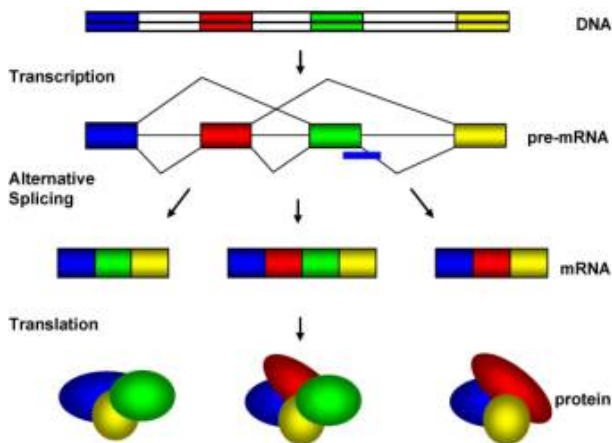


## Joint work with...



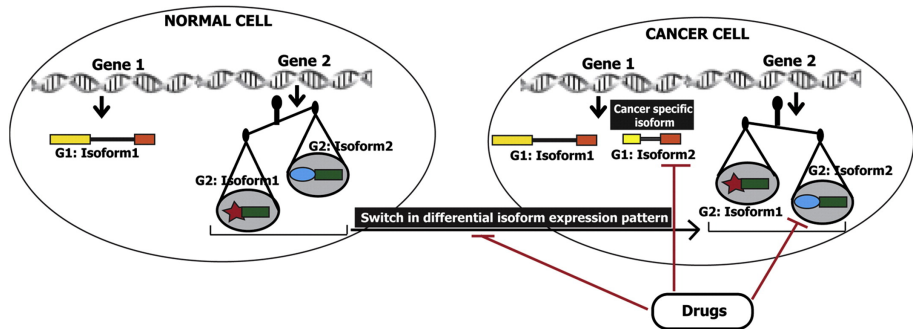
Elsa Bernard (Mines ParisTech / Institut Curie), Laurent Jacob (CNRS / LBBE), Julien Mairal (INRIA)

# Alternative splicing: 1 gene = many proteins



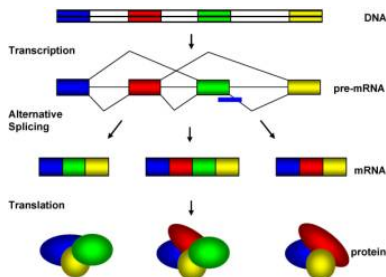
In human, 28k genes give 120k known transcripts (*Pal et al., 2012*)

# Opportunities for drug developments...



(Pal et al., 2012)

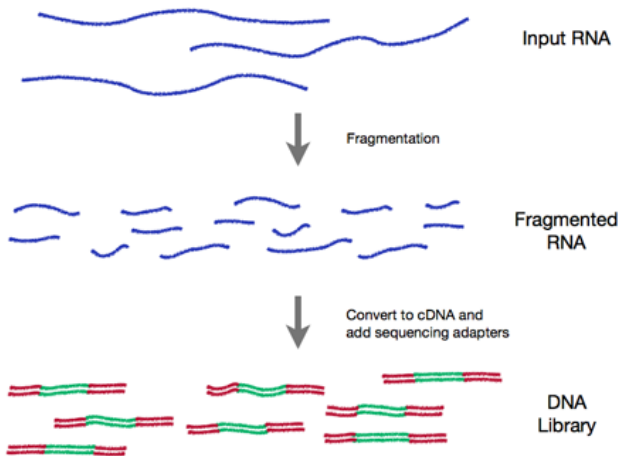
# The isoform identification and quantification problem



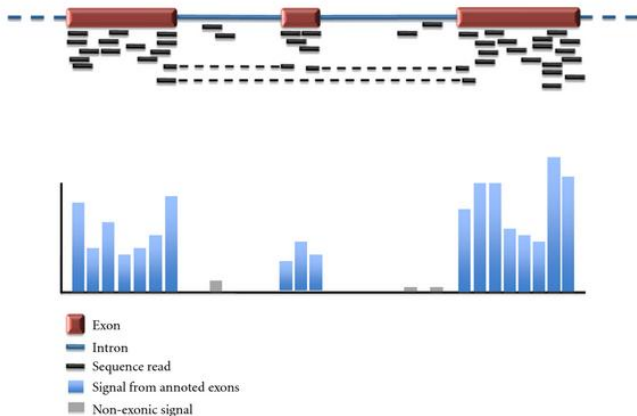
Given a biological sample (e.g., cancer tissue), can we:

- 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

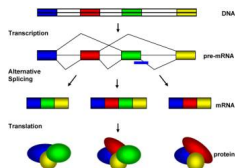


# RNA-seq and alternative splicing



(Costa et al., 2011)

# Lasso-based estimation of isoforms

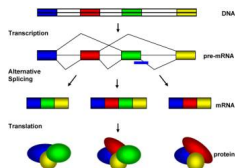


- Let a gene with  $e$  exons
- Suppose there are  $c$  candidate isoform ( $c$  large, up to  $2^e$ )
- Let  $\phi \in \mathbb{R}^c$  the unknown  $c$ -dimensional vector of abundance
- Let  $L(\phi)$  quantify whether  $\phi$  explains well the observed read counts (e.g., minus log-likelihood)
- Find a sparse vector of abundances by solving (e.g., IsoLasso, SLIDE, NSMAP...)

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

- Computational problem: Lasso problem with  $2^e$  variables

# Lasso-based estimation of isoforms



- Let a gene with  $e$  exons
- Suppose there are  $c$  candidate isoform ( $c$  large, up to  $2^e$ )
- Let  $\phi \in \mathbb{R}^c$  the unknown  $c$ -dimensional vector of abundance
- Let  $L(\phi)$  quantify whether  $\phi$  explains well the observed read counts (e.g., minus log-likelihood)
- Find a sparse vector of abundances by solving (e.g., IsoLasso, SLIDE, NSMAP...)

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

- Computational problem: Lasso problem with  $2^e$  variables



# Fast isoform deconvolution with the Lasso (FlipFlop)

Theorem (Bernard, Mairal, Jacob and V., 2014)

The isoform deconvolution problem

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

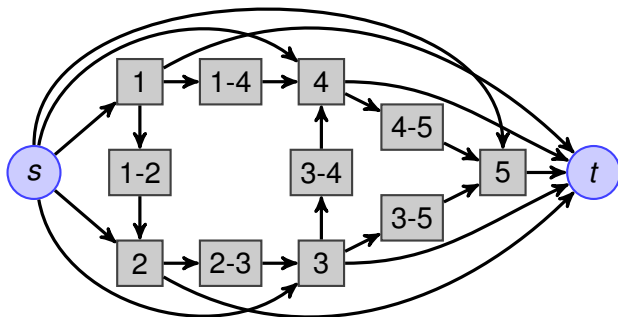
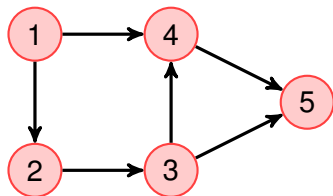
can be solved in **polynomial time** in the number of exon.

