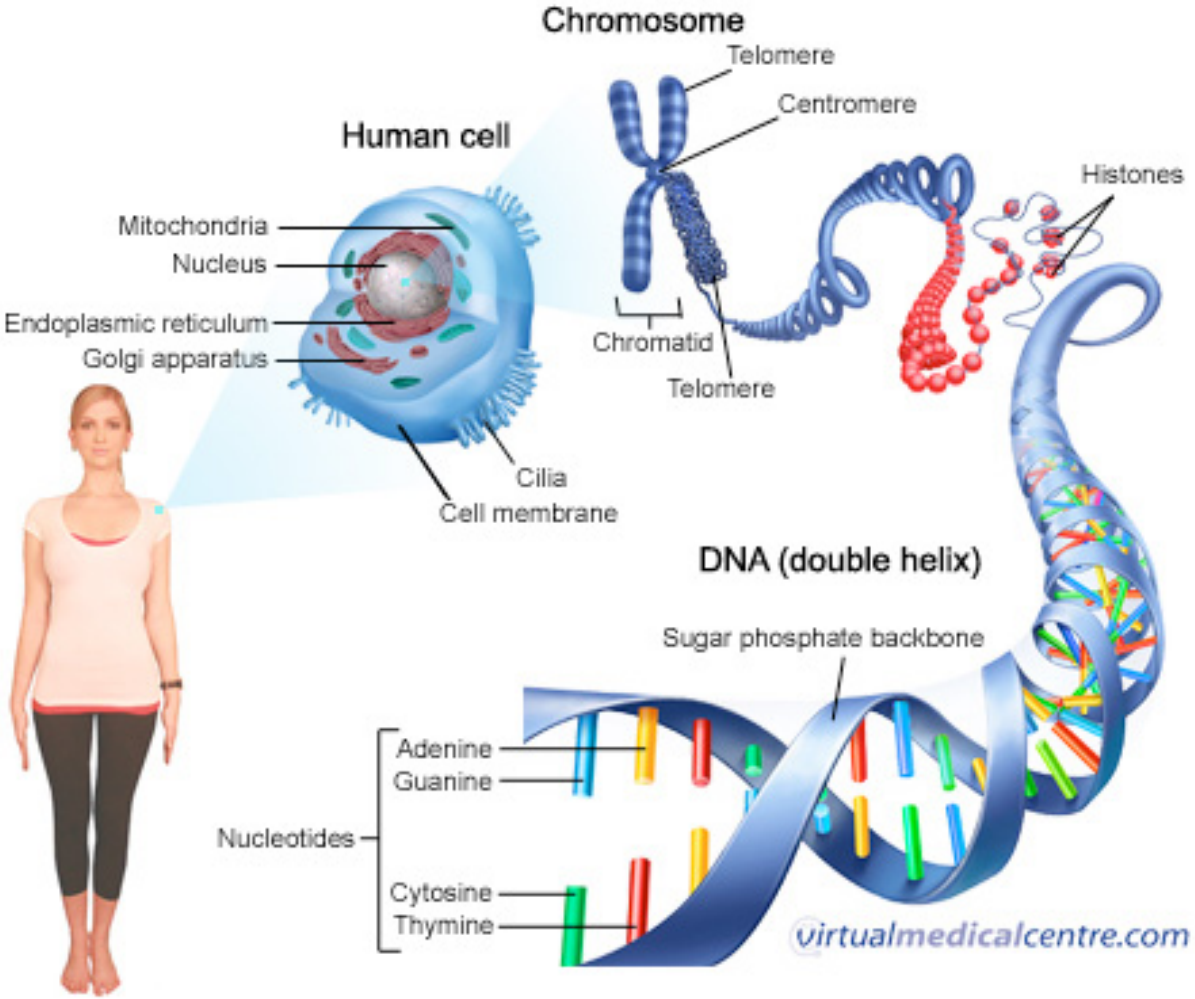
Machine learning for computational genomics and precision medicine

Jean-Philippe Vert
jean-philippe.vert@ens.fr

KAIST, October 7, 2016

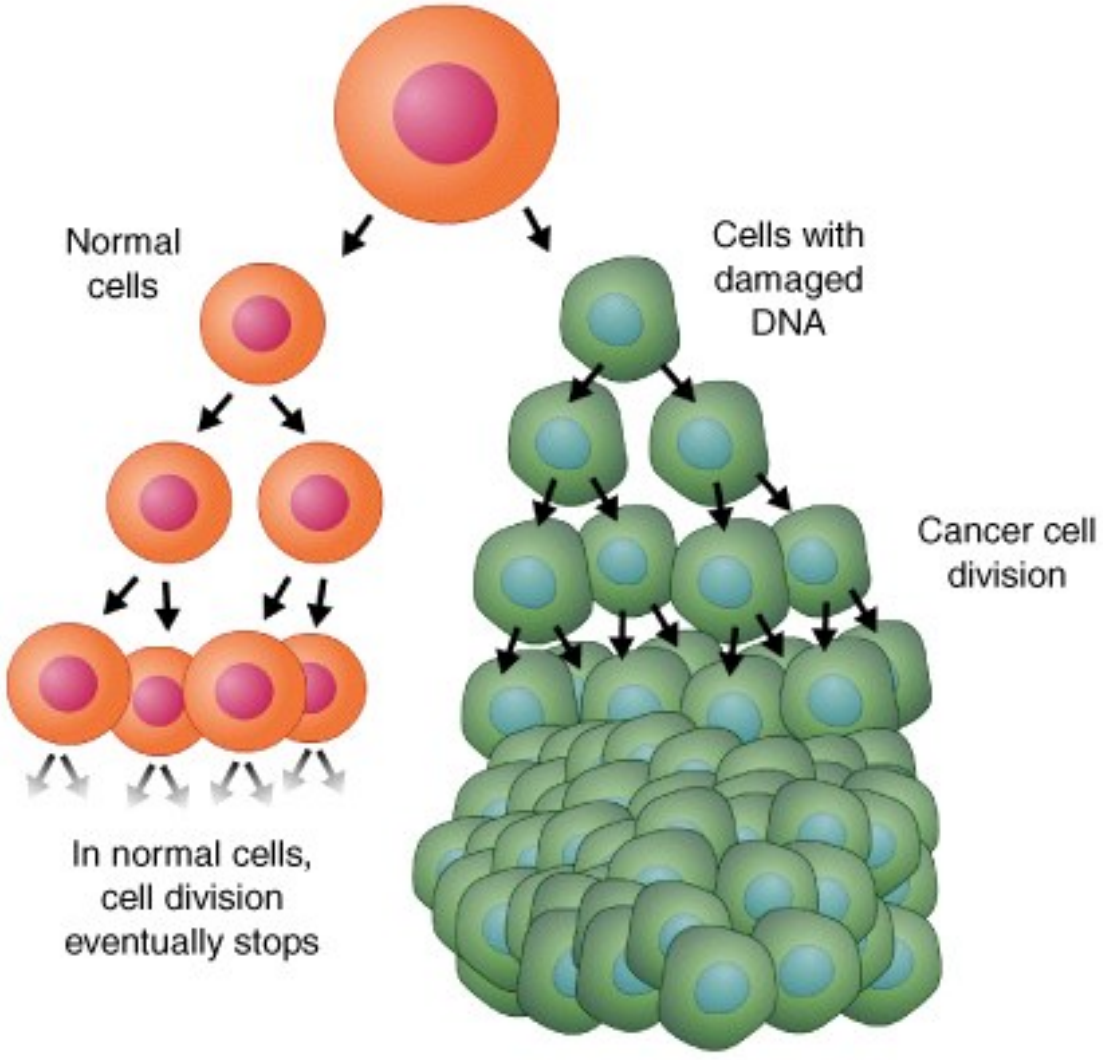
A complex system



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

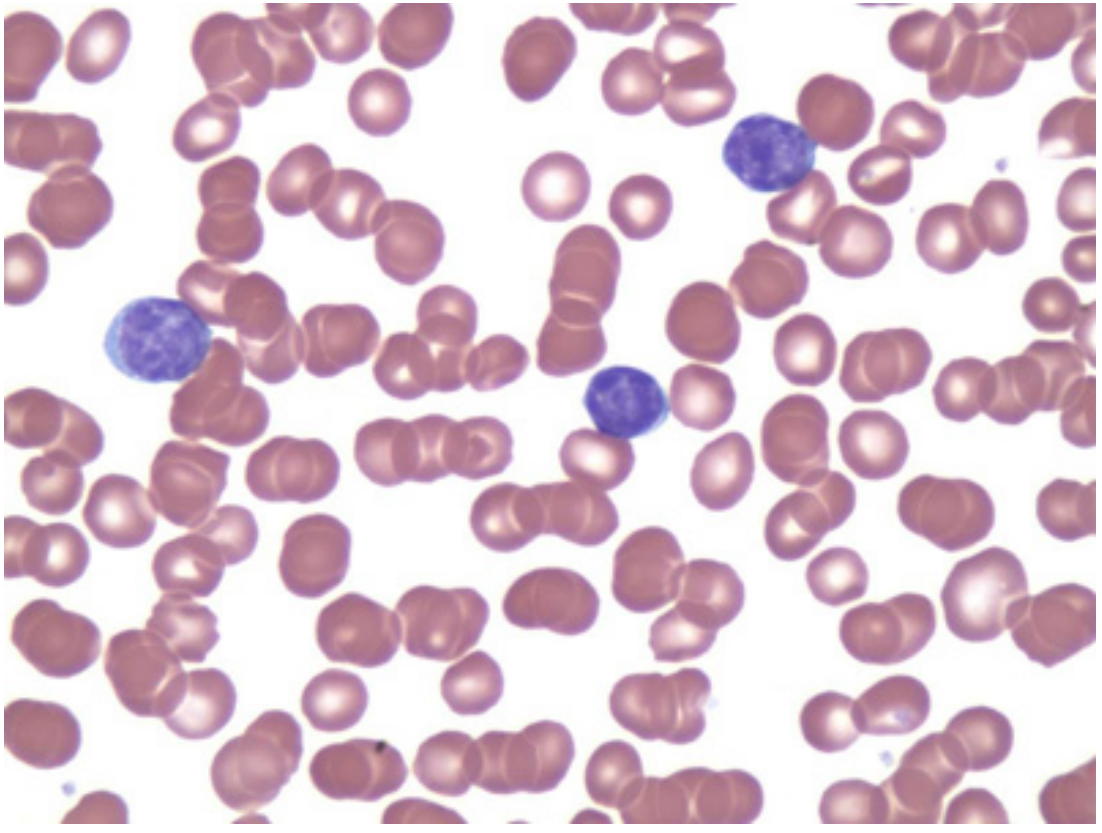
1 cell = 6×10^9 ACGT coding for $20,000$ genes

Cancer

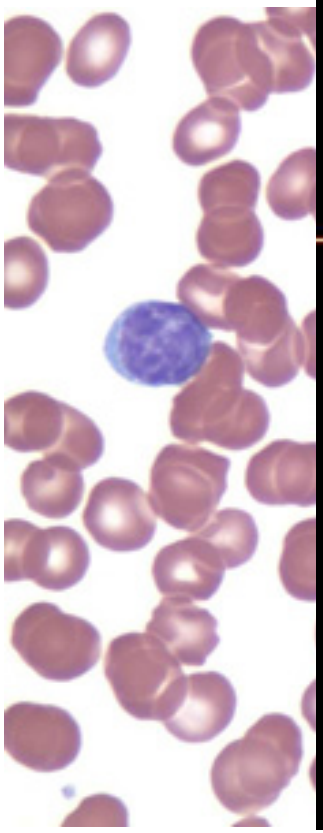


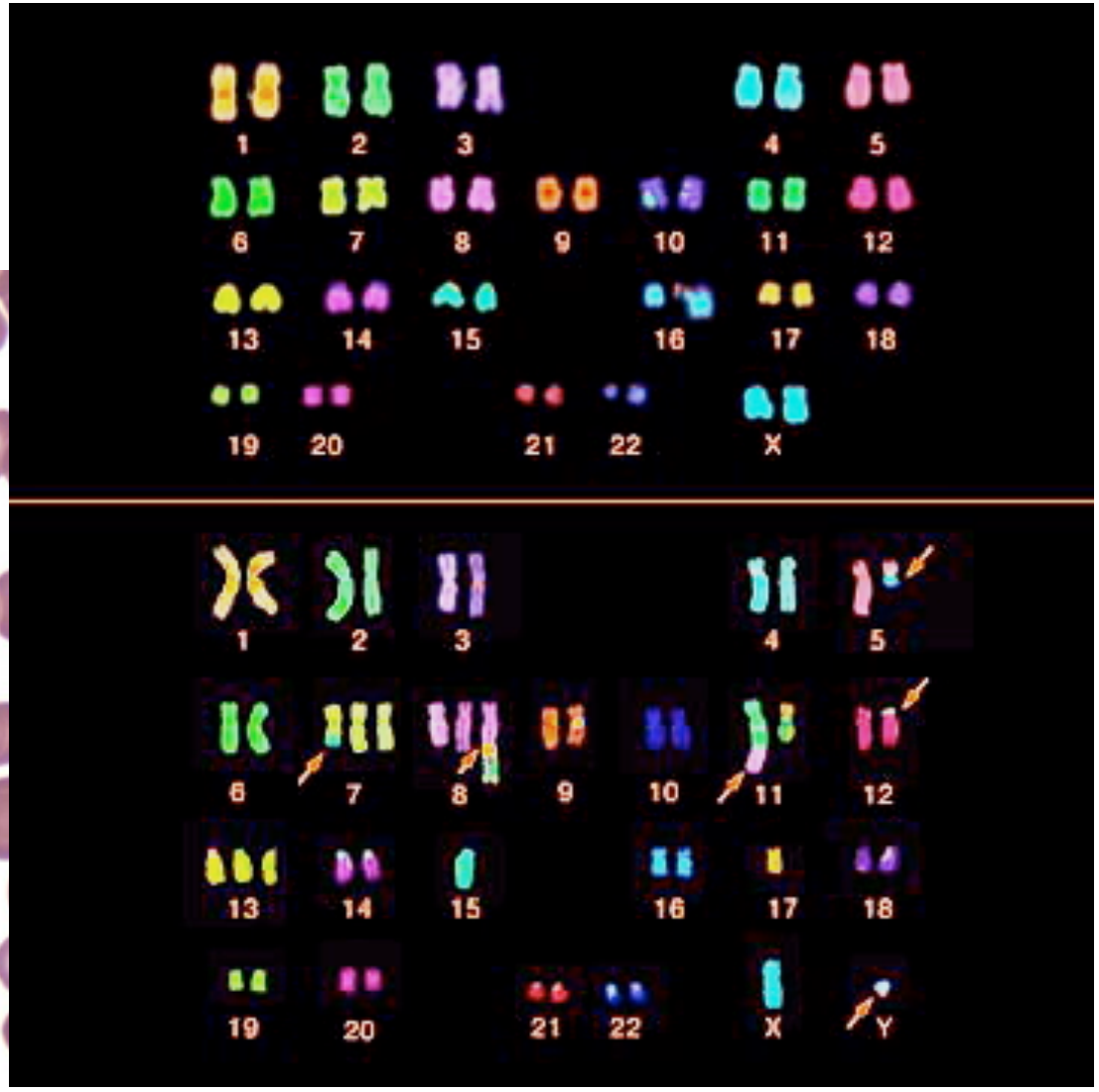
http://rise.duke.edu/seek/pages/page.html?0205

A cancer cell (1900)

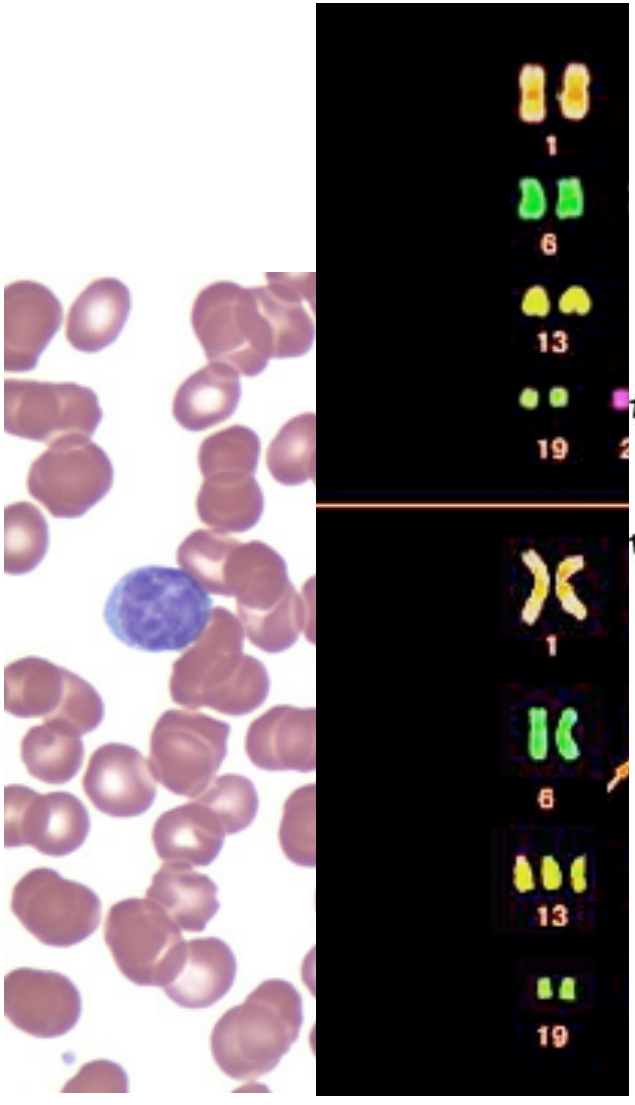


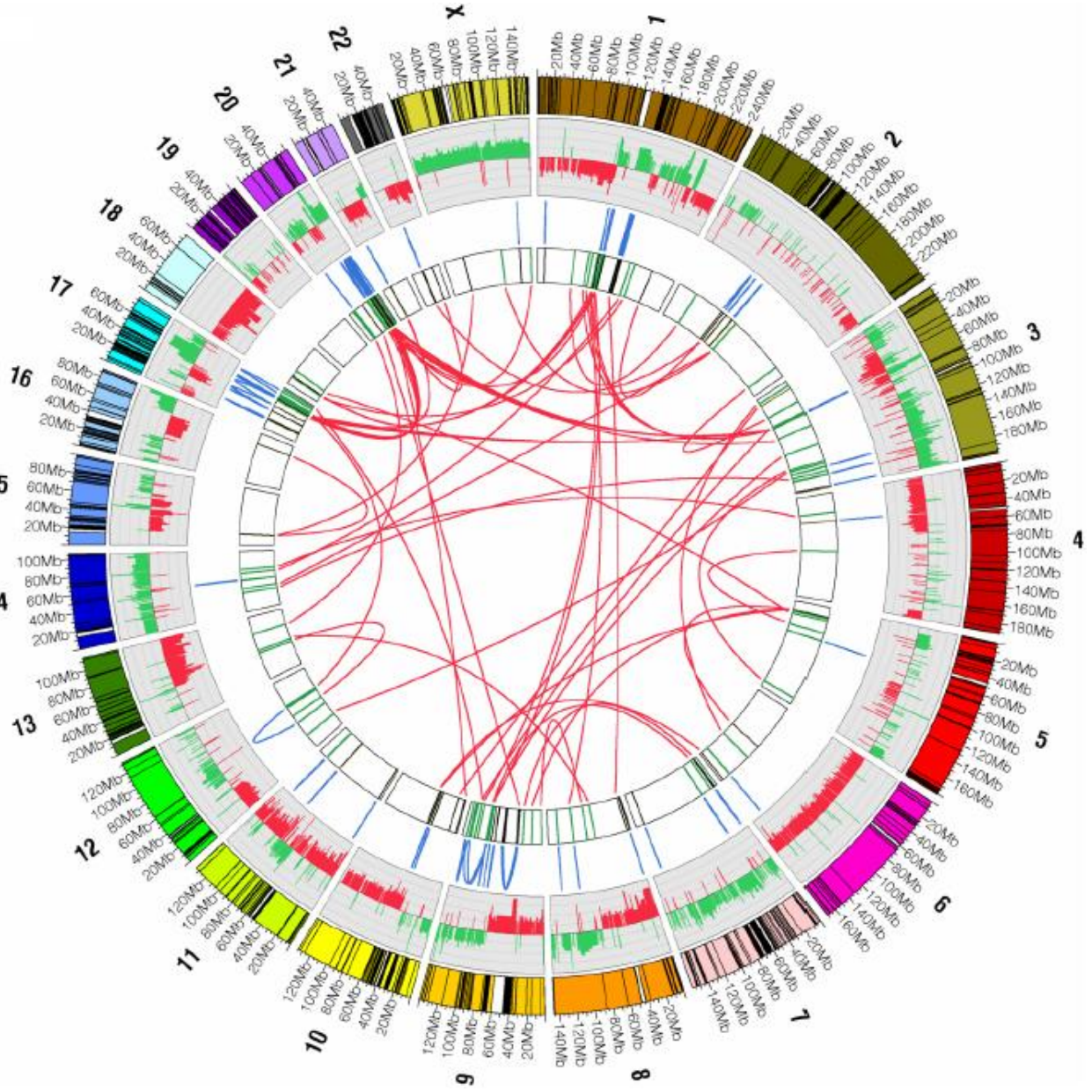
A cancer cell (1960)



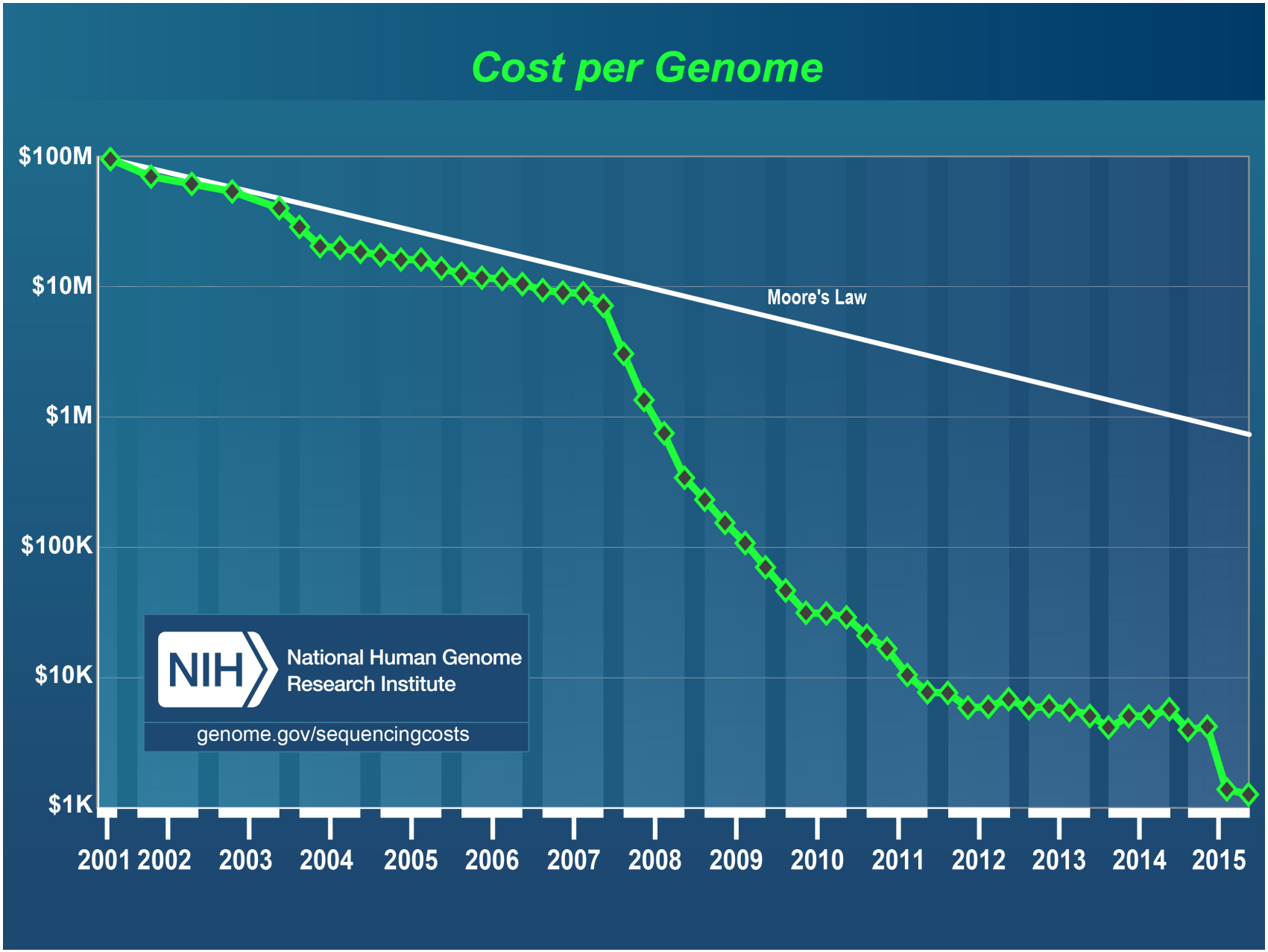


A cancer cell (2010)

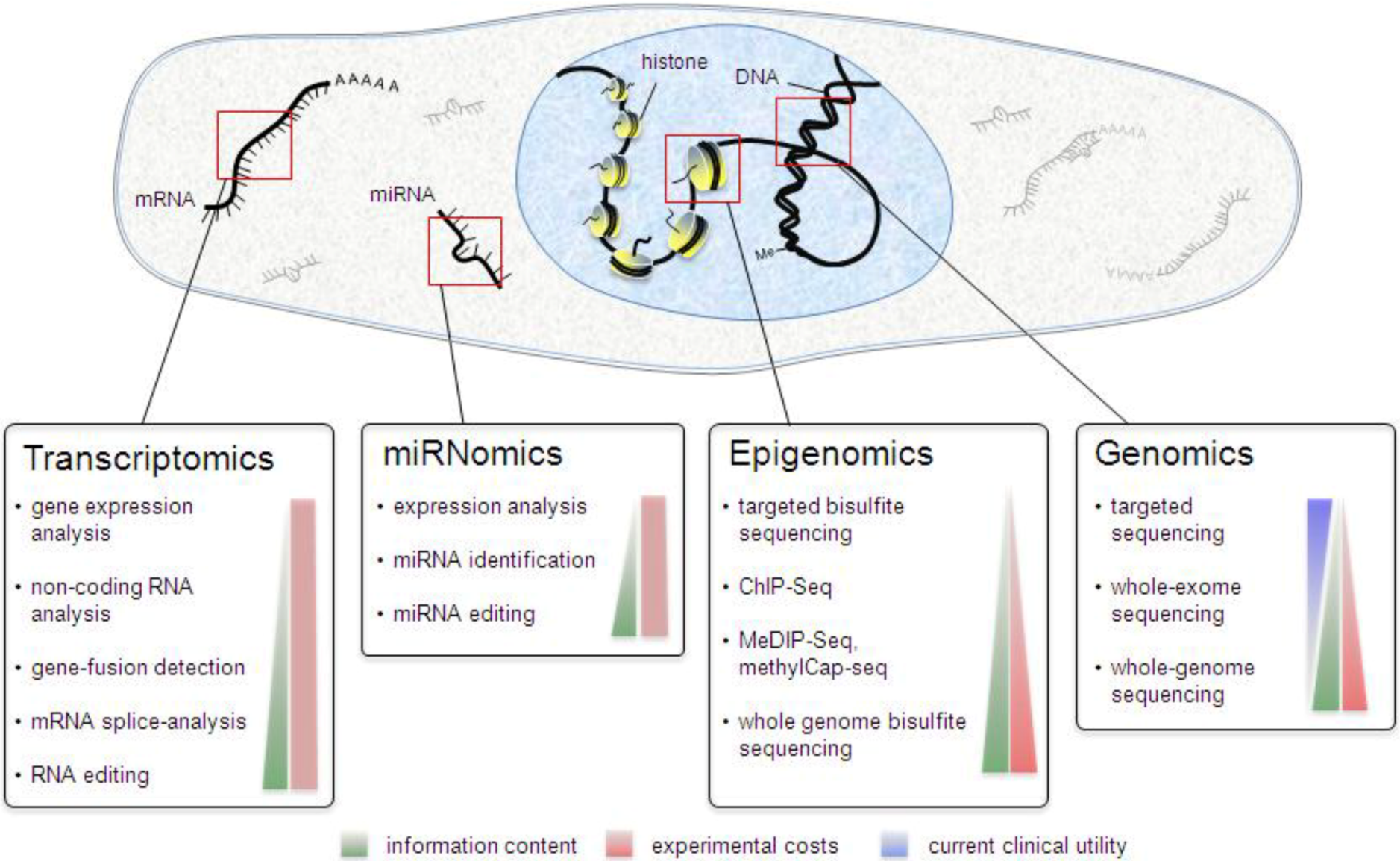




What happened?

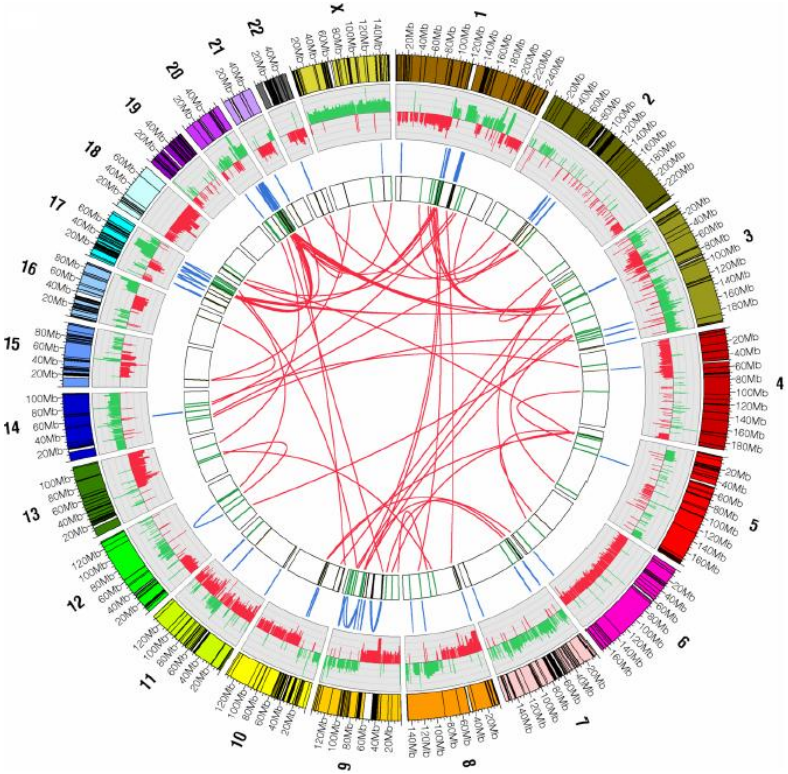


Sequencing has many applications

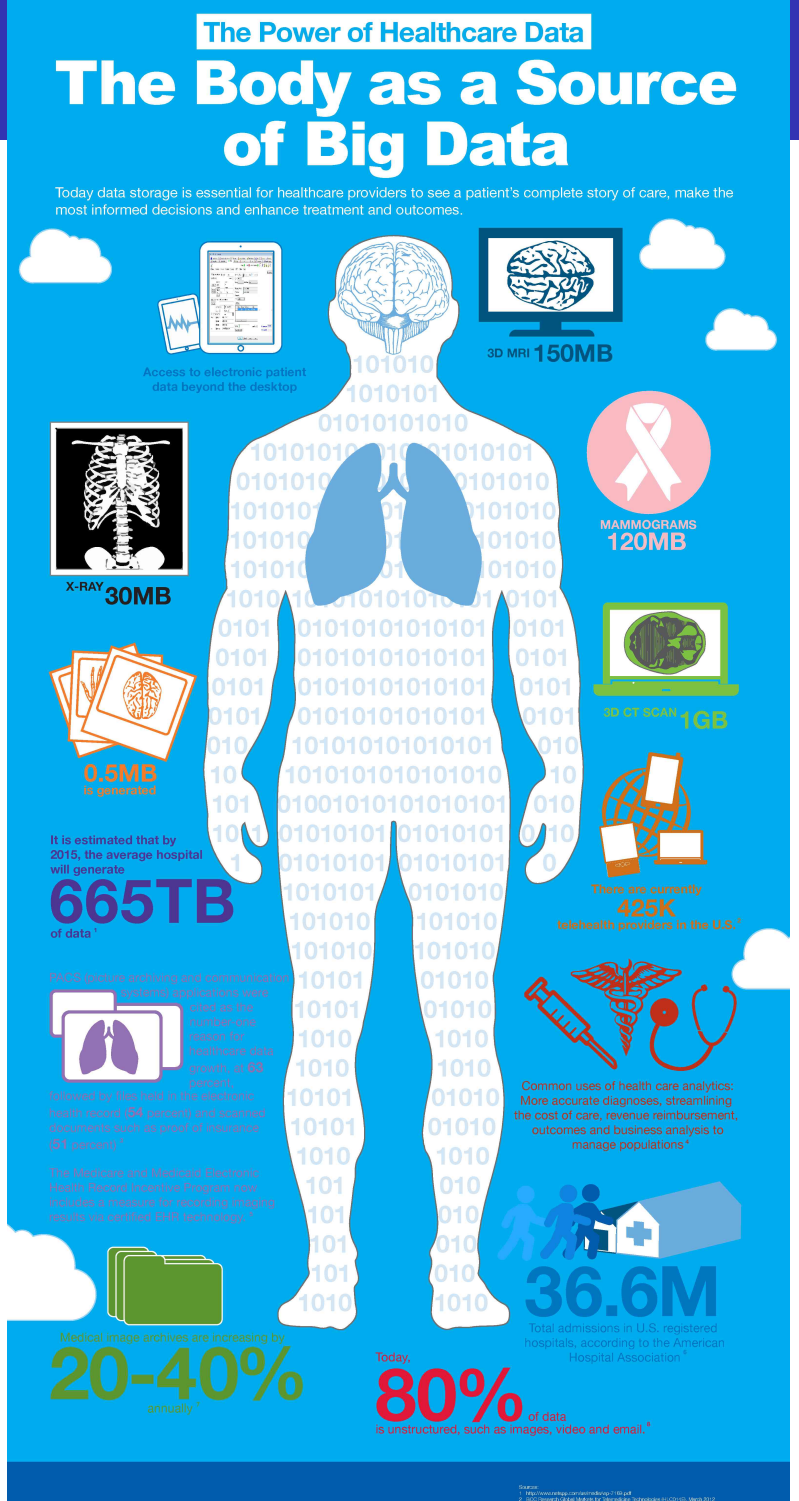


(Frese et al., 2013)

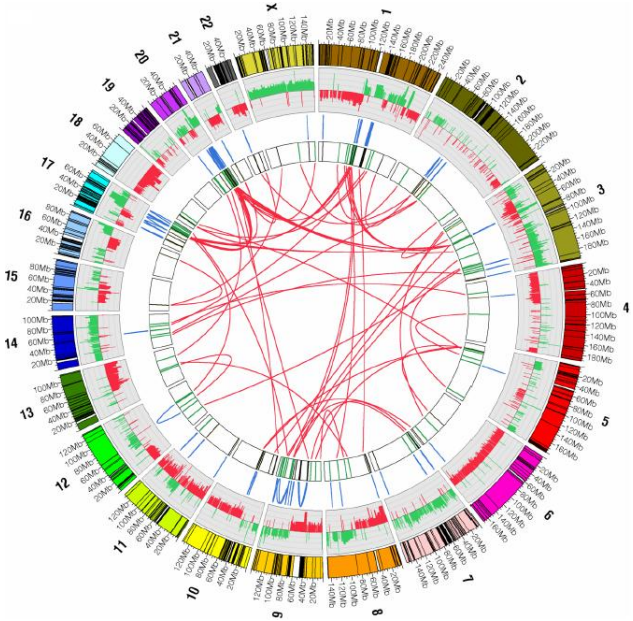
More data to come

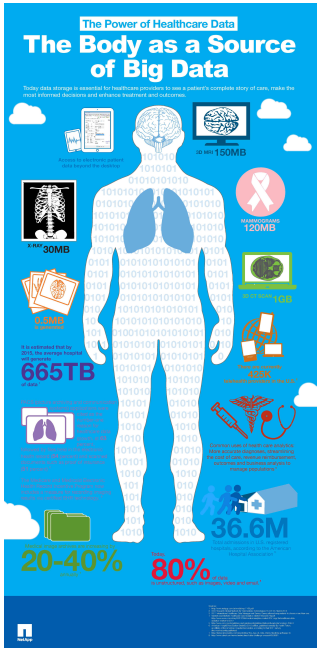


http://ihealthtran.com/wordpress/2013/03/infographic-friday-the-body-as-a-source-of-big-data/



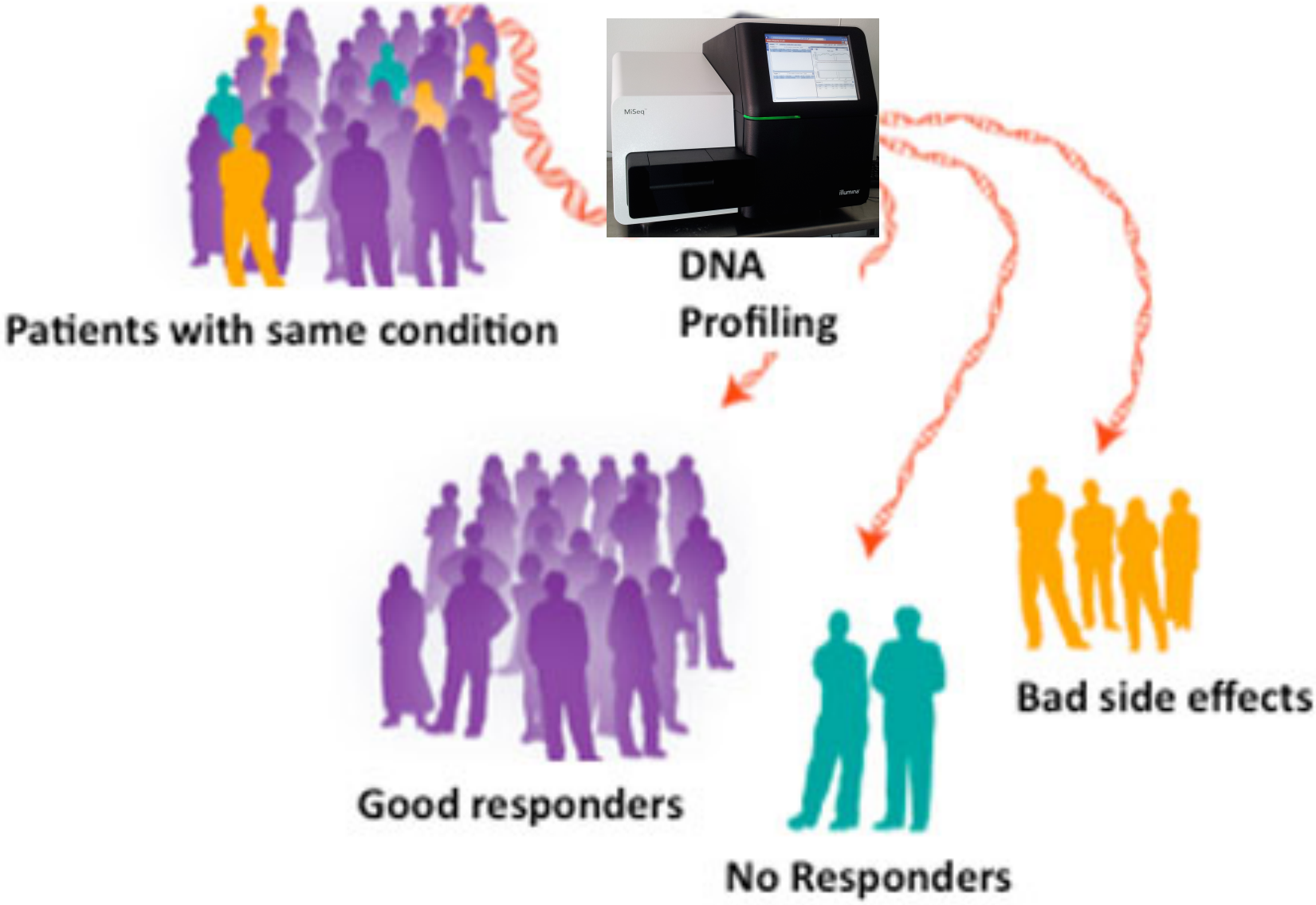
Opportunities





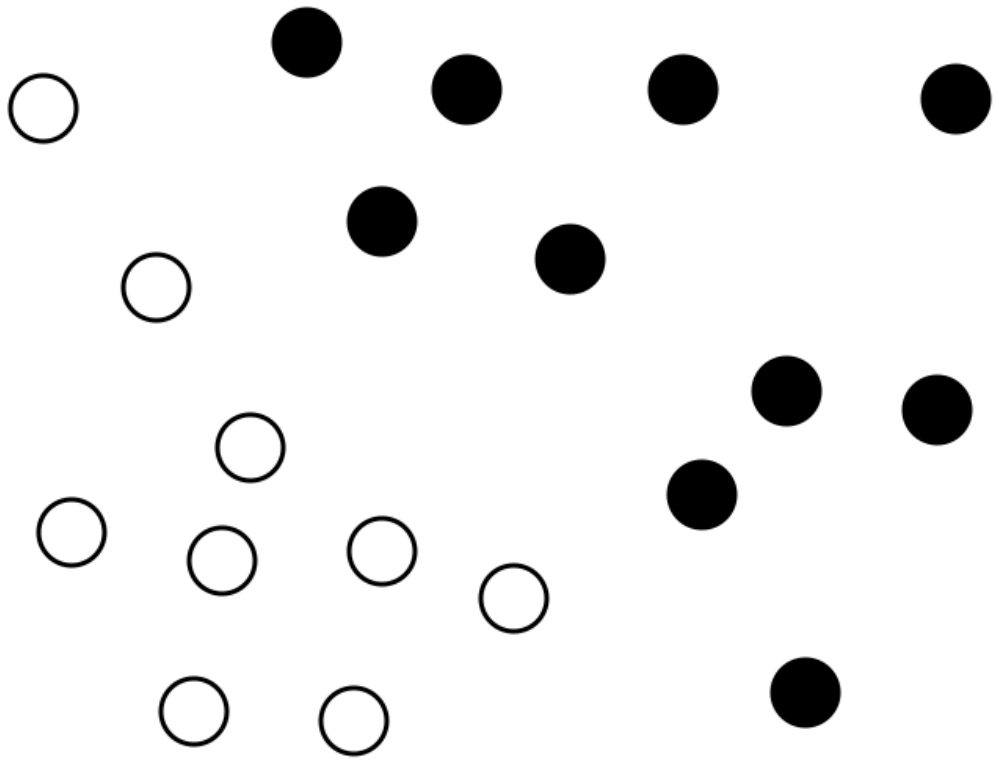
- What is your risk of developing a cancer? (*prevention*)
- Once detected, what precisely is your cancer (*diagnosis*)
- After treatment, what is your risk of relapse? (*prognosis*)
- What is the best therapy for your cancer? (*precision medicine*)

