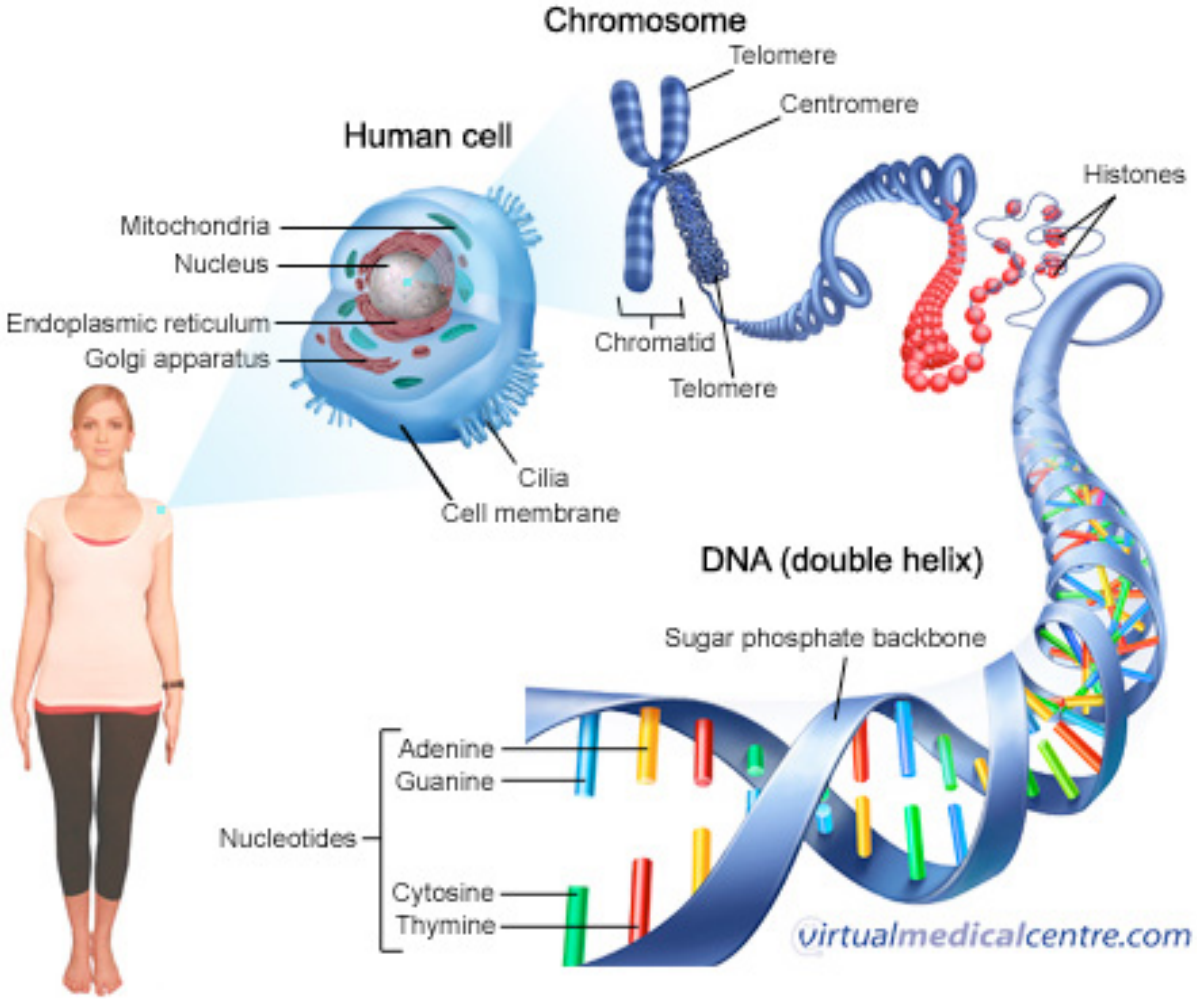
Can "Big Data" cure cancer?