Key ideas

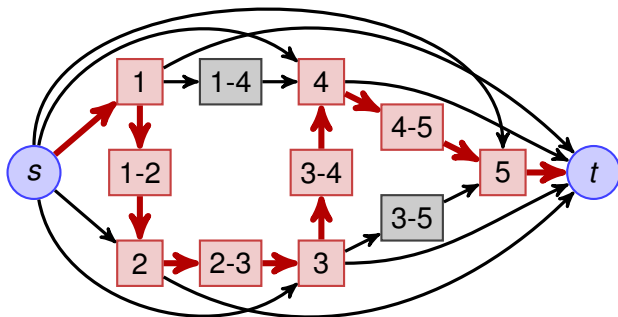
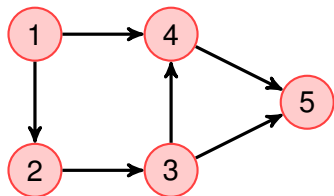
- 1 Reformulation as a **convex cost flow problem** (Mairal and Yu, 2012)
- 2 Recover isoforms by flow decomposition algorithm

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

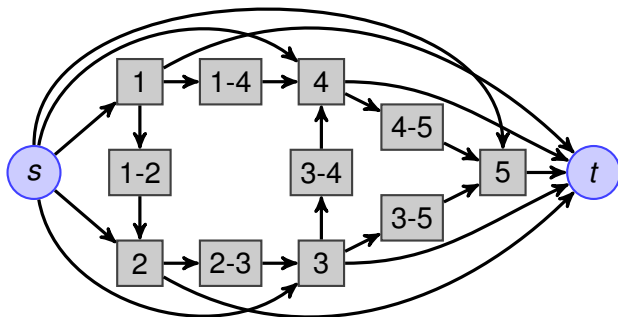
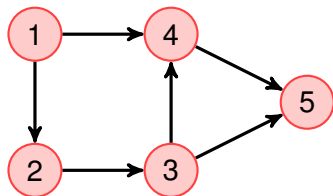
# Isoforms are Paths in a Graph



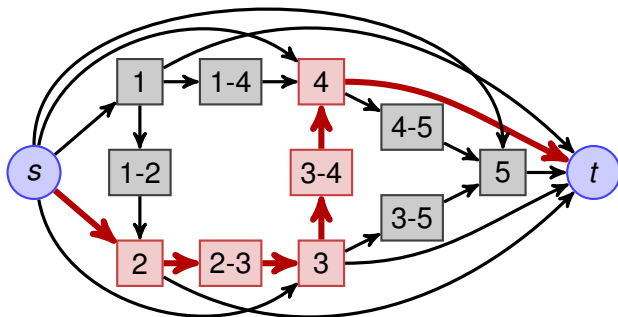
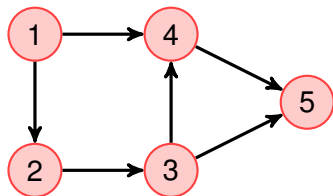
# Isoforms are Paths in a Graph



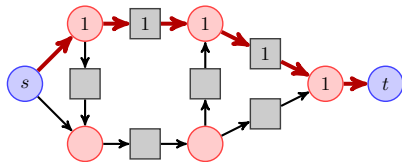
# Isoforms are Paths in a Graph



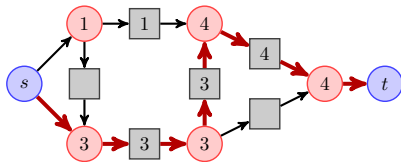
# Isoforms are Paths in a Graph



# Combinations of isoforms are flows



(a) Reads at every node corresponding to one isoform.



(b) Reads at every node after adding another isoform.

- $L(\phi)$  depends only on the values of the flow on the vertices
- $\|\phi\|_1 = f_t$

Therefore,

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

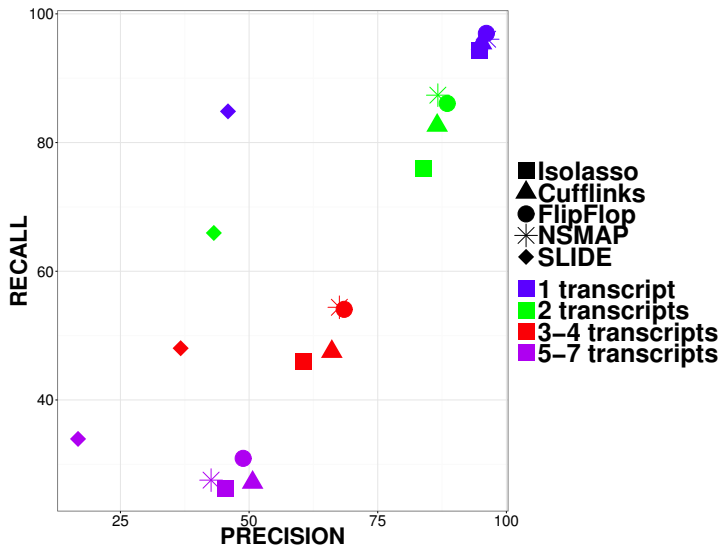
is equivalent to

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

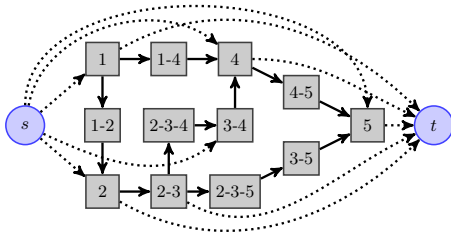
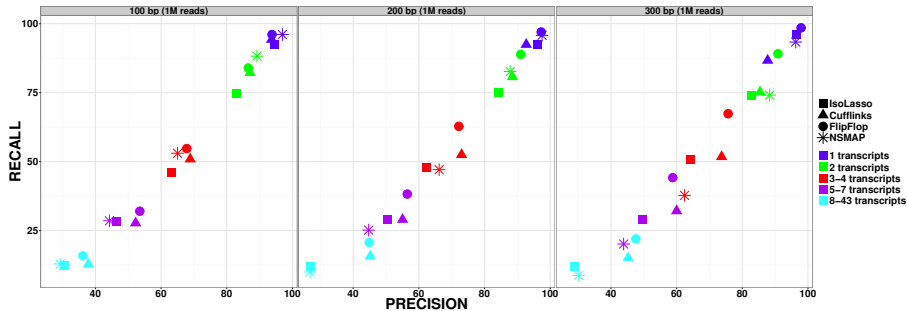
# Human Simulation: Precision/Recall