Example: precision medicine



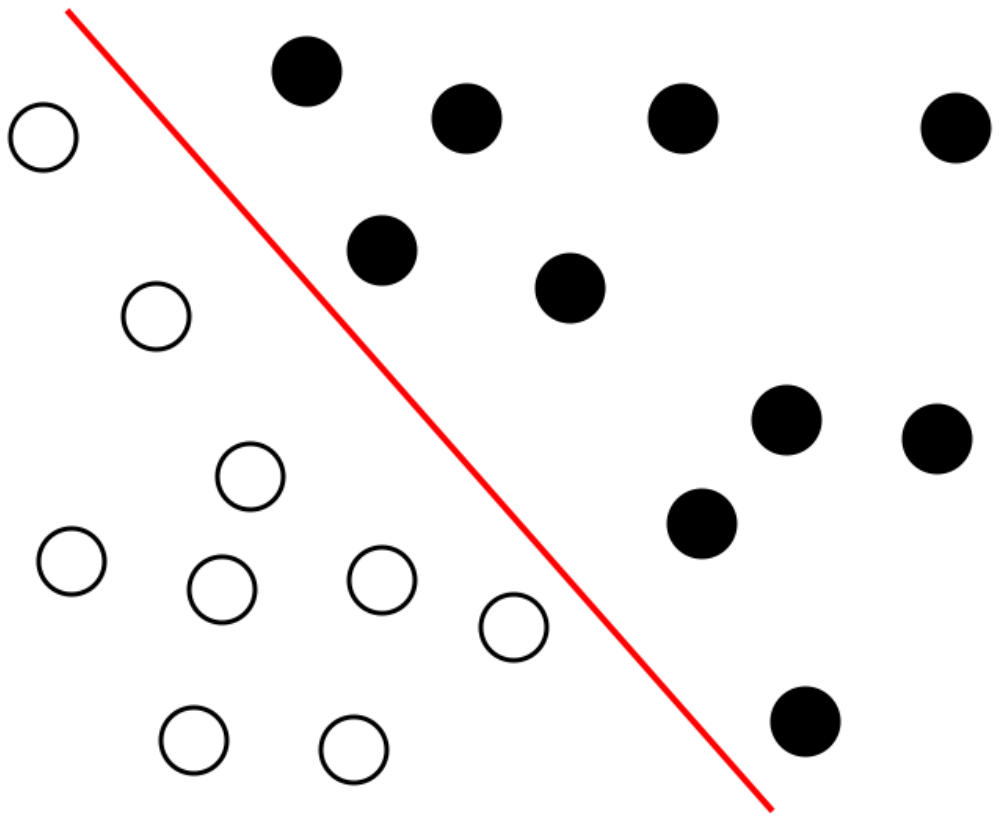
Learning from data (EASY case)

$n(= 19)$ patients $>>$ $p(= 2)$ genes



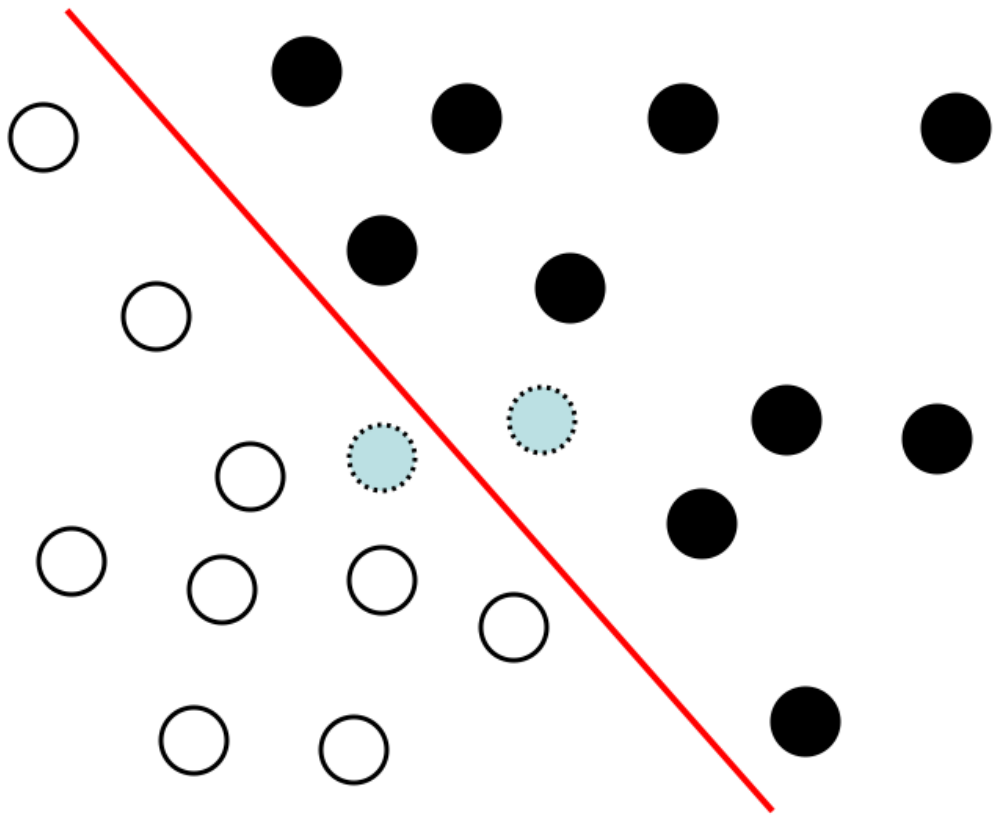
Learning from data (EASY case)

$n(= 19)$ patients $>>$ $p(= 2)$ genes



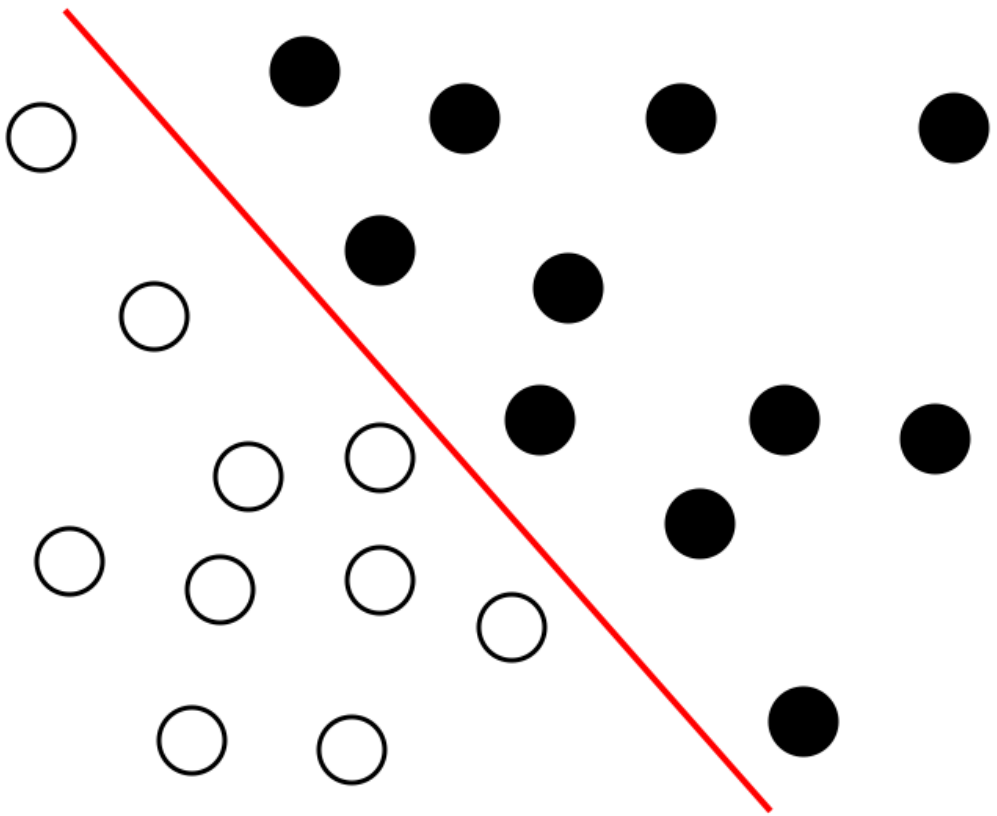
Learning from data (EASY case)

$n(= 19)$ patients $>>$ $p(= 2)$ genes

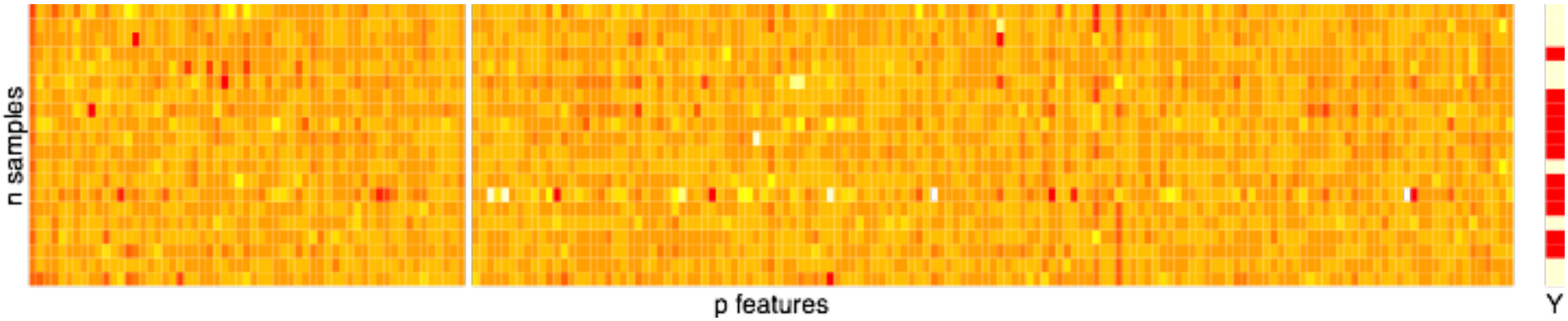


Learning from data (EASY case)

$n(= 19)$ patients $>>$ $p(= 2)$ genes



*-omics challenge: $n << p$

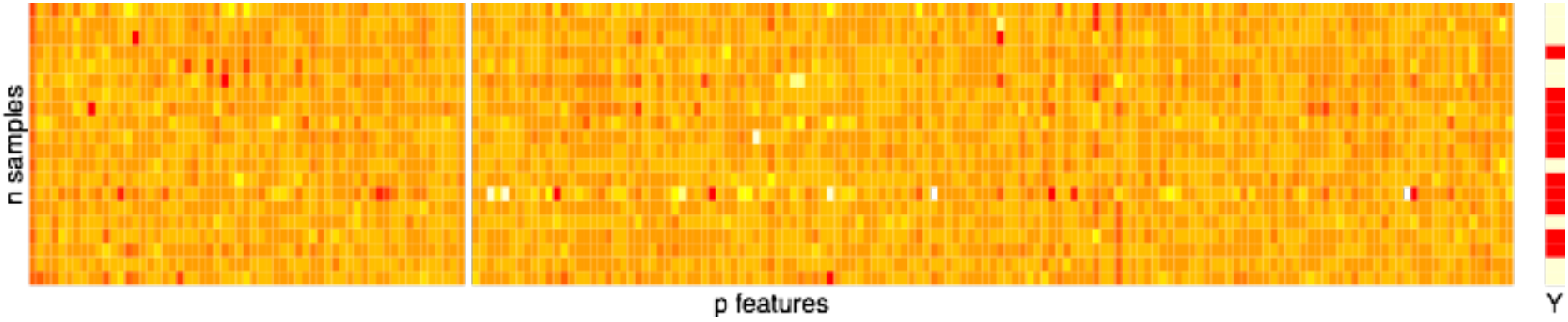


- $n = 10^2 \sim 10^4$ (patients)
- $p = 10^4 \sim 10^7$ (genes, mutations, copy number, ...)
- Data of various nature (continuous, discrete, structured, ...)
- Data of variable quality (technical/batch variations, noise, ...)

Consequences:

- Accuracy drops
- Biomarker selection unstable
- Speed and scalability can become an issue

Some general ideas



- How to represent the data?
- How adapt ML algorithms to specific problems, e.g., by including prior knowledge?
- How scale algorithms by, e.g., reformulations, relaxations or tricks?

Outline

1. Learning with regularization and prior knowledge
2. Cancer patient stratification from somatic mutations
3. Learning from rankings through pairwise comparisons
4. FlipFlop: fast isoform prediction from RNA-seq data
5. Conclusion

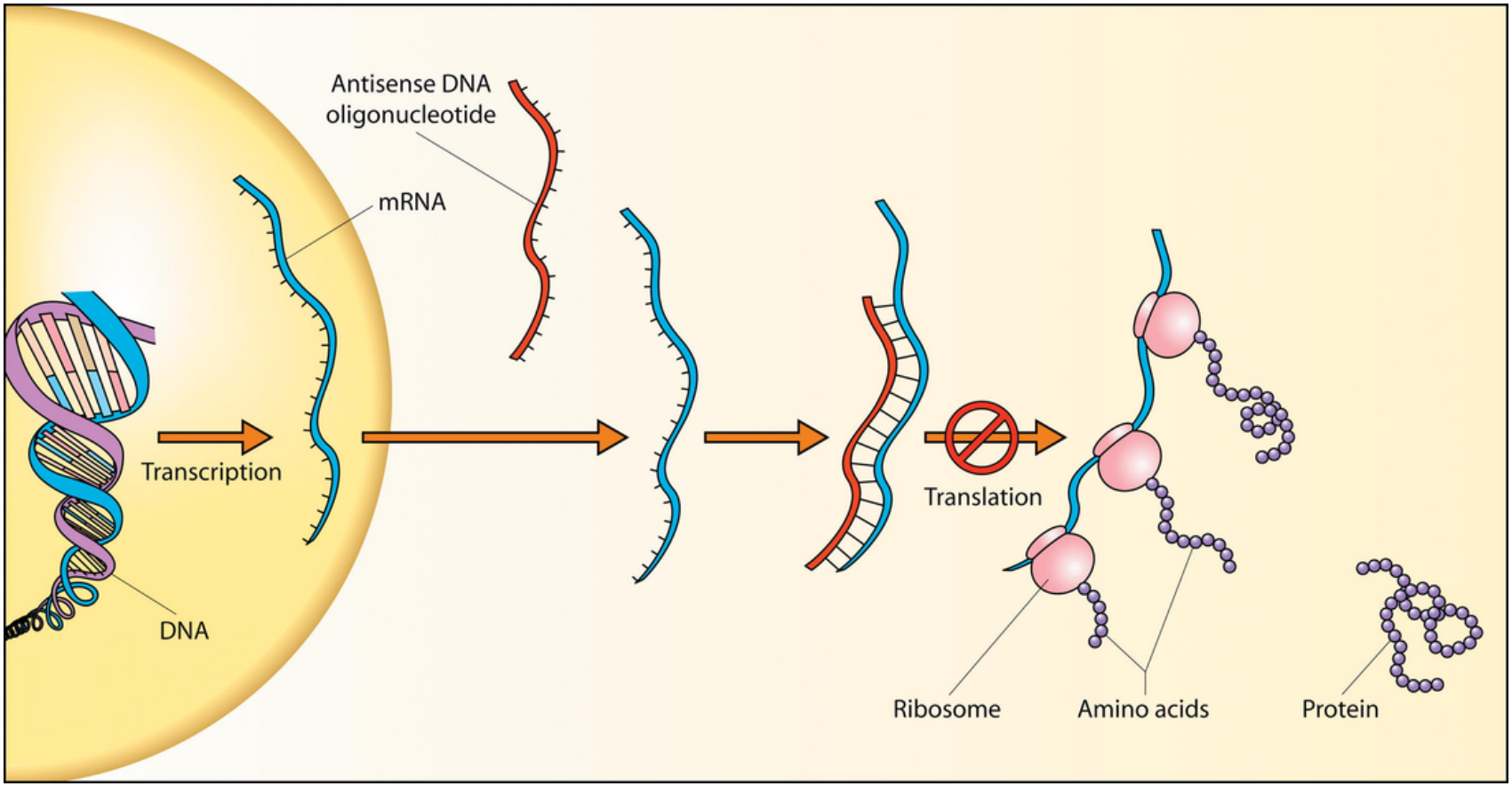
Outline

1. Learning with regularization and prior knowledge
2. Cancer patient stratification from somatic mutations
3. Learning from rankings through pairwise comparisons
4. FlipFlop: fast isoform prediction from RNA-seq data
5. Conclusion

Joint work with...

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

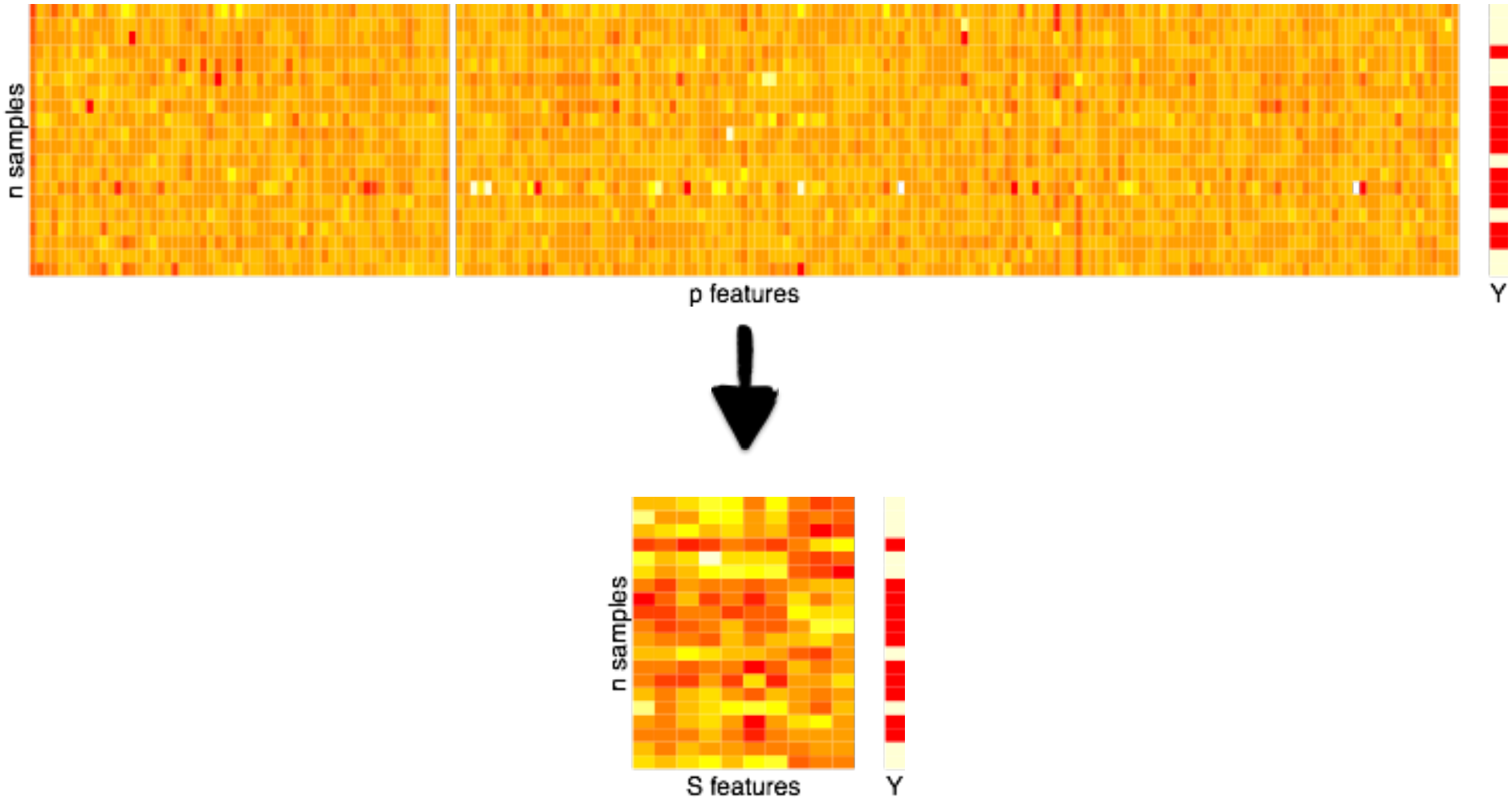
Gene expression



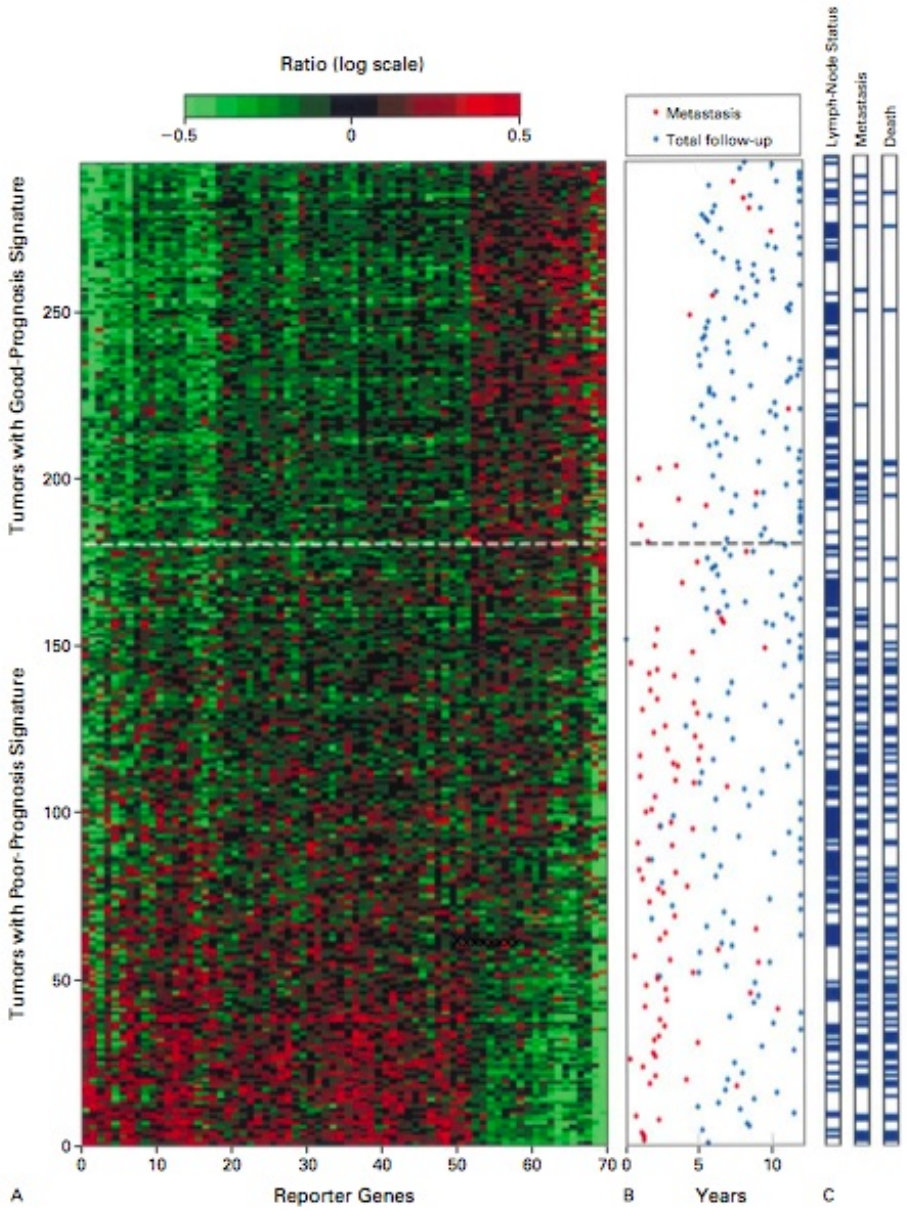
http://mrsbabbkv.weebly.com/rna--protein.html

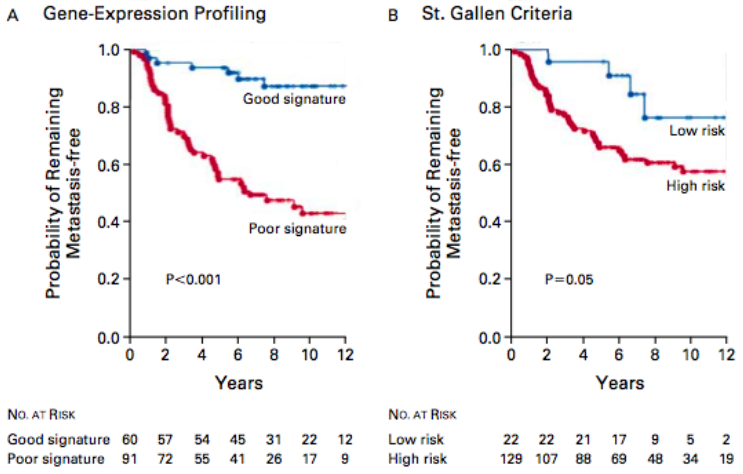
- About 22,000 genes encoded in DNA (same for all cells)
- Expression of each gene (= RNA synthesis) varies between cells
- Can be measured for all genes simultaneously with sequencing

Feature selection (a.k.a. *molecular signature*)



Example: 70-gene breast cancer prognostic signature





van ’t Veer et al. (2002);
van de Vijver et al. (2002)

But...

Gene expression profiling predicts clinical outcome of breast cancer

Laura J. van 't Veer*†, Hongyue Dai†‡, Marc J. van de Vijver*†, Yudong D. He‡, Augustinus A. M. Hart*, Mao Mao‡, Hans L. Peterse*, Karin van der Kooy*, Matthew J. Marton‡, Anke T. Witteveen*, George J. Schreiber‡, Ron M. Kerkhoven*, Chris Roberts‡, Peter S. Linsley‡, René Bernards* & Stephen H. Friend‡

** Divisions of Diagnostic Oncology, Radiotherapy and Molecular Carcinogenesis and Center for Biomedical Genetics, The Netherlands Cancer Institute, 121 Plesmanlaan, 1066 CX Amsterdam, The Netherlands*
‡ Rosetta Inpharmatics, 12040 115th Avenue NE, Kirkland, Washington 98034.

70 genes (Nature, 2002)

Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer

Yixin Wang, Jan G M Klijn, Yi Zhang, Anieta M Sieuwerts, Maxime P Look, Fei Yang, Dmitri Talantov, Mieke Timmermans, Marion E Meijer-van Gelder, Jack Yu, Tim Jatkoe, Els M J J Berns, David Atkins, John A Foekens

76 genes (Lancet, 2005)

3 genes in common

van 't Veer et al. (2002); Wang et al. (2005)

3 genes is the best you can expect given *n* and *p*

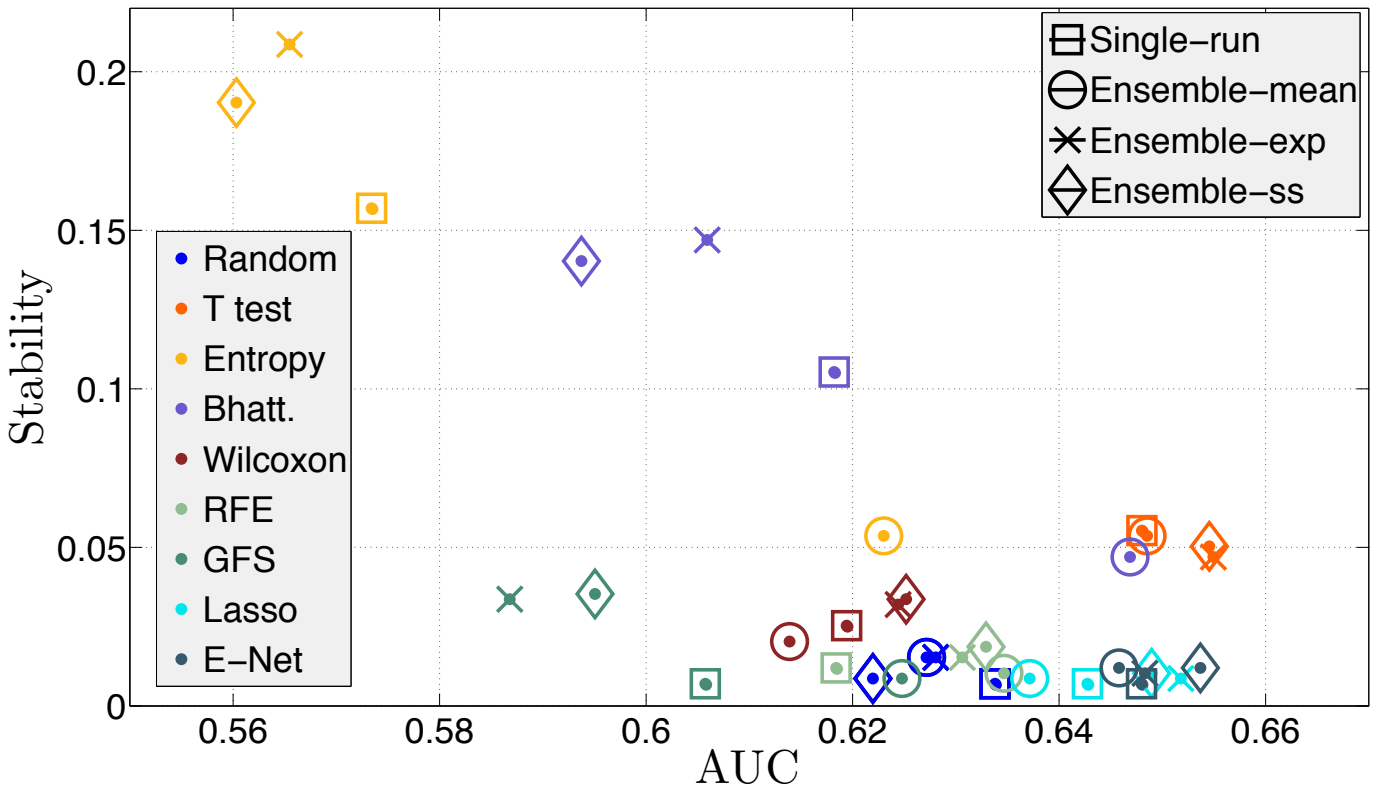
OPEN ACCESS Freely available online

PLoS one

The Influence of Feature Selection Methods on Accuracy, Stability and Interpretability of Molecular Signatures

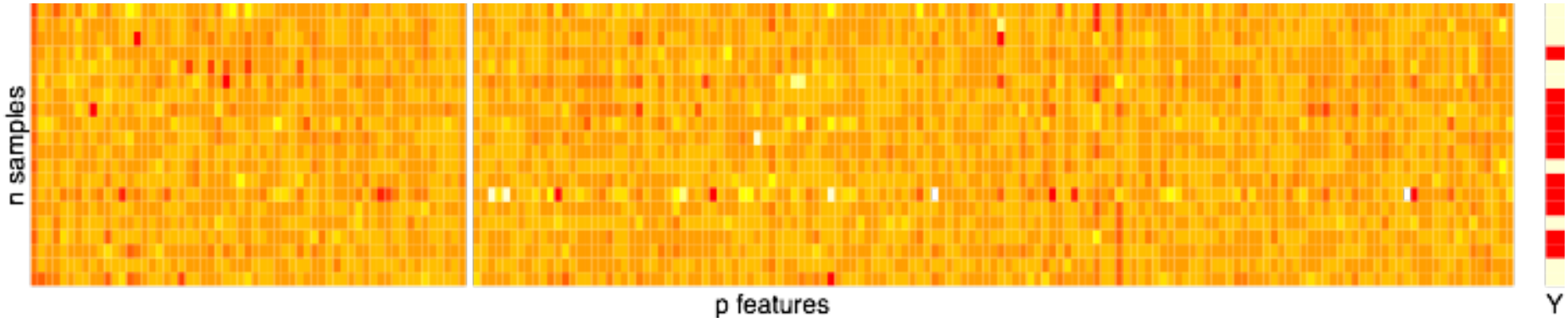
Anne-Claire Haury[1,2,3]*, Pierre Gestraud[1,2,3], Jean-Philippe Vert[1,2,3]

1 Mines ParisTech, Centre for Computational Biology, Fontainebleau, France, **2** Institut Curie, Paris, France, **3** Institut National de la Santé et de la Recherche Médicale, Paris, France



Haury et al. (2011)

Ideas

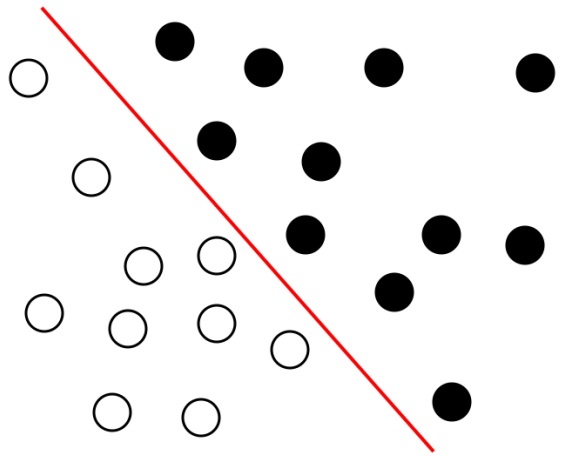


Can we improve the $p << n$ situation,

- either explicitly (reduce p)
- or implicitly (change the metric / the learning algorithm)

using prior knowledge we may have about the genes?

Learning with regularization



For a sample $x \in \mathbb{R}^p$, learn a linear decision function:

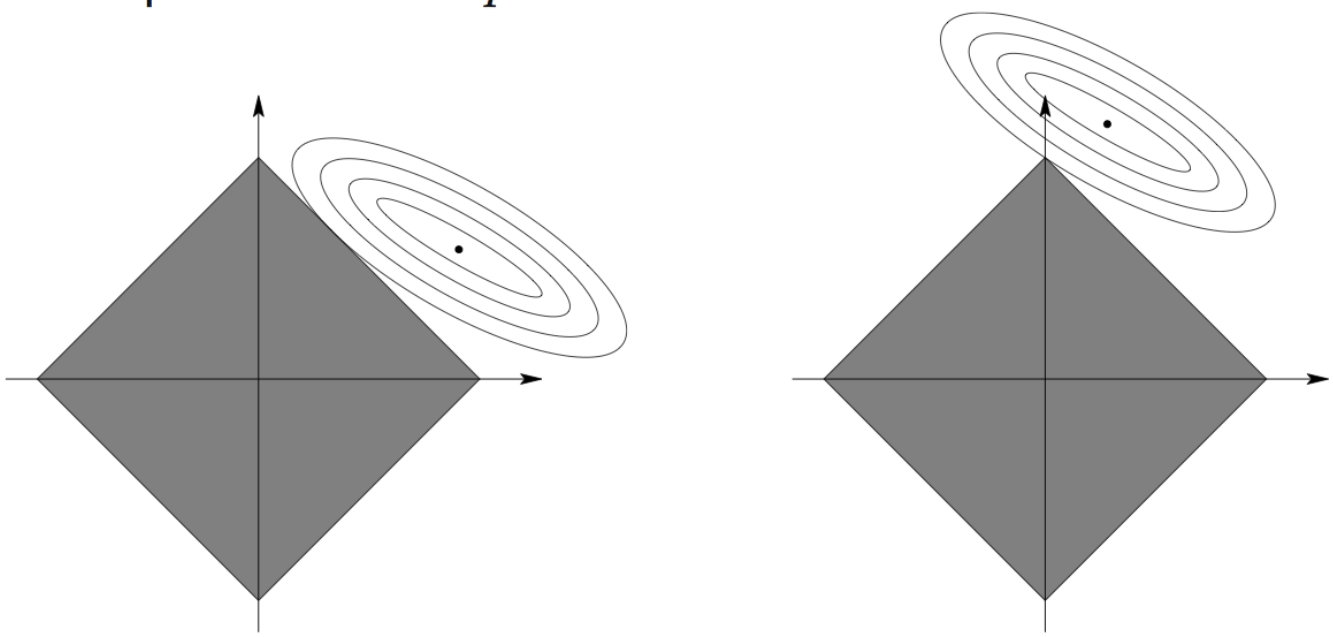
$$f_\beta(x) = \beta^\top x \qquad \min_{\beta \in \mathbb{R}^p} R(f_\beta) + \lambda \Omega(\beta)$$

- $R(f_\beta)$ empirical risk, e.g., $R(f_\beta) = \frac{1}{n} \sum_{i=1}^{n} (f_\beta(x_i) - y_i)^2$
- $\Omega(\beta)$ penalty, to control overfitting in high dimension, e.g.:
  - $\Omega(\beta) = \sum_{i=1}^{p} \beta_i^2$ (ridge regression, SVM,...)
  - $\Omega(\beta) = \sum_{i=1}^{p} |\beta_i|$ (lasso, boosting,...)

Example: ℓ_1 regularization

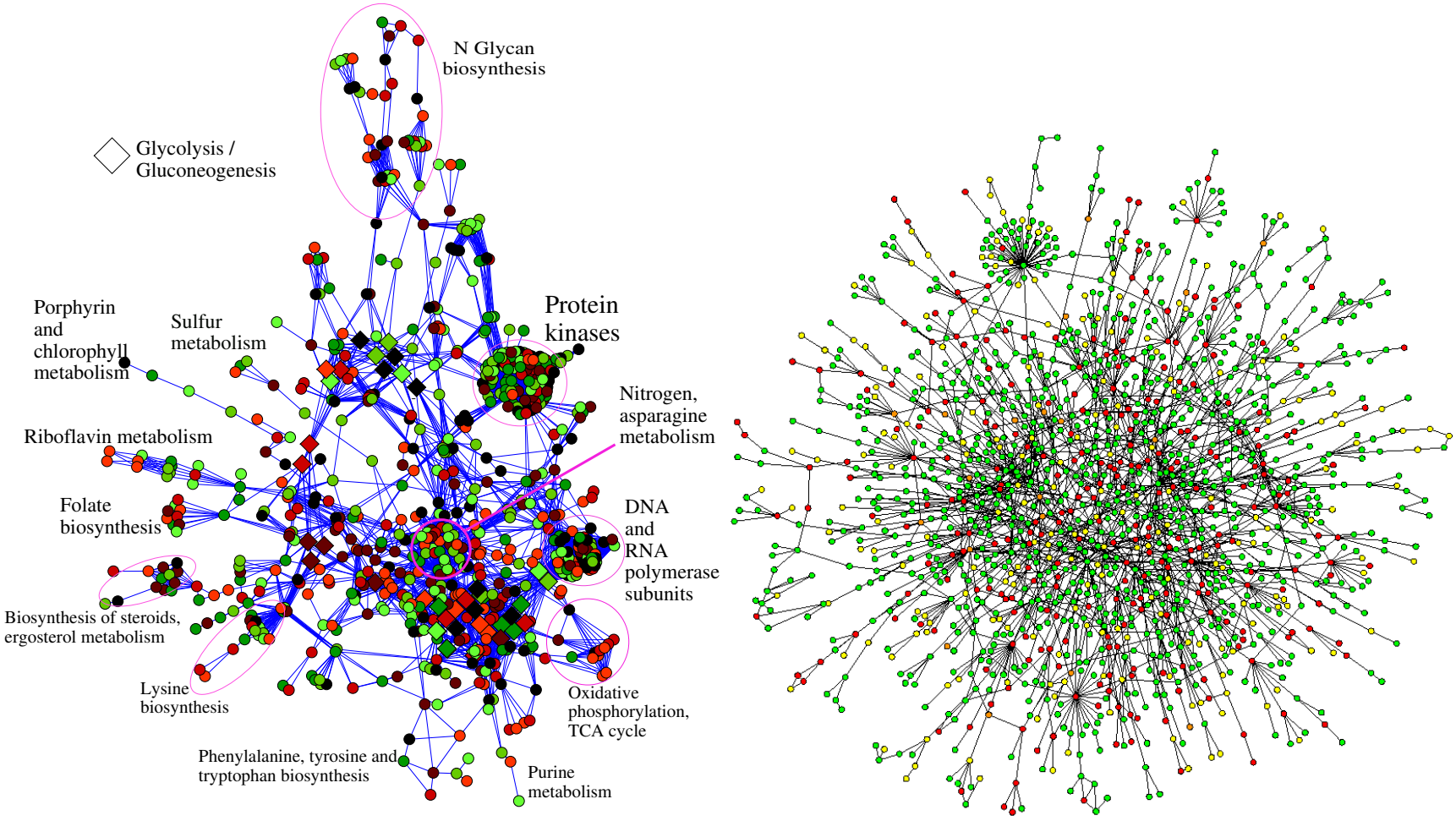
$$\min_{\beta} R(f_\beta) + \lambda \sum_{i=1}^{p} |\beta_i| \quad \Leftrightarrow \quad \min_{\beta} R(f_\beta) \text{ such that } \sum_{i=1}^{p} |\beta_i| \leq C$$

Geometric interpretation with $p = 2$



Leads to sparse models (feature selection)

Gene networks as prior knowledge



Let’s force the signatures to be "coherent" with a known gene network?