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

ENS Data Science Colloquim, October 10, 2016

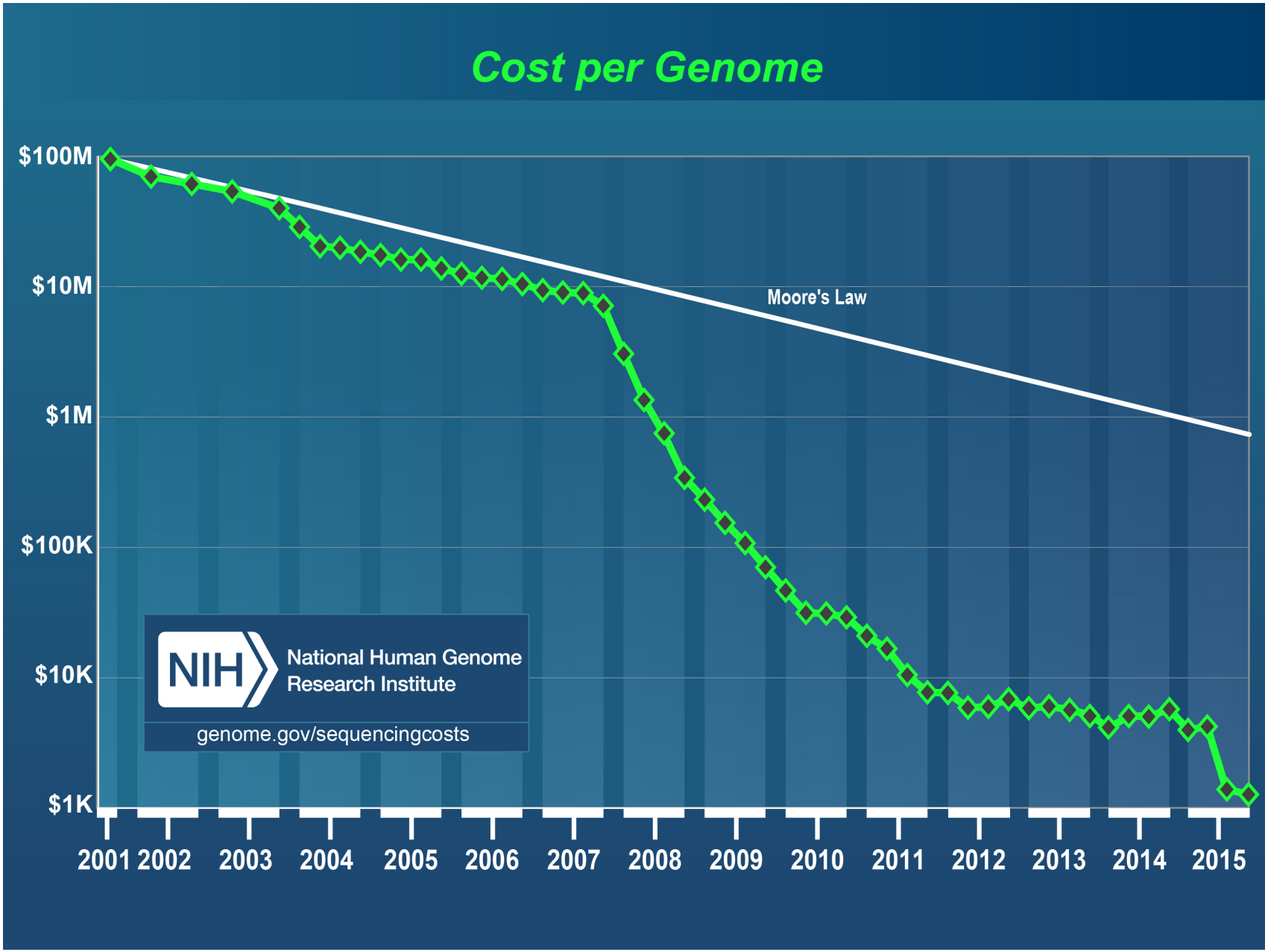
A complex system



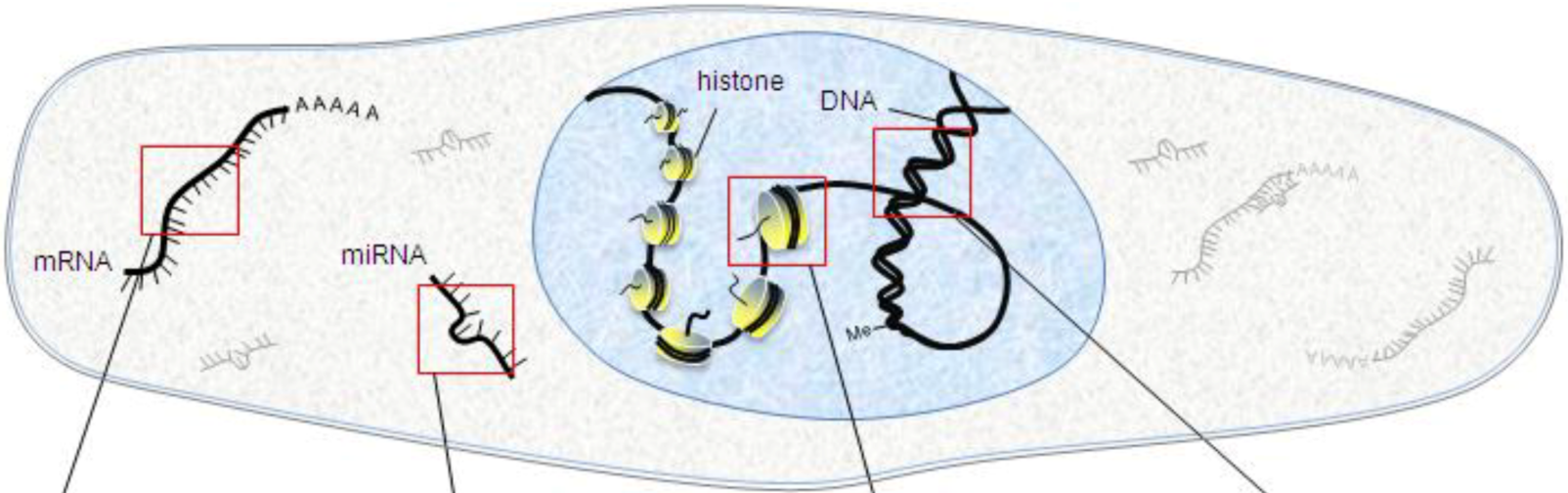
1 body = 10^{14} human cells (and 100x more non-human cells)

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

The sequencing revolution



Sequencing is a swiss knife



Transcriptomics

- gene expression analysis
- non-coding RNA analysis
- gene-fusion detection
- mRNA splice-analysis
- RNA editing

miRNomics

- expression analysis
- miRNA identification
- miRNA editing

Epigenomics

- targeted bisulfite sequencing
- ChIP-Seq
- MeDIP-Seq, methylCap-seq
- whole genome bisulfite sequencing