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

Simulator: <http://alumni.cs.ucr.edu/~liw/rnaseqreadsimulator.html>

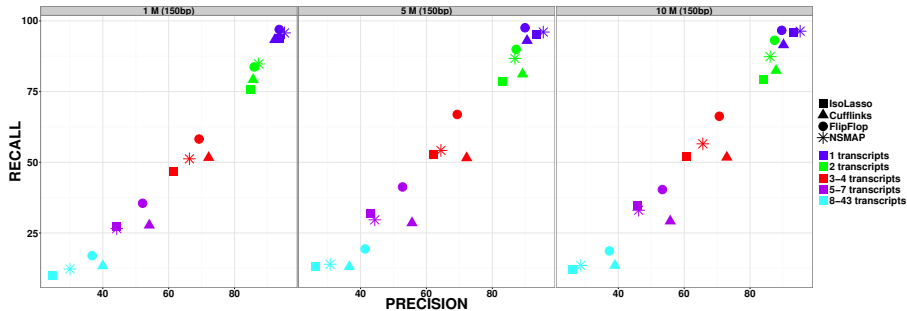


# Performance increases with read length

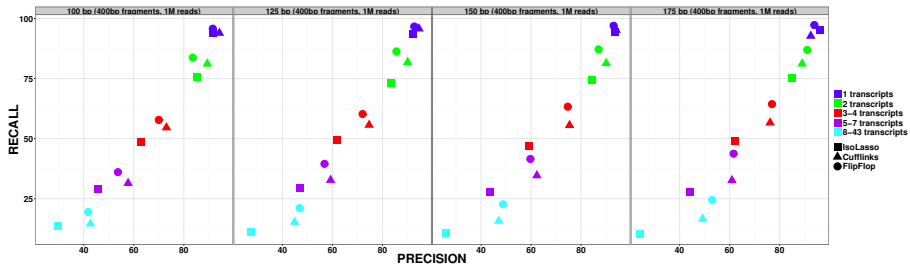




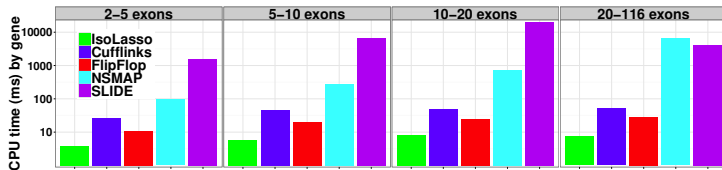
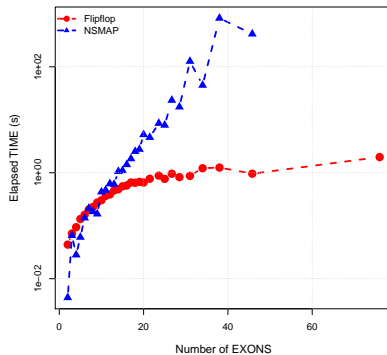
# Performance increases with coverage



# Extension to paired-end reads OK.



# Speed trial



# FlipFlop summary

- Fast method for exact Lasso-based isoform detection and quantification
- <http://cbio.mines-paristech.fr/flipflop>
- Available as an R package

```
> source("http://bioconductor.org/biocLite.R")
> biocLite("flipflop")
```
- Reference: E. Bernard, L. Jacob, J. Mairal and J.-P. Vert. Efficient RNA isoform identification and quantification from RNA-seq data with network flows. *Bioinformatics*, 2014.
- Ongoing: extension to **multiple samples** and **differential analysis**

# Outline

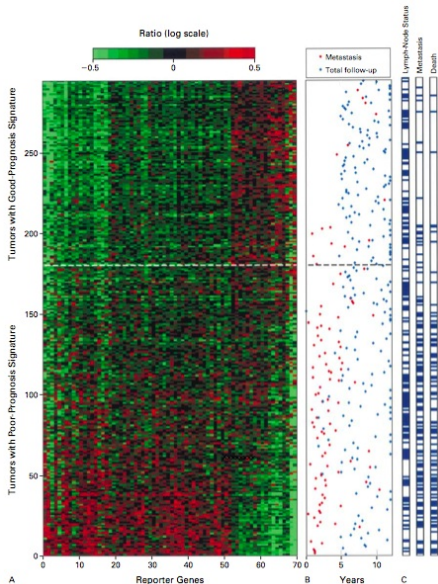
- 1 FlipFlop: fast isoform prediction from RNA-seq data
- 2 Learning molecular classifiers with network information
- 3 Kernel bilinear regression for toxicogenomics

## Joint work with...

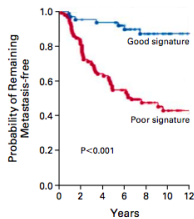


Franck Rapaport, Emmanuel Barillot, Andrei Zinovyev, Anne-Claire Haury, Laurent Jacob, Guillaume Obozinski

# Breast cancer prognosis



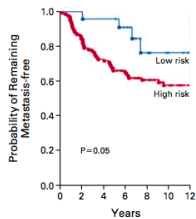
**A Gene-Expression Profiling**



NO. AT RISK

Good signature	60	57	54	45	31	22	12
Poor signature	91	72	55	41	26	17	9

**B St. Gallen Criteria**



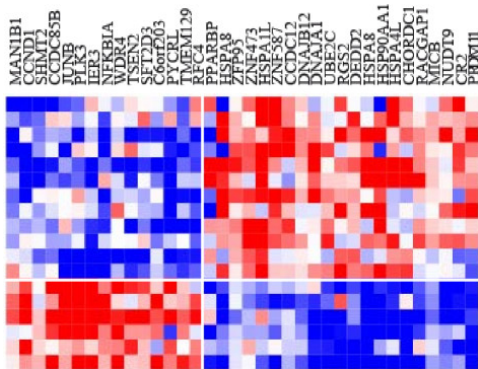
NO. AT RISK

Low risk	22	22	21	17	9	5	2
High risk	129	107	88	69	48	34	19

# Gene selection, molecular signature

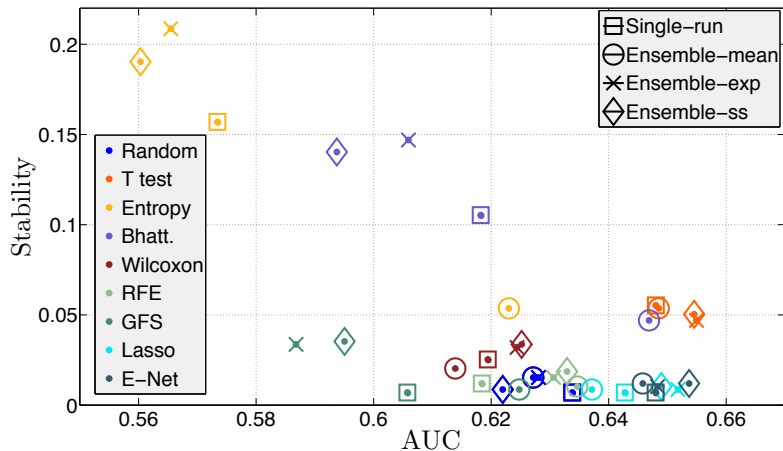
## The idea

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



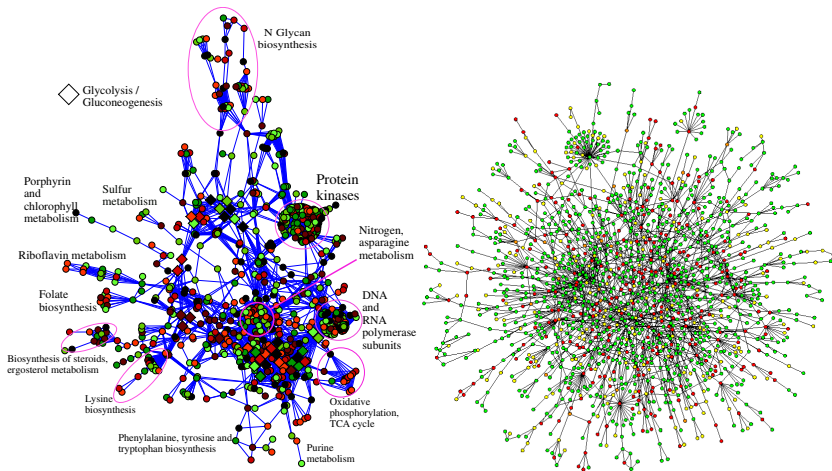


# Lack of stability of signatures



Haury et al. (2011)