Graph based penalty

$$f_\beta(x) = \beta^\top x \qquad \min_\beta R(f_\beta) + \lambda\Omega(\beta)$$

Prior hypothesis

Genes near each other on the graph should have similar weigths.

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

$$f_\beta(x) = \beta^\top x \qquad \min_\beta R(f_\beta) + \lambda\Omega(\beta)$$

Prior hypothesis

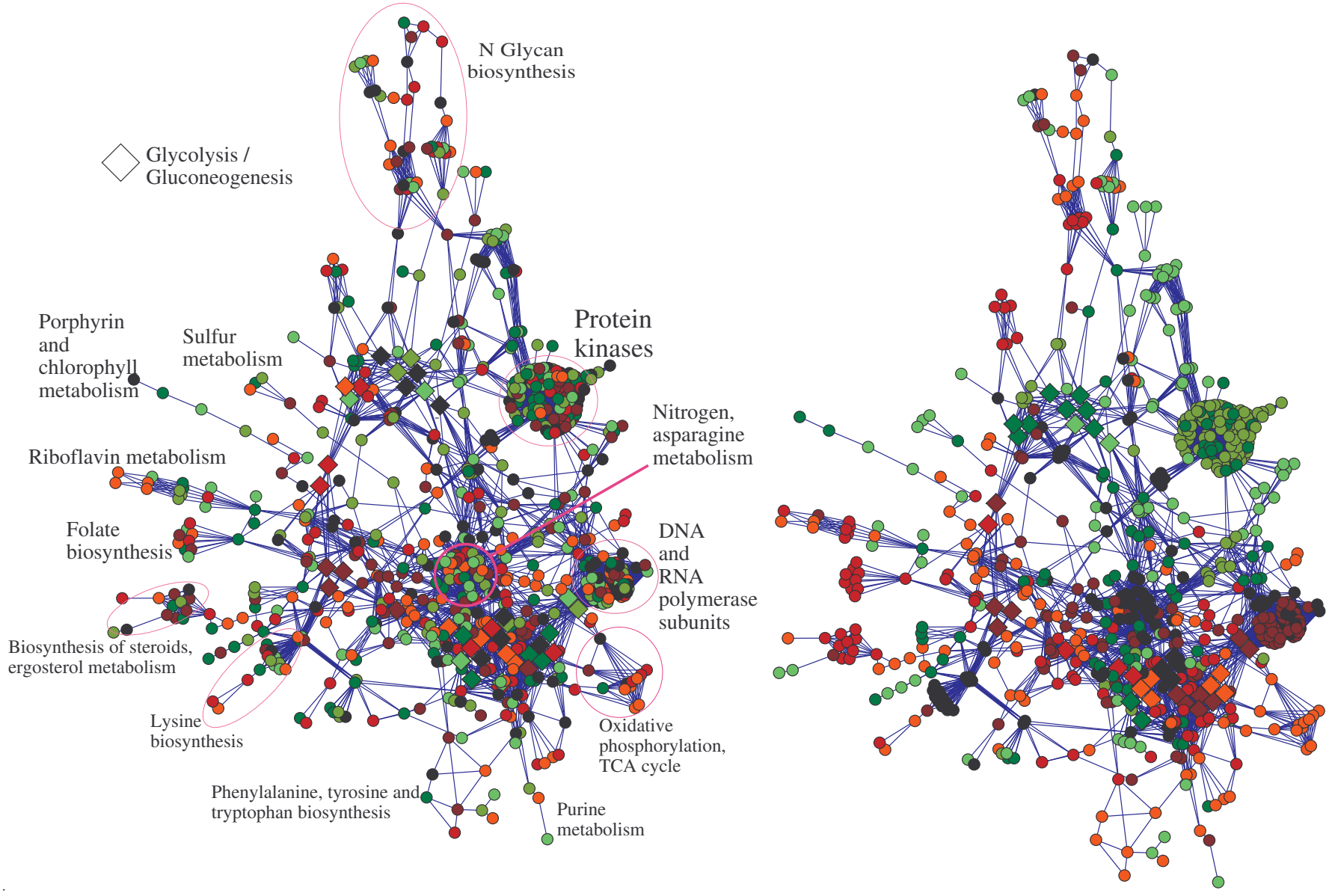
Genes near each other on the graph should have similar weigths.

An idea (Rapaport et al., 2007)

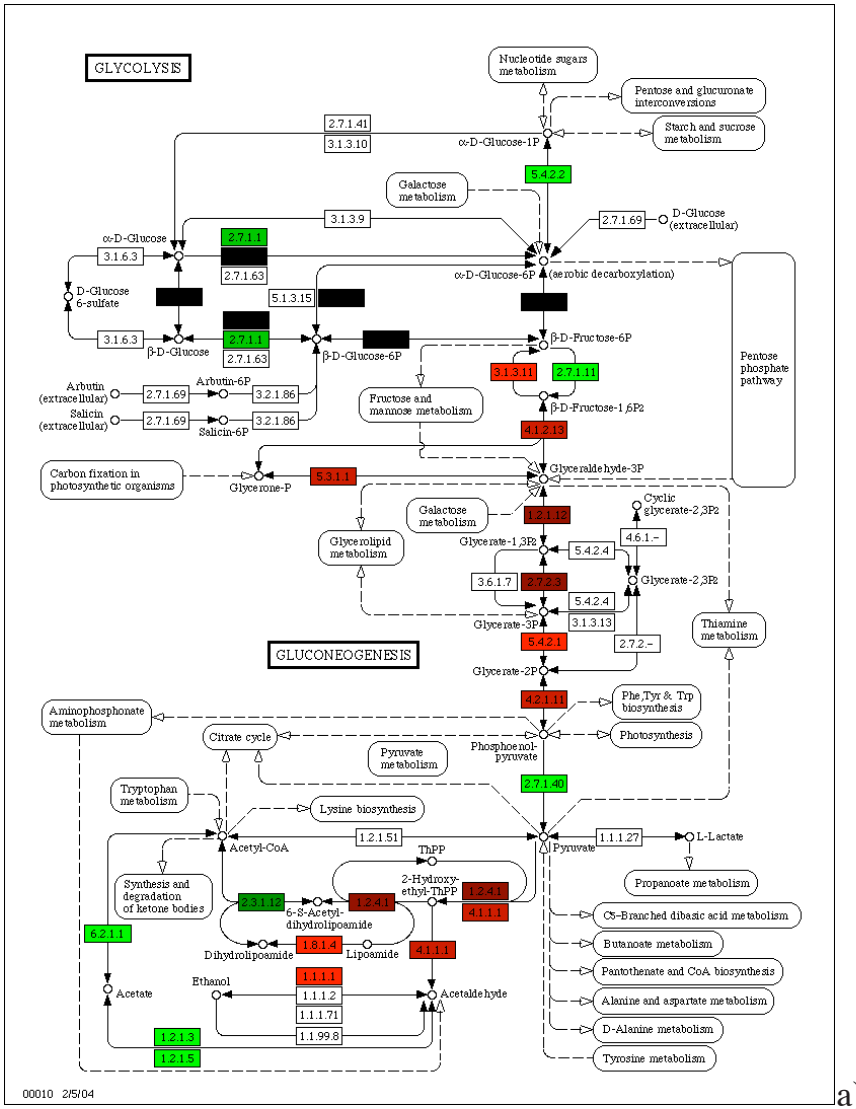
$$\Omega(\beta) = \sum_{i\sim j}(\beta_i - \beta_j)^2 \,,$$

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

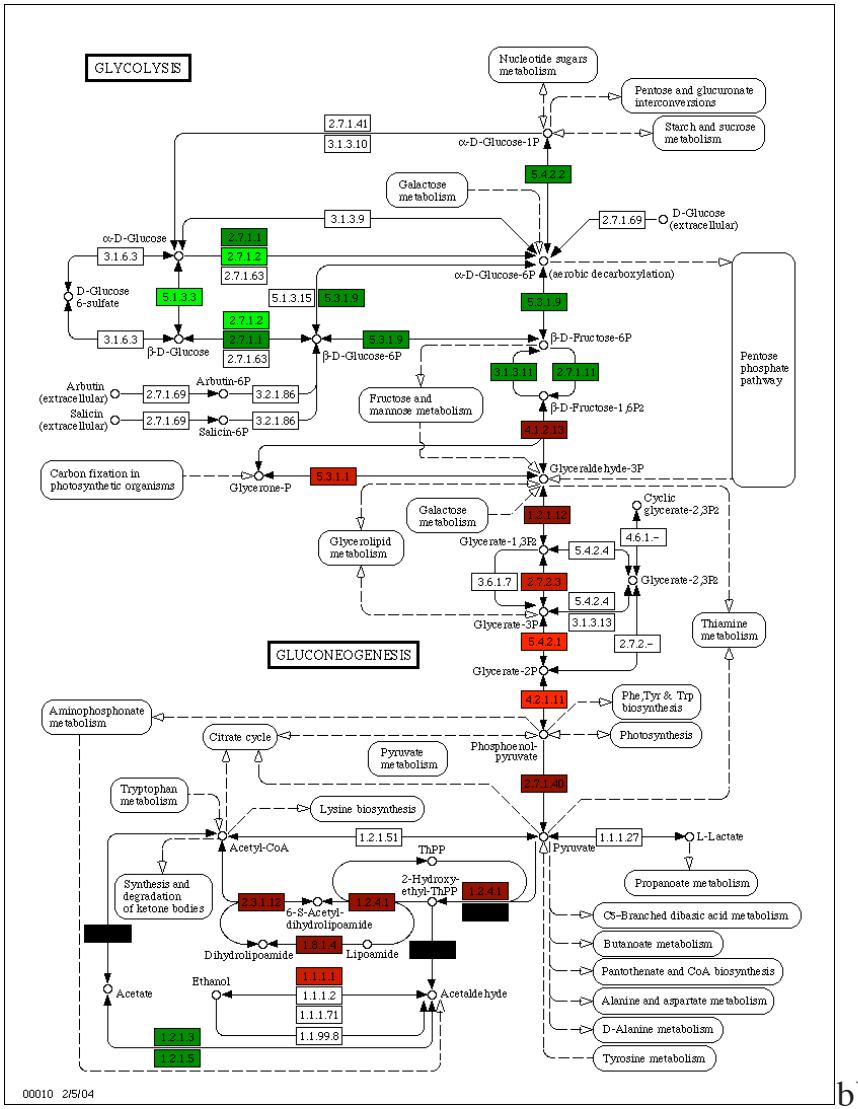
Classifiers



Classifier



a)



b)

Graph-based penalty as change of representation

Theorem

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

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

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

$$\min_{\gamma \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \ell\left(\gamma^\top \Phi(x_i), y_i\right) + \lambda \sum_{j=1}^{p} \gamma_j^2 \,,$$

and where

$$\Phi(x) = L^{-1/2} x$$

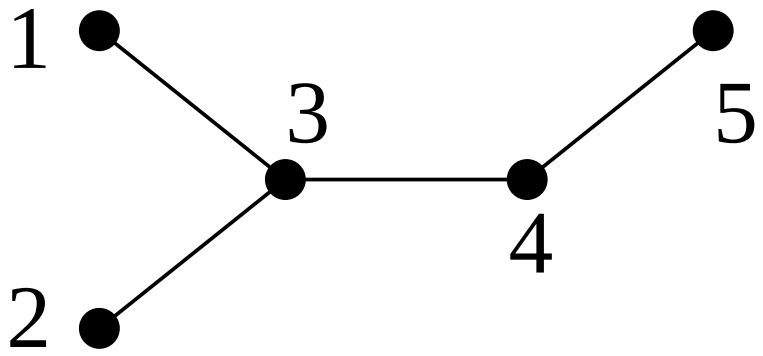
with L the graph Laplacian.

$L^{-1/2}$ is the square root of the pseudo-inverse of L.
Assuming each sample is centered on each connected component of the graph.

Graph Laplacian

Definition

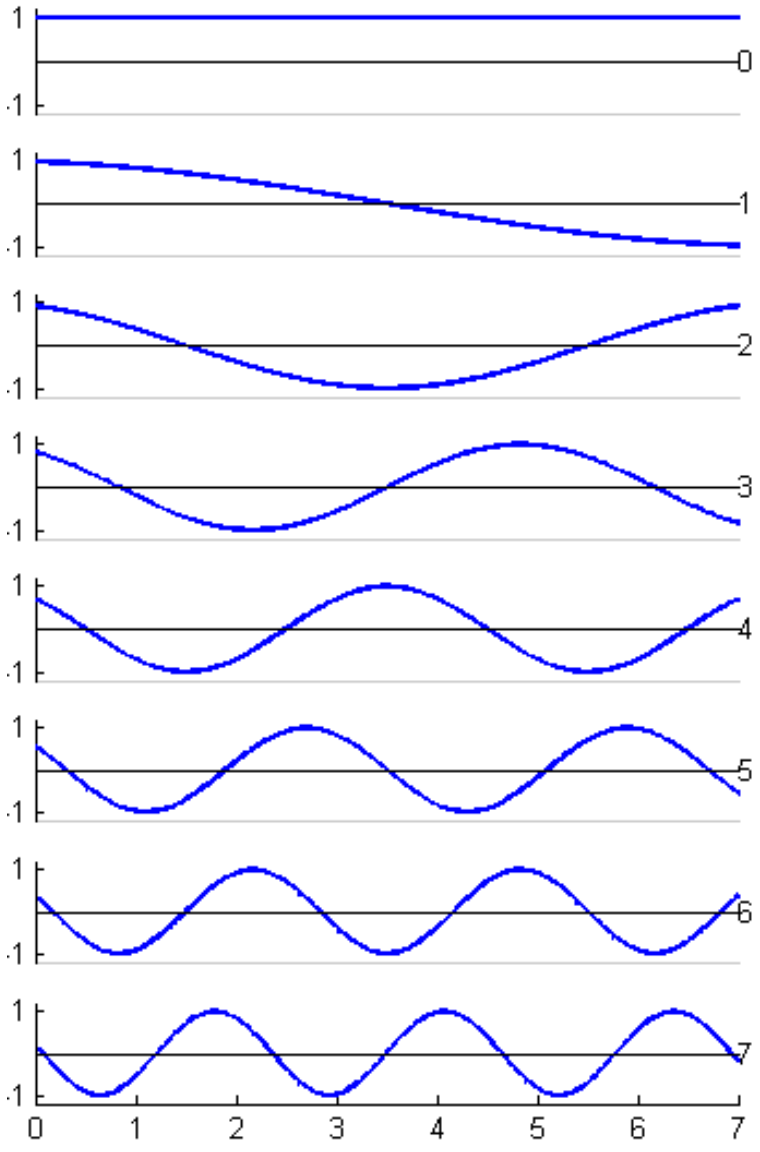
The Laplacian of the graph is the matrix $L = D - A$.



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

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

Fourier analysis on graphs



- Eigenvectors of $(e_i)_{i=1,\dots,p}$ of L form the Fourier basis on the graph
- Eigenvalue $(\lambda_i)_{i=1,\dots,p}$ the "frequencies"
- $\Phi(x) = L^{-1/2}x$ smoothes x:

$$\Phi(x) = \sum_{i:\lambda_i>0} \frac{1}{\sqrt{\lambda_i}} (x^\top e_i) e_i$$

while

$$x = \sum_{i:\lambda_i>0} (x^\top e_i) e_i$$

Other penalties with kernels

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

with:

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

$$\Omega(\beta) = c \sum_{i=1}^{p} \beta_i^2 + \sum_{i \sim j} (\beta_i - \beta_j)^2 \,, \quad \Phi(x) = \sum_i \frac{1}{\sqrt{c + \lambda_i}} (x^\top e_i) e_i$$

- The diffusion kernel:

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

penalizes high frequencies of β in the Fourier domain:

$$\Phi(x) = \sum_i e^{-t\lambda_i} (x^\top e_i) e_i$$

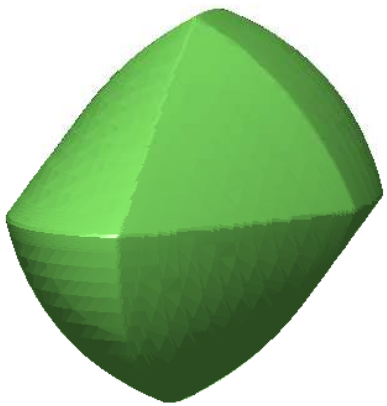
Fused lasso and generalized fused lasso

- Gene selection + Piecewise constant on the graph (fused lasso, Tibshirani et al., 2005).

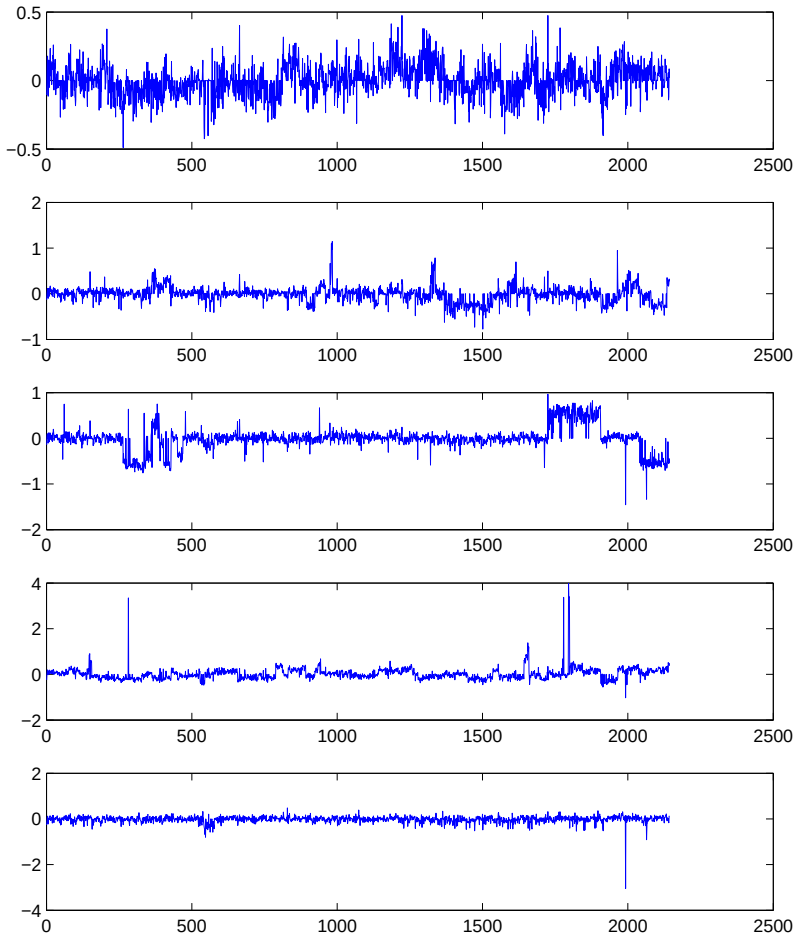
$$\Omega(\beta) = \sum_{i \sim j} |\beta_i - \beta_j| + \sum_{i=1}^{p} |\beta_i|$$

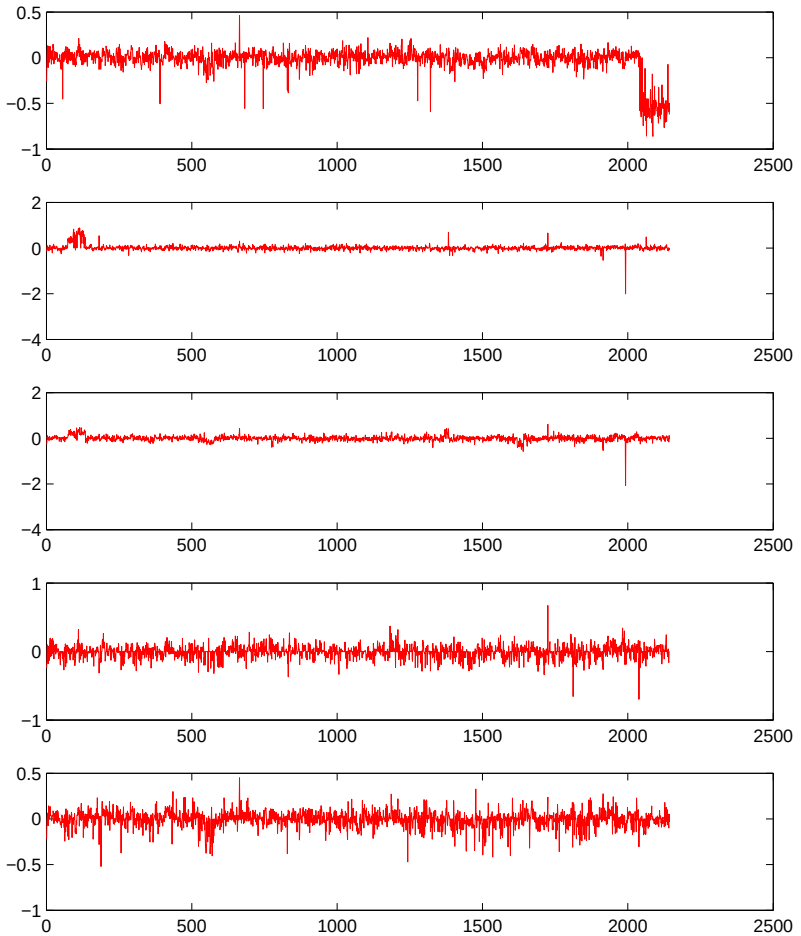
- Gene selection + smooth on the graph

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



Example: classification of DNA copy number profiles

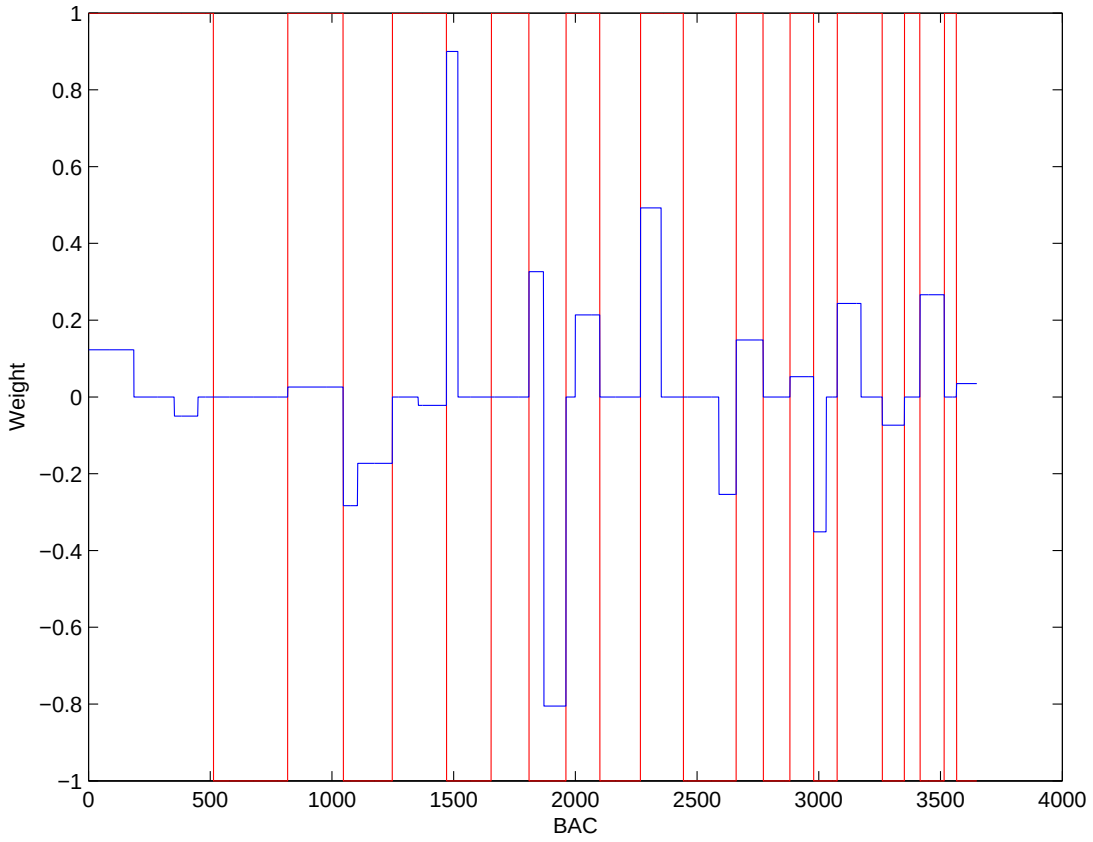




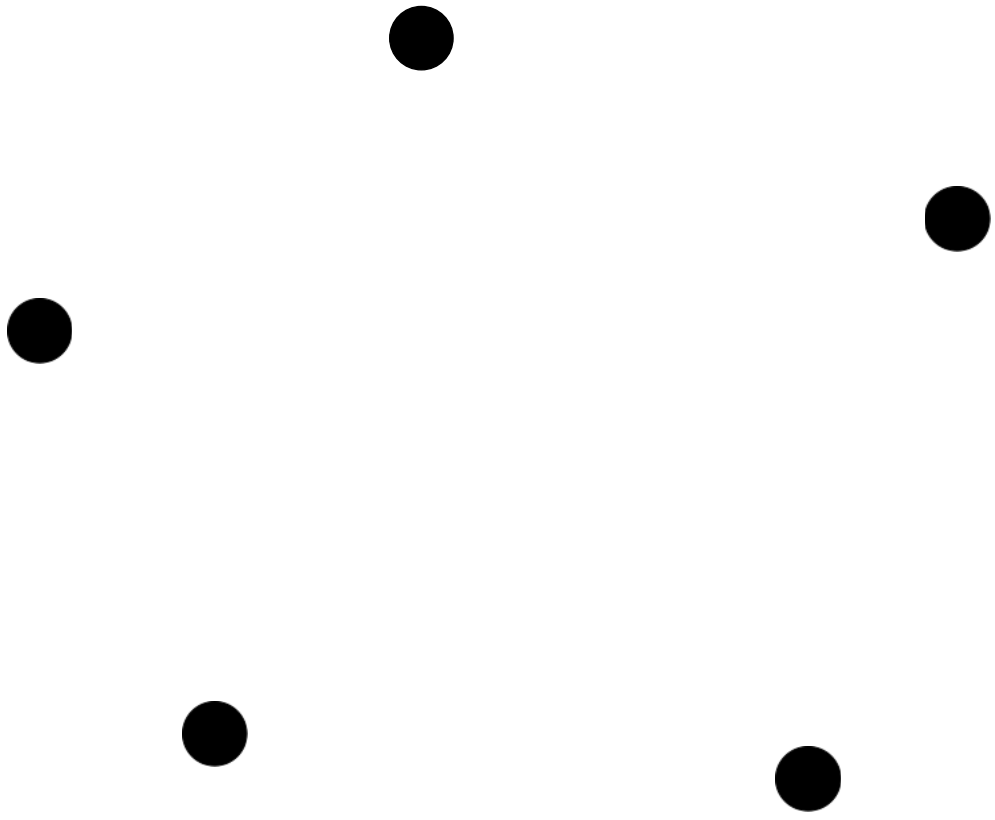
Aggressive (left) vs non-aggressive (right) melanoma

Fused lasso solution (Rapaport et al., 2008)

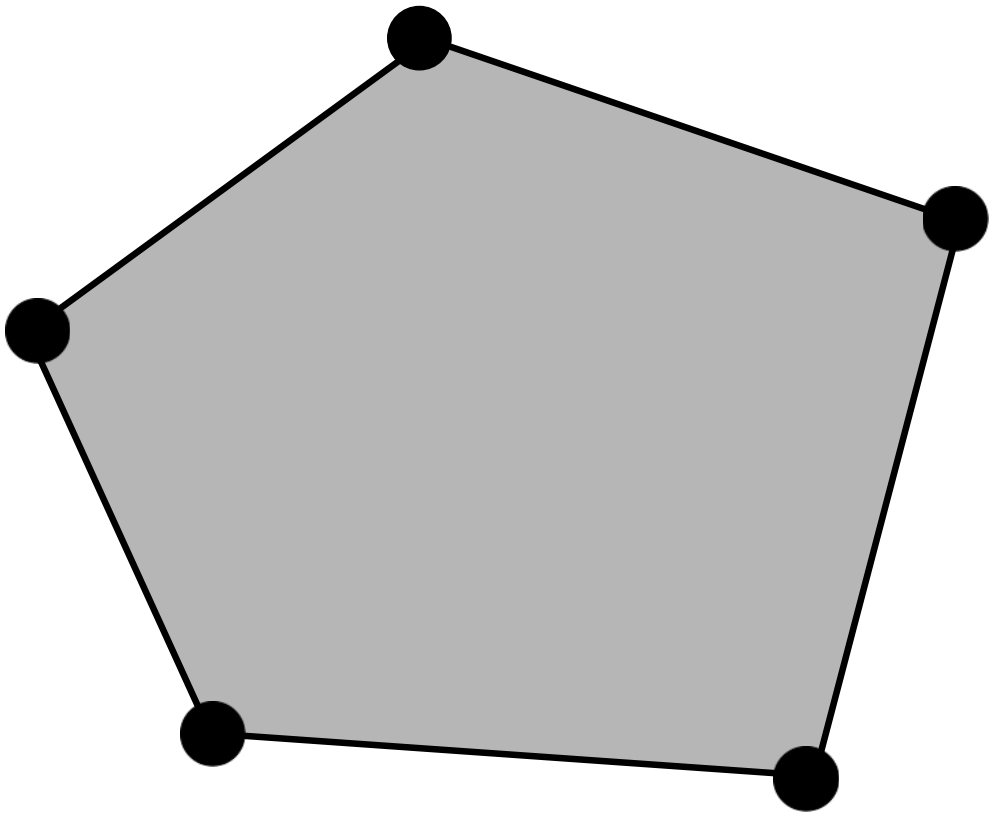
$$\min_{\beta} \left\{ R(f_\beta) + \lambda_1 \sum_{i \sim j} |\beta_i - \beta_j| + \lambda_2 \sum_{i=1}^{p} |\beta_i| \right\}$$



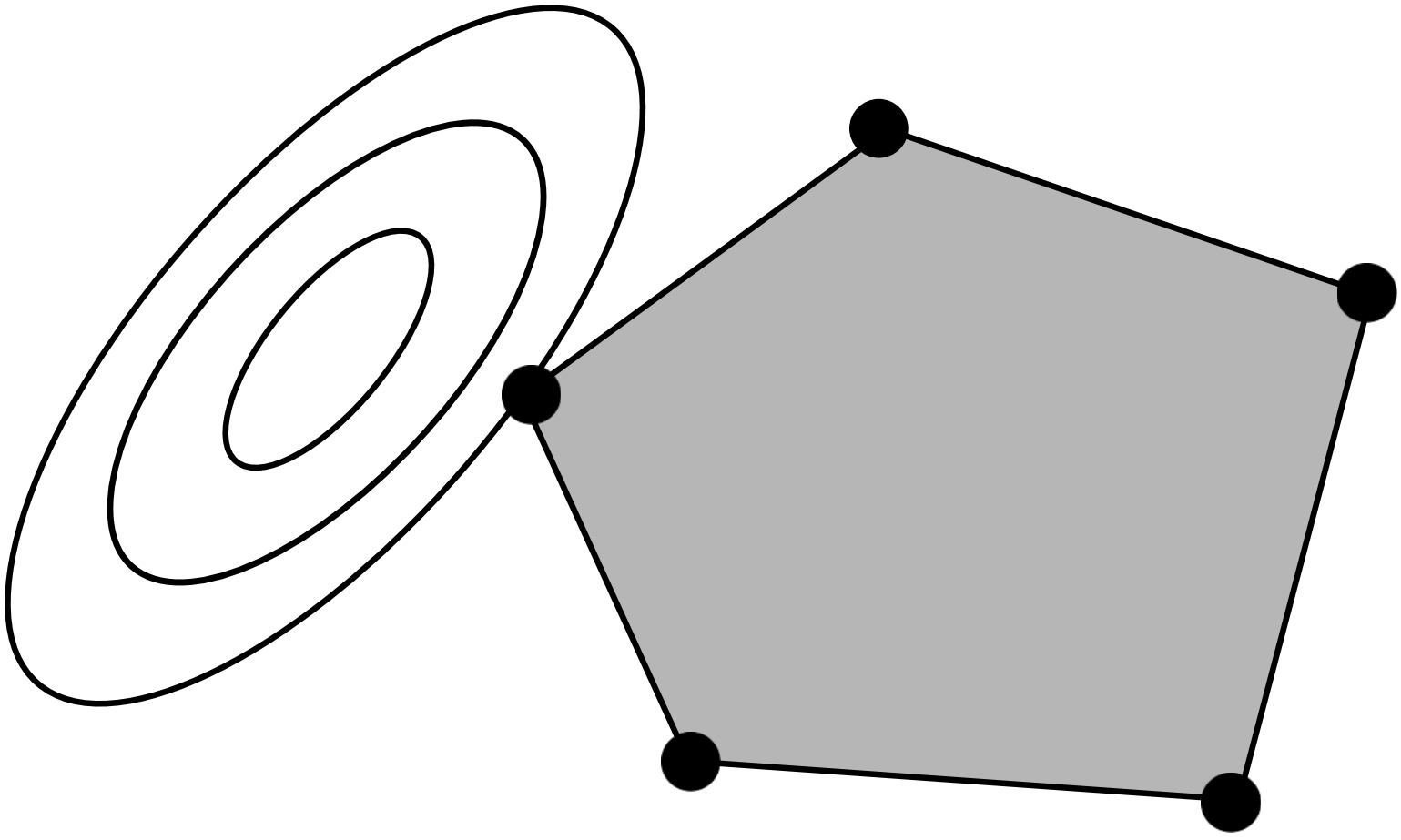
Generalization: atomic norms



Generalization: atomic norms



Generalization: atomic norms



Atomic Norm (Chandrasekaran et al., 2012)

Definition

Given a set of atoms $\mathcal{A}$, the associated atomic norm is

$$\|x\|_{\mathcal{A}} = \inf\{t > 0 \mid x \in t\,\mathrm{conv}(\mathcal{A})\}.$$

NB: This is really a norm if $\mathcal{A}$ is centrally symmetric and spans $\mathbb{R}^p$

Primal and dual form of the norm

$$\|x\|_{\mathcal{A}} = \inf\left\{ \sum_{a\in\mathcal{A}} c_a \mid x = \sum_{a\in\mathcal{A}} c_a\, a, \quad c_a > 0,\ \forall a \in \mathcal{A} \right\}$$

$$\|x\|_{\mathcal{A}}^* = \sup_{a\in\mathcal{A}} \langle a, x\rangle$$

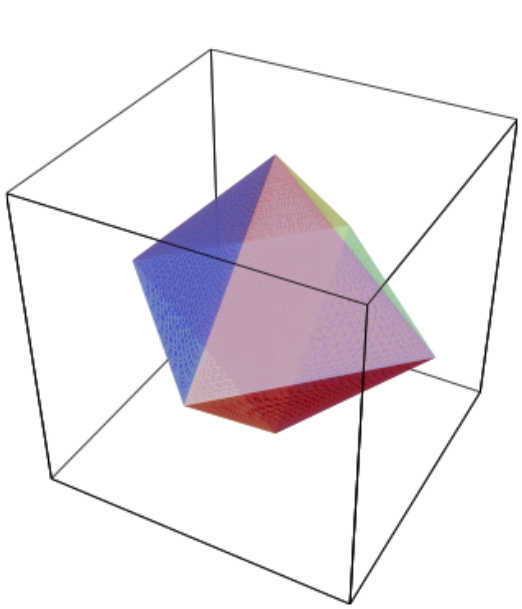
Examples

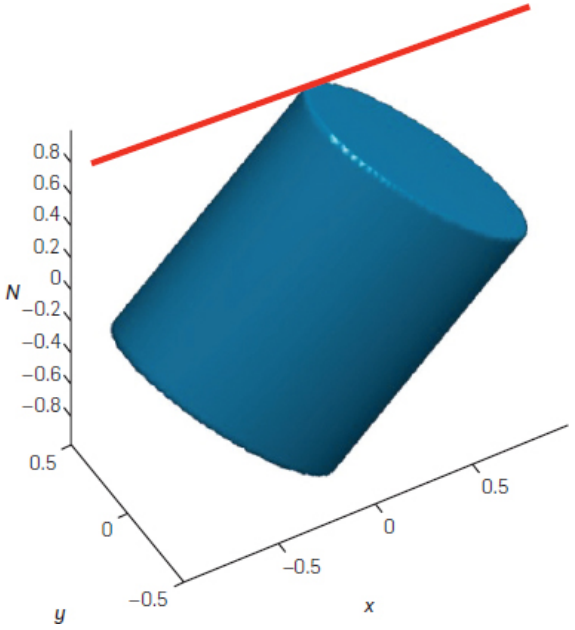
- Vector ℓ_1-norm: $x \in \mathbb{R}^p \mapsto \|x\|_1$

$$\mathcal{A} = \left\{ \pm e_k \;\middle|\; 1 \leq k \leq p \right\}$$

- Matrix trace norm: $Z \in \mathbb{R}^{m_1 \times m_2} \mapsto \|Z\|_*$ (sum of singular value)

$$\mathcal{A} = \left\{ ab^\top \;:\; a \in \mathbb{R}^{m_1}, b \in \mathbb{R}^{m_2}, \| a \|_2 = \| b \|_2 = 1 \right\}$$





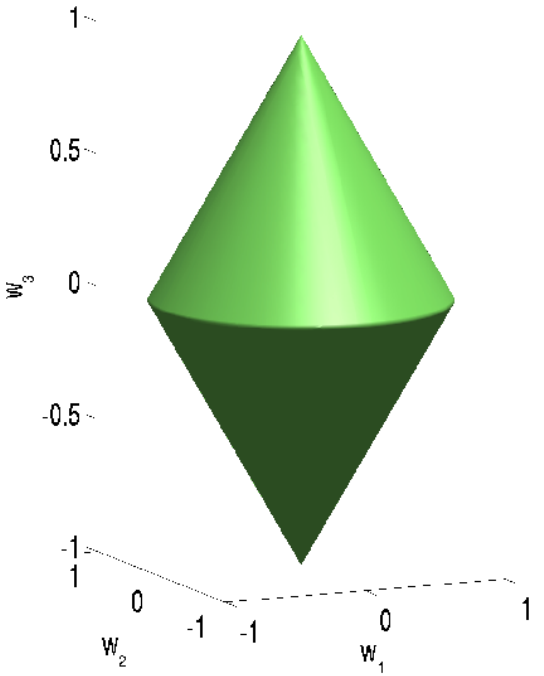
Group lasso (Yuan and Lin, 2006)

For $x \in \mathbb{R}^p$ and $\mathcal{G} = \{g_1, \ldots, g_G\}$ a partition of $[1, p]$:

$$\| x \|_{1,2} = \sum_{g \in \mathcal{G}} \| x_g \|_2$$

is the atomic norm associated to the set of atoms

$$\mathcal{A}_{\mathcal{G}} = \bigcup_{g \in \mathcal{G}} \{u \in \mathbb{R}^p \ : \ \text{supp}(u) = g, \| u \|_2 = 1\}$$



$$\mathcal{G} = \{\{1, 2\}, \{3\}\}$$

$$\begin{aligned} \| x \|_{1,2} &= \| (x_1, x_2)^\top \|_2 + \| x_3 \|_2 \\ &= \sqrt{x_1^2 + x_2^2} + \sqrt{x_3^2} \end{aligned}$$

Group lasso with overlaps

How to generalize the group lasso when the groups overlap?