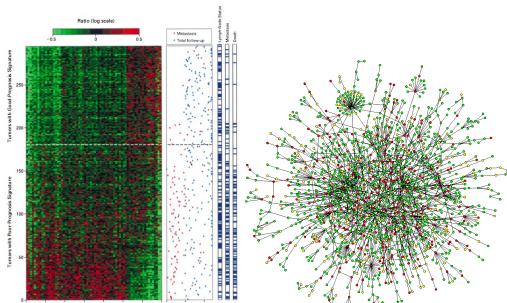
# Gene networks



# Gene networks and expression data

## Motivation

- Basic biological functions usually involve the **coordinated action of several proteins**:
  - Formation of **protein complexes**
  - Activation of metabolic, signalling or regulatory **pathways**
- Many pathways and protein-protein interactions are **already known**
- **Hypothesis**: the weights of the classifier should be “coherent” with respect to this **prior knowledge**



# Graph based penalty

$$f_{\beta}(x) = \beta^T x \quad \min_{\beta} R(f_{\beta}) + \lambda \Omega(\beta)$$

## Prior hypothesis

Genes near each other on the graph should have **similar weights**.

An idea (Rapaport et al., 2007)

$$\Omega(\beta) = \sum_{i \sim j} (\beta_i - \beta_j)^2,$$

$$\min_{\beta \in \mathbb{R}^p} R(f_{\beta}) + \lambda \sum_{i \sim j} (\beta_i - \beta_j)^2.$$

# Graph based penalty

$$f_{\beta}(x) = \beta^T x \quad \min_{\beta} R(f_{\beta}) + \lambda \Omega(\beta)$$

## Prior hypothesis

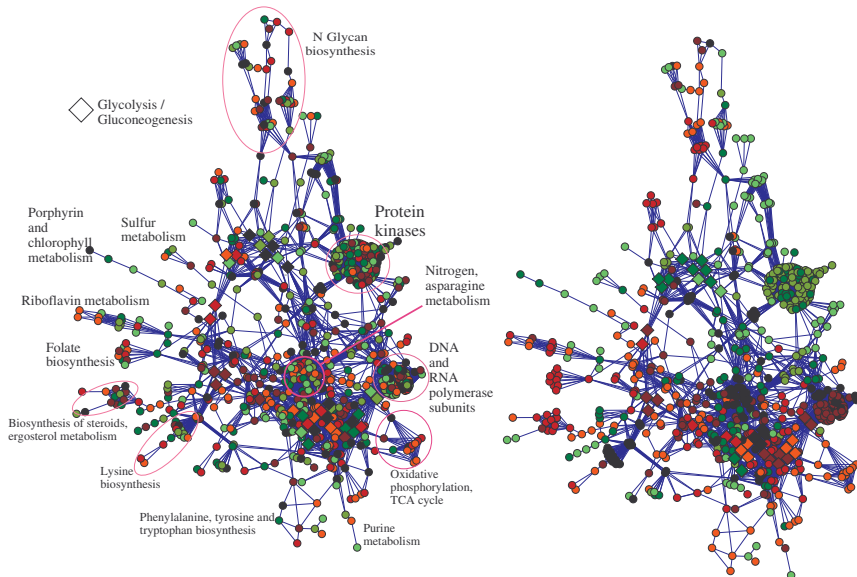
Genes near each other on the graph should have **similar weights**.

## An idea (Rapaport et al., 2007)

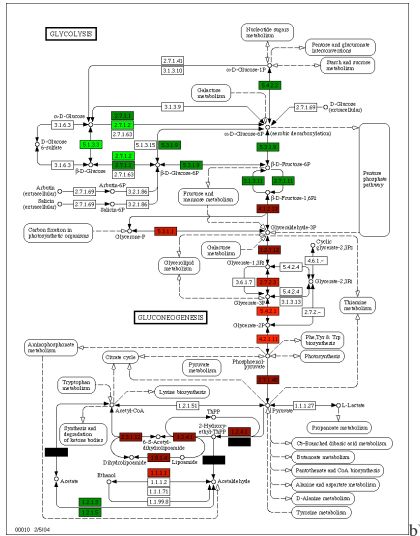
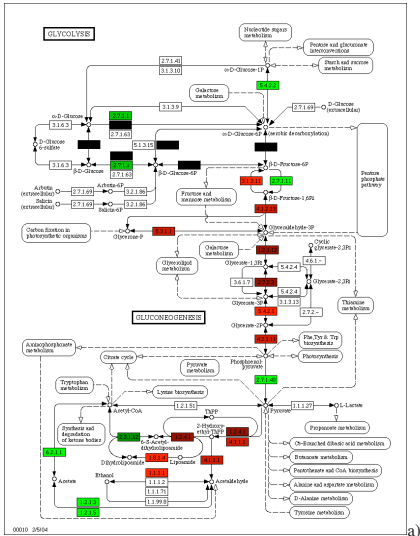
$$\Omega(\beta) = \sum_{i \sim j} (\beta_i - \beta_j)^2,$$

$$\min_{\beta \in \mathbb{R}^p} R(f_{\beta}) + \lambda \sum_{i \sim j} (\beta_i - \beta_j)^2.$$

# Classifiers



# Classifier



# Spectral penalty as a kernel

## Theorem

The function  $f(x) = \beta^\top x$  where  $\beta$  is solution of

$$\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n \ell(\beta^\top x_i, y_i) + \lambda \sum_{i \sim j} (\beta_i - \beta_j)^2$$

is equal to  $g(x) = \gamma^\top \Phi(x)$  where  $\gamma$  is solution of

$$\min_{\gamma \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n \ell(\gamma^\top \Phi(x_i), y_i) + \lambda \gamma^\top \gamma,$$

and where

$$\Phi(x)^\top \Phi(x') = x^\top K_G x'$$

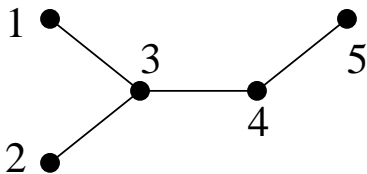
for  $K_G = L^*$ , the pseudo-inverse of the graph Laplacian.



# Graph Laplacian

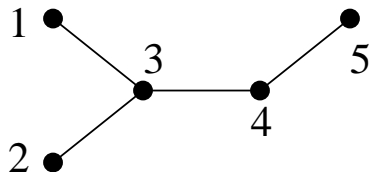
## Definition

The Laplacian of the graph is the matrix  $L = D - A$ .



$$L = D - A = \begin{pmatrix} 1 & 0 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 \\ -1 & -1 & 3 & -1 & 0 \\ 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & 1 & 1 \end{pmatrix}$$

# Pseufo-inverse of the Laplacian



$$L^* = \begin{pmatrix} 0.88 & -0.12 & 0.08 & -0.32 & -0.52 \\ -0.12 & 0.88 & 0.08 & -0.32 & -0.52 \\ 0.08 & 0.08 & 0.28 & -0.12 & -0.32 \\ -0.32 & -0.32 & -0.12 & 0.48 & 0.28 \\ -0.52 & -0.52 & -0.32 & 0.28 & 1.08 \end{pmatrix}$$

## Other penalties with kernels

$$\Phi(x)^\top \Phi(x') = x^\top K_G x'$$

with:

- $K_G = (c + L)^{-1}$  leads to

$$\Omega(\beta) = c \sum_{i=1}^p \beta_i^2 + \sum_{i \sim j} (\beta_i - \beta_j)^2 .$$

- The diffusion kernel:

$$K_G = \exp_M(-2tL) .$$

penalizes high frequencies of  $\beta$  in the Fourier domain.