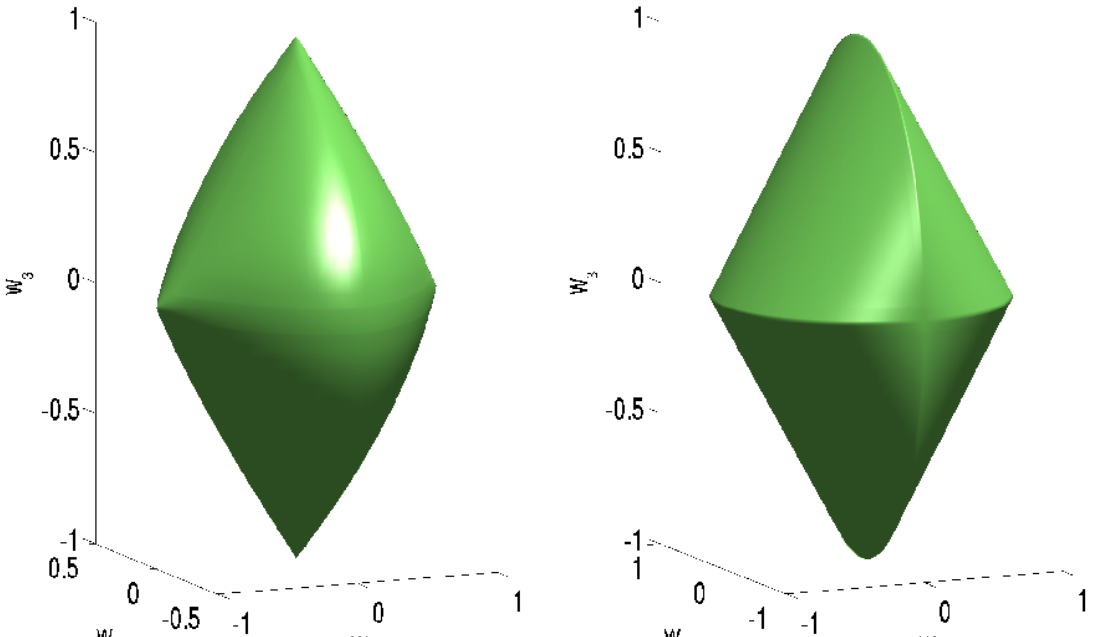
- Set features to zero by groups (Jenatton et al., 2011)

$$\| x \|_{1,2} = \sum_{g \in \mathcal{G}} \| x_g \|_2$$

- Select support as a union of groups (Jacob et al., 2009)

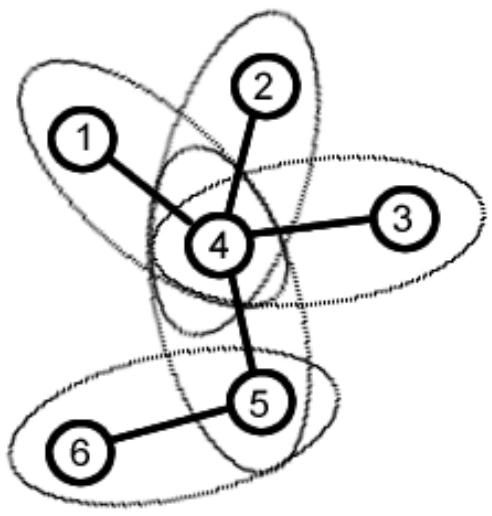
$$\| x \|_{\mathcal{A}_{\mathcal{G}}}$$

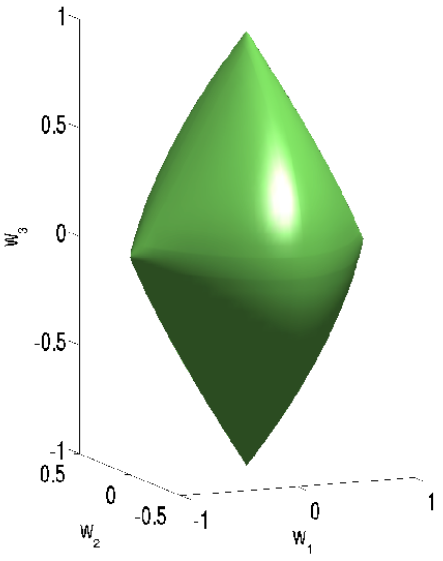
see also MKL (Bach et al., 2004)

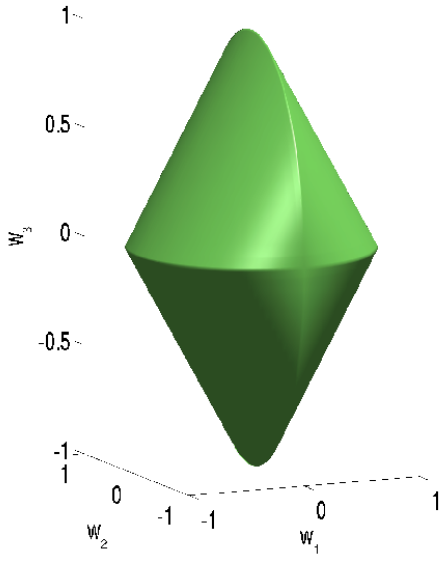


$$\mathcal{G} = \{\{1,2\},\{2,3\}\}$$

Graph-based structured feature selection





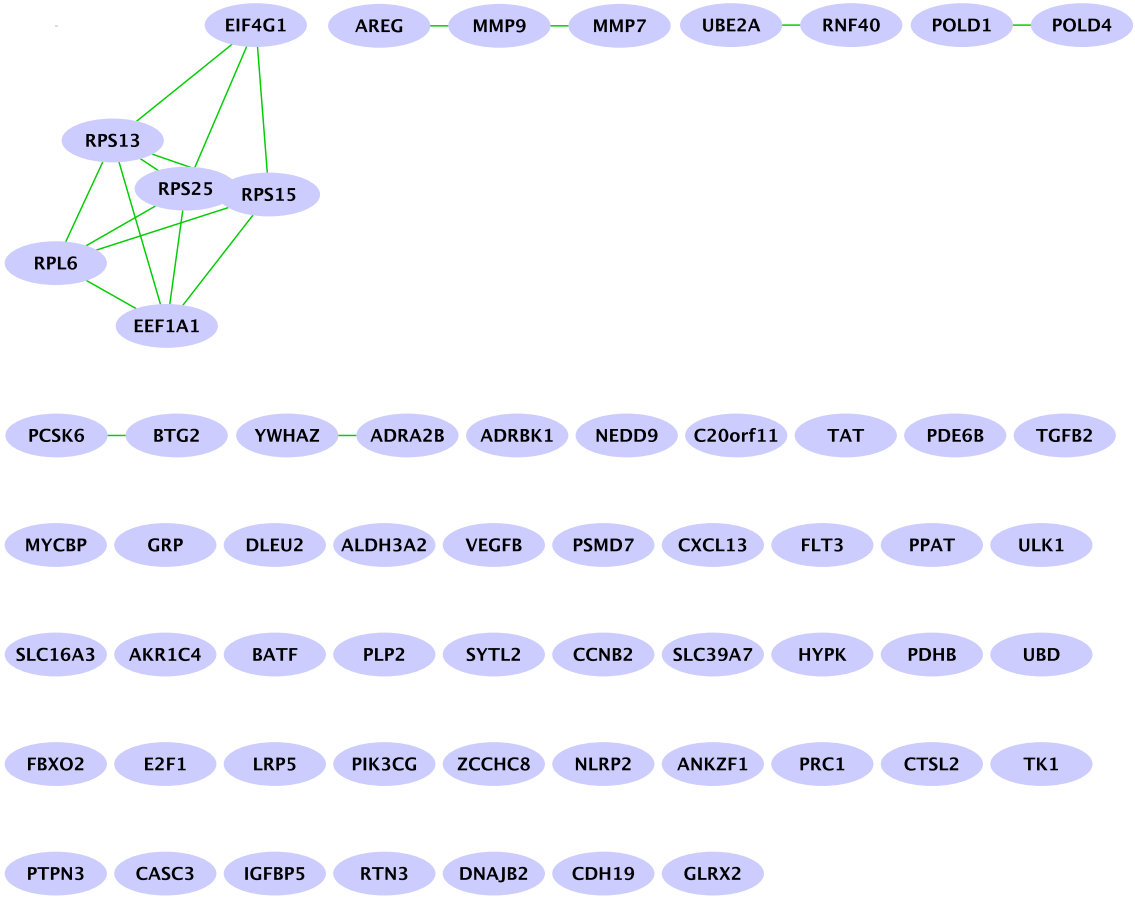


Graph lasso(s)

$$\Omega_1(\beta) = \sum_{i \sim j} \sqrt{\beta_i^2 + \beta_j^2} \qquad \text{(Jenatton et al., 2011)}$$

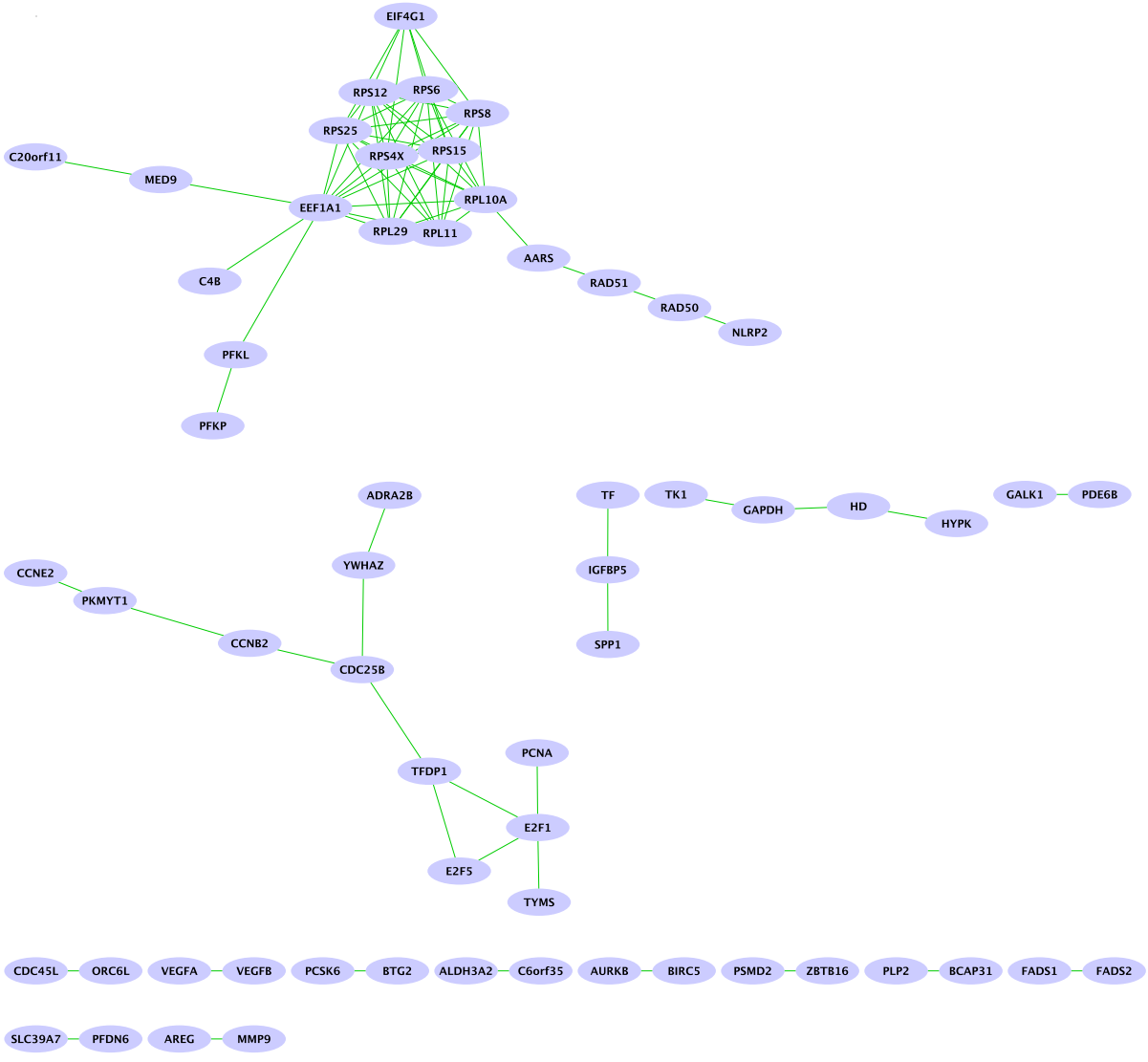
$$\Omega_2(\beta) = \sup_{\alpha \in \mathbb{R}^p : \forall i \sim j, \|\alpha_i^2 + \alpha_j^2\| \leq 1} \alpha^\top \beta \qquad \text{(Jacob et al., 2009)}$$

Lasso signature (accuracy 0.61)



Breast cancer prognosis, Jacob et al. (2009)

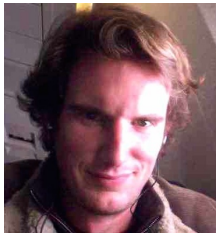
Graph Lasso signature (accuracy 0.64)



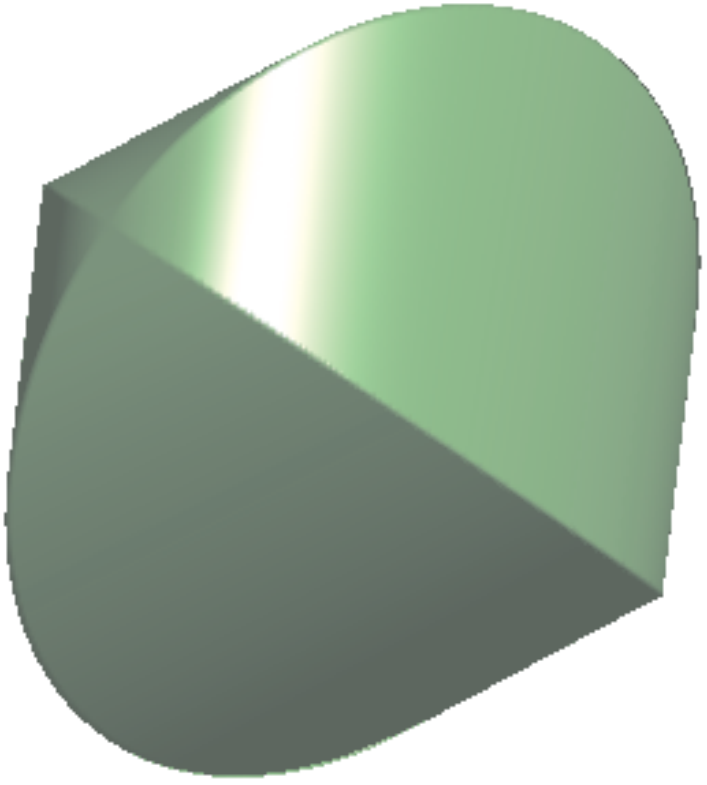
Breast cancer prognosis, Jacob et al. (2009)

Disjoint feature selection

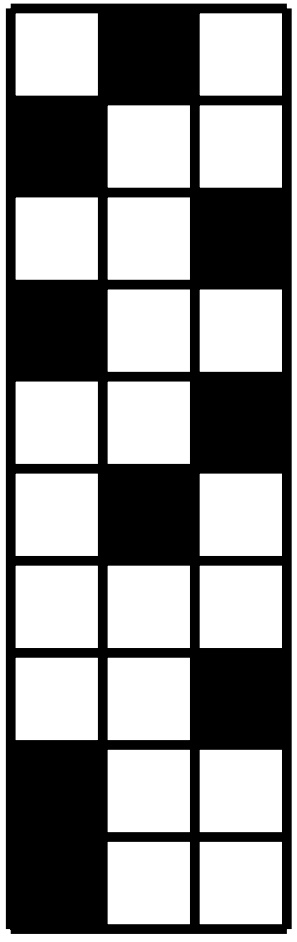
K. Vervier



A. d'Aspremont



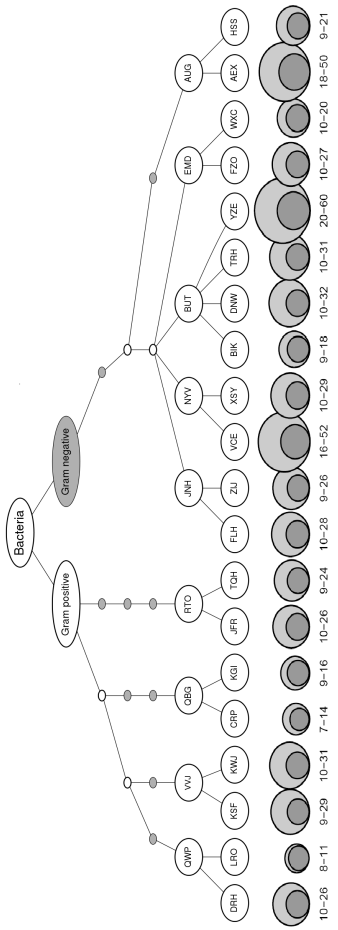
$$X =$$

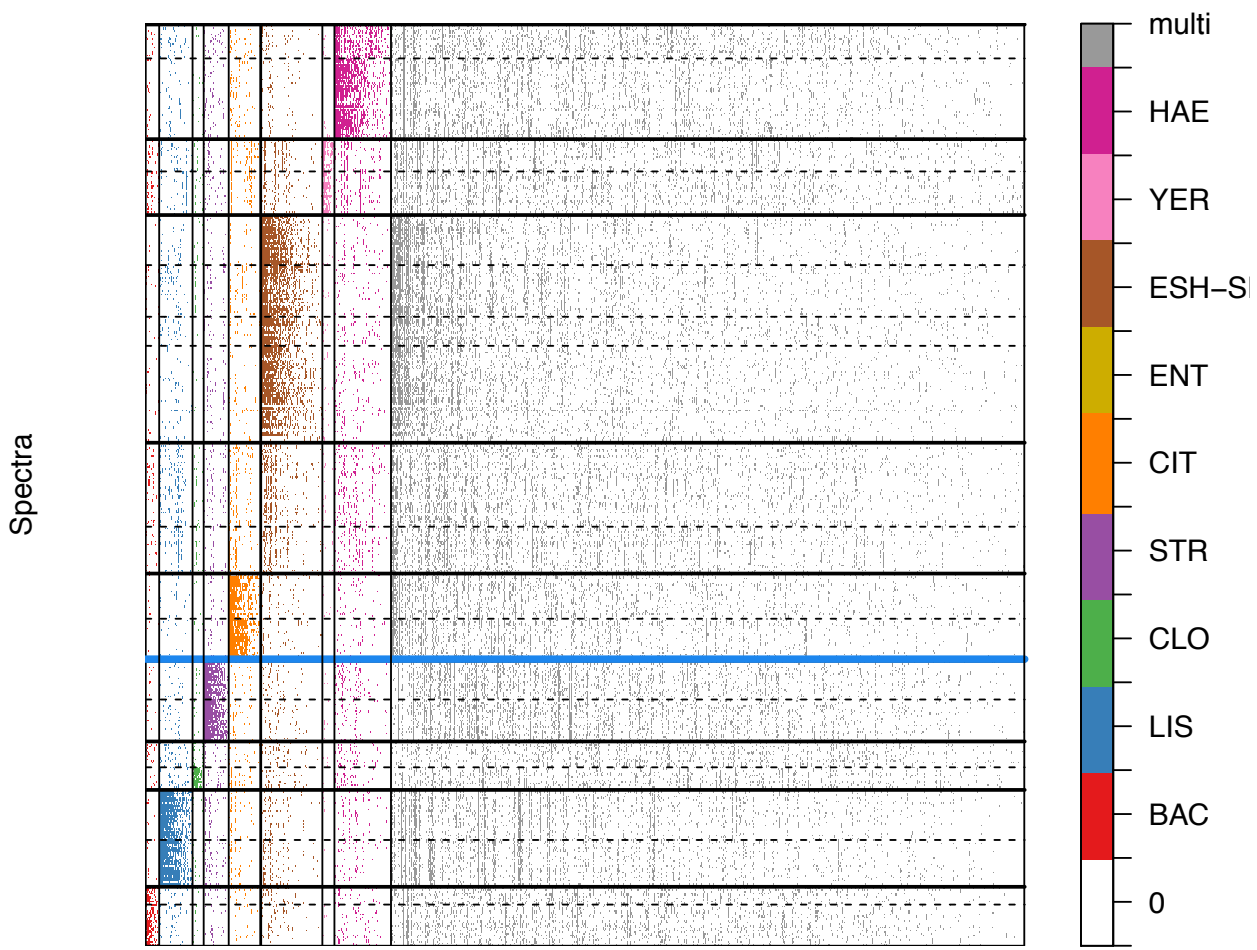


$$\Omega_K(X) = \sum_{i=1}^{p} K_{ii} \| x_i \|^2 + \sum_{i \neq j} K_{ij} \left| x_i^\top x_j \right|$$

(Vervier et al., 2014)

Example: microbial identification from MS spectra





(Vervier et al., 2014)

Summary

$$\min_{\beta} R(f_\beta) + \lambda\Omega(\beta)$$

- Regularization helps learning when $n << p$
- The penalty Ω is a good place to put prior knowledge (related to Bayesian priors)
- A lot of research on positive definite kernels
- Atomic norms offers a general toolbox
  - Structured sparsity
  - Efficient algorithms (convex optimization)
  - Theoretical results

Outline

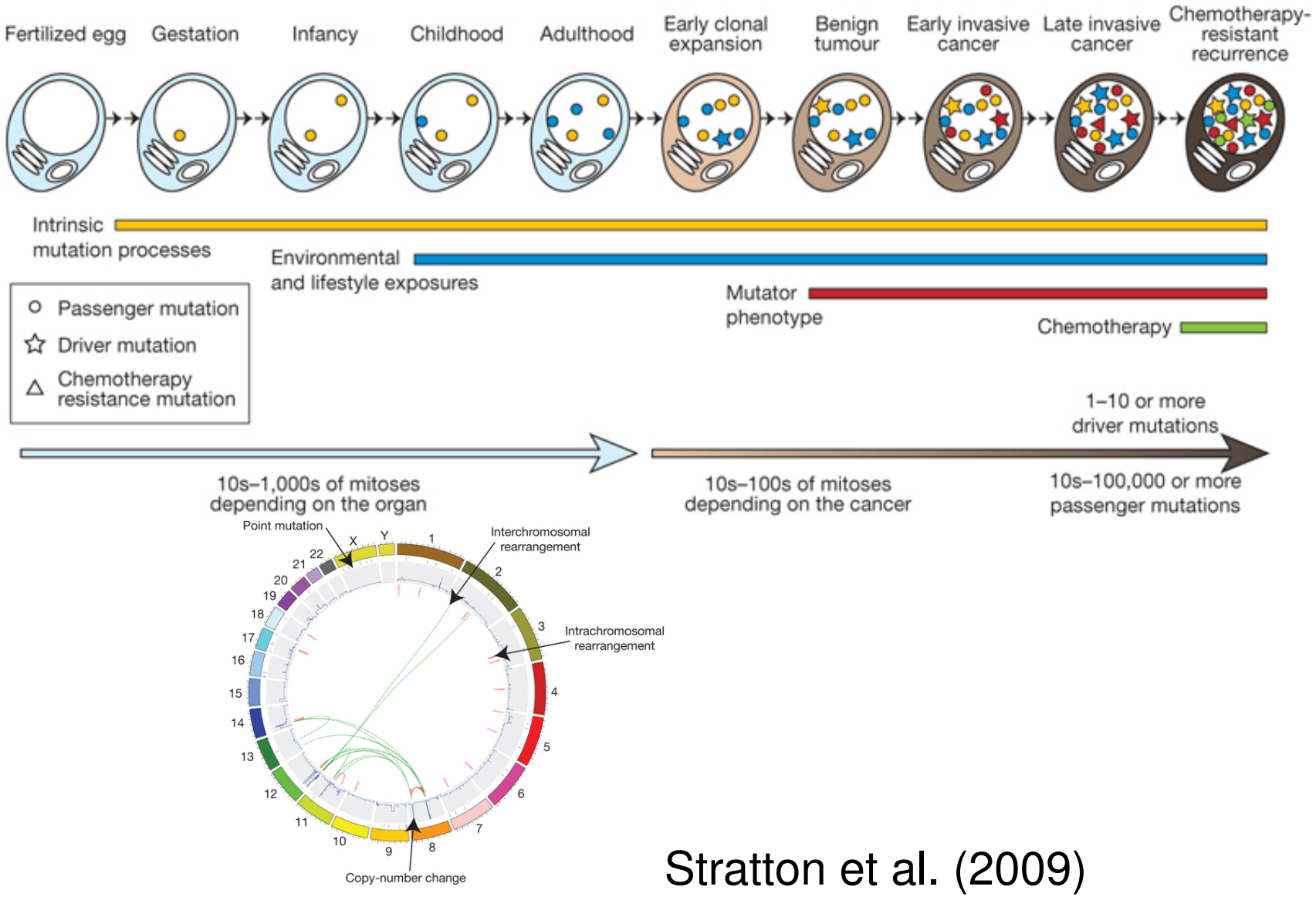
1. Learning with regularization and prior knowledge
2. Cancer patient stratification from somatic mutations
3. Learning from rankings through pairwise comparisons
4. FlipFlop: fast isoform prediction from RNA-seq data
5. Conclusion

Joint work with

Marine Le Morvan

Andrei Zinovyev

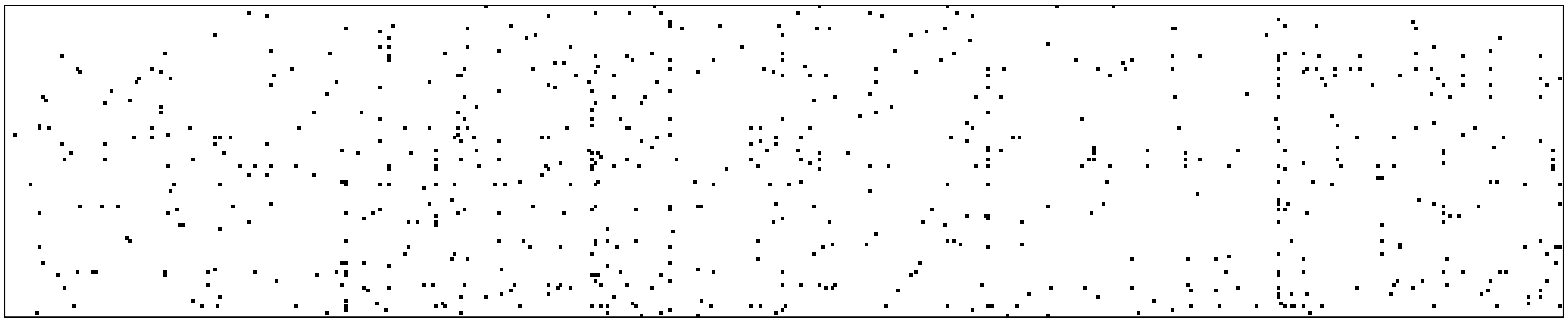
Somatic mutations in cancer



Stratton et al. (2009)

Large-scale efforts to collect somatic mutations

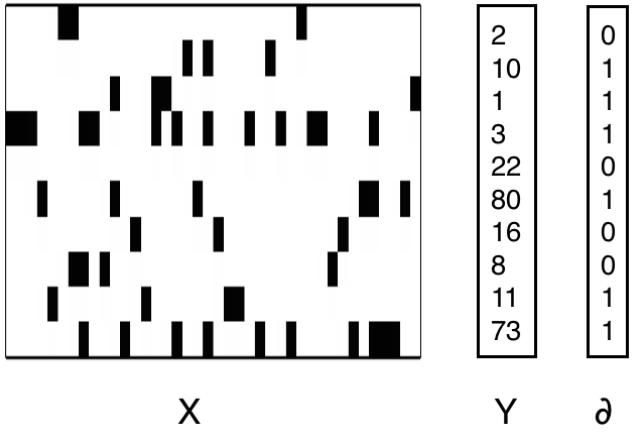
- 3,378 samples with survival information from 8 cancer types
- downloaded from the TCGA / cBioPortal portals.

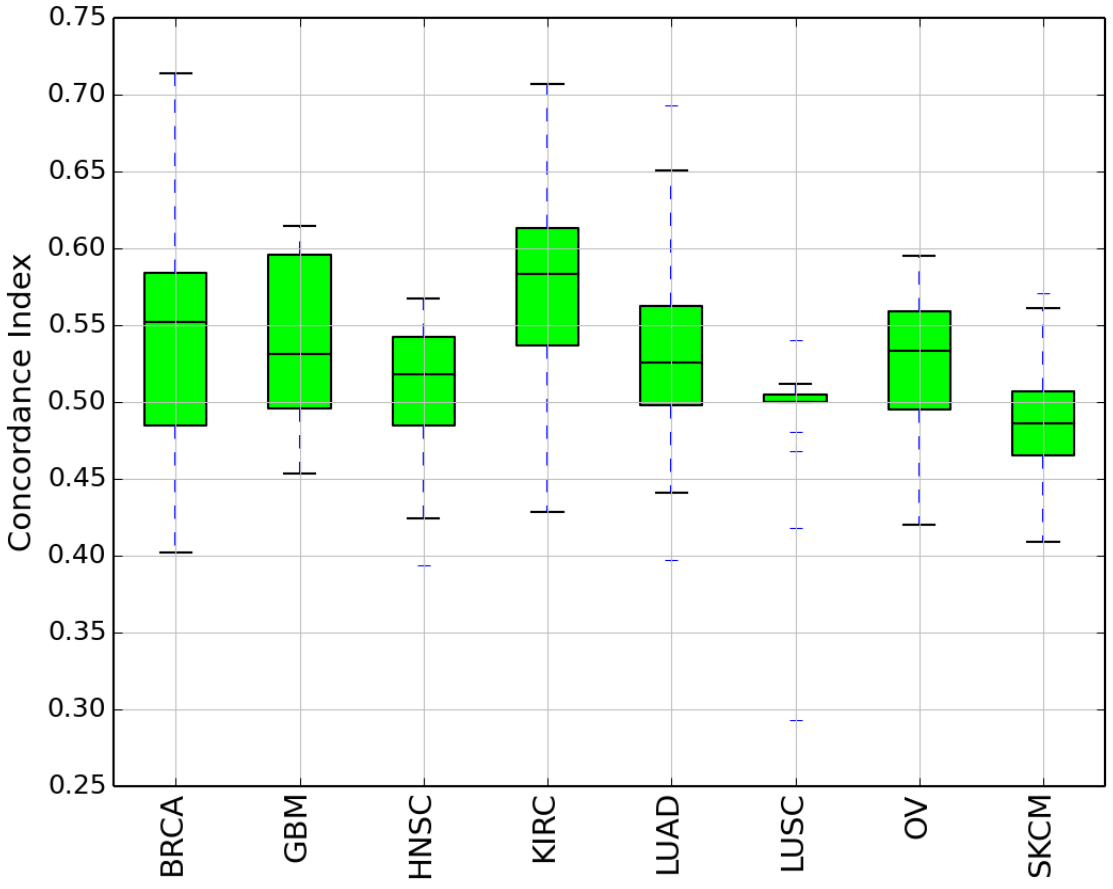


Cancer type	Patients	Genes
LUAD (Lung adenocarcinoma)	430	20 596
SKCM (Skin cutaneous melanoma)	307	17 463
GBM (Glioblastoma multiforme)	265	14 750
BRCA (Breast invasive carcinoma)	945	16 806
KIRC (Kidney renal clear cell carcinoma)	411	10 609
HNSC (Head and Neck squamous cell carcinoma)	388	17 022
LUSC (Lung squamous cell carcinoma)	169	13 590
OV (Ovarian serous cystadenocarcinoma)	363	10 195

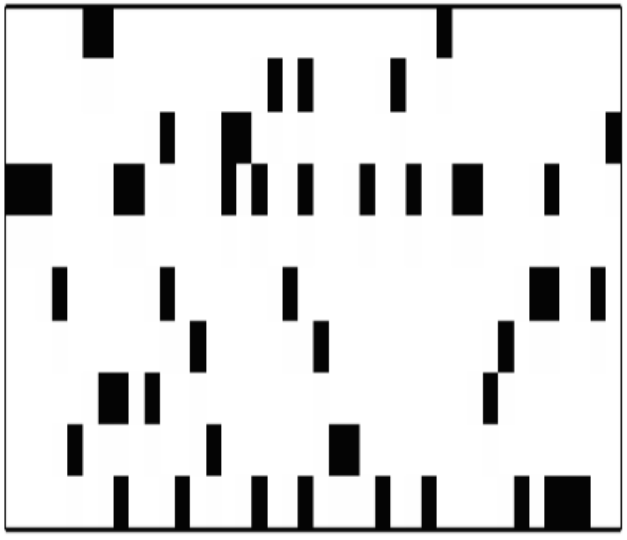
Survival prediction from raw mutation profiles

- Each patient is a binary vector: each gene is mutated (1) or not (2)
- Silent mutations are removed
- Survival model estimated with sparse survival SVM
- Results on 5-fold cross-validation repeated 4 times



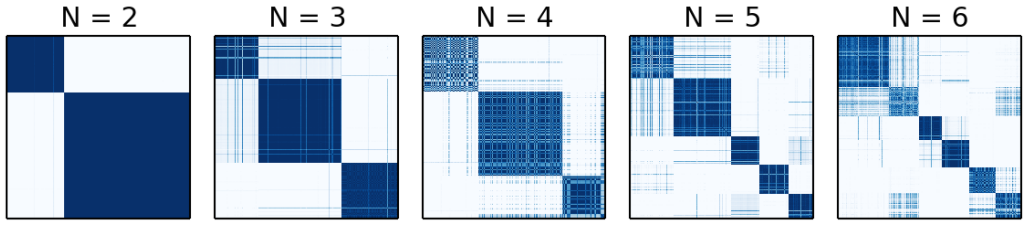


Patient stratification (unsupervised) from raw mutation profiles

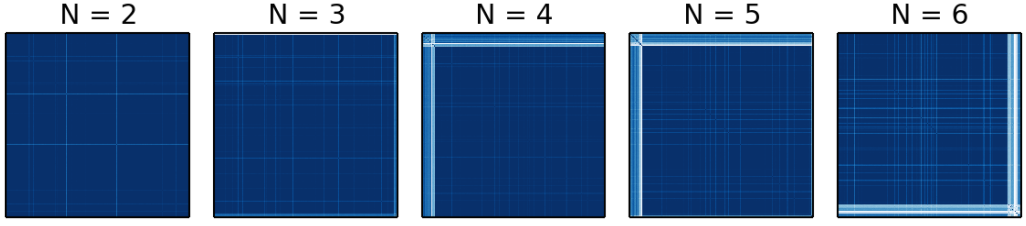


✓ Non-Negative matrix factorisation (NMF)

✓ Desired behaviour:

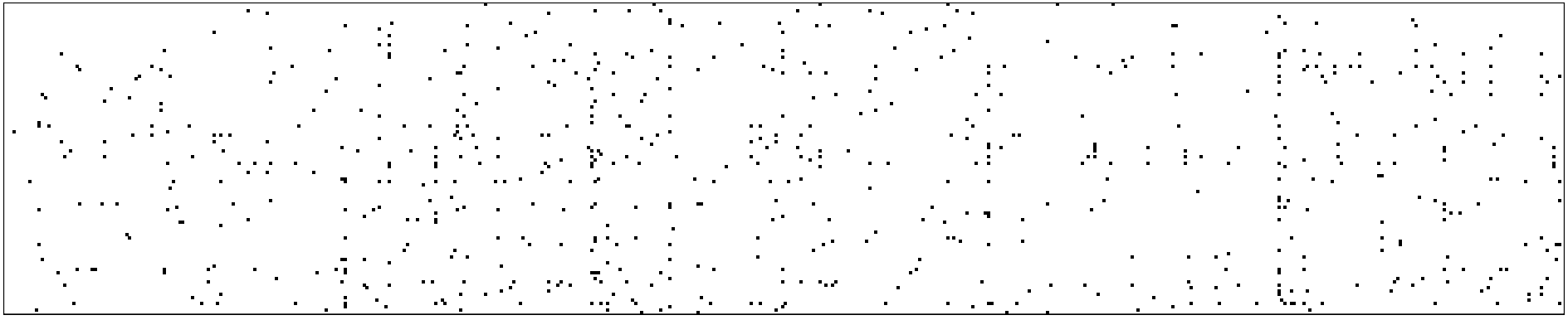


✓ Observed behaviour:



Patients share very few mutated genes!

Changing the representation?



Can we replace

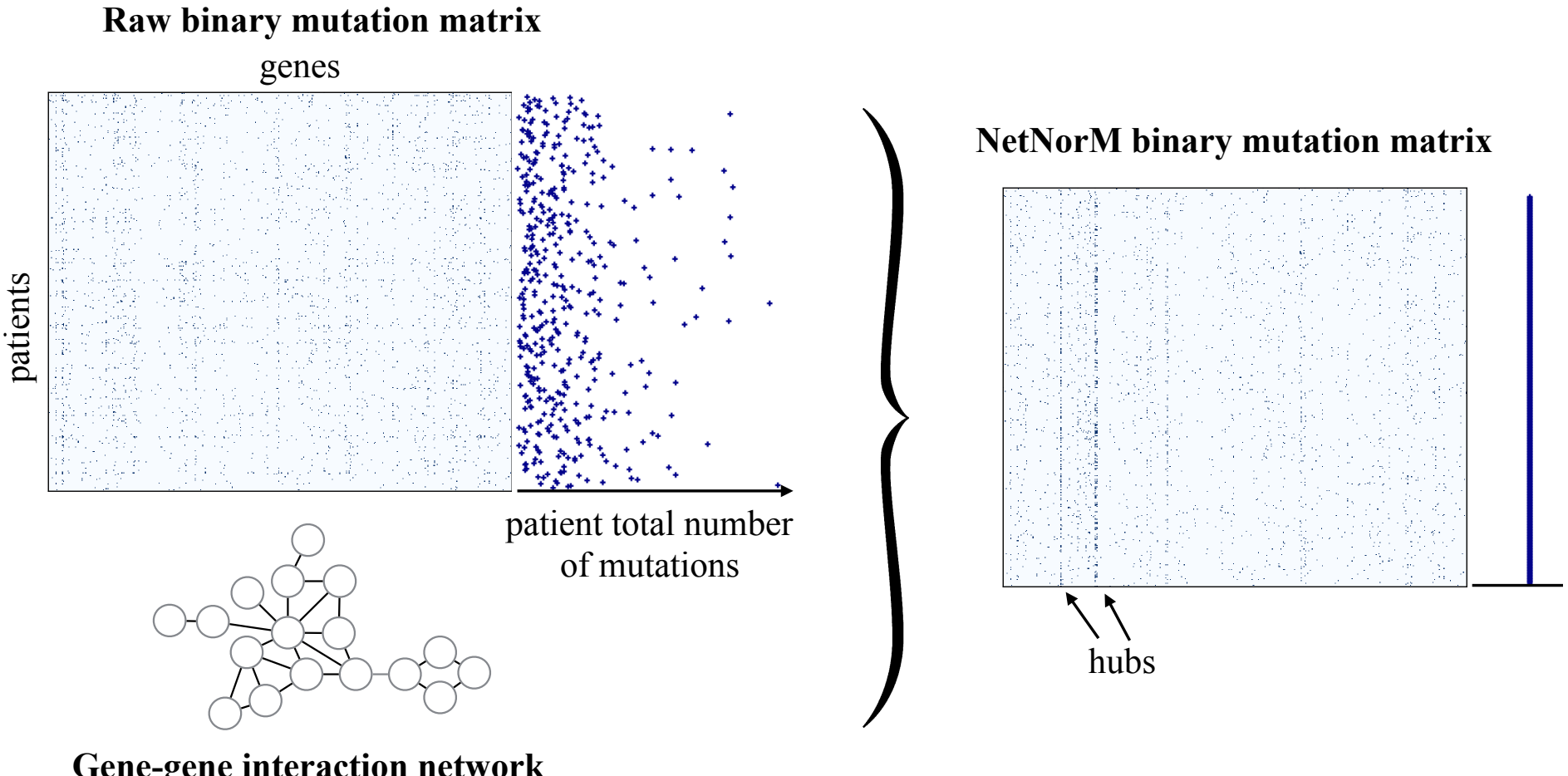
$$x \in \{0,1\}^p \quad \text{with } p \text{ very large, very sparse}$$

by a representation with more information shared between samples

$$\Phi(x) \in \mathcal{H} \quad ?$$

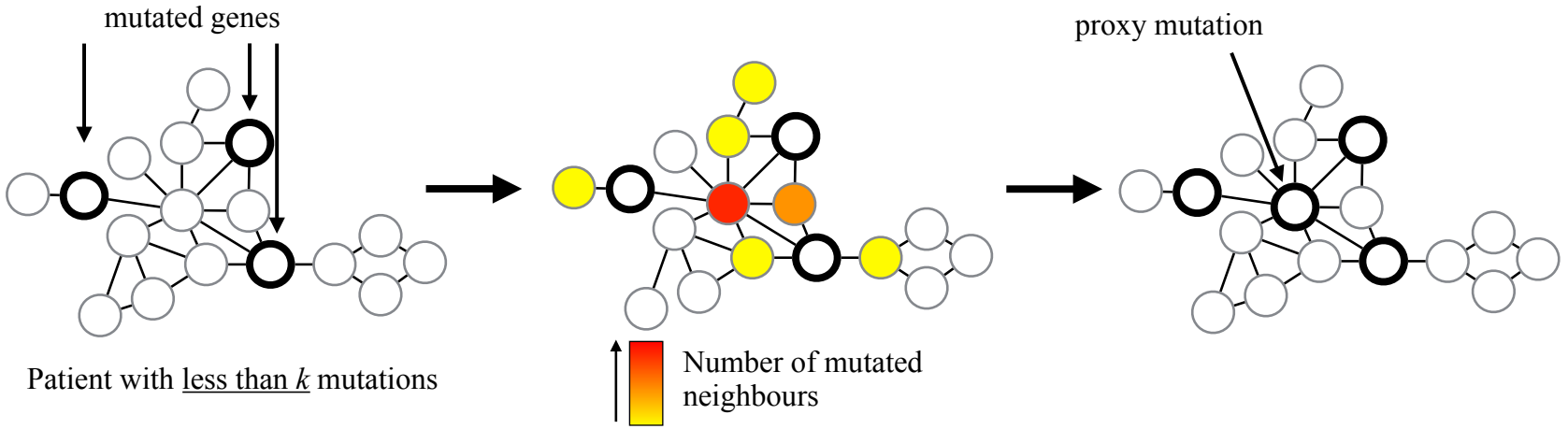
NetNorm Overview (Le Morvan et al., 2016)

- Modify the binary vector $x \in \{0,1\}^p$ of each patient by adding or removing mutations, using a gene network as prior knowledge
- After Netnorm, all patients $\Phi(x) \in \{0,1\}^p$ have the same number of (pseudo-)mutations

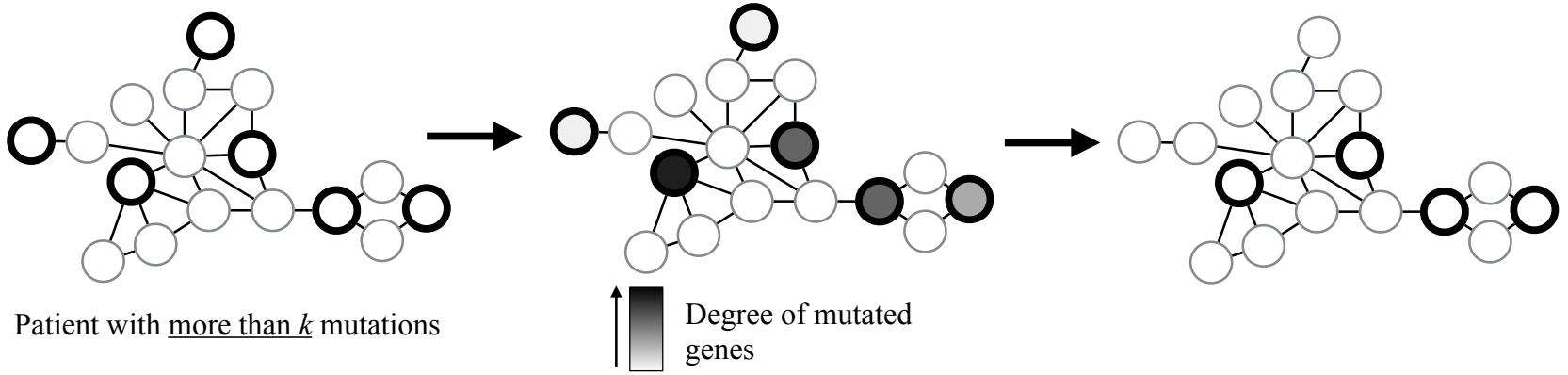


NetNorm detail (k=4)

1. Add mutations for patients with few (less than *k*) mutations



2. Remove mutations for patients for many (more than *k*) mutations



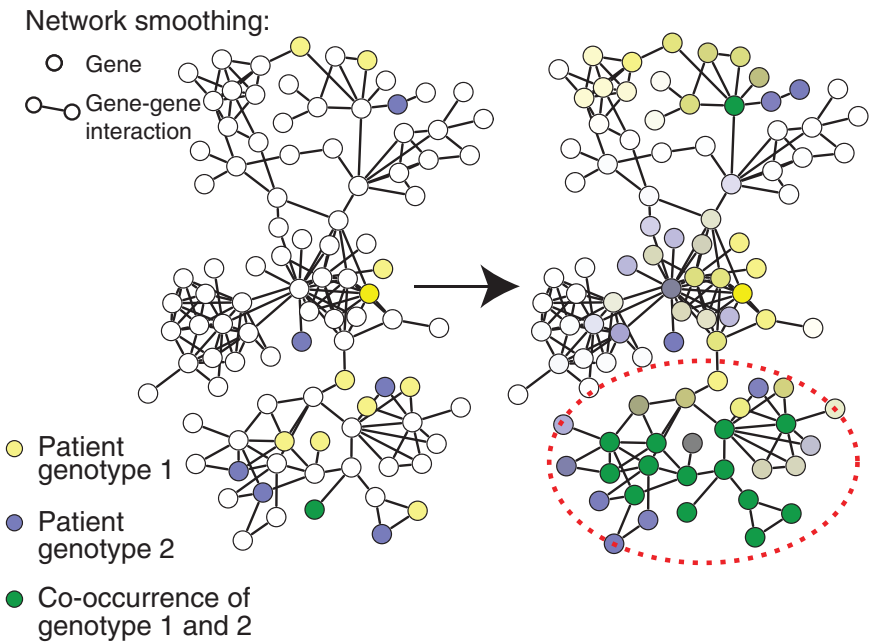
Related work (Hofree et al., 2013)

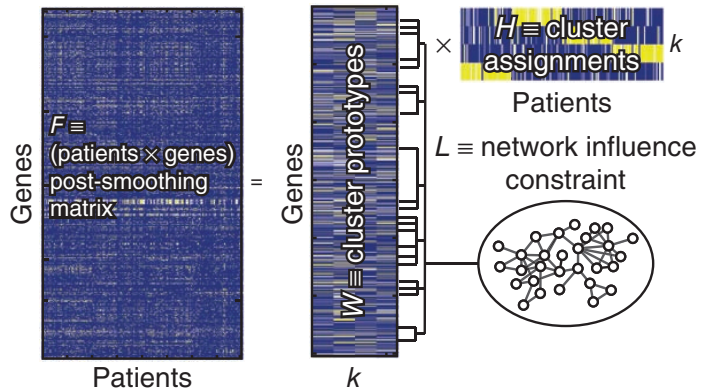
Network-based stratification of tumor mutations

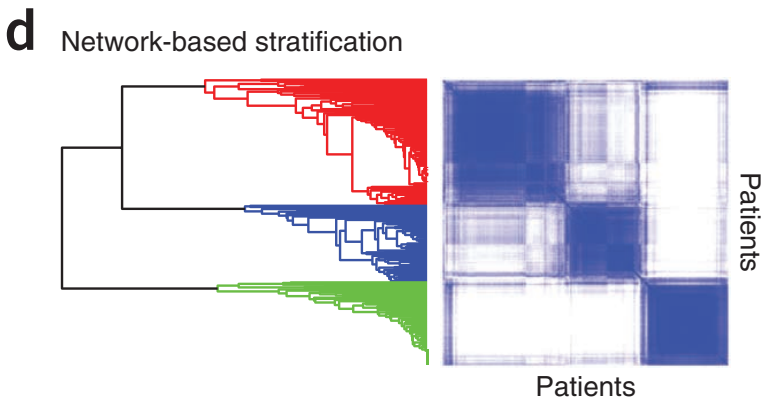
Matan Hofree[1], John P Shen[2], Hannah Carter[2], Andrew Gross[3] & Trey Ideker[1–3]

[1]Department of Computer Science and Engineering, University of California, San Diego, La Jolla, California, USA. [2]Department of Medicine, University of California, San Diego, La Jolla, California, USA. [3]Department of Bioengineering, University of California, San Diego, La Jolla, California, USA. Correspondence should be addressed to T.I. (tideker@ucsd.edu).

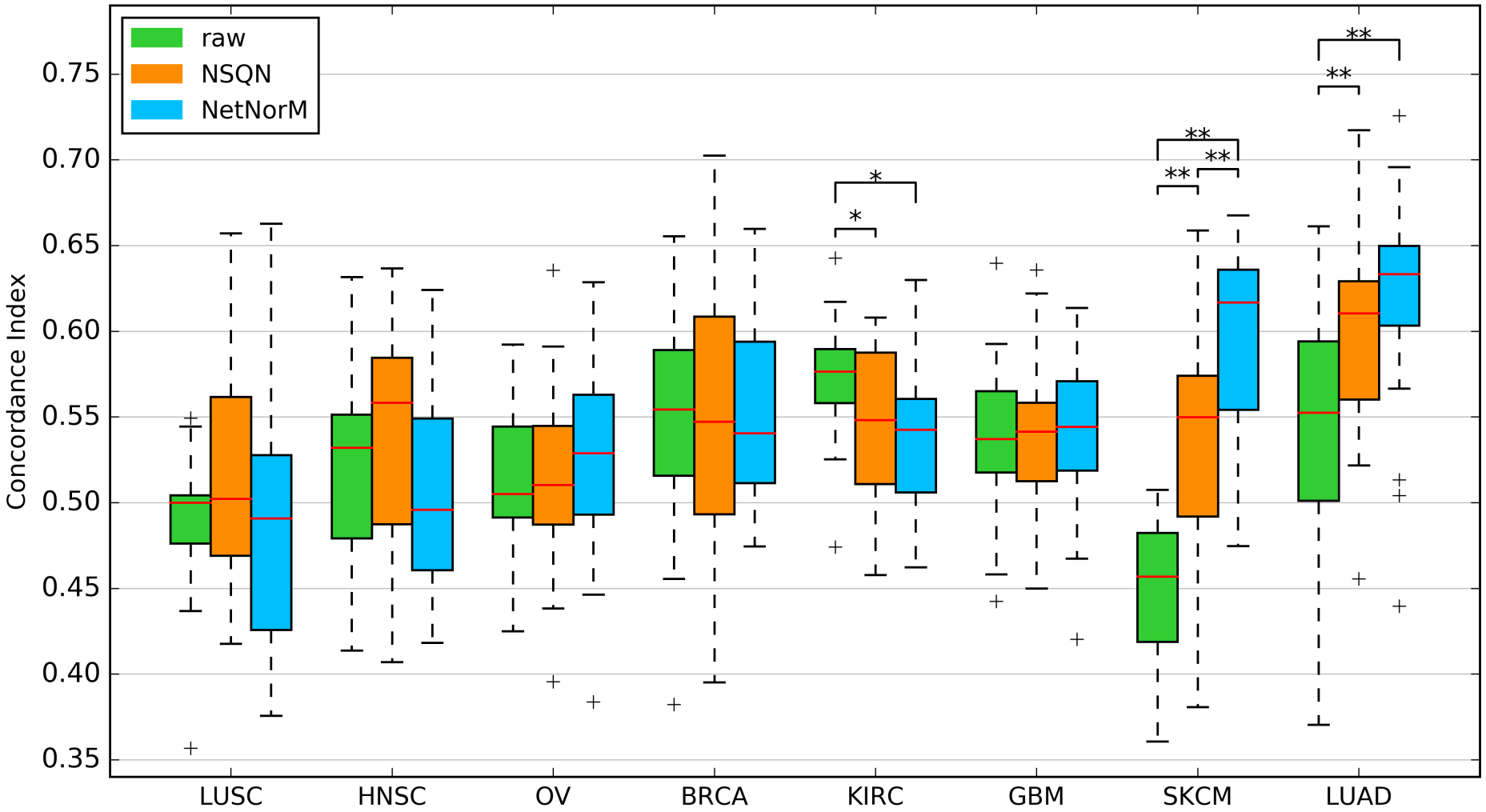
RECEIVED 14 FEBRUARY; ACCEPTED 12 AUGUST; PUBLISHED ONLINE 15 SEPTEMBER 2013; DOI:10.1038/NMETH.2651







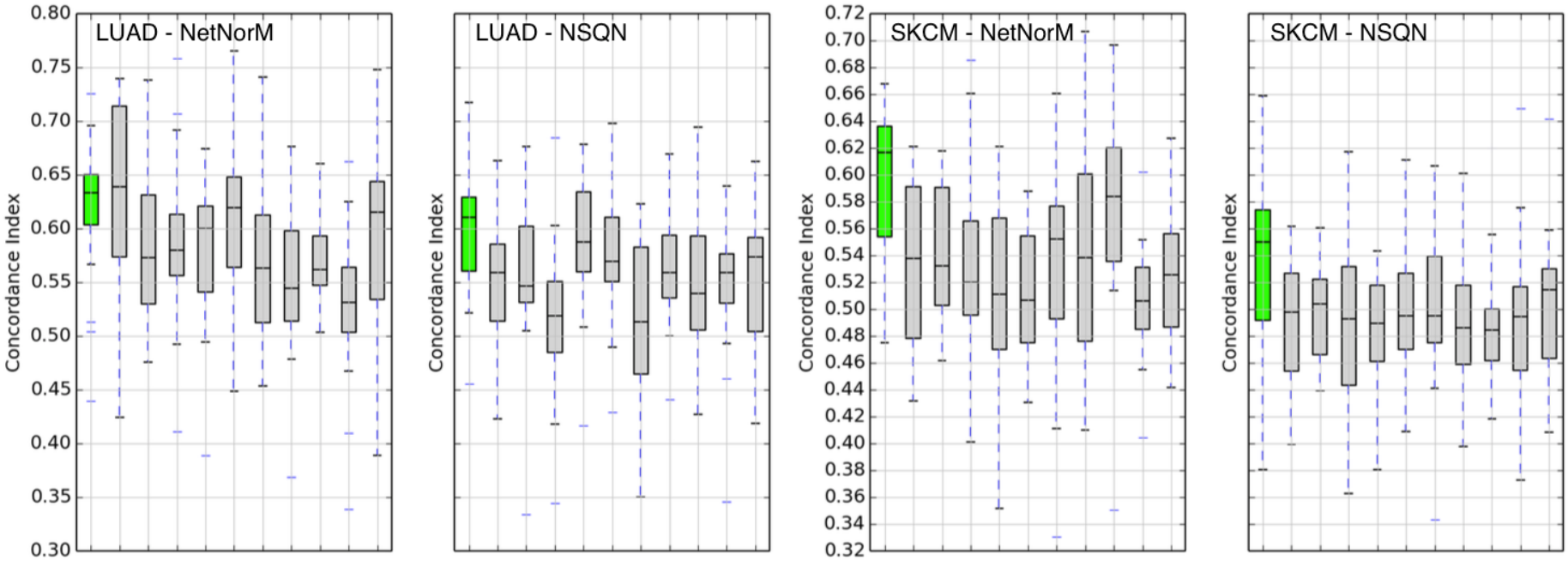
Performance on survival prediction



Use Pathway Commons as gene network.
NSQN = Network Smoothing / Quantile Normalization (Hofree et al., 2013)

NetNorM and NSQN benefit from biological information in the gene network

Comparison with 10 randomly permuted networks:



P-values (Welch *t*-test):

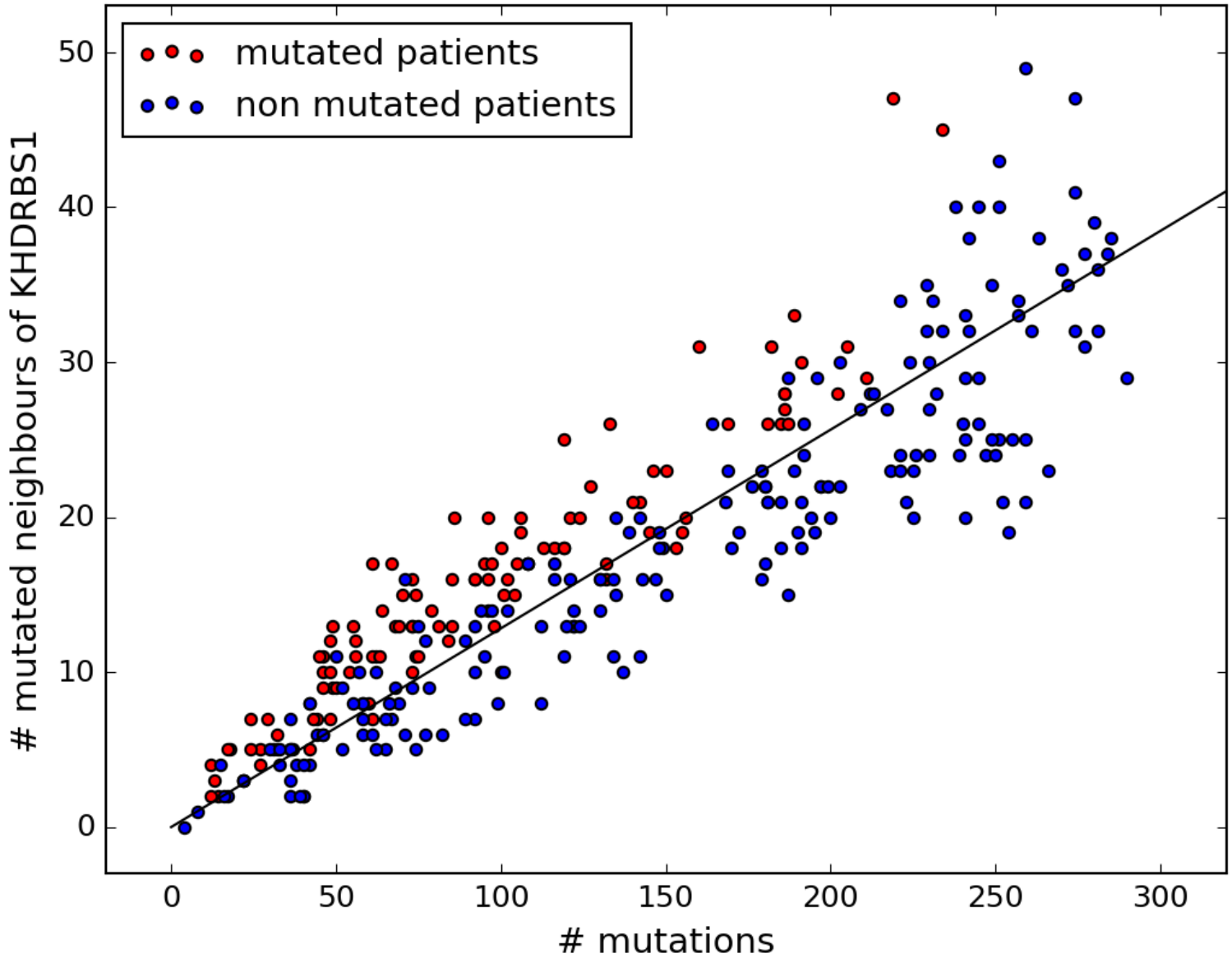
	NSQN	NetNorM
LUAD	2×10^{-3}	3.5×10^{-2}
SKCM	1.2×10^{-2}	1×10^{-4}

Selected genes represent "true" or "proxy" mutations

	freq	coef	m_{all}		$m_{<k_{med}}$		$m_{\geq k_{med}}$		Log-rank test (p-value)		Welsh t-test (p-value)	
			raw	NetNorM	raw	NetNorM	raw	NetNorM	raw	NetNorM	raw	NetNorM
TP53	19	-0.16	238	274	123	159	115	115	7.6×10^{-2}	9.4×10^{-2}	5.2×10^{-22}	1.2×10^{-13}
CRB1	18	-0.4	44	38	22	22	22	16	1.6×10^{-4}	1.4×10^{-6}	9.9×10^{-4}	6.9×10^{-2}
NOTCH4	17	-0.23	42	26	14	14	28	12	9.3×10^{-1}	3.3×10^{-2}	1.9×10^{-6}	2.6×10^{-1}
ANK2	17	0.1	90	90	33	33	57	57	1.2×10^{-2}	1.2×10^{-2}	6.3×10^{-10}	6.3×10^{-10}
RPS9	16	0.38	0	106	0	106	0	0	-	1.8×10^{-1}	-	4.2×10^{-47}
LAMA2	15	0.16	52	38	14	15	38	23	1.5×10^{-2}	2.3×10^{-2}	6.3×10^{-9}	2.6×10^{-3}
RYR2	14	0.07	165	161	70	70	95	91	1.4×10^{-2}	2.1×10^{-2}	6.7×10^{-19}	1×10^{-15}
IGF2BP2	14	-0.15	6	67	2	63	4	4	1.4×10^{-5}	3.6×10^{-3}	1×10^{-1}	6.8×10^{-7}
SMARCA5	14	-0.09	5	137	1	133	4	4	2.1×10^{-1}	5.3×10^{-3}	1.3×10^{-1}	1×10^{-27}
KHDRBS1	13	0.11	7	117	2	112	5	5	7.1×10^{-1}	9.7×10^{-1}	6.5×10^{-2}	1.3×10^{-18}
YWHAZ	13	-0.18	2	241	0	239	2	2	2.5×10^{-31}	6.1×10^{-4}	4.7×10^{-1}	4.4×10^{-37}
HRNR	13	-0.12	62	64	20	22	42	42	1.1×10^{-1}	1.1×10^{-1}	6×10^{-10}	2.9×10^{-9}
CSNK2A2	11	0.06	2	129	1	128	1	1	9×10^{-1}	8.8×10^{-1}	5.9×10^{-1}	4.2×10^{-27}
MED12L	11	0.04	27	27	8	8	19	19	5.5×10^{-2}	5.5×10^{-2}	1.7×10^{-4}	1.7×10^{-4}

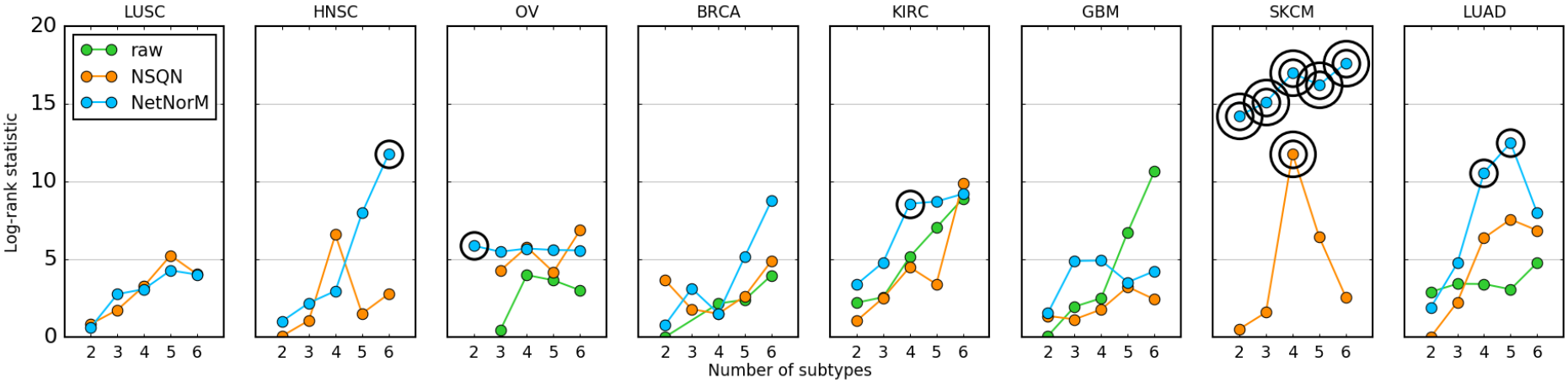
- 14 genes are selected at least 50% of the time
- 6/14 are "proxy" genes (in blue)
  - big hubs in the network
  - get mutated by NetNorm in patients with few mutations $\Longrightarrow$ they encode the mutation rate
- 8/14 are "normal" prognostic genes

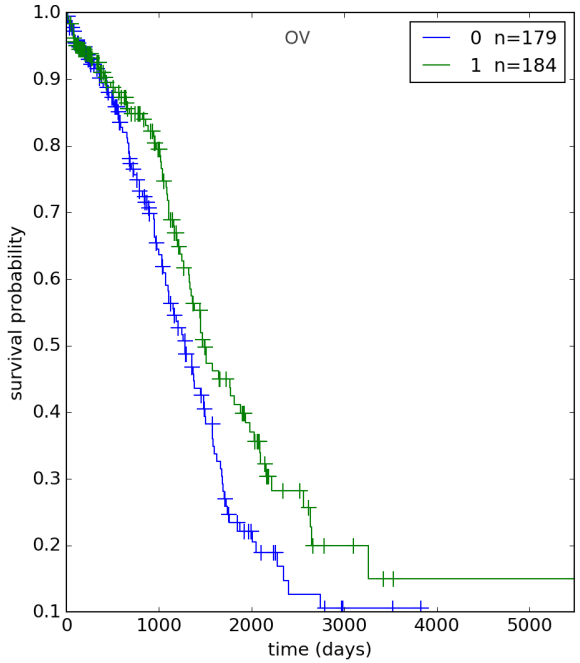
Proxy mutations encode local mutational burden

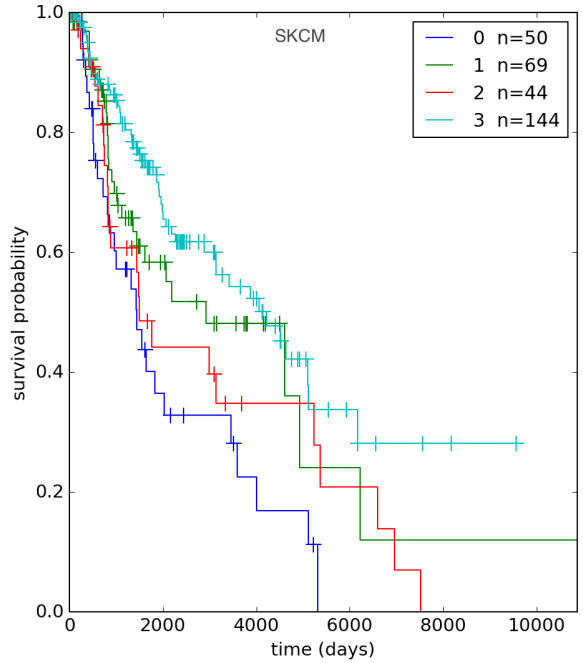


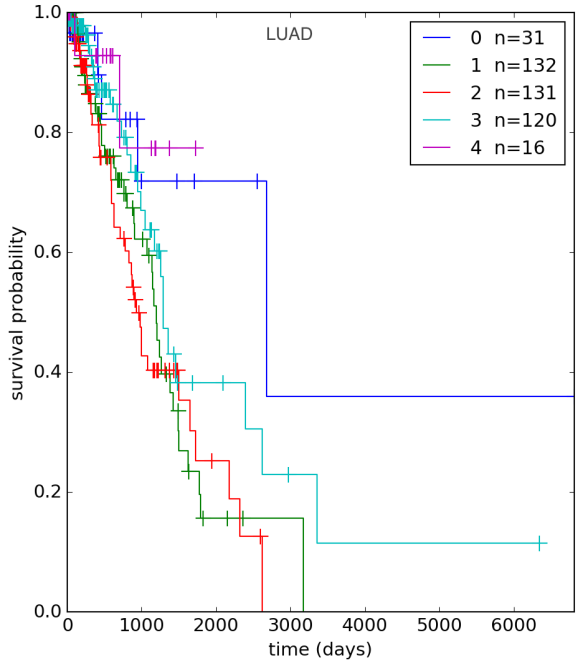
KHDRBS1: a member of the K homology domain-containing, RNA-binding, signal transduction-associated protein family

Performance on unsupervised patient stratification









Summary

- Somatic mutation profiles are challenging because
  - Little overlap between patients
  - Large variability in number of mutations
- Network smoothing / local averaging sometimes helps
  - but with current methods, looking at the direct neighbors is good enough
- Normalizing for total number of mutations is important
  - through QN or NetNorm, for example
  - this is not for biological reasons, but for mathematical reasons
  - probably room for improvement to find a good representation $\Phi(x)$
- References
  - `https://hal.archives-ouvertes.fr/hal-01341856`
  - `https://github.com/marineLM/NetNorM`

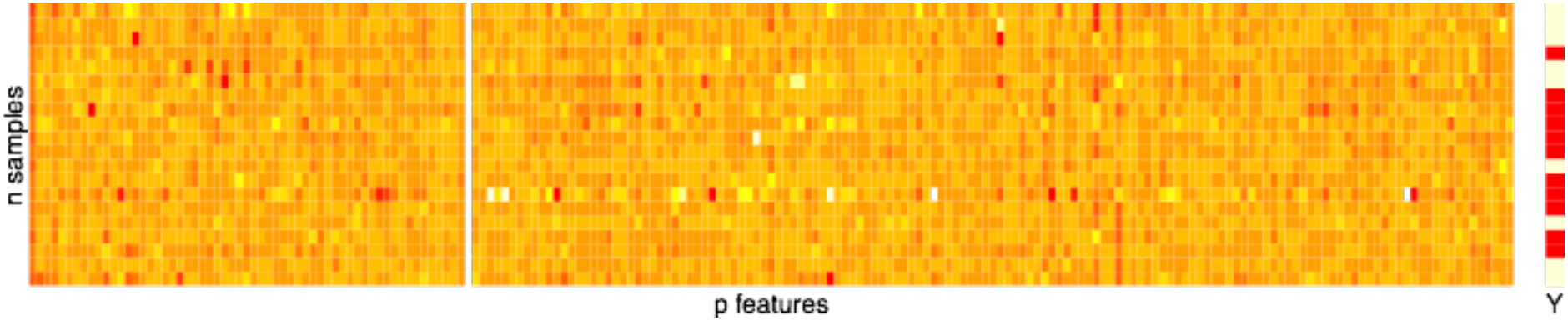
Outline

1. Learning with regularization and prior knowledge
2. Cancer patient stratification from somatic mutations
3. Learning from rankings through pairwise comparisons
4. FlipFlop: fast isoform prediction from RNA-seq data
5. Conclusion

Joint work with

Yunlong Jiao

Back to the $n << p$ problem



Can we replace

$$x \in \mathbb{R}^p$$

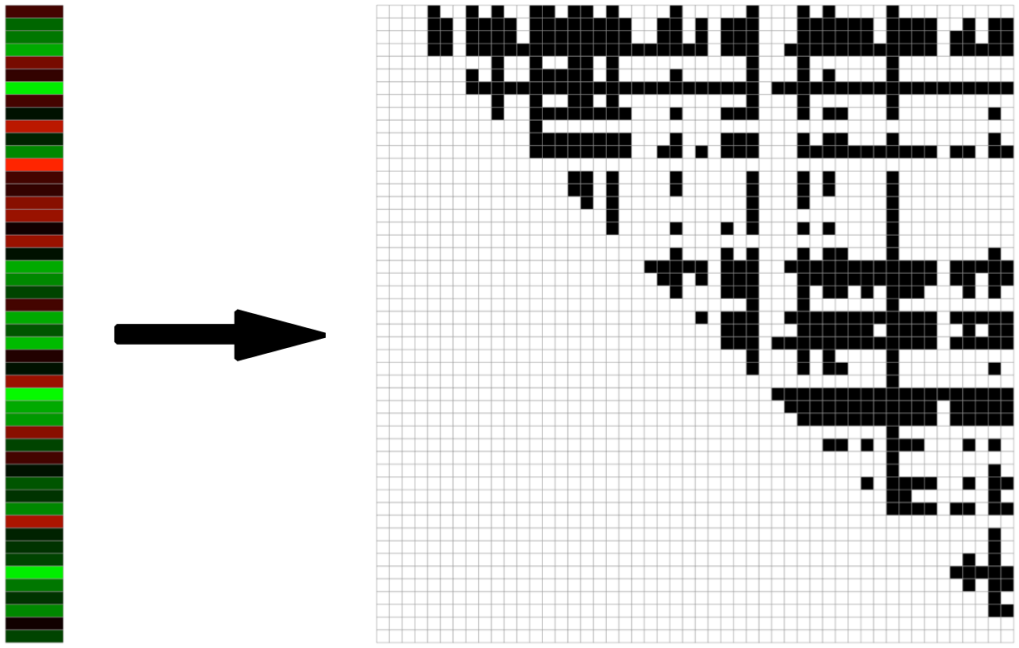
by a "simpler" representation

$$\Phi(x) \in \mathcal{H} \quad ?$$

An idea: all pairwise comparisons

Replace $x \in \mathbb{R}^p$ by $\Phi(x) \in \{0,1\}^{p(p-1)/2}$:

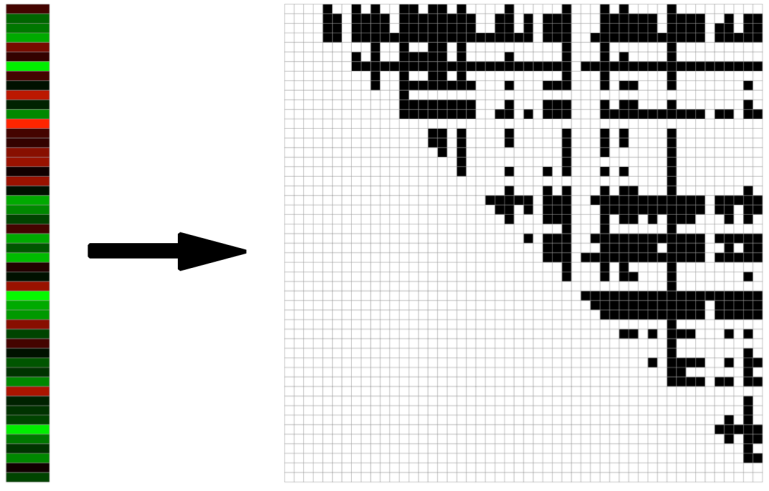
$$\Phi_{i,j}(x) = \begin{cases} 1 & \text{if } x_i \leq x_j\,, \\ 0 & \text{otherwise.} \end{cases}$$



One sample x
p features

Mapping f(x)
p(p-1)/2 bits

Remark: representation of the symmetric group

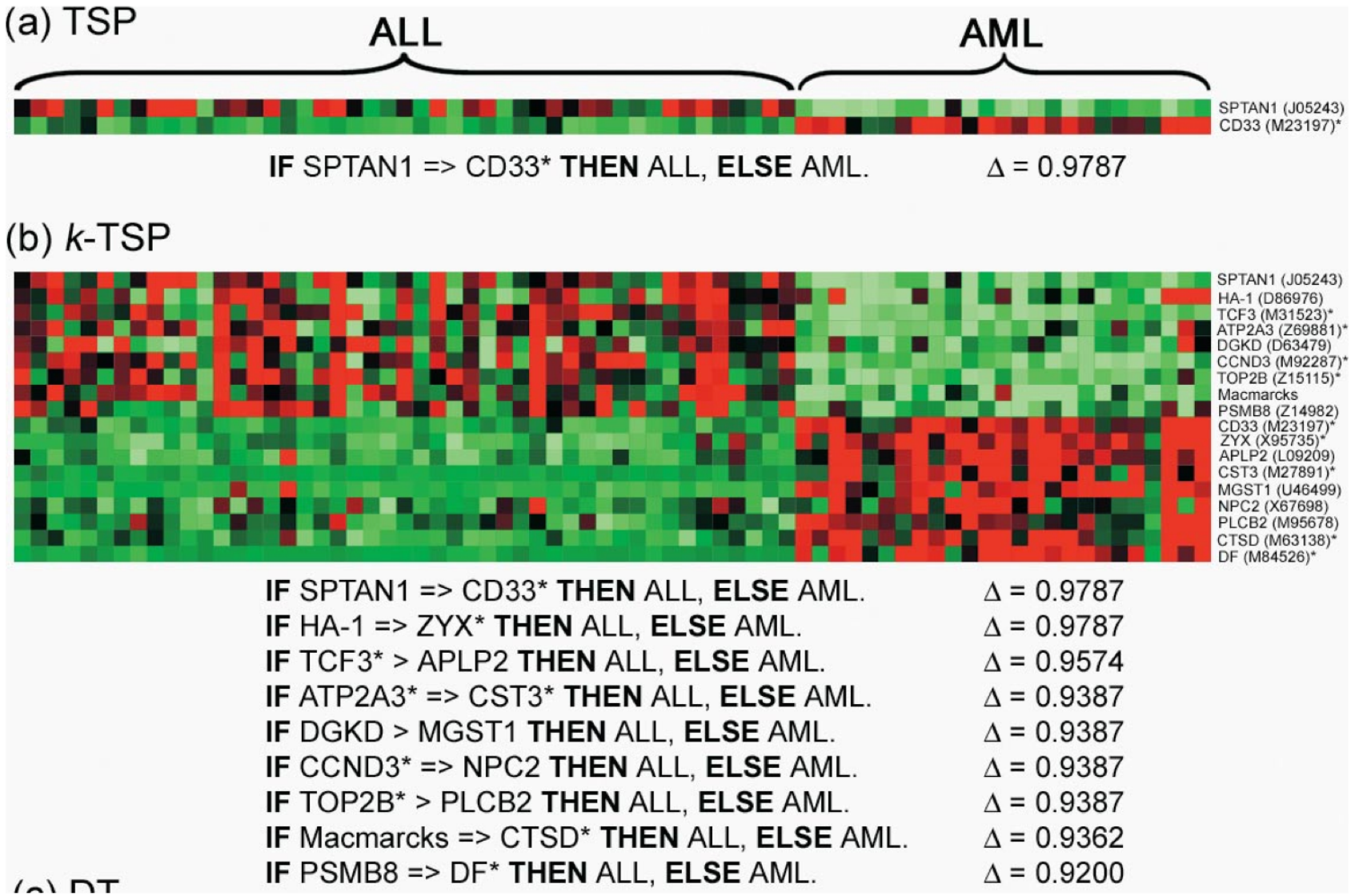


One sample x
p features

Mapping f(x)
p(p-1)/2 bits

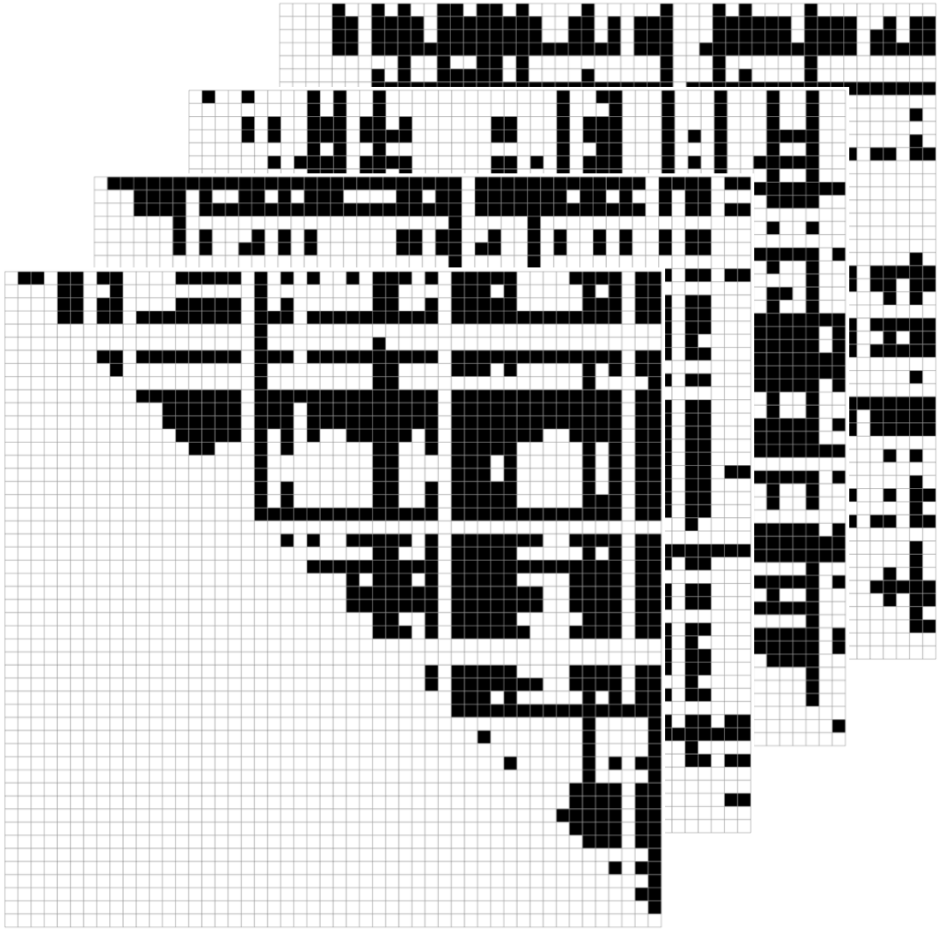
- Obviously, this representation as $O(p^2)$ bits exists for any ranking or permutation of p items
- Many other applications in learning over rankings, learning to rank, learning permutations etc...
- We are interested particularly in practical solutions when p is large

Related work: Top scoring pairs (TSP)



(Geman et al., 2004; Tan et al., 2005; Leek, 2009)

Practical challenge



- Need to store $O(p^2)$ bits per sample
- Need to train a model in $O(p^2)$ dimensions

Kernel trick

Theorem (Wahba, Schölkopf, ...)

Training a linear model over a representation $\Phi(x) \in \mathbb{R}^Q$ of the form:

$$\min_{w \in \mathbb{R}^Q} \frac{1}{n} \sum_{i=1}^{n} \ell(w^\top \Phi(x_i), y_i) + \lambda ||w||^2$$

can be done efficiently, independently of Q, if the kernel

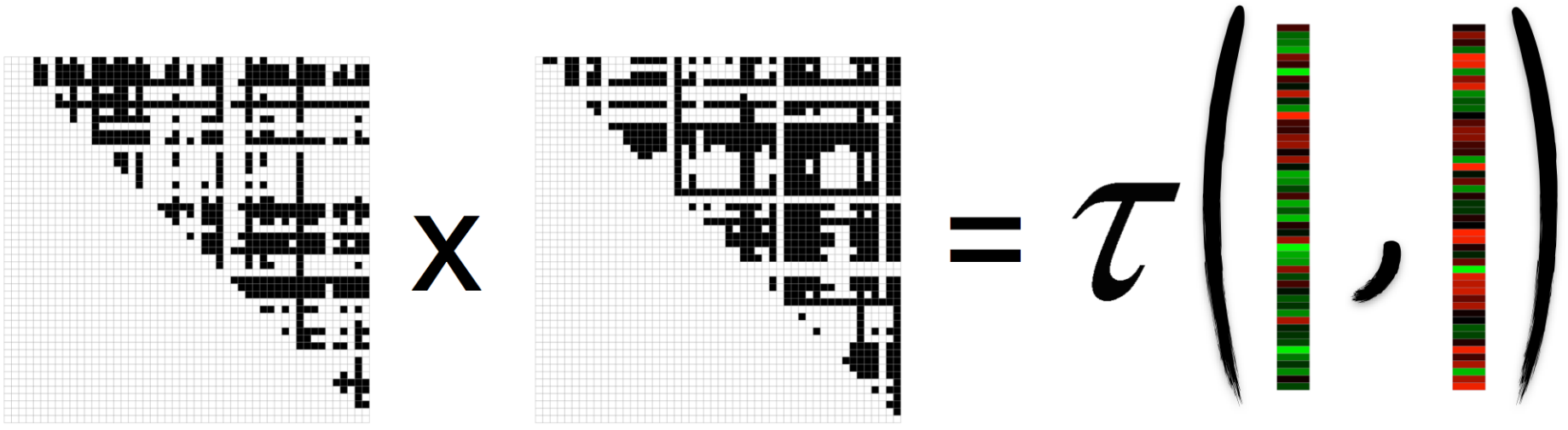
$$K(x, x') = \Phi(x)^\top \Phi(x')$$

can be computed efficiently.

Ex: ridge regression, $O(Q^3 + nQ^2)$ becomes $O(n^3 + n^2 T)$
Other: SVM, logistic regression, Cox model, survival SVM, ...

Kernel trick for us: Kendall's τ

$$\Phi(x)^\top \Phi(x') = \tau(x, x') \quad \text{(up to a scaling)}$$



O(p^2)

O(p log(p))

Good news for SVM and kernel methods!