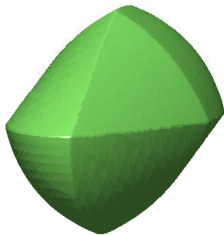
## Other penalties without kernels

- Gene selection + Piecewise constant on the graph

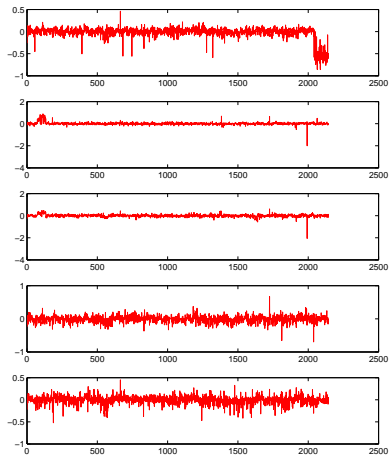
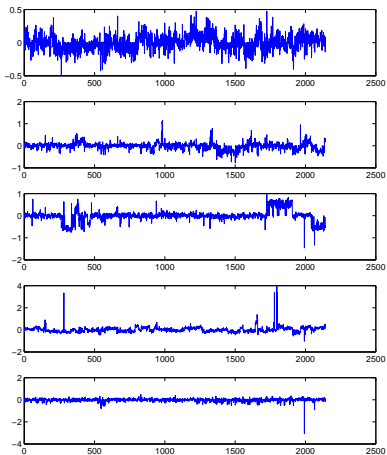
$$\Omega(\beta) = \sum_{i \sim j} |\beta_i - \beta_j| + \sum_{i=1}^p |\beta_i|$$

- Gene selection + smooth on the graph

$$\Omega(\beta) = \sum_{i \sim j} (\beta_i - \beta_j)^2 + \sum_{i=1}^p |\beta_i|$$



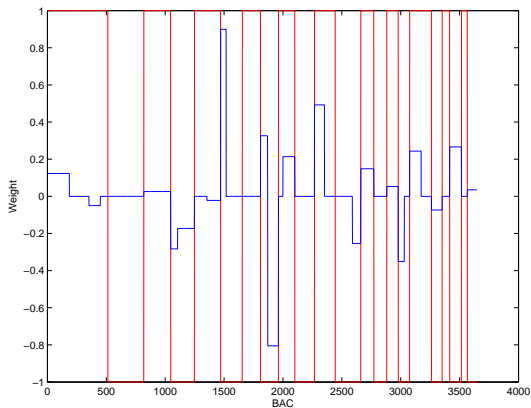
# Example: classification of DNA copy number profiles



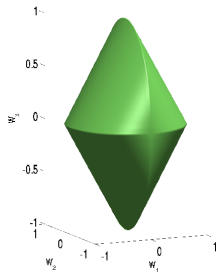
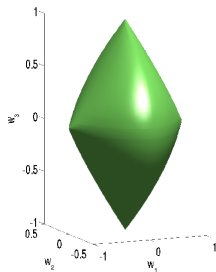
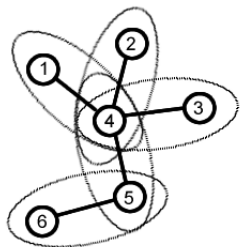
*Aggressive (left) vs non-aggressive (right) melanoma*

# Fused lasso solution (Rapaport et al., 2008)

$$\Omega(\beta) = \sum_{i \sim j} |\beta_i - \beta_j| + \sum_{i=1}^p |\beta_i|$$



# Graph-based structured feature selection

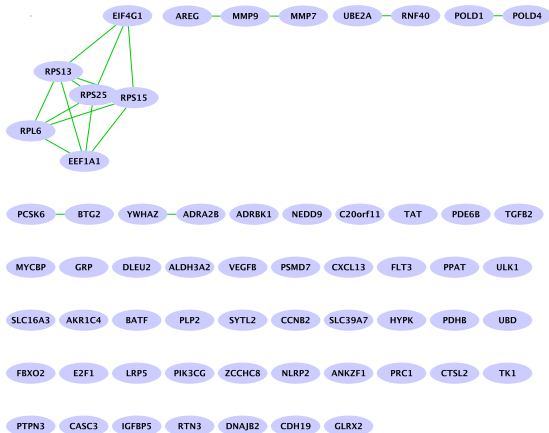


## Graph lasso(s)

$$\Omega_1(\beta) = \sum_{i \sim j} \sqrt{\beta_i^2 + \beta_j^2}, \quad (\text{Jenatton et al., 2009})$$

$$\Omega_2(\beta) = \sup_{\alpha \in \mathbb{R}^p: \forall i \sim j, \|\alpha_i^2 + \alpha_j^2\| \leq 1} \alpha^\top \beta. \quad (\text{Jacob et al., 2008})$$

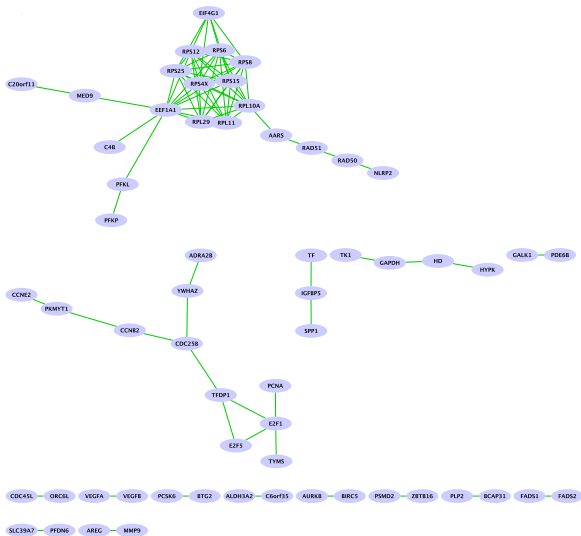
# Lasso signature (accuracy 0.61)



*Breast cancer prognosis*



# Graph Lasso signature (accuracy 0.64)

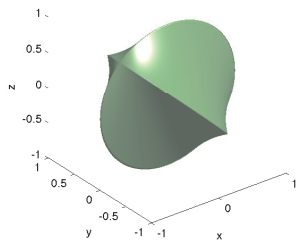
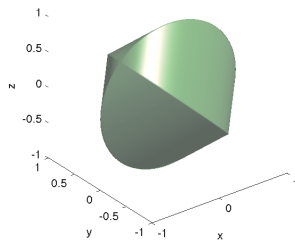
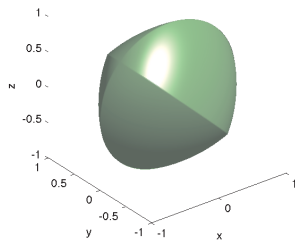


*Breast cancer prognosis*

# Disjoint feature selection

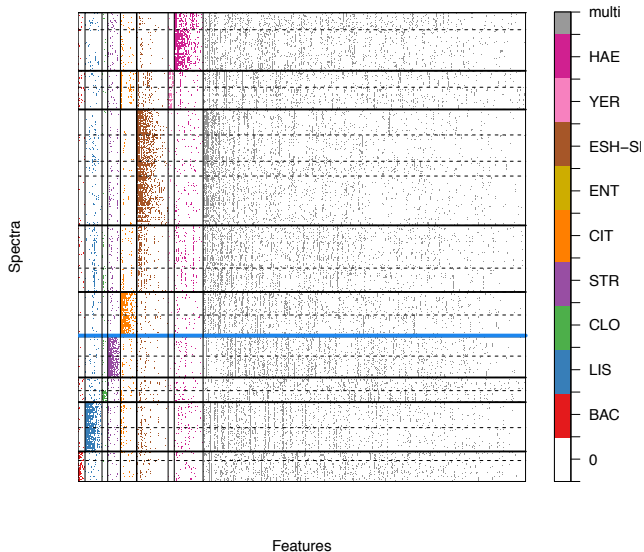
$$W = (w_i)_{i \in V} \in \mathbb{R}^{p \times V}$$

$$\Omega(W) = \min_{-H \leq W \leq H} \sum_{i \sim j} K_{ij} |h_i^\top h_j|$$



(Vervier et al, 2014)

# Example: multiclass classification of MS spectra



*(Vervier et al, 2013, unpublished)*

# Outline

- 1 FlipFlop: fast isoform prediction from RNA-seq data
- 2 Learning molecular classifiers with network information
- 3 Kernel bilinear regression for toxicogenomics**

## Joint work with...

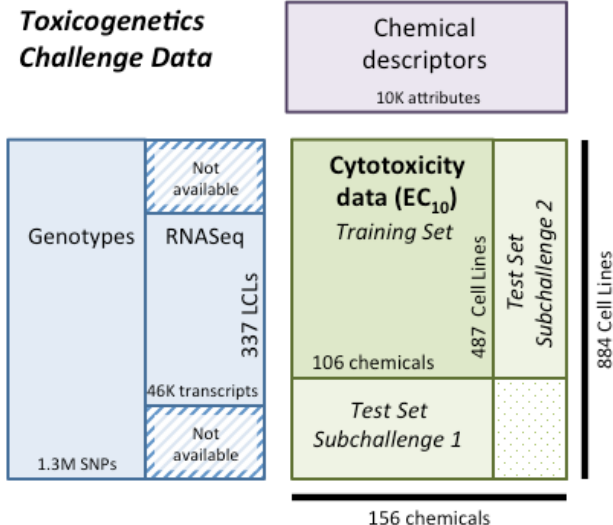


Elsa Bernard, Erwan Scornet, Yunlong Jiao, Véronique Stoven,  
Thomas Walter

# Pharmacogenomics / Toxicogenomics



# DREAM8 Toxicogenetics challenge



Genotypes from the 1000 genome project  
RNASeq from the Geuvadis project

# Bilinear regression

- Cell line  $X$ , chemical  $Y$ , toxicity  $Z$ .
- Bilinear regression model:

$$Z = f(X, Y) + b(Y) + \epsilon,$$

- Estimation by kernel ridge regression:

$$\min_{f \in \mathcal{H}, b \in \mathbb{R}^p} \sum_{i=1}^n \sum_{j=1}^p (f(x_i, y_j) + b_j - z_{ij})^2 + \lambda \|f\|^2,$$



# Solving in $O(\max(n, p)^3)$

**Theorem 1.** Let  $Z \in \mathbb{R}^{n \times p}$  be the response matrix, and  $K_X \in \mathbb{R}^{n \times n}$  and  $K_Y \in \mathbb{R}^{p \times p}$  be the kernel Gram matrices of the  $n$  cell lines and  $p$  chemicals, with respective eigenvalue decompositions  $K_X = U_X D_X U_X^\top$  and  $K_Y = U_Y D_Y U_Y^\top$ . Let  $\gamma = U_X^\top \mathbf{1}_n$  and  $S \in \mathbb{R}^{n \times p}$  be defined by  $S_{ij} = 1 / (\lambda + D_X^i D_Y^j)$ , where  $D_X^i$  (resp.  $D_Y^j$ ) denotes the  $i$ -th diagonal term of  $D_X$  (resp.  $D_Y$ ). Then the solution  $(f^*, b^*)$  of (2) is given by

$$b^* = U_Y \text{Diag} \left( S^\top \gamma^{\circ 2} \right)^{-1} \left( S^\top \circ \left( U_Y^\top Z^\top U_X \right) \right) \gamma \quad (3)$$

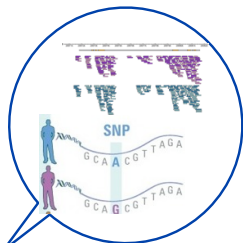
and

$$\forall (x, y) \in \mathcal{X} \times \mathcal{Y}, \quad f^*(x, y) = \sum_{i=1}^n \sum_{j=1}^p \alpha_{i,j}^* K_X(x_i, x) K_Y(y_i, y), \quad (4)$$

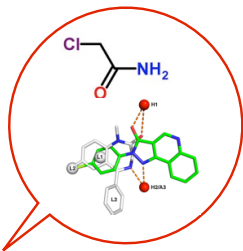
where

$$\alpha^* = U_X \left( S \circ \left( U_X^\top \left( Z - \mathbf{1}_n b^{*\top} \right) U_Y \right) \right) U_Y^\top. \quad (5)$$