More formally

- For two permutations σ, σ' let $n_c(\sigma, \sigma')$ (resp. $n_d(\sigma, \sigma')$) the number of concordant (resp. discordant) pairs.
- The Kendall kernel (a.k.a. Kendall tau coefficient) is defined as

$$K_\tau(\sigma, \sigma') = \frac{n_c(\sigma, \sigma') - n_d(\sigma, \sigma')}{\binom{p}{2}} .$$

- The Mallows kernel is defined for any $\lambda \geq 0$ by

$$K_M^\lambda(\sigma, \sigma') = e^{-\lambda n_d(\sigma, \sigma')} .$$

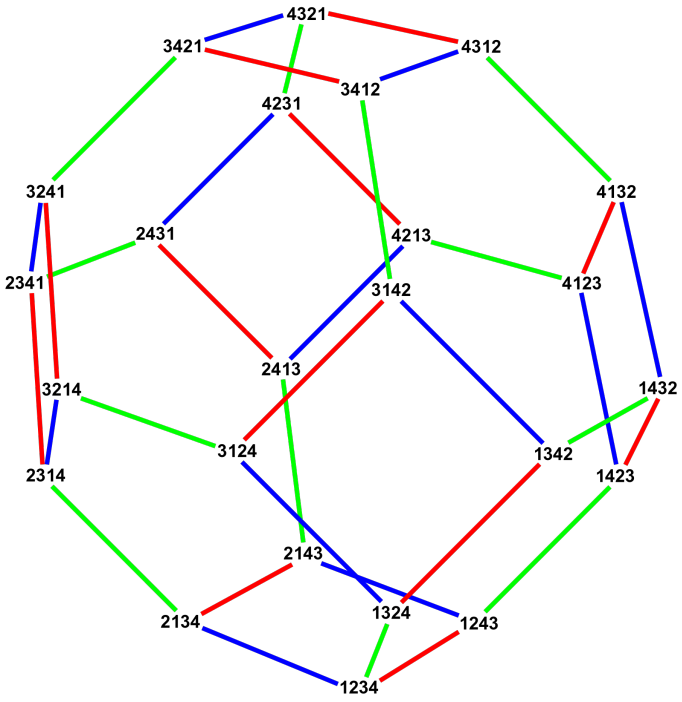
Theorem ((Jiao and Vert, 2015))

The Kendall and Mallows kernels are positive definite.

Theorem ((Knight, 1966))

These two kernels for permutations can be evaluated in $O(p \log p)$ time.

Related work



Cayley graph of $\mathbb{S}_4$

- Kondor and Barbarosa (2010) proposed the diffusion kernel on the Cayley graph of the symmetric group generated by adjacent transpositions.
- Computationally intensive ($O(p^p)$)
- Mallows kernel is written as

$$K_M^{\lambda}(\sigma, \sigma') = e^{-\lambda n_d(\sigma,\sigma')},$$

where $n_d(\sigma, \sigma')$ is the shortest path distance on the Cayley graph.

- It can be computed in $O(p \log p)$

Application: supervised classification

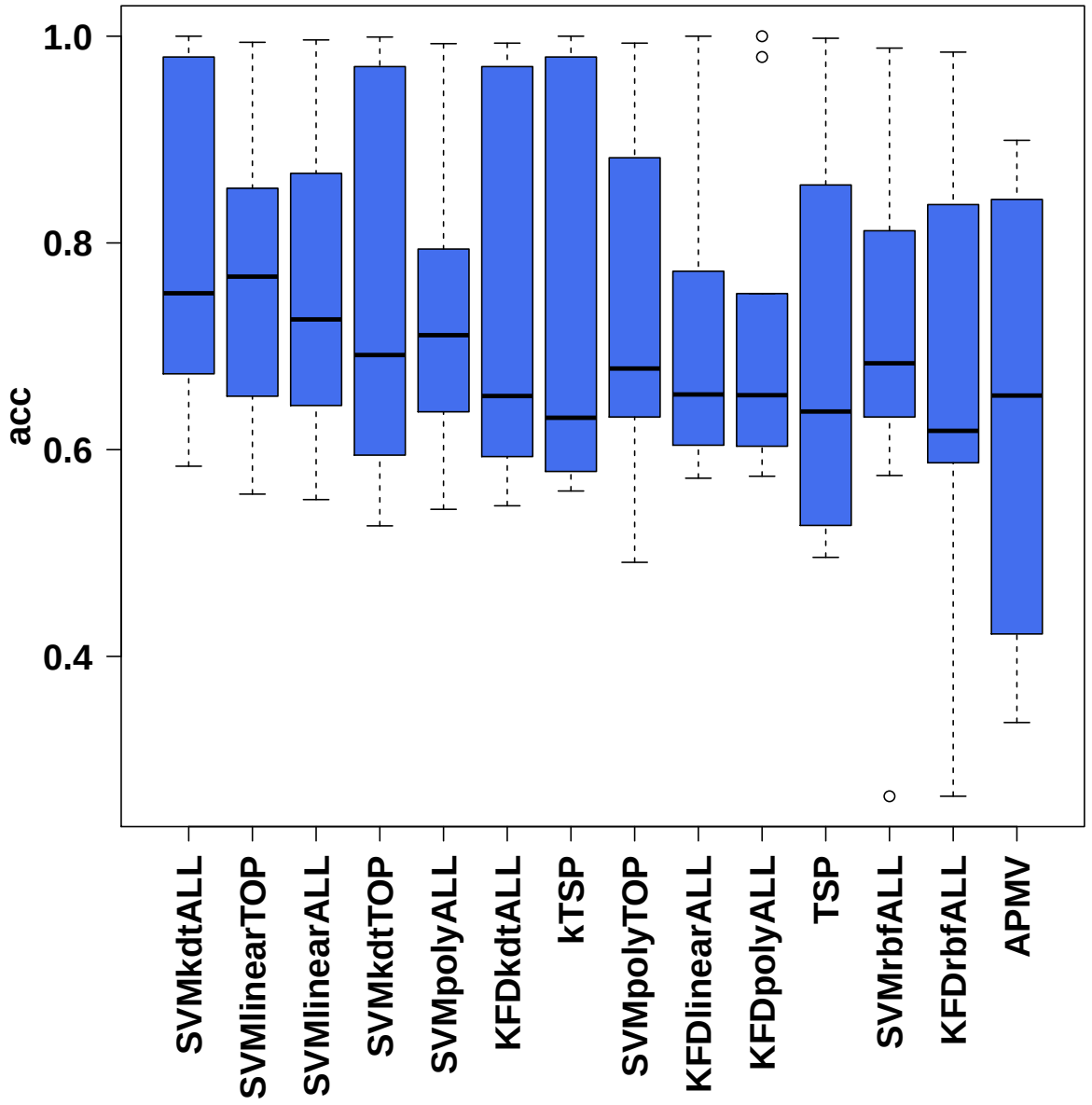
Datasets

Dataset	No. of features	No. of samples (training/test)	
		C_1	C_2
Breast Cancer 1	23624	44/7 (Non-relapse)	32/12 (Relapse)
Breast Cancer 2	22283	142 (Non-relapse)	56 (Relapse)
Breast Cancer 3	22283	71 (Poor Prognosis)	138 (Good Prognosis)
Colon Tumor	2000	40 (Tumor)	22 (Normal)
Lung Cancer 1	7129	24 (Poor Prognosis)	62 (Good Prognosis)
Lung Cancer 2	12533	16/134 (ADCA)	16/15 (MPM)
Medulloblastoma	7129	39 (Failure)	21 (Survivor)
Ovarian Cancer	15154	162 (Cancer)	91 (Normal)
Prostate Cancer 1	12600	50/9 (Normal)	52/25 (Tumor)
Prostate Cancer 2	12600	13 (Non-relapse)	8 (Relapse)

Methods

- Kernel machines Support Vector Machines (SVM) and Kernel Fisher Discriminant (KFD) with Kendall kernel, linear kernel, Gaussian RBF kernel, polynomial kernel.
- Top Scoring Pairs (TSP) classifiers Tan et al. (2005).
- Hybrid scheme of SVM + TSP feature selection algorithm.

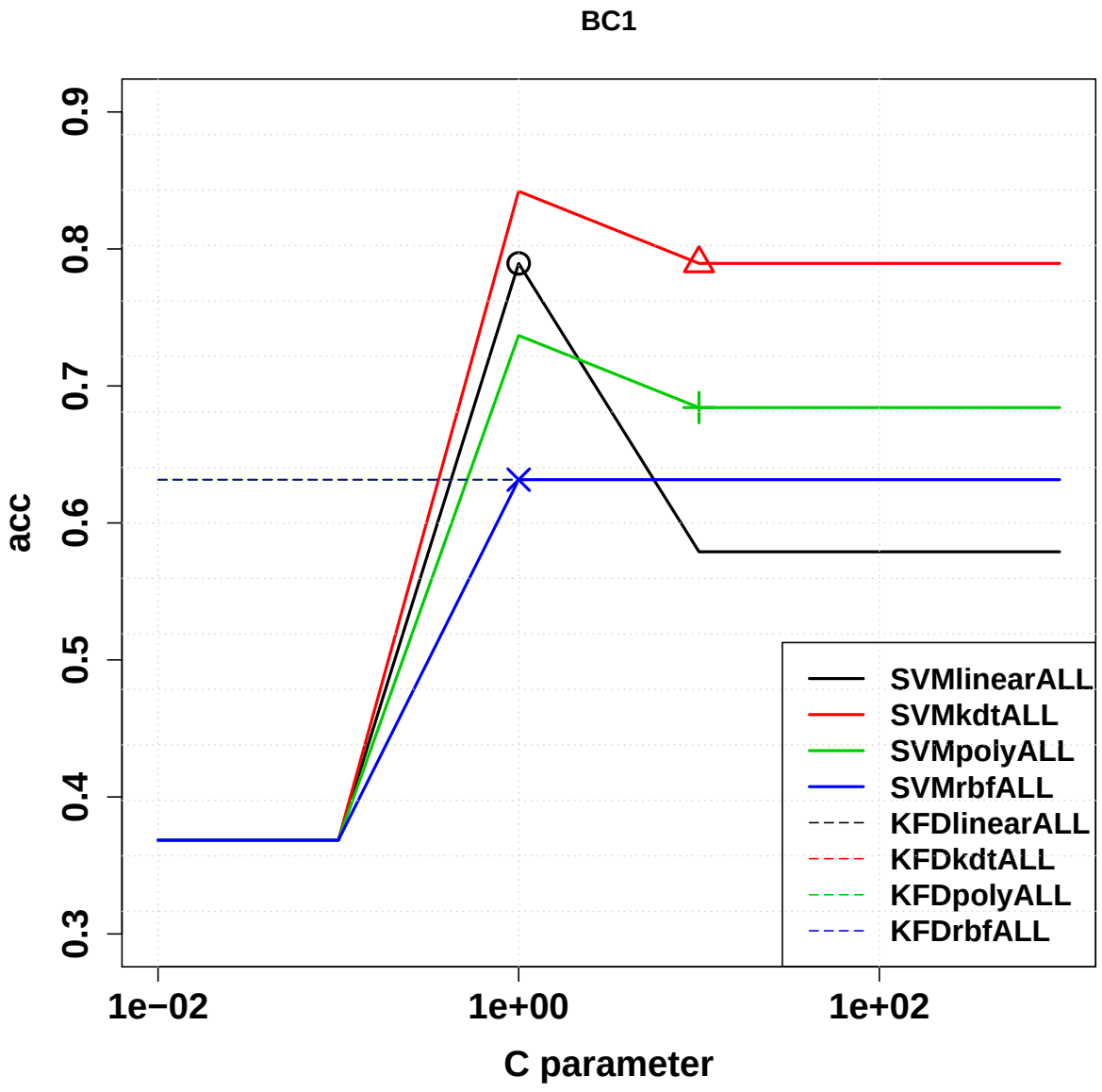
Results



Kendall kernel SVM

- **Competitive accuracy!**
- Less sensitive to regularization parameter!
- No need for feature selection!

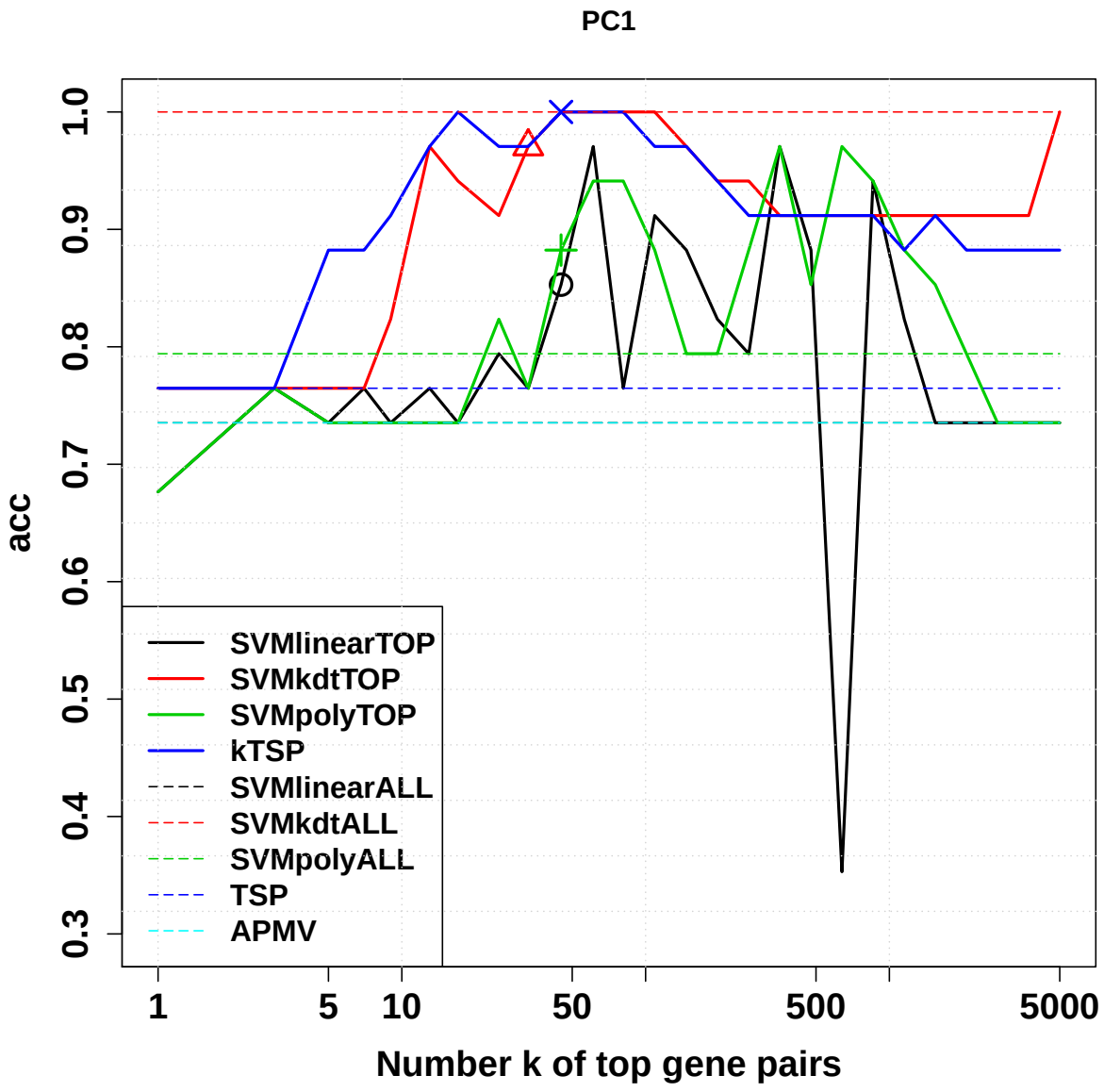
Results



Kendall kernel SVM

- Competitive accuracy!
- **Less sensitive to regularization parameter!**
- No need for feature selection!

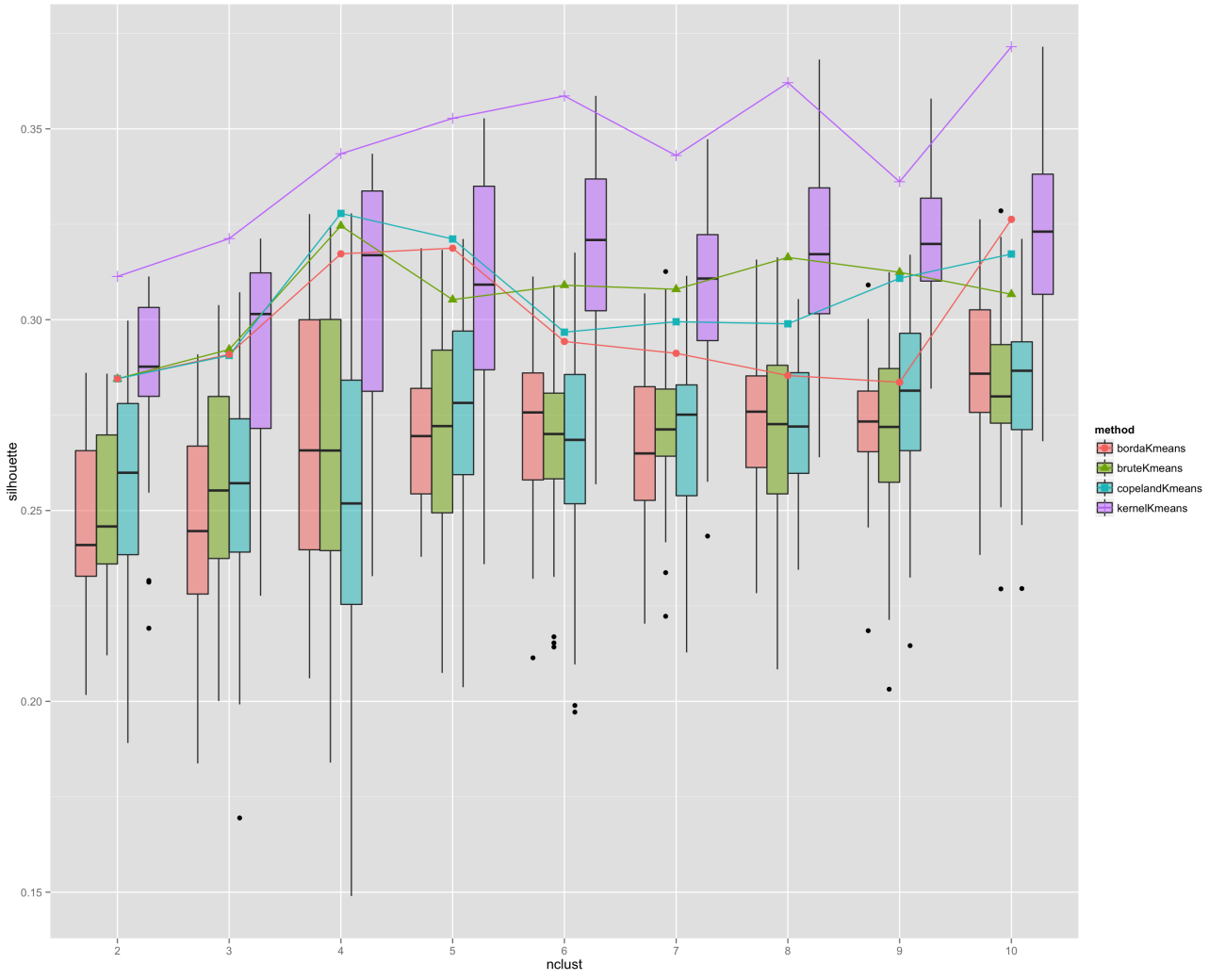
Results



Kendall kernel SVM

- Competitive accuracy!
- Less sensitive to regularization parameter!
- **No need for feature selection!**

Application: clustering



- APA data (full rankings)
- $n = 5738$, $p = 5$
- (new) Kernel k-means vs (standard) k-means in $\mathbb{S}_5$
- Show silhouette as a function of number of clusters (higher better)

Extension to partial rankings

- Two interesting types of partial rankings are interleaving partial ranking

$$x_{i_1} \succ x_{i_2} \succ \cdots \succ x_{i_k}, \quad k \leq n.$$

  and top-k partial ranking

$$x_{i_1} \succ x_{i_2} \succ \cdots \succ x_{i_k} \succ X_{\text{rest}}, \quad k \leq n.$$

- Partial rankings can be uniquely represented by a set of permutations compatible with all the observed partial orders.

Theorem

For these two particular types of partial rankings, the convolution kernel (Haussler, 1999) induced by Kendall kernel

$$K_\tau^\star(R, R') = \frac{1}{|R||R'|} \sum_{\sigma \in R} \sum_{\sigma' \in R'} K_\tau(\sigma, \sigma')$$

can be evaluated in $O(k \log k)$ time.

Extension to partial rankings

- Two interesting types of partial rankings are interleaving partial ranking

$$x_{i_1} \succ x_{i_2} \succ \cdots \succ x_{i_k}, \quad k \leq n.$$

  and top-k partial ranking

$$x_{i_1} \succ x_{i_2} \succ \cdots \succ x_{i_k} \succ X_{\text{rest}}, \quad k \leq n.$$

- Partial rankings can be uniquely represented by a set of permutations compatible with all the observed partial orders.

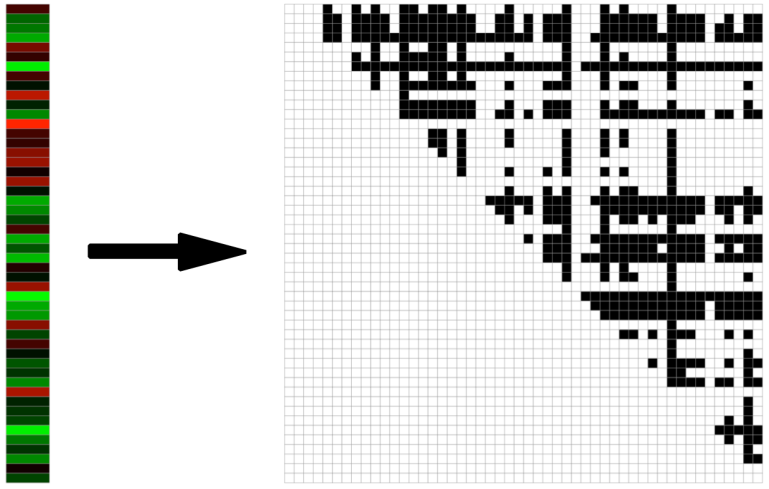
Theorem

For these two particular types of partial rankings, the convolution kernel (Haussler, 1999) induced by Kendall kernel

$$K_\tau^\star(R, R') = \frac{1}{|R||R'|} \sum_{\sigma \in R} \sum_{\sigma' \in R'} K_\tau(\sigma, \sigma')$$

can be evaluated in $O(k \log k)$ time.

Extension to smoother, continuous representations



One sample x
p features

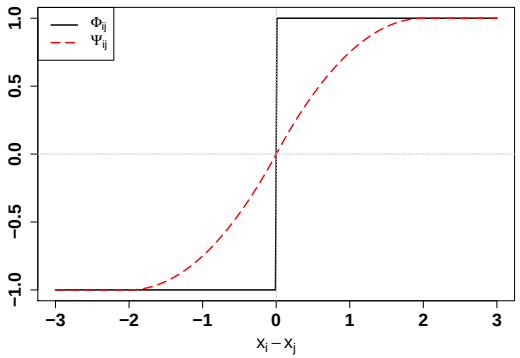
Mapping f(x)
p(p-1)/2 bits

- Instead of $\Phi : \mathbb{R}^p \to \{0,1\}^{p(p-1)/2}$, consider the continuous mapping $\Psi_a : \mathbb{R}^p \to \mathbb{R}^{p(p-1)/2}$:

$$\Psi_a(x) = \mathbb{E}\Phi(x+\epsilon) \quad \text{with} \quad \epsilon \sim (\mathcal{U}[-\frac{a}{2}, \frac{a}{2}])^n$$

- Corresponding kernel $G_a(x, x') = \Psi_a(x)^\top \Psi_a(x')$

Computation of $G(x, x')$



- $G_a(x, x')$ can be computed exactly in $O(p^2)$ by explicit computation of $\Psi_a(x)$ in $\mathbb{R}^{p(p-1)/2}$

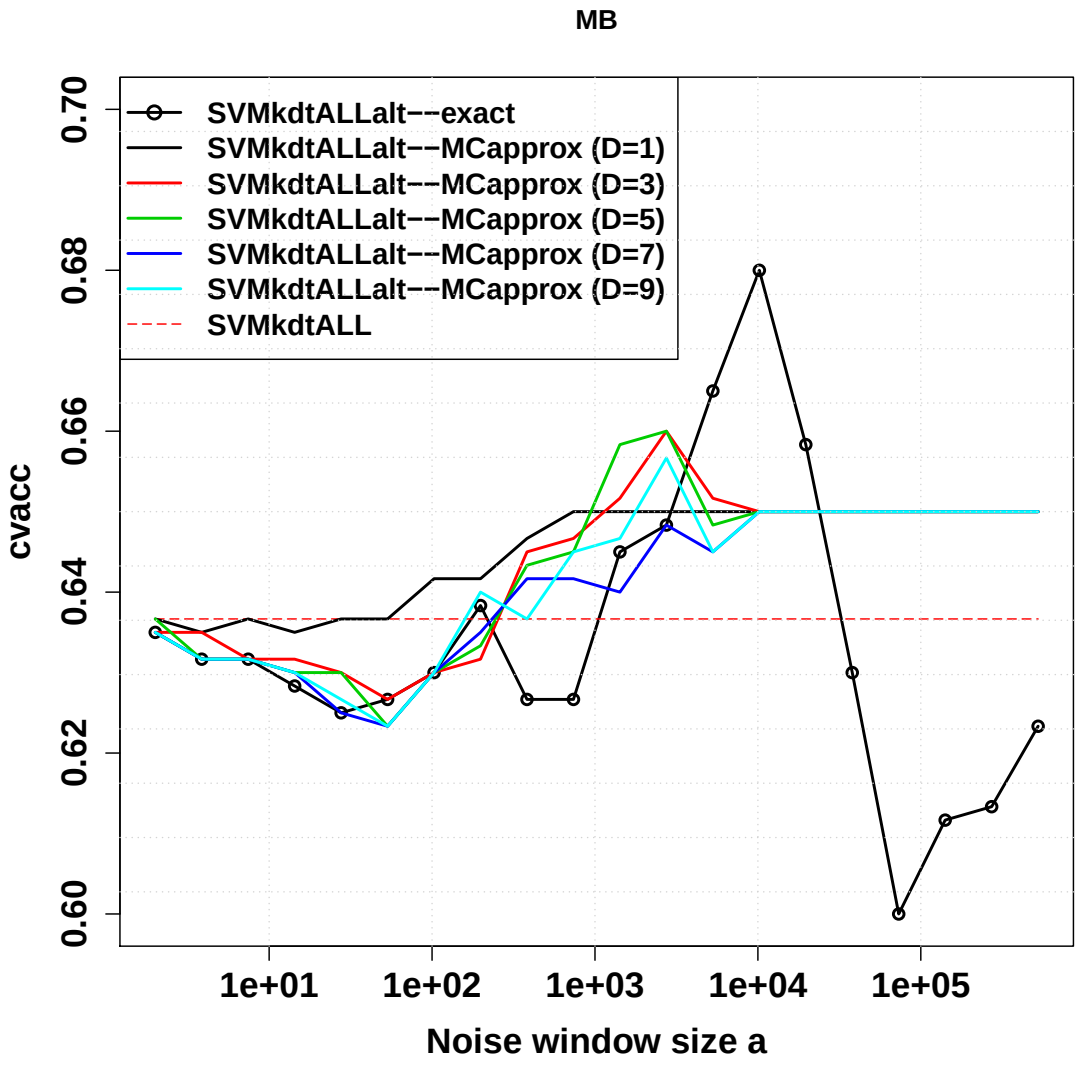
- $G_a(x, x')$ can be computed approximately in $O(D^2 p \log p)$ by Monte-Carlo approximation:

$$\tilde{G}_a(x, x') = \frac{1}{D^2} \sum_{i,j=1}^{D} K(x + \epsilon_i, x' + \epsilon'_j)$$

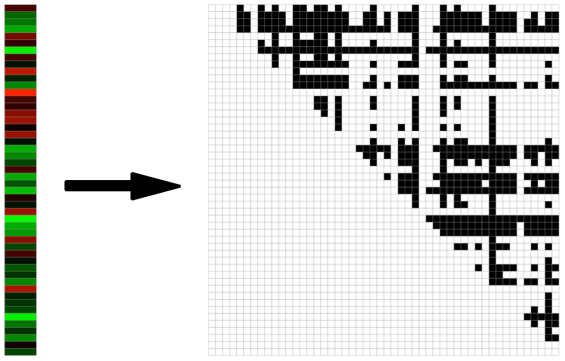
- Theorem: for supervised learning, Monte-Carlo approximation is better[1] than exact computation when $n = o(p^{1/3})$

[1] faster for the same accuracy

Performance of $G_a(x, x)$



Summary



- A representation adapted to data with monotonic noise
- Equivalent to learning over the symmetric group of permutations
- Kernel trick allows to work with large p / small n
- Available as an R package

```
> install.packages("devtools")
> devtools::install_github("YunlongJiao/kernrank")
```

- More details in Jiao and Vert (2015)

Outline

1. Learning with regularization and prior knowledge
2. Cancer patient stratification from somatic mutations
3. Learning from rankings through pairwise comparisons
4. FlipFlop: fast isoform prediction from RNA-seq data
5. Conclusion

Joint work with...

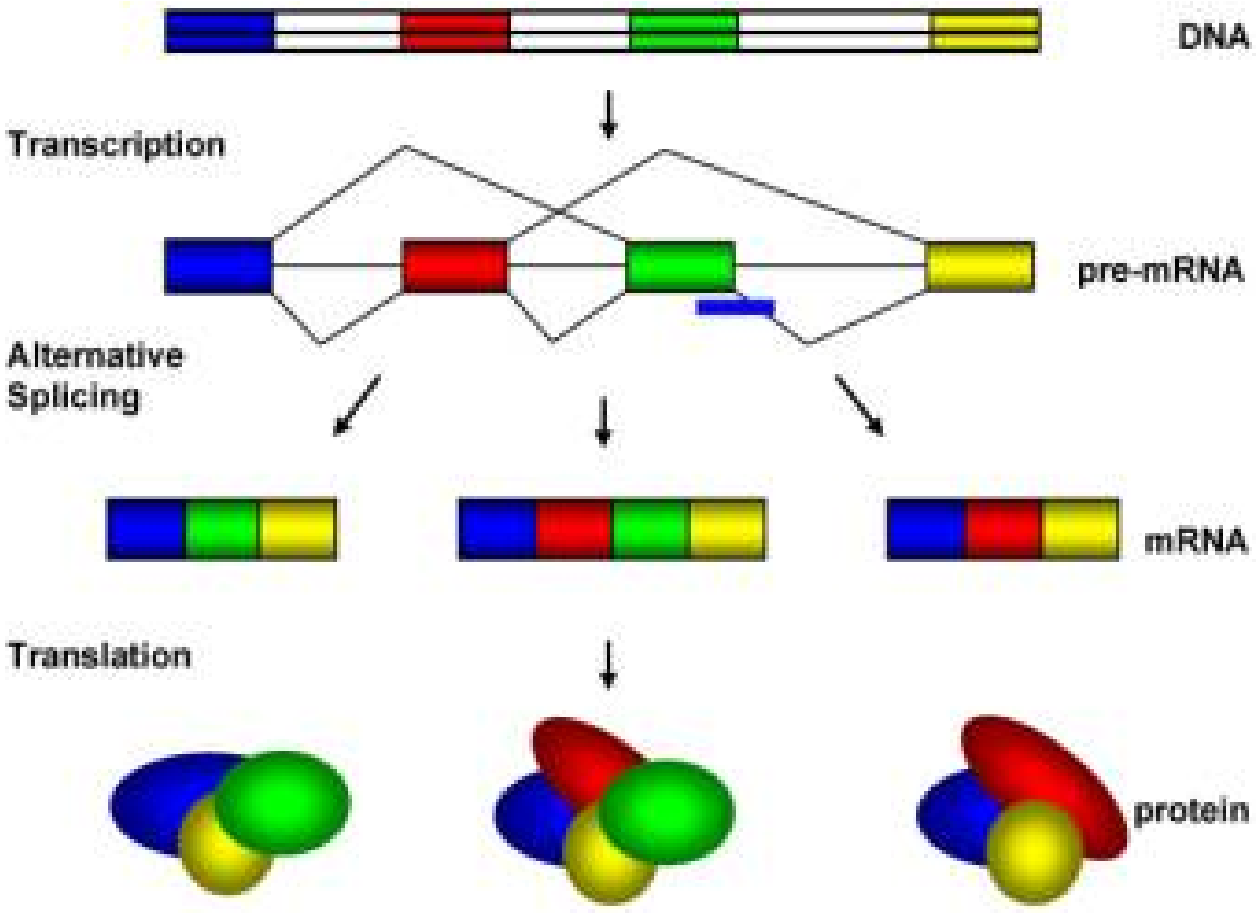
Elsa Bernard

Laurent Jacob

Julien Mairal

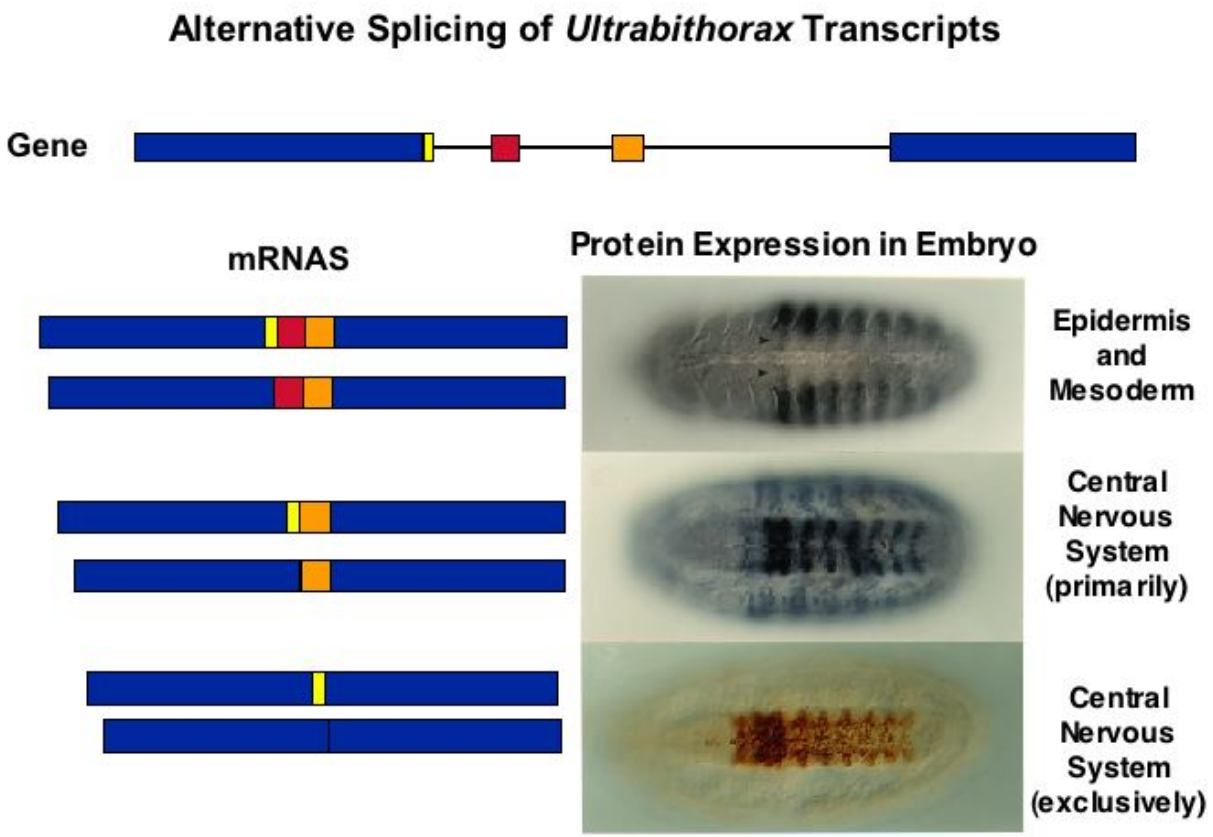
Eric Viara

Alternative splicing: 1 gene = many proteins



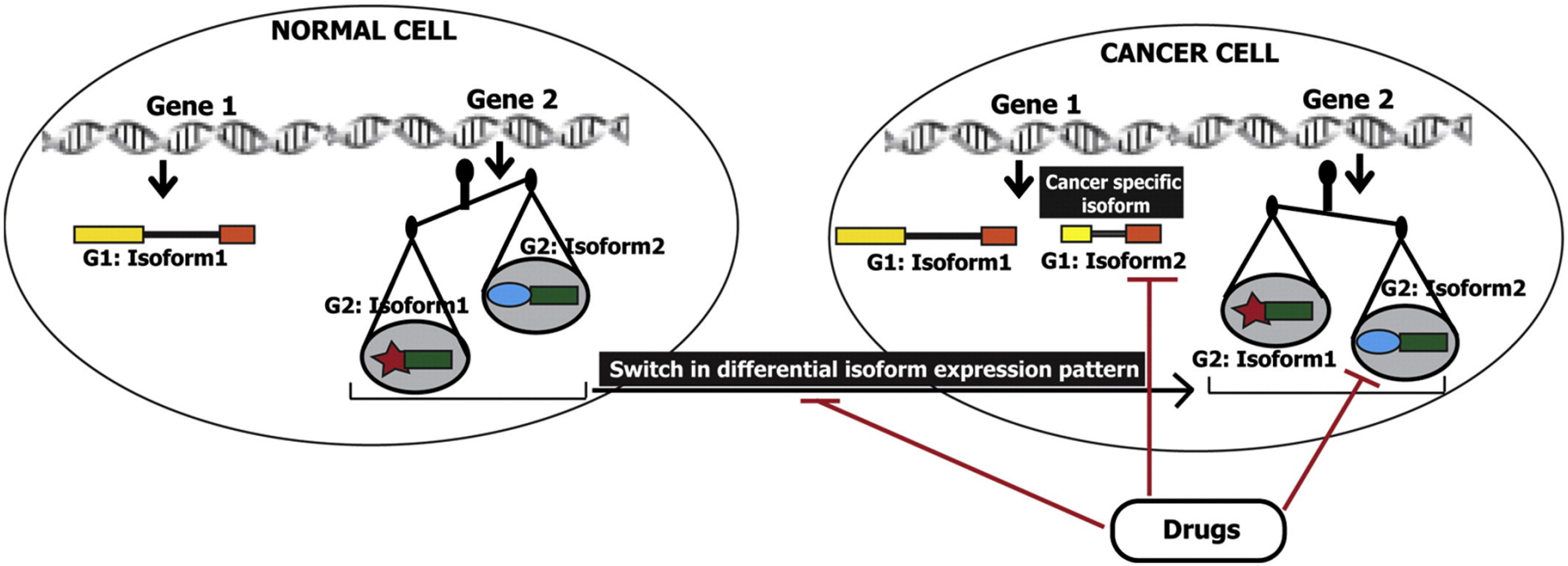
In human, 28k genes give 120k known transcripts (Pal et al., 2012))

Alternative splicing matters: developmental regulation in Drosophila



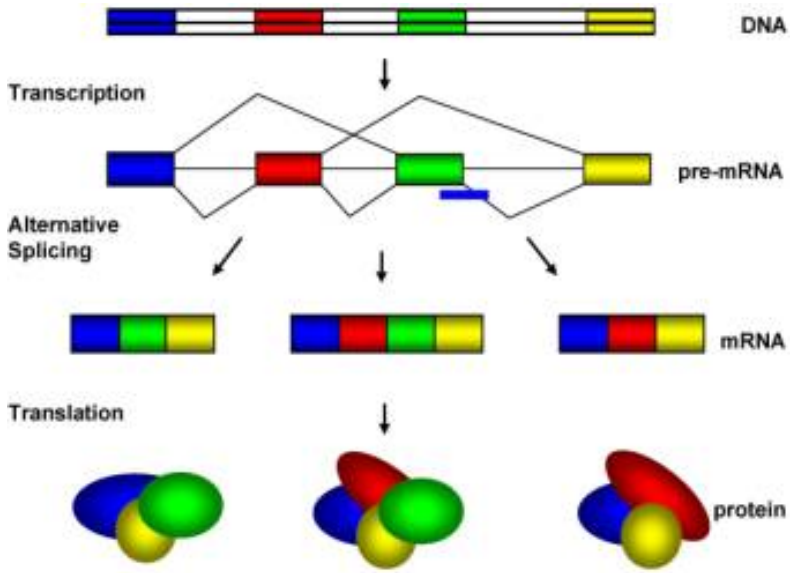
http://orchid.bio.cmu.edu/research.html

Alternative splicing matters: drug targets



(Pal et al., 2012)

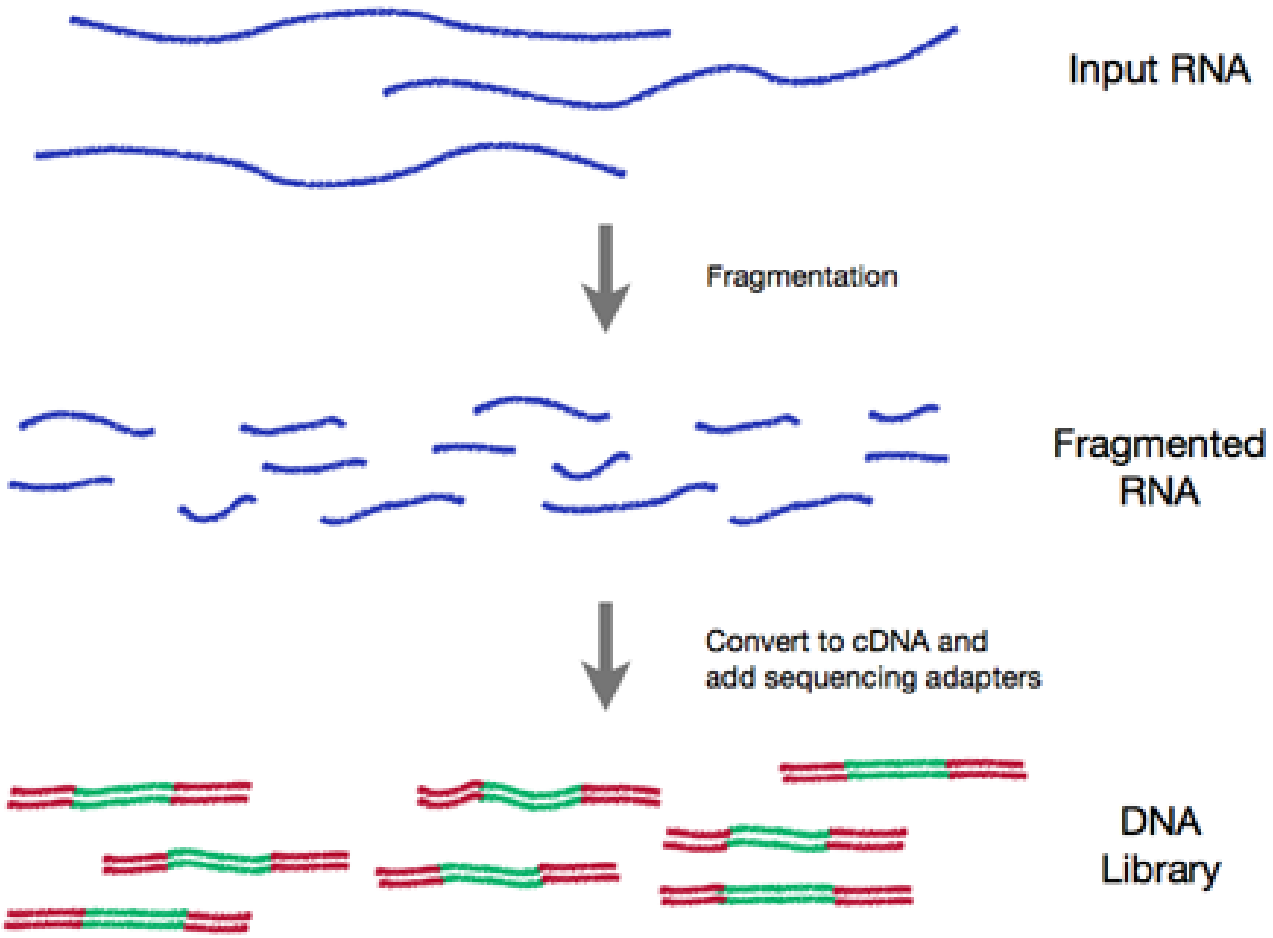
The isoform identification and quantification problem



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

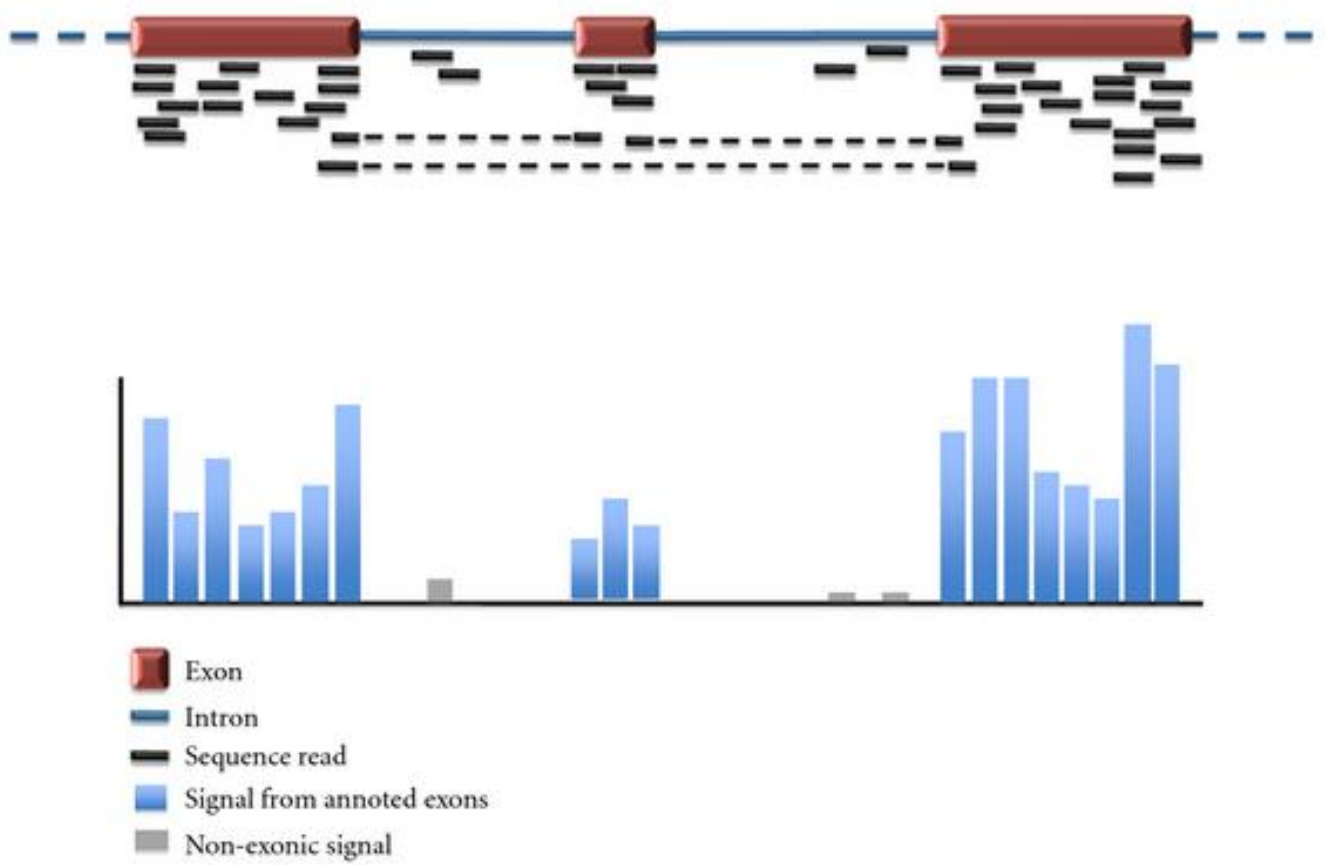
1. identify the isoform(s) of each gene present in the sample?
2. quantify their abundance?

RNA-seq measures mRNA abundance by sequencing short fragments



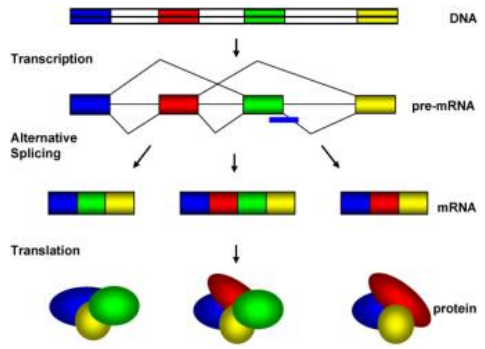
http://rnaseq.uoregon.edu

RNA-seq and alternative splicing



(Costa et al., 2011)

Lasso-based estimation of isoforms

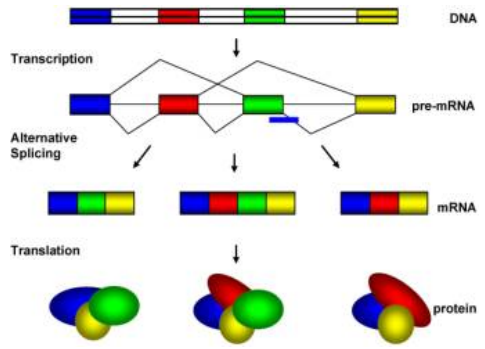


- Let a gene with e exons
- Suppose there are c candidate isoform (c large, up to 2^e)
- Let $\phi \in \mathbb{R}^c$ the unknown c-dimensional vector of abundance
- Let $L(\phi)$ quantify whether ϕ explains well the observed read counts (e.g., minus log-likelihood)
- Find a sparse vector of abundances 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 ϕ 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 et al., 2013)

The isoform deconvolution problem

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

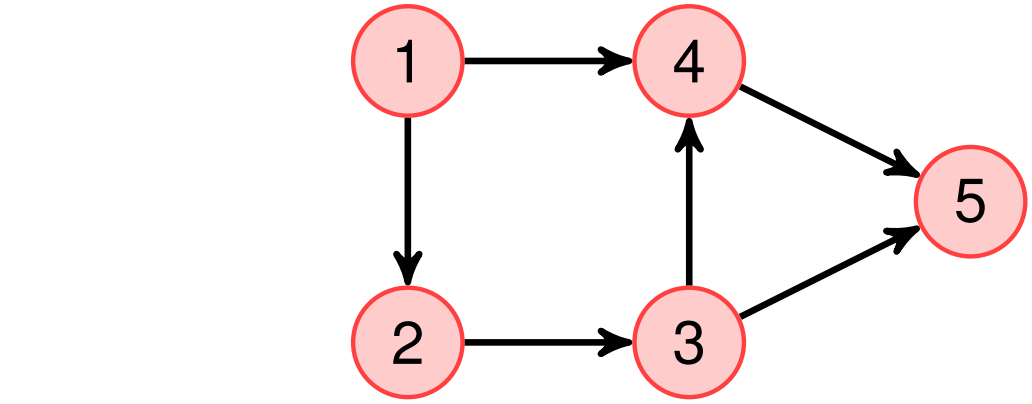
can be solved in polynomial time in the number of exon.

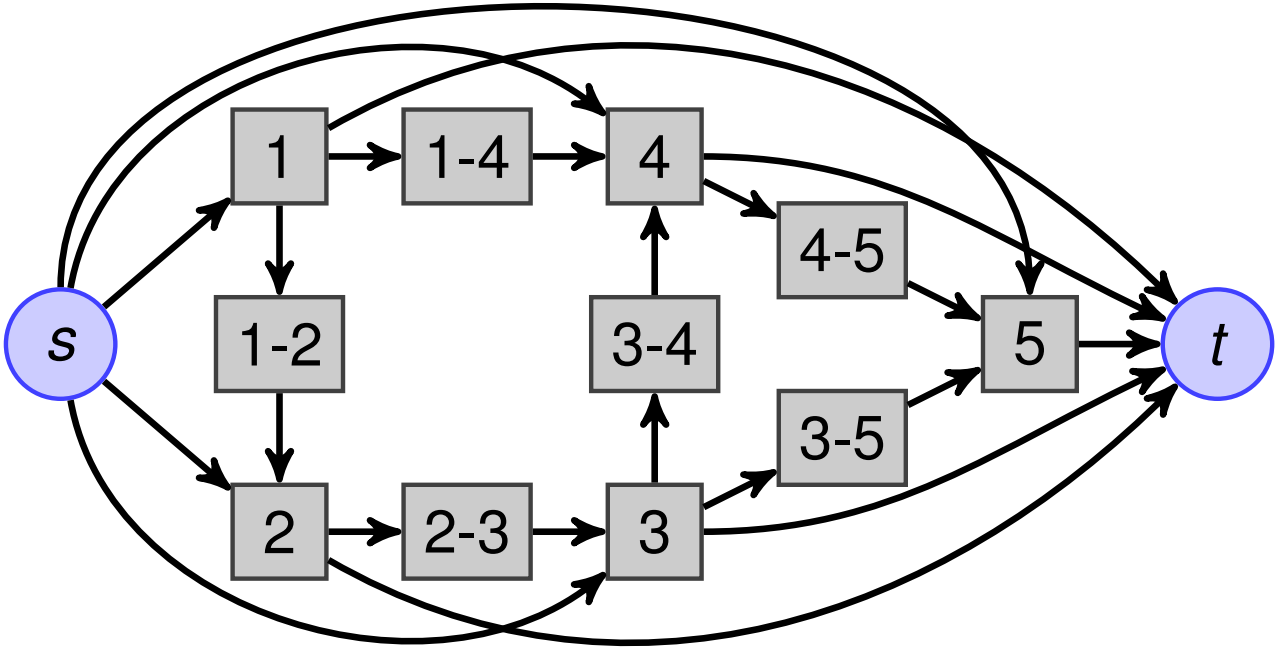
Key ideas

1. Reformulation as a convex cost flow problem (Mairal and Yu, 2013)
2. Recover isoforms by flow decomposition algorithm

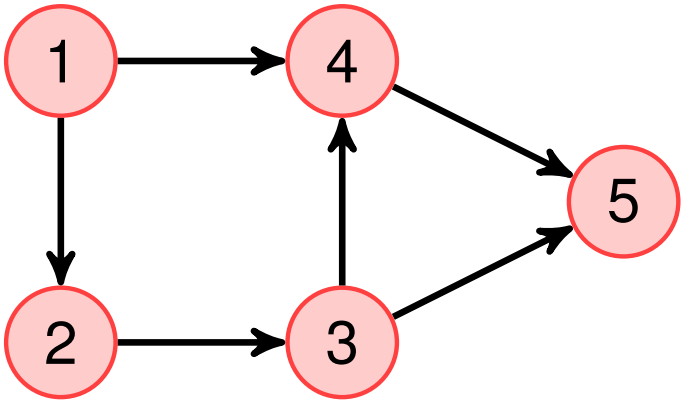
"Feature selection on an exponential number of features in polynomial time"

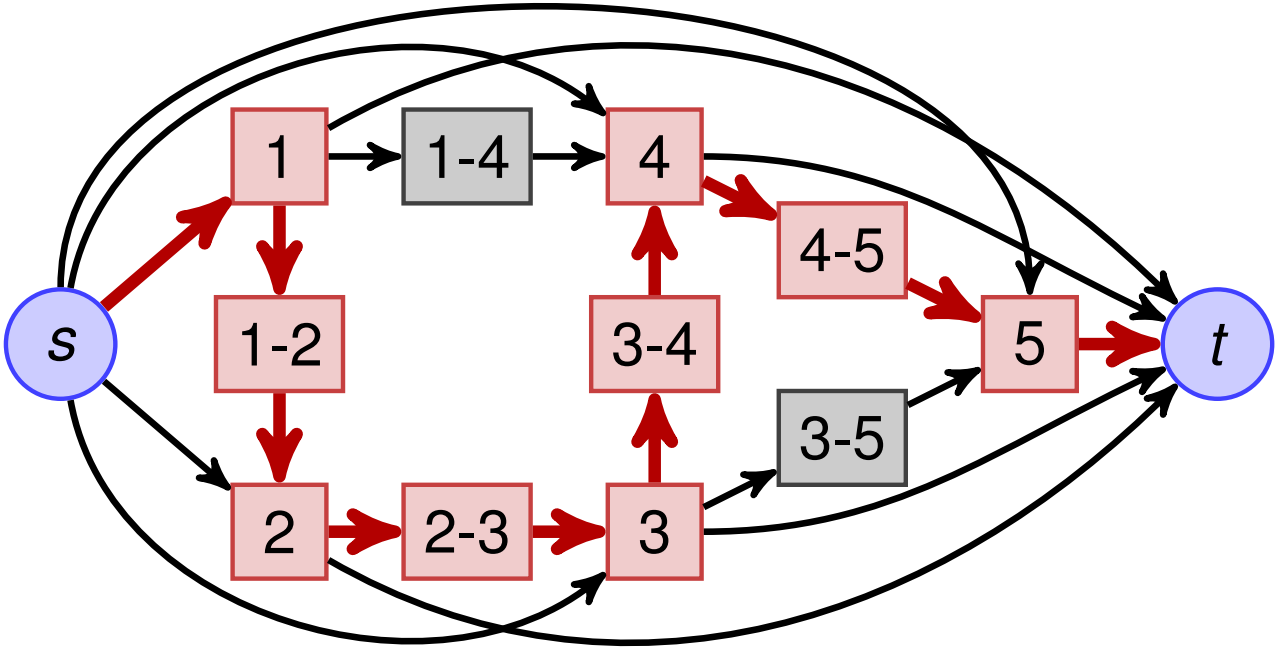
Isoforms are Paths in a Graph



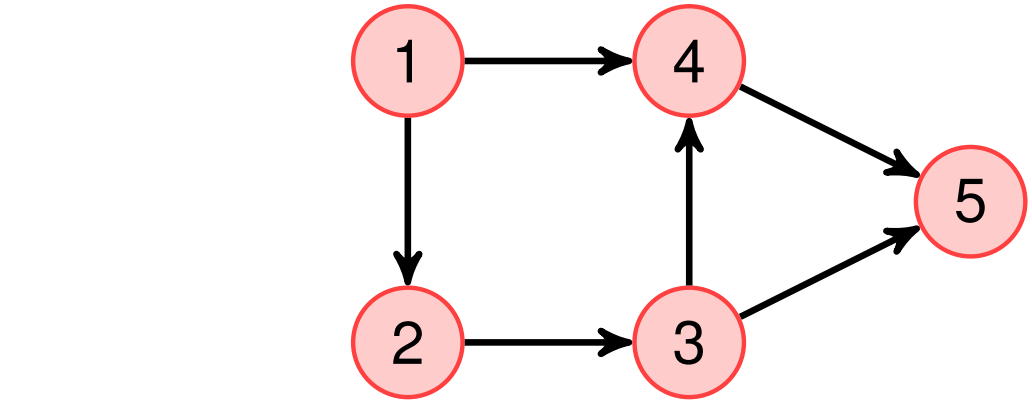


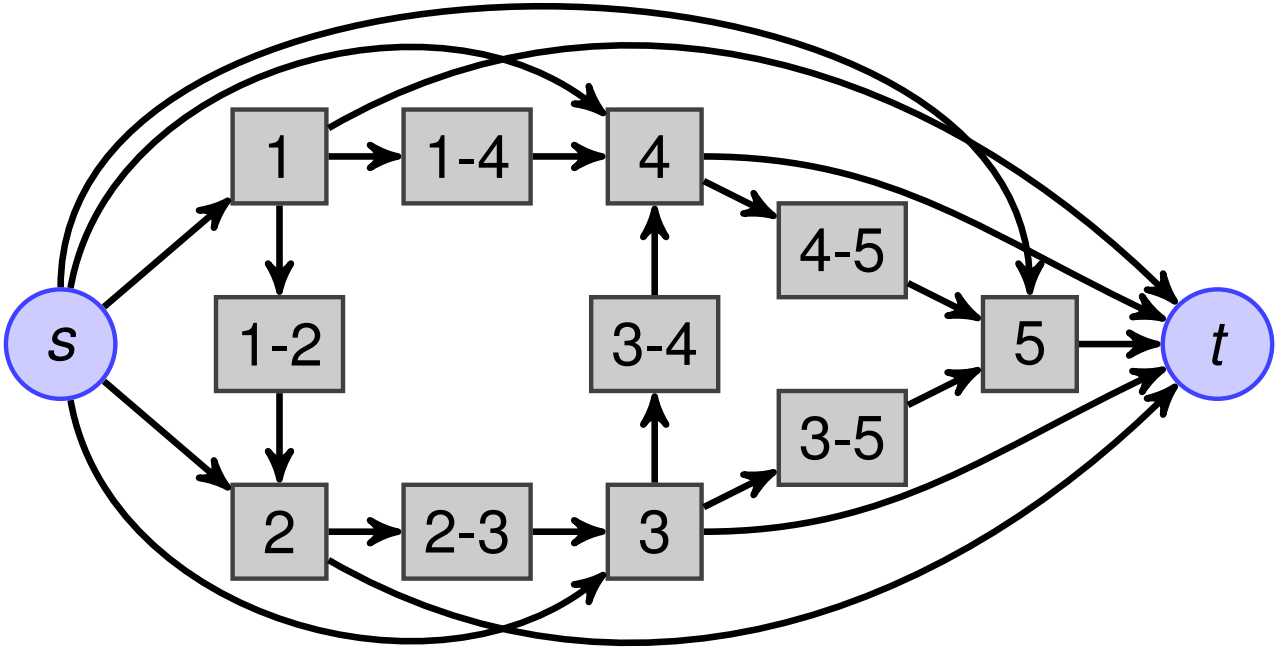
Isoforms are Paths in a Graph



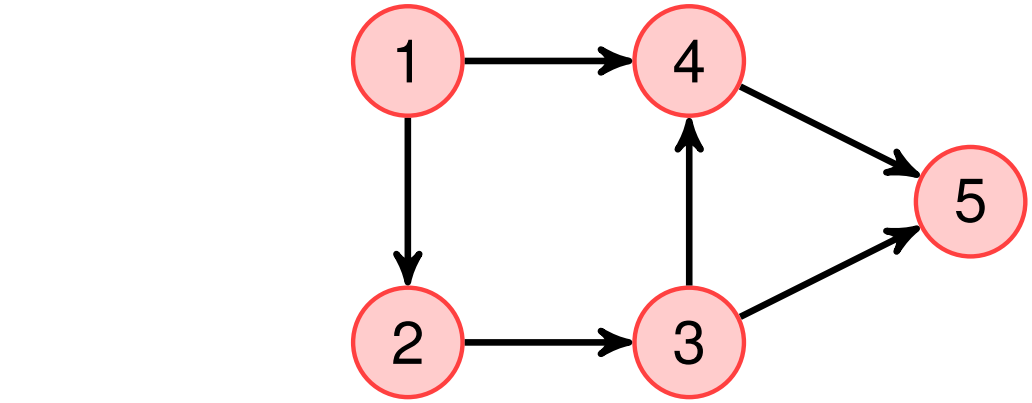


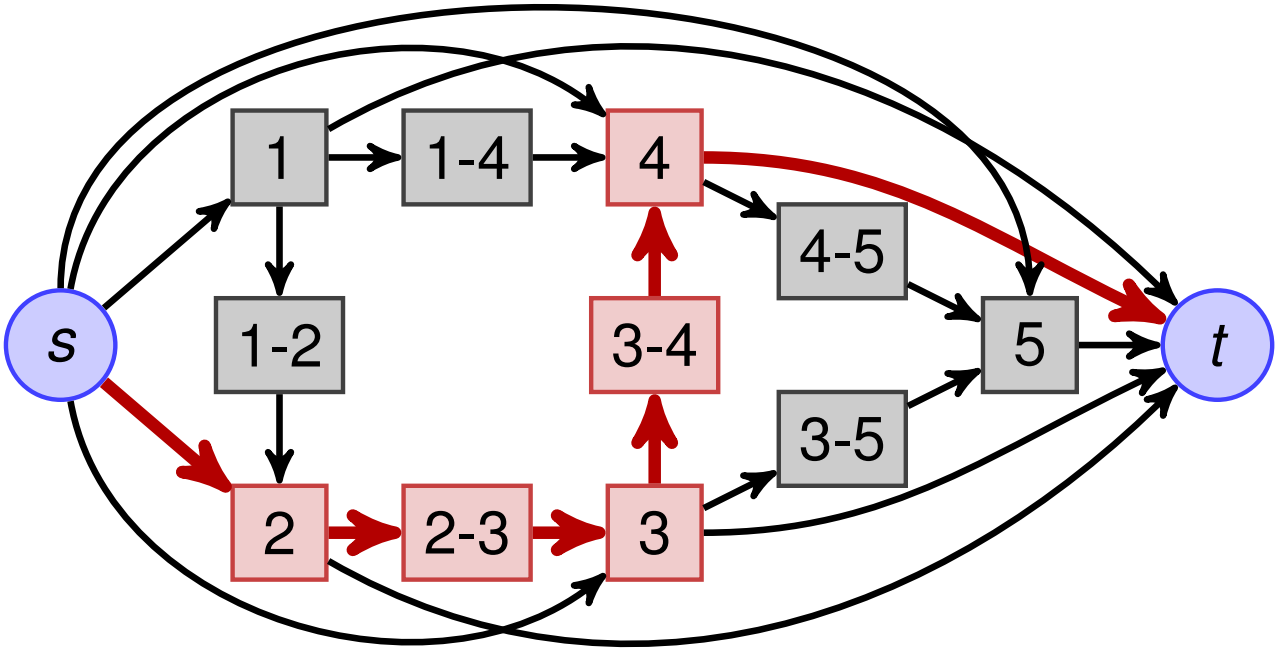
Isoforms are Paths in a Graph



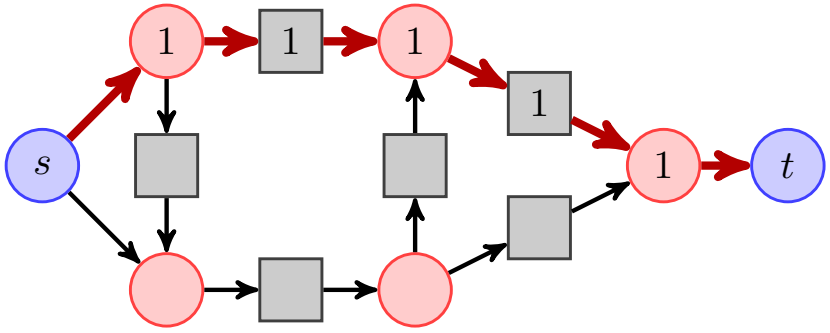


Isoforms are Paths in a Graph

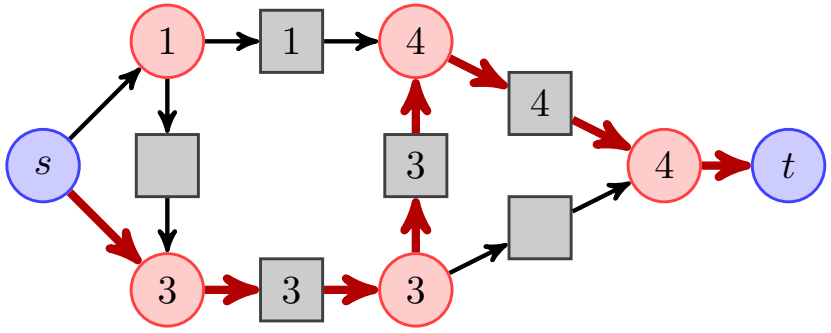




Combinations of isoforms are flows



(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,

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

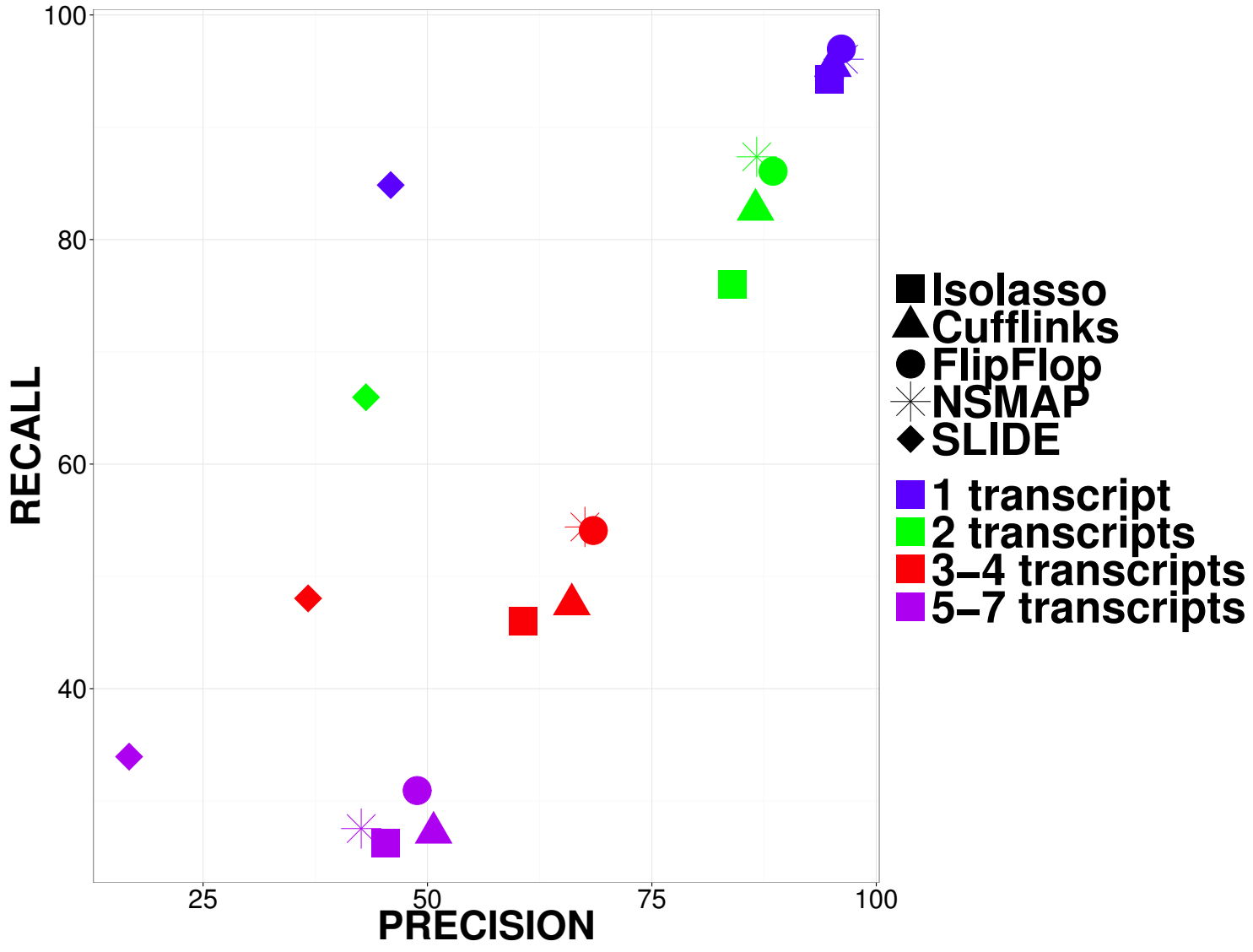
is equivalent to

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

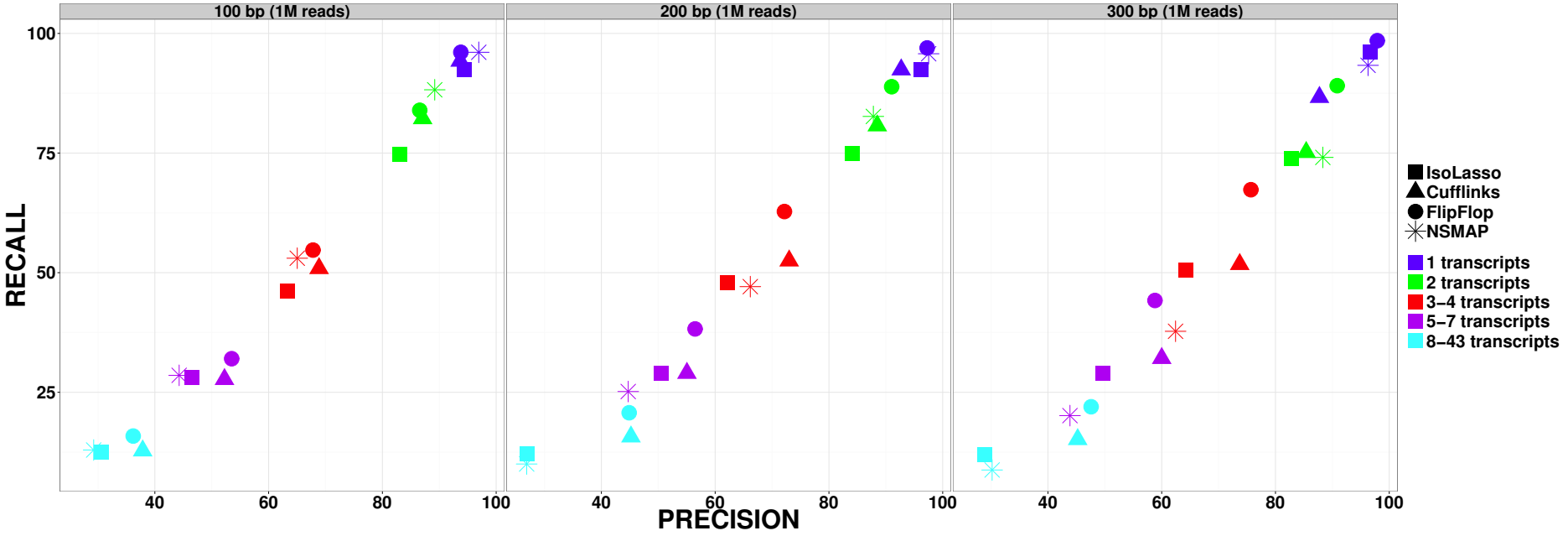
Human Simulation: Precision/Recall

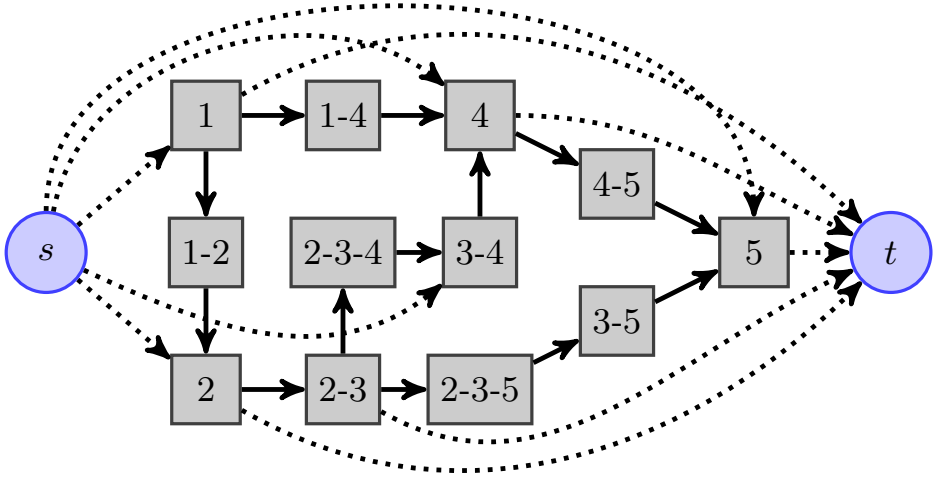
hg19, 1137 genes on chr1, 1million 75 bp single-end reads by transcript levels.

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

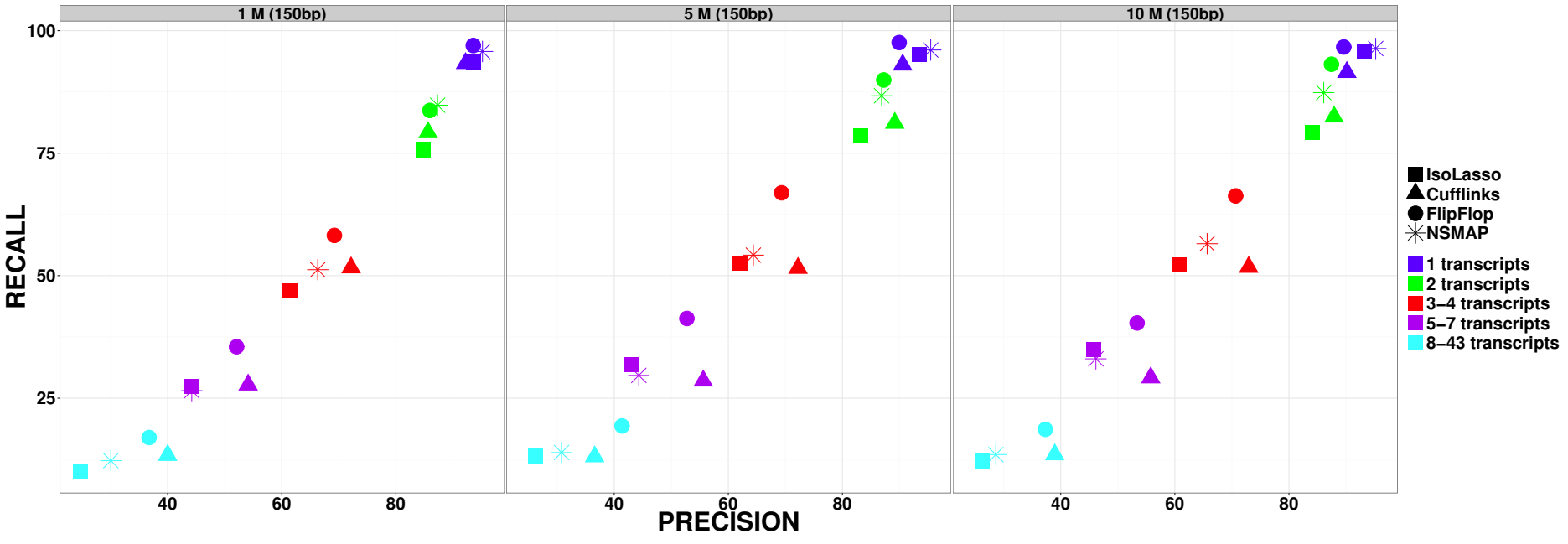


Performance increases with read length

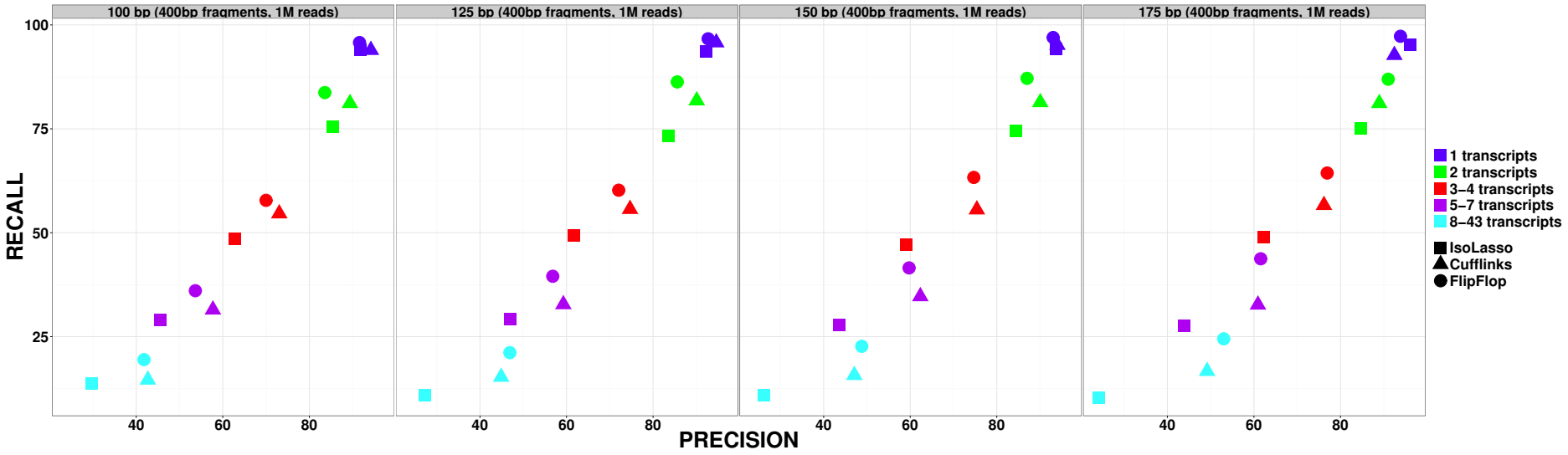




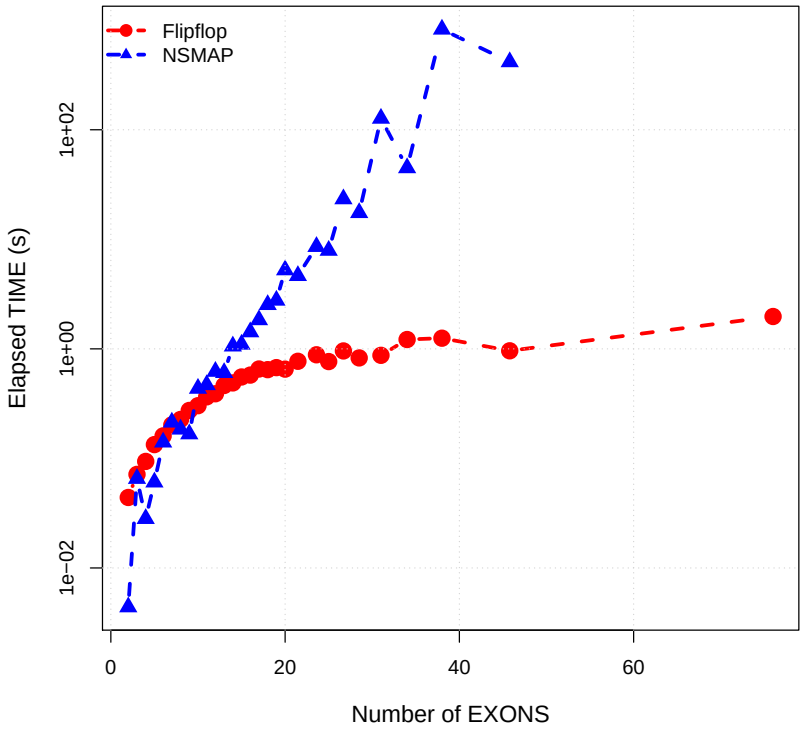
Performance increases with coverage

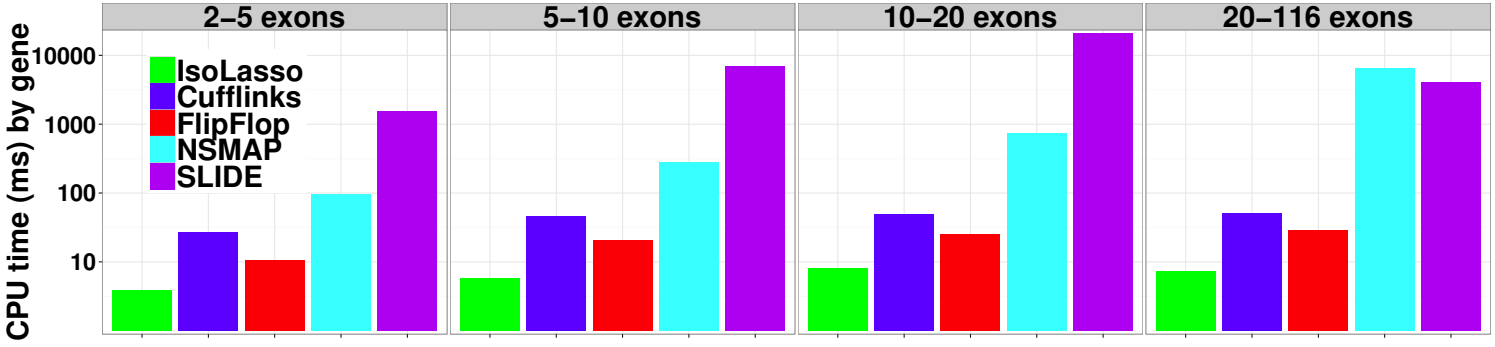


Extension to paired-end reads OK.