# Kernel Trick

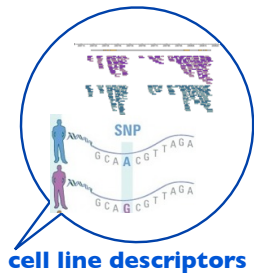


**cell line descriptors**

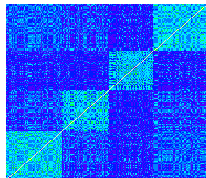


**drug descriptors**

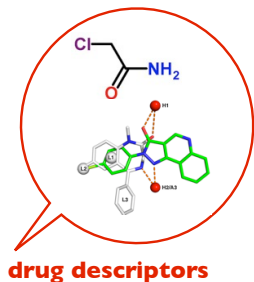
# Kernel Trick



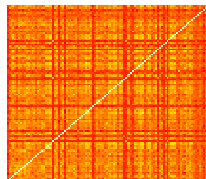
**Kcell**



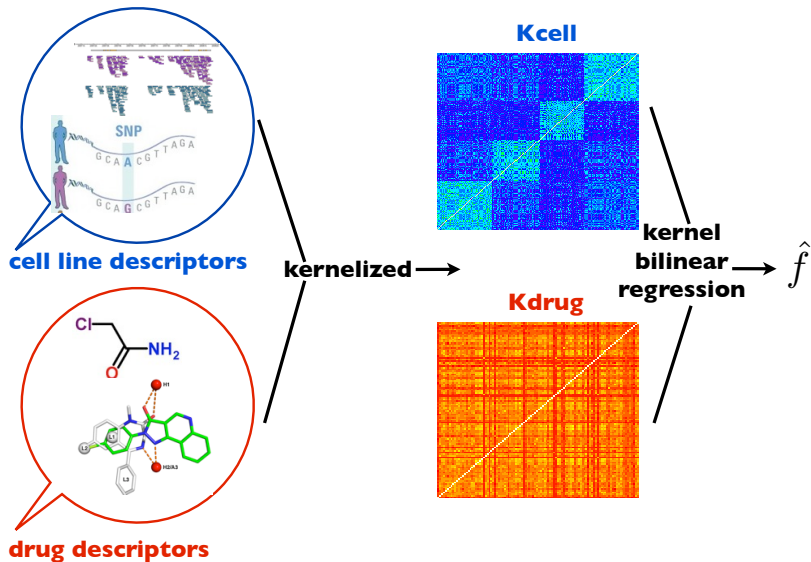
kernelized →



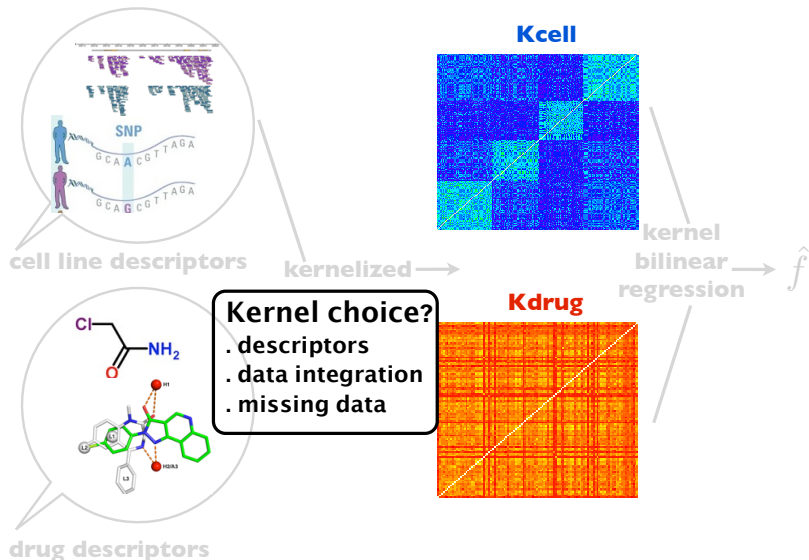
**Kdrug**



# Kernel Trick



# Kernel Trick



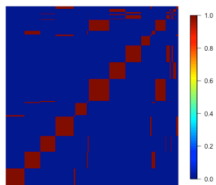
- 1  $K_{\text{cell}}$  :
  - ⇒ 29 cell line kernels tested
  - ⇒ 1 kernel that *integrate all information*
  - ⇒ deal with missing data
- 2  $K_{\text{drug}}$  :
  - ⇒ 48 drug kernels tested
  - ⇒ multi-task kernels

- 1  $K_{\text{cell}}$  :
  - ⇒ 29 cell line kernels tested
  - ⇒ 1 kernel that *integrate all information*
  - ⇒ deal with missing data
- 2  $K_{\text{drug}}$  :
  - ⇒ 48 drug kernels tested
  - ⇒ **multi-task** kernels

# Cell line data integration

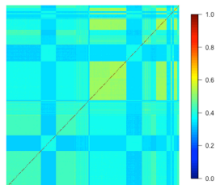
## Covariates

. linear kernel



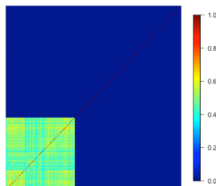
## SNPs

. 10 gaussian kernels



## RNA-seq

. 10 gaussian kernels

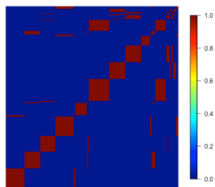




# Cell line data integration

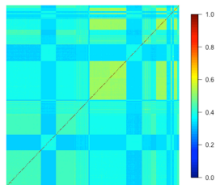
## Covariates

. linear kernel



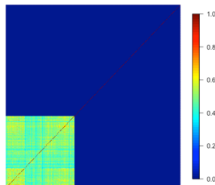
## SNPs

. 10 gaussian kernels

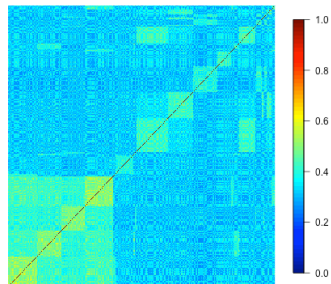


## RNA-seq

. 10 gaussian kernels

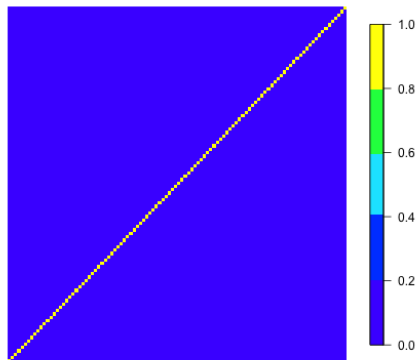


## Integrated kernel



# Multi-task drug kernels

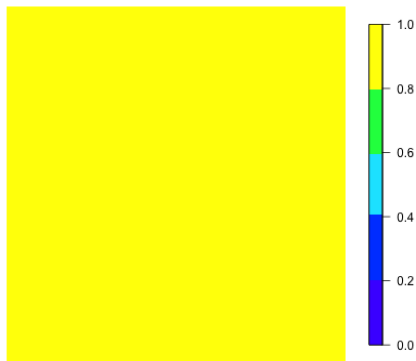
- 1 **Dirac**
- 2 Multi-Task
- 3 Feature-based
- 4 Empirical
- 5 Integrated



independent regression for each drug

# Multi-task drug kernels

- 1 Dirac
- 2 **Multi-Task**
- 3 Feature-based
- 4 Empirical
- 5 Integrated



sharing information across drugs

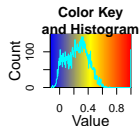
# Multi-task drug kernels

- 1 Dirac
- 2 Multi-Task
- 3 **Feature-based**
- 4 Empirical
- 5 Integrated

Linear kernel and 10 gaussian kernels based on features:

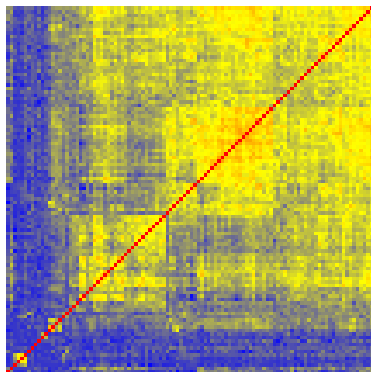
- CDK (160 descriptors) and SIRMS (9272 descriptors)
- Graph kernel for molecules (2D walk kernel)
- Fingerprint of 2D substructures (881 descriptors)
- Ability to bind human proteins (1554 descriptors)

# Multi-task drug kernels



## Empirical correlation

- 1 Dirac
- 2 Multi-Task
- 3 Feature-based
- 4 **Empirical**
- 5 Integrated



# Multi-task drug kernels

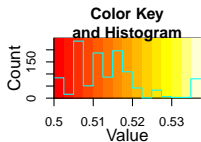
- 1 Dirac
- 2 Multi-Task
- 3 Feature-based
- 4 Empirical
- 5 **Integrated**

$$K_{int} = \sum_i K_i$$

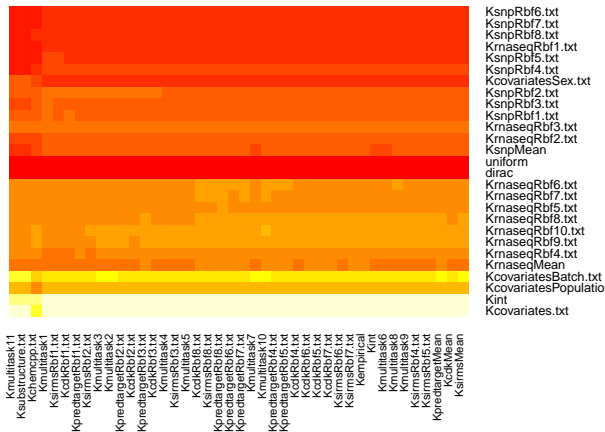
Integrated kernel:

- Combine all information on drugs

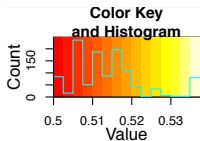
# 29x48 kernel combinations: CV results



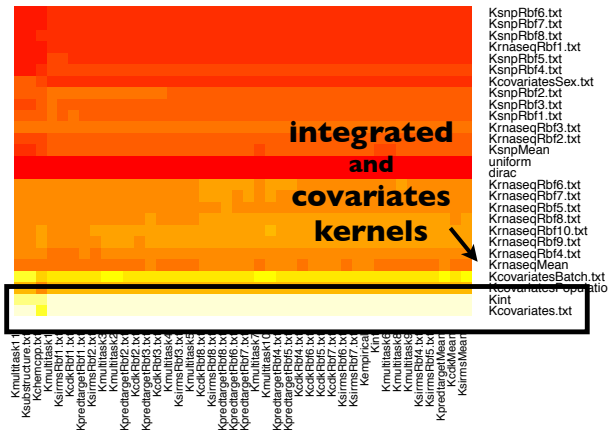
CI



# 29x48 kernel combinations: CV results

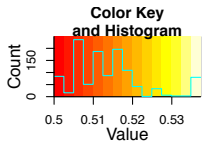


CI

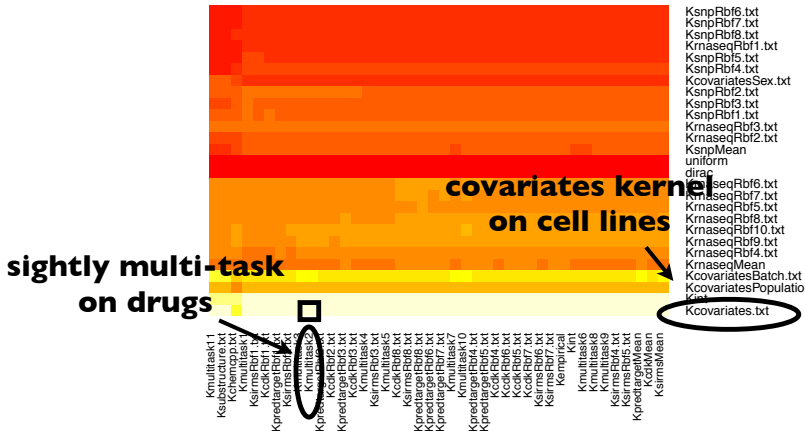




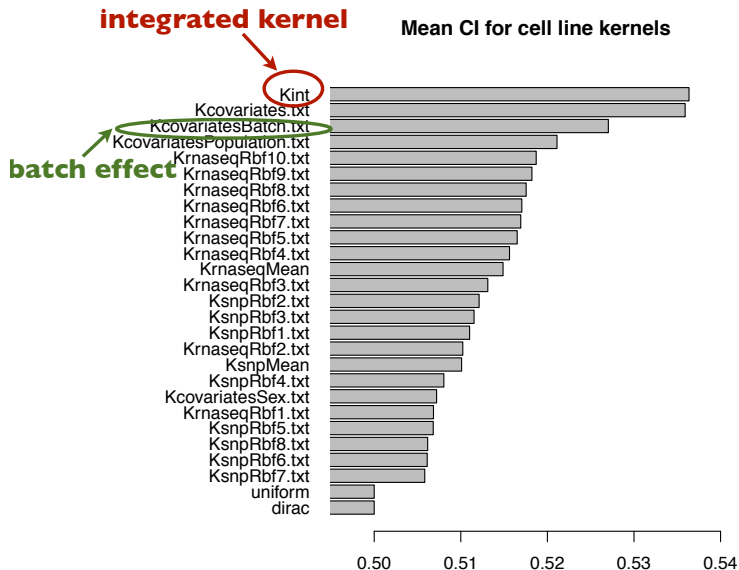
# 29x48 kernel combinations: CV results



CI

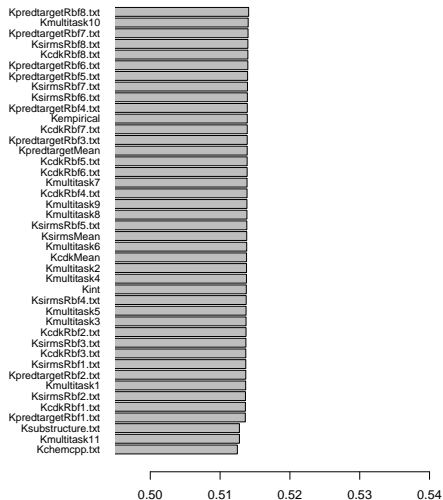


# Kernel on cell lines: CV results



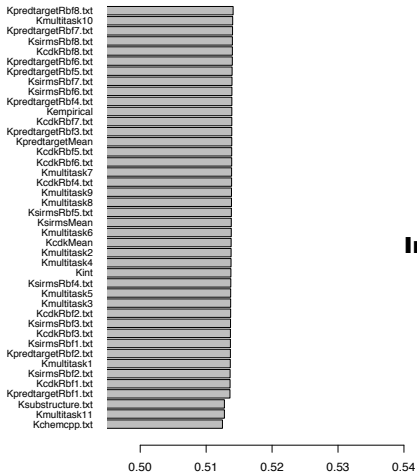
# Kernel on drugs: CV results

Mean CI for chemicals kernels

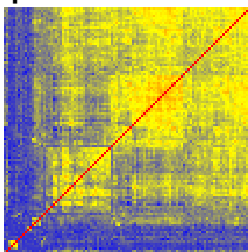


# Final Submission (ranked 2nd)

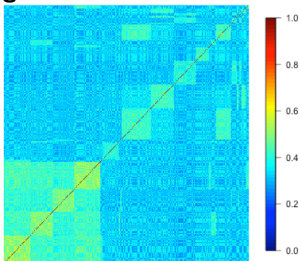
Mean CI for chemicals kernels



Empirical kernel on drugs



Integrated kernel on cell lines



# Conclusion

- Many new problems and lots of data in computational genomics
- Computational constraints  $\implies$  fast sparse models (FlipFlop)
- Small  $n$  large  $p$   $\implies$  regularized models with prior knowledge
- Heterogeneous data integration  $\implies$  kernel methods
- Personalized medicine promising but difficult!

# Thanks

Alexandre d'Aspremont, Emmanuel Barillot, Anne-Claire Haury,  
Laurent Jacob, Pierre Mahé, Julien Mairal, Guillaume Obozinski,  
Franck Rapaport, Jean-Baptiste Veyrieras, Andrei Zynoviev  
... and all CBIO@Mines