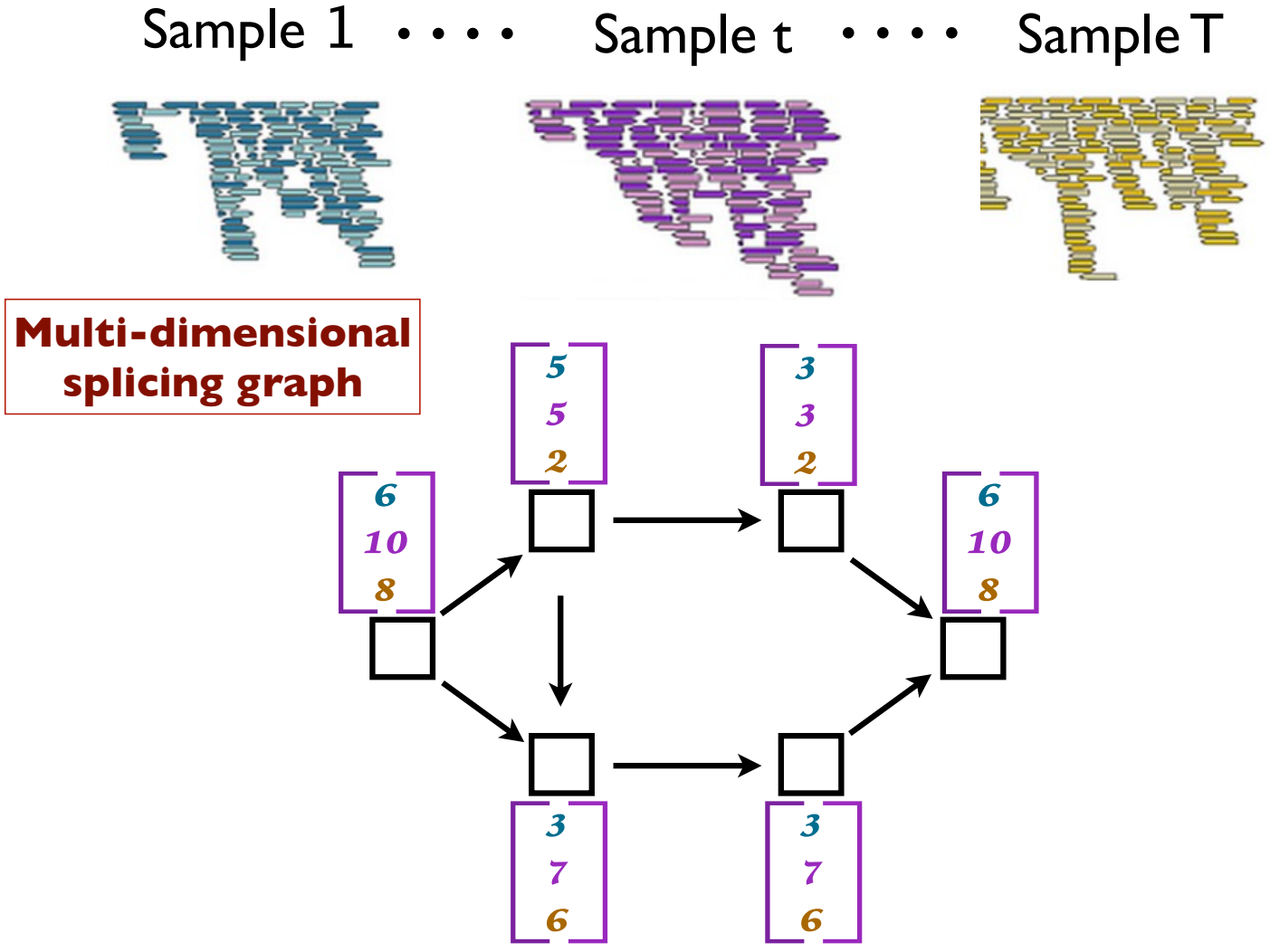


Speed trial


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

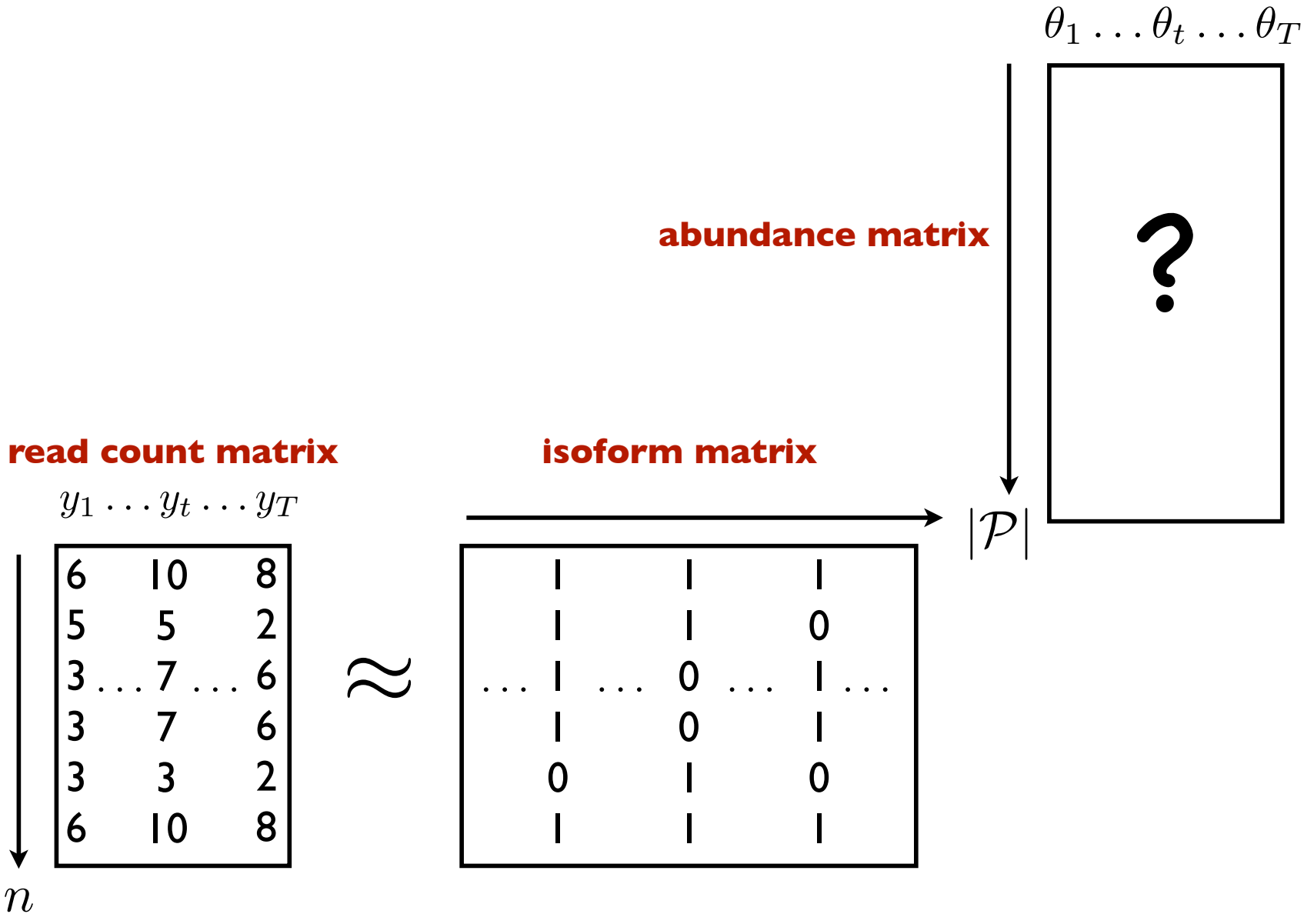

2–5 exons
5–10 exons
10–20 exons
20–116 exons
CPU time (ms) by gene
10000
1000
100
10
IsoLasso
Cufflinks
FlipFlop
NSMAP
SLIDE

Multiple samples

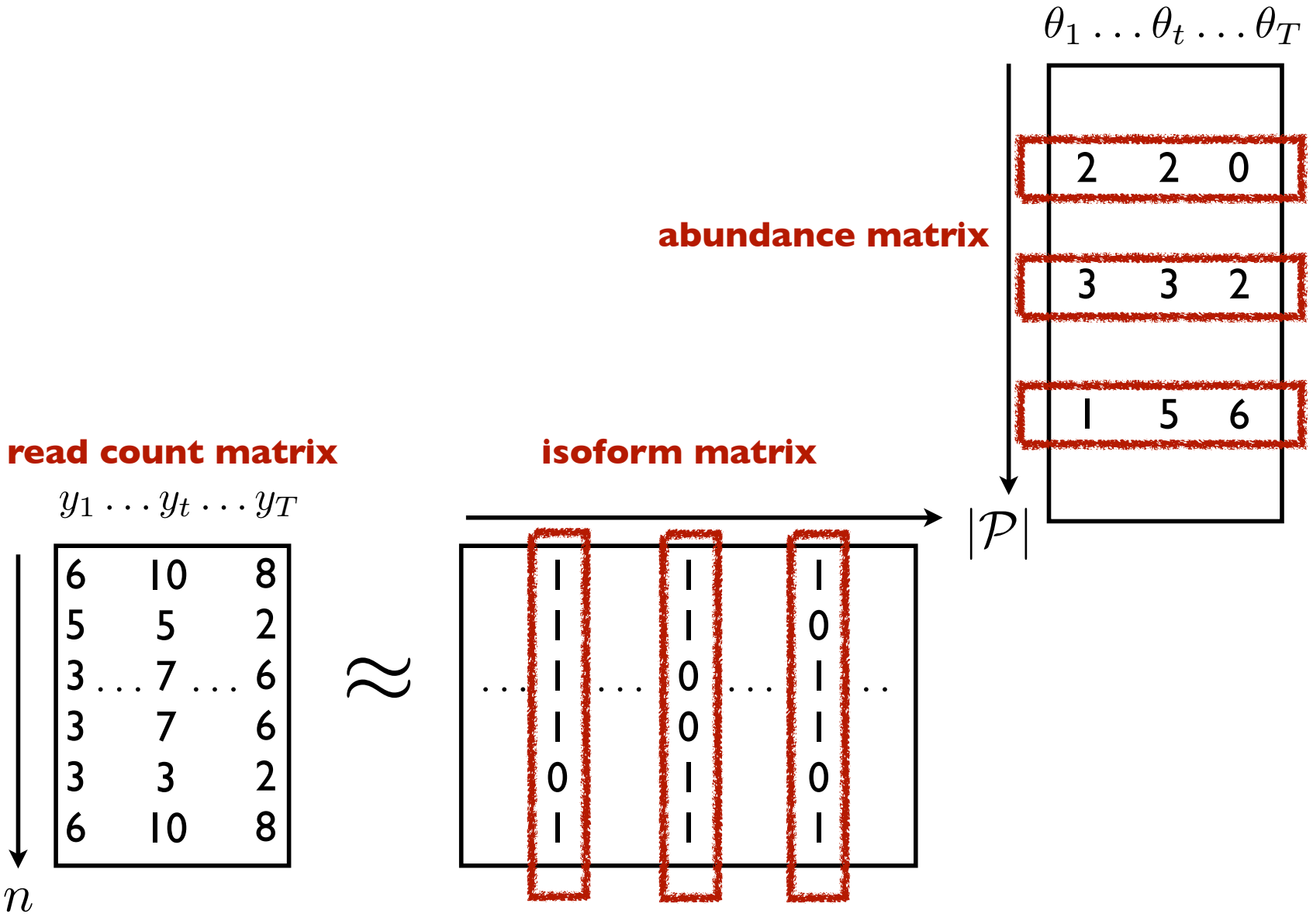


Can we find a sparse set of paths that explains the multi-dimensional read counts?

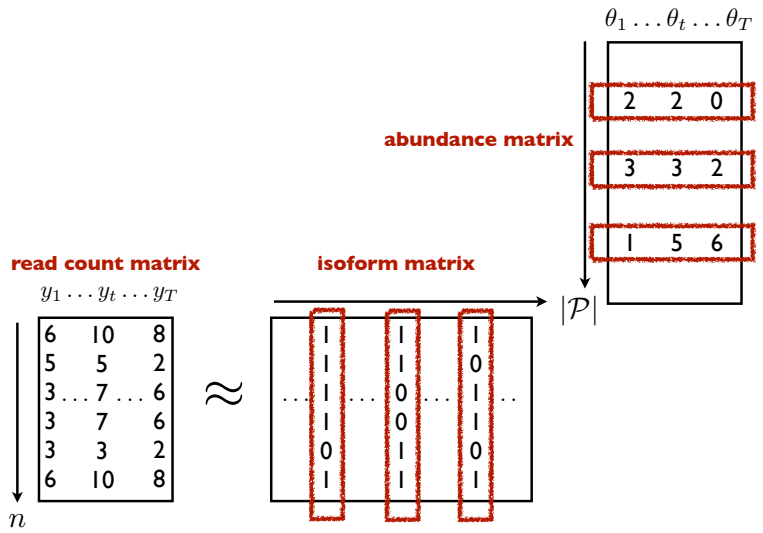
Formulation as multivariate regression problem



Formulation as multivariate regression problem



More formally



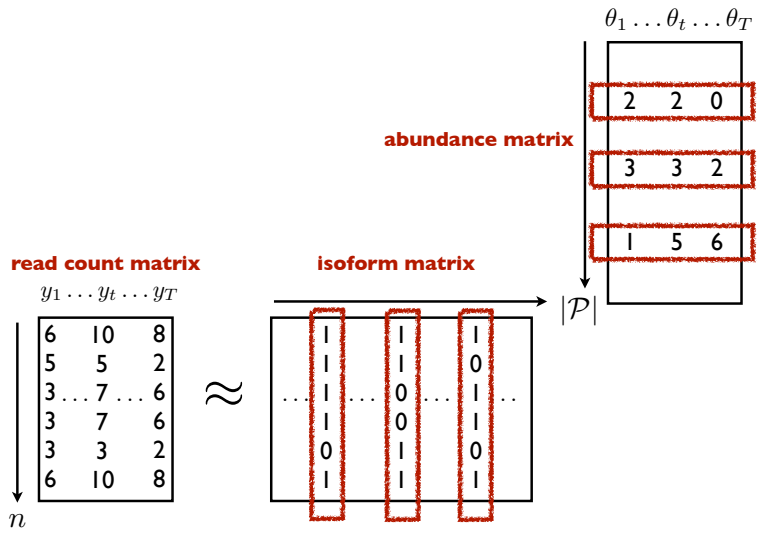
- each isoform defines a group $\boldsymbol{\theta}_p = \{\theta_p^t, t \in [\![1, T]\!]\}$
- the multi-samples loss is the sum of the independent losses

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{t=1}^{T} \text{loss}(y_t, \theta_t)$$

- Ideally we want to solve the NP-hard L0 problem

$$\min_{\{\boldsymbol{\theta}_p\}_{p \in 1,\dots,|\mathcal{P}|}} \mathcal{L}(\boldsymbol{\theta}) + \lambda \sum_{p \in \mathcal{P}} \mathbf{1}_{\{\boldsymbol{\theta}_p \neq \mathbf{0}\}}$$

More formally



- each isoform defines a group $\boldsymbol{\theta}_p = \{\theta_p^t, t \in [\![1, T]\!]\}$
- the multi-samples loss is the sum of the independent losses

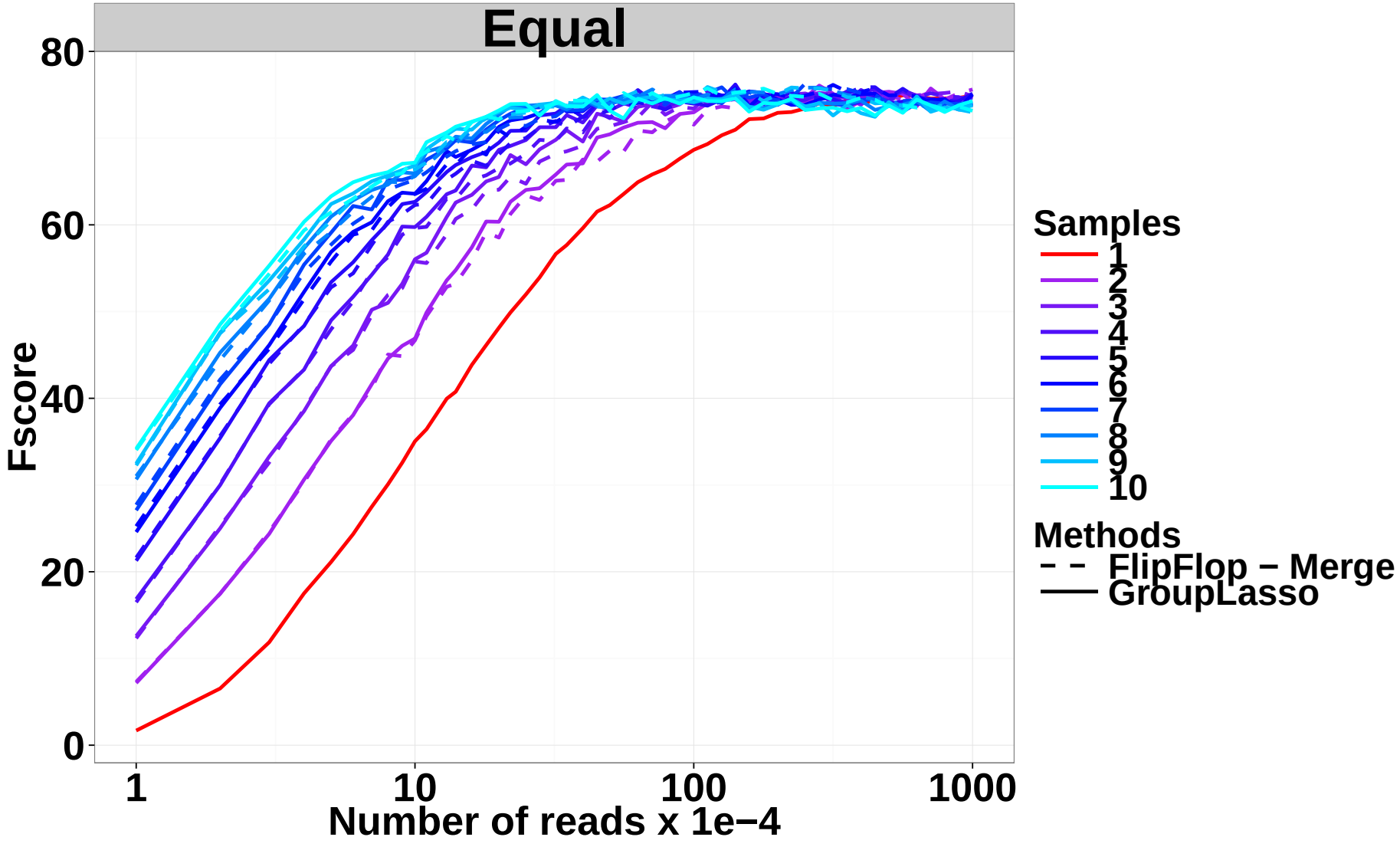
$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{t=1}^{T} \text{loss}(y_t, \theta_t)$$

- Instead we solve the group-lasso convex relaxation

$$\min_{\{\boldsymbol{\theta}_p\}_{p \in 1,\dots,|\mathcal{P}|}} \mathcal{L}(\boldsymbol{\theta}) + \lambda \sum_{p \in \mathcal{P}} \|\boldsymbol{\theta}_p\|_2$$

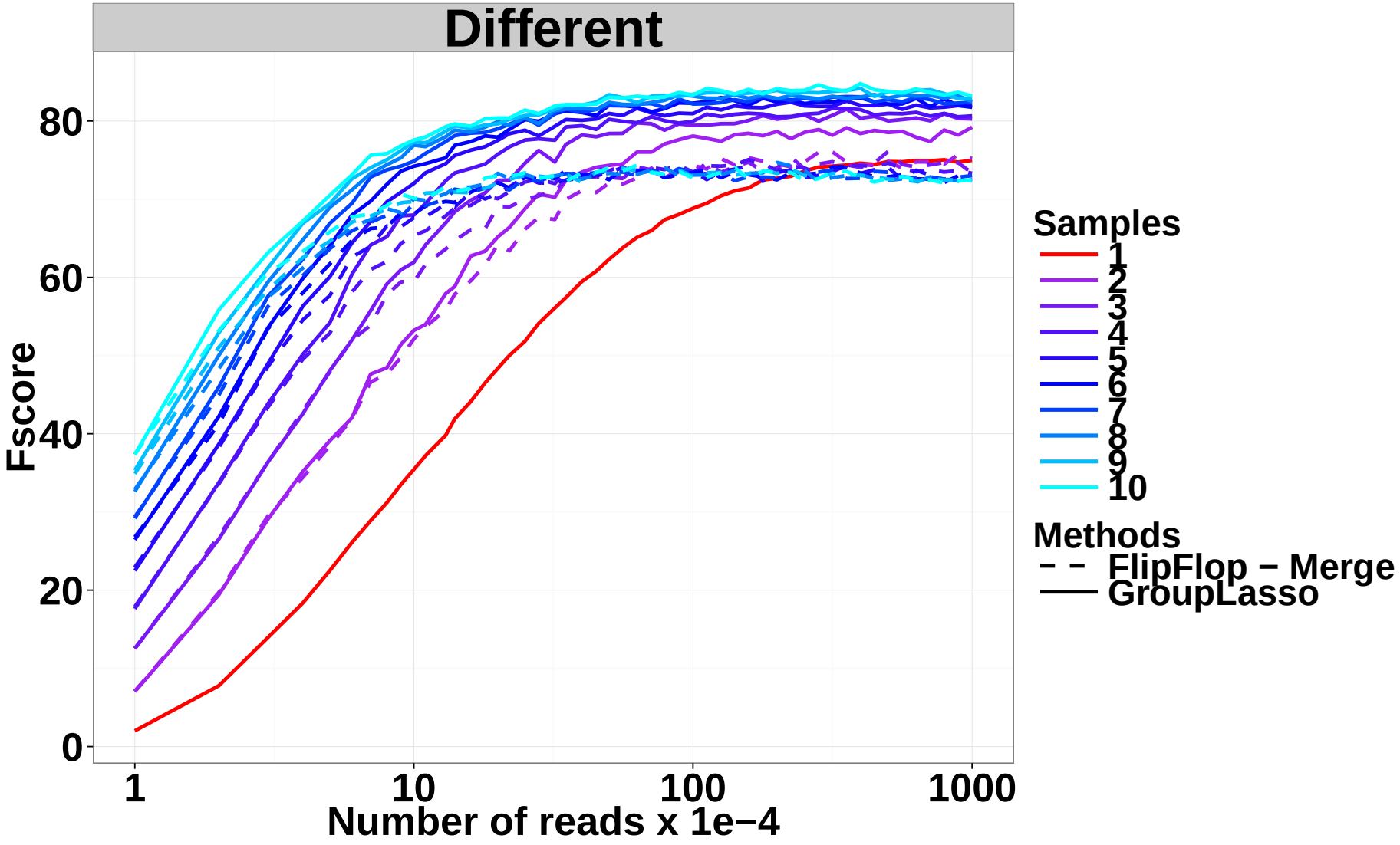
Toy simulation

$$\forall t \in \{1, \ldots, T\}, \theta_t = \theta_o + \epsilon$$



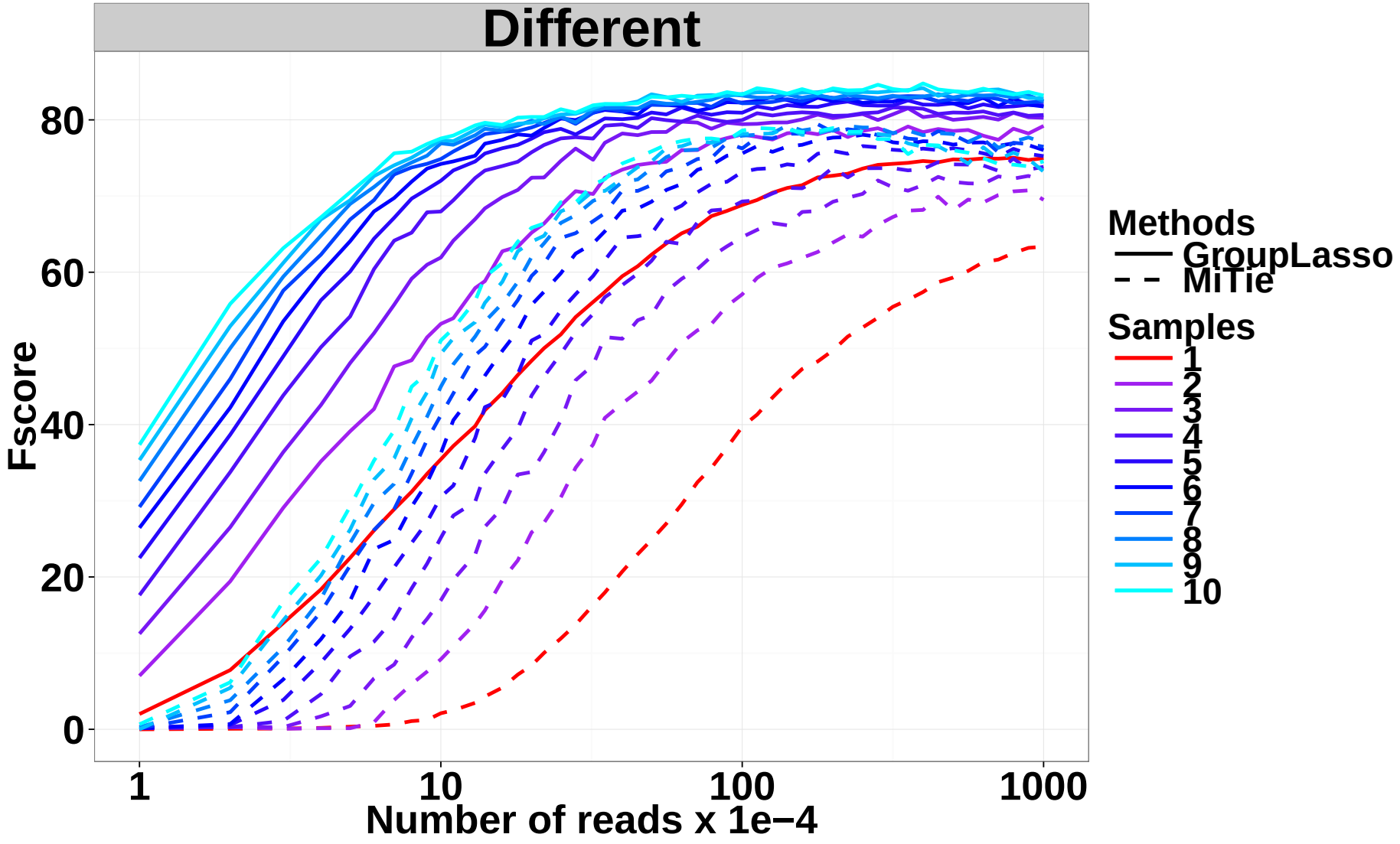
More realistic simulation

$$\forall t \in \{1, \ldots, T\}, \text{supp}\theta_t = \text{supp}\theta_o$$

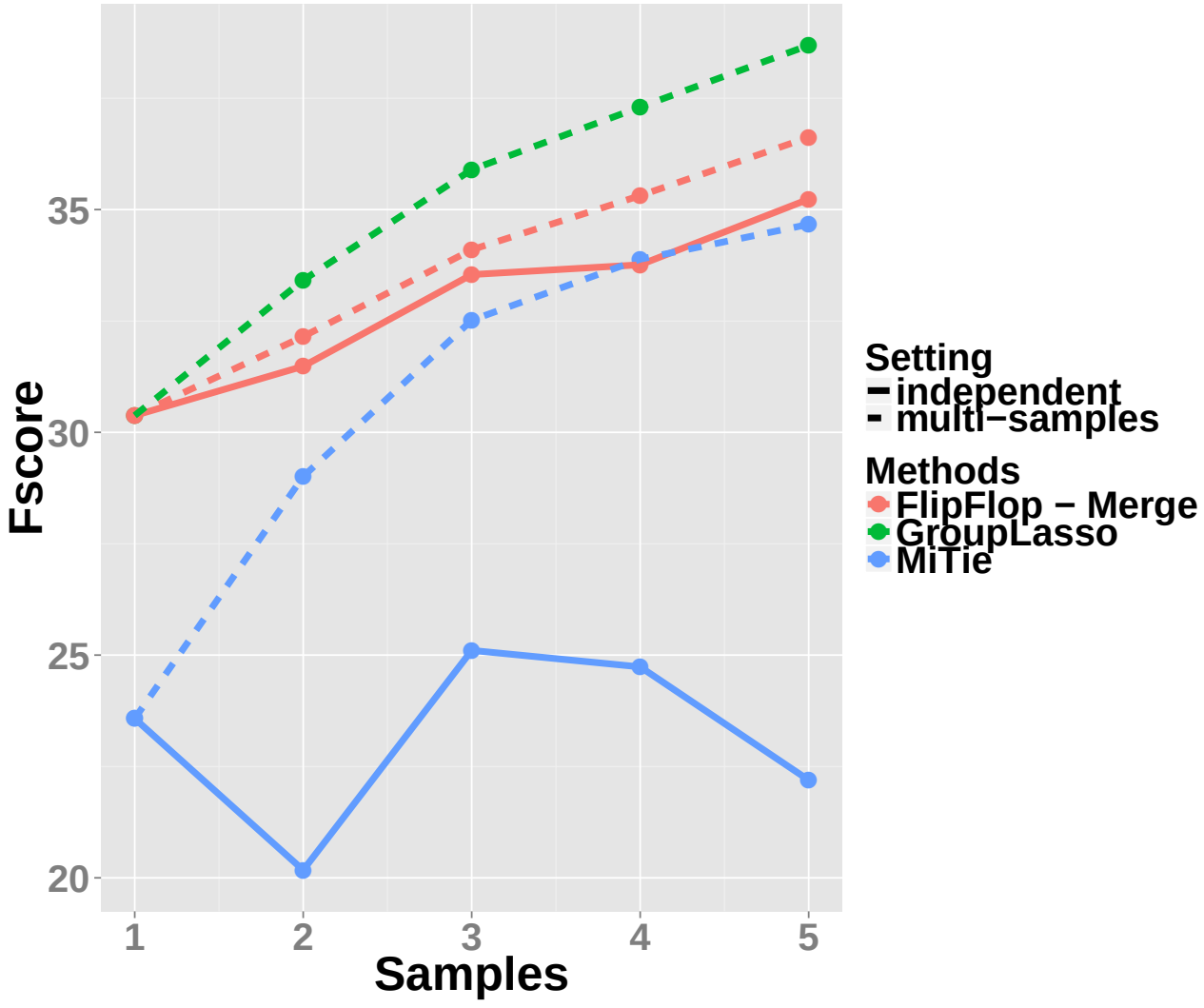


GroupLasso vs State-of-Art

$$\forall t \in \{1, \ldots, T\}, \text{supp}\theta_t = \text{supp}\theta_o$$



modENCODE data
Time course development of D.melanogaster



FlipFlop summary

- Fast method for exact Lasso-based isoform detection and quantification, with the "flow trick"
- Extension to multiple samples with structured sparsity
- `http://cbio.mines-paristech.fr/flipflop`
- Available as an R package

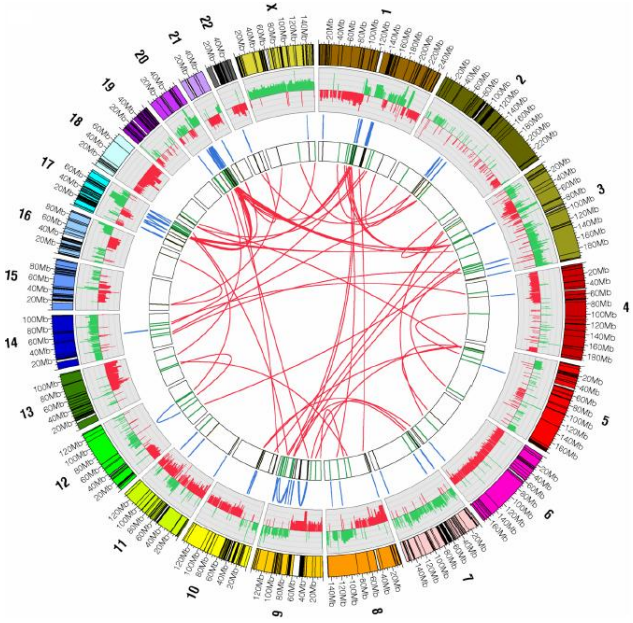
```
> source("http://bioconductor.org/biocLite.R")
> biocLite("flipflop")
```

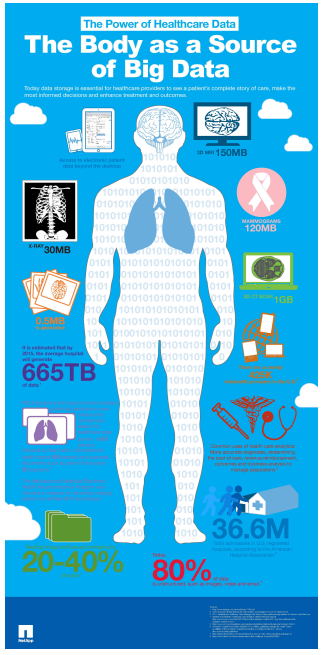
- More details in Bernard et al. (2014, 2015)

Outline

1. Learning with regularization and prior knowledge
2. Cancer patient stratification from somatic mutations
3. Learning from rankings through pairwise comparisons
4. FlipFlop: fast isoform prediction from RNA-seq data
5. Conclusion

Conclusion





- Many new problems and lots of data in computational genomics and precision medicine
- $n << p$ problem requires dedicated methods
  - new representations $x \rightarrow \Phi(x)$
  - new learning techniques (structured sparsity, regularization)
  - scalable algorithms

Thanks

References

F. R. Bach, G. R. G. Lanckriet, and M. I. Jordan. Multiple kernel learning, conic duality, and the SMO algorithm. In *Proceedings of the Twenty-First International Conference on Machine Learning*, page 6, New York, NY, USA, 2004. ACM. doi: 10.1145/1015330.1015424. URL http://doi.acm.org/10.1145/1015330.1015424.

E. Bernard, L. Jacob, J. Mairal, and J.-P. Vert. Efficient rna isoform identification and quantification from rna-seq data with network flows. Technical Report 00803134, HAL, 2013.

E. Bernard, L. Jacob, J. Mairal, and J.-P. Vert. Efficient RNA isoform identification and quantification from RNA-Seq data with network flows. *Bioinformatics*, 30(17):2447–2455, Sep 2014. doi: 10.1093/bioinformatics/btu317. URL http://dx.doi.org/10.1093/bioinformatics/btu317.

E. Bernard, L. Jacob, J. Mairal, E. Viara, and J.-P. Vert. A convex formulation for joint rna isoform detection and quantification from multiple rna-seq samples. *BMC bioinformatics*, 16:262, 2015. ISSN 1471-2105. doi: 10.1186/s12859-015-0695-9. URL http://dx.doi.org/10.1186/s12859-015-0695-9.

V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky. The convex geometry of linear inverse problems. *Found. Comput. Math.*, 12(6):805–849, 2012. doi: 10.1007/s10208-012-9135-7. URL http://dx.doi.org/10.1007/s10208-012-9135-7.

K. S. Frese, H. A. Katus, and B. Meder. Next-generation sequencing: from understanding biology to personalized medicine. *Biology*, 2:378–398, 2013. ISSN 2079-7737. doi: 10.3390/biology2010378. URL http://dx.doi.org/10.3390/biology2010378.

References (cont.)

A.-C. Haury, P. Gestraud, and J.-P. Vert. The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures. *PLoS One*, 6(12):e28210, 2011. doi: 10.1371/journal.pone.0028210. URL http://dx.doi.org/10.1371/journal.pone.0028210.

M. Hofree, J. P. Shen, H. Carter, A. Gross, and T. Ideker. Network-based stratification of tumor mutations. *Nat Methods*, 10(11):1108–1115, Nov 2013. doi: 10.1038/nmeth.2651. URL http://dx.doi.org/10.1038/nmeth.2651.

L. Jacob, G. Obozinski, and J.-P. Vert. Group lasso with overlap and graph lasso. In *ICML '09: Proceedings of the 26th Annual International Conference on Machine Learning*, pages 433–440, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-516-1. doi: 10.1145/1553374.1553431. URL http://dx.doi.org/10.1145/1553374.1553431.

R. Jenatton, J.-Y. Audibert, and F. Bach. Structured variable selection with sparsity-inducing norms. *J. Mach. Learn. Res.*, 12:2777–2824, 2011. URL http://www.jmlr.org/papers/volume12/jenatton11b/jenatton11b.pdf.

Y. Jiao and J.-P. Vert. The Kendall and Mallows kernels for permutations. In *Proceedings of The 32nd International Conference on Machine Learning*, volume 37 of *JMLR:W&CP*, pages 1935–1944, 2015. URL http://jmlr.org/proceedings/papers/v37/jiao15.html.

W. R. Knight. A computer method for calculating Kendall's tau with ungrouped data. *J. Am. Stat. Assoc.*, 61(314):436–439, 1966. URL http://www.jstor.org/stable/2282833.

References (cont.)

M. Le Morvan, A. Zinovyev, and J.-P. Vert. Netnorm: capturing cancer-relevant information in somatic exome mutation data with gene networks for cancer stratification and prognosis. Technical Report 01341856, HAL, 2016. URL http://hal.archives-ouvertes.fr/hal-01341856.

J. Mairal and B. Yu. Supervised feature selection in graphs with path coding penalties and network flows. *J. Mach. Learn. Res.*, 14:2449–2485, 2013.

S. Pal, R. Gupta, and R. V. Davuluri. Alternative transcription and alternative splicing in cancer. *Pharmacology and Therapeutics*, 136:283–294, 2012. doi: 10.1016/j.pharmthera.2012.08.005. URL http://dx.doi.org/10.1016/j.pharmthera.2012.08.005.

F. Rapaport, A. Zinovyev, M. Dutreix, E. Barillot, and J.-P. Vert. Classification of microarray data using gene networks. *BMC Bioinformatics*, 8:35, 2007. doi: 10.1186/1471-2105-8-35. URL http://dx.doi.org/10.1186/1471-2105-8-35.

F. Rapaport, E. Barillot, and J.-P. Vert. Classification of arrayCGH data using fused SVM. *Bioinformatics*, 24(13):i375–i382, Jul 2008. doi: 10.1093/bioinformatics/btn188. URL http://dx.doi.org/10.1093/bioinformatics/btn188.

M. R. Stratton, P. J. Campbell, and P. A. Futreal. The cancer genome. *Nature*, 458(7239): 719–724, Apr 2009. doi: 10.1038/nature07943. URL http://dx.doi.org/10.1038/nature07943.

A. C. Tan, D. Q. Naiman, L. Xu, R. L. Winslow, and D. Geman. Simple decision rules for classifying human cancers from gene expression profiles. *Bioinformatics*, 21(20):3896–3904, Oct 2005. doi: 10.1093/bioinformatics/bti631. URL http://dx.doi.org/10.1093/bioinformatics/bti631.

References (cont.)

R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 67(1):91–108, 2005. URL http://ideas.repec.org/a/bla/jorssb/v67y2005i1p91-108.html.

M. J. van de Vijver, Y. D. He, L. J. van't Veer, H. Dai, A. A. M. Hart, D. W. Voskuil, G. J. Schreiber, J. L. Peterse, C. Roberts, M. J. Marton, M. Parrish, D. Atsma, A. Witteveen, A. Glas, L. Delahaye, T. van der Velde, H. Bartelink, S. Rodenhuis, E. T. Rutgers, S. H. Friend, and R. Bernards. A gene-expression signature as a predictor of survival in breast cancer. *N. Engl. J. Med.*, 347(25):1999–2009, Dec 2002. doi: 10.1056/NEJMoa021967. URL http://dx.doi.org/10.1056/NEJMoa021967.

L. J. van 't Veer, H. Dai, M. J. van de Vijver, Y. D. He, A. A. M. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards, and S. H. Friend. Gene expression profiling predicts clinical outcome of breast cancers. *Nature*, 415(6871):530–536, Jan 2002. doi: 10.1038/415530a. URL http://dx.doi.org/10.1038/415530a.

K. Vervier, P. Mahé, A. DâĂŹAspremont, J.-B. Veyrieras, and J.-P. Vert. On learning matrices with orthogonal columns or disjoint supports. In T. Calders, F. Esposito, E. Hüllermeier, and R. Meo, editors, *Machine Learning and Knowledge Discovery in Databases*, volume 8726 of *Lecture Notes in Computer Science*, pages 274–289. Springer Berlin Heidelberg, 2014. doi: 10.1007/978-3-662-44845-8_18. URL http://dx.doi.org/10.1007/978-3-662-44845-8_18.

References (cont.)

Y. Wang, J. Klijn, Y. Zhang, A. Sieuwerts, M. Look, F. Yang, D. Talantov, M. Timmermans, M. Meijer-van Gelder, J. Yu, T. Jatkoe, E. Berns, D. Atkins, and J. Foekens. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancers. *Lancet*, 365(9460):671–679, 2005. doi: 10.1016/S0140-6736(05)17947-1. URL http://dx.doi.org/10.1016/S0140-6736(05)17947-1.

M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B*, 68(1):49–67, 2006. doi: 10.1111/j.1467-9868.2005.00532.x. URL http://dx.doi.org/10.1111/j.1467-9868.2005.00532.x.