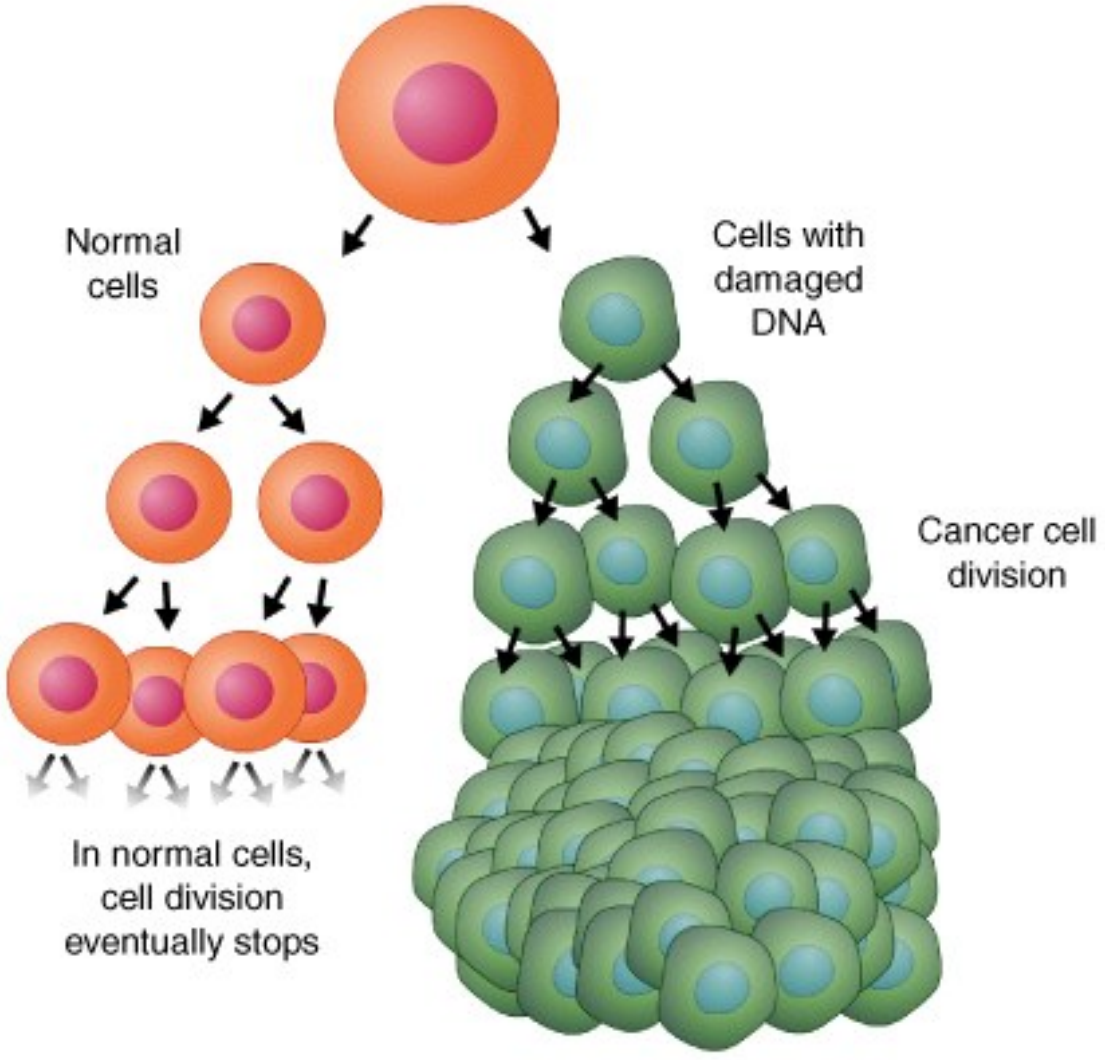
Genomics

- targeted sequencing
- whole-exome sequencing
- whole-genome sequencing

information content | experimental costs | current clinical utility

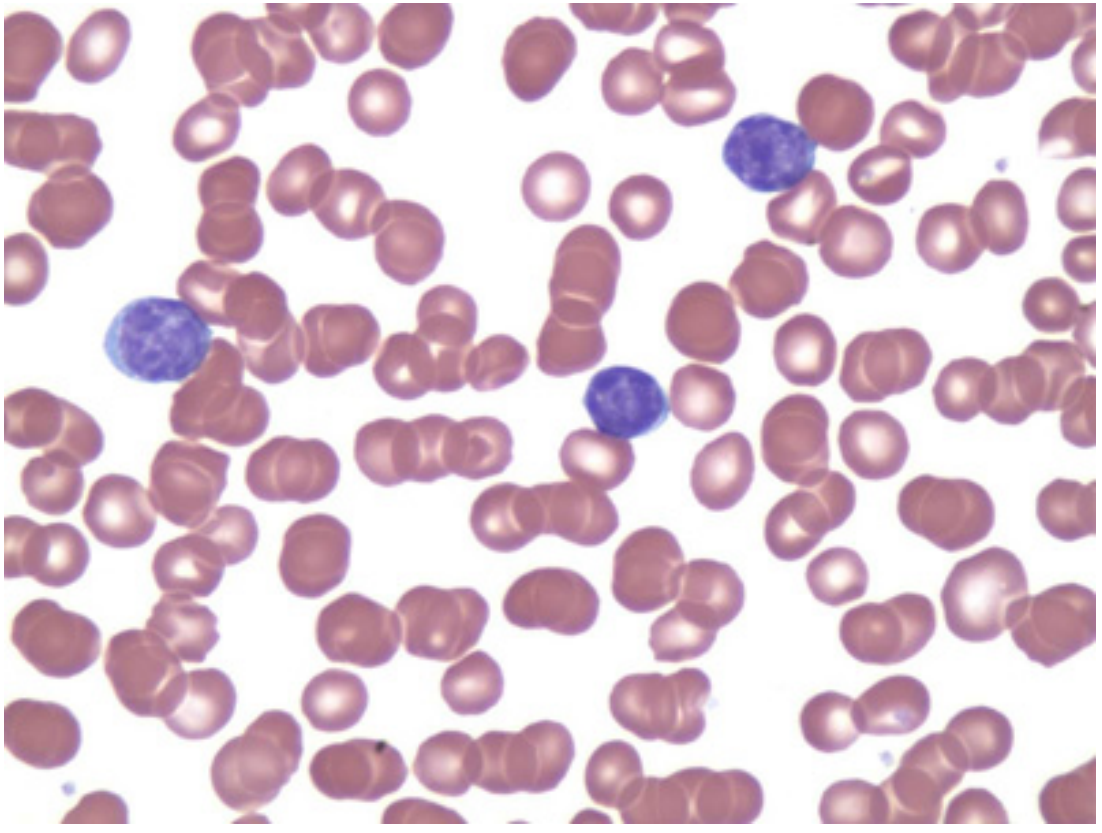
(Frese et al., 2013)

Cancer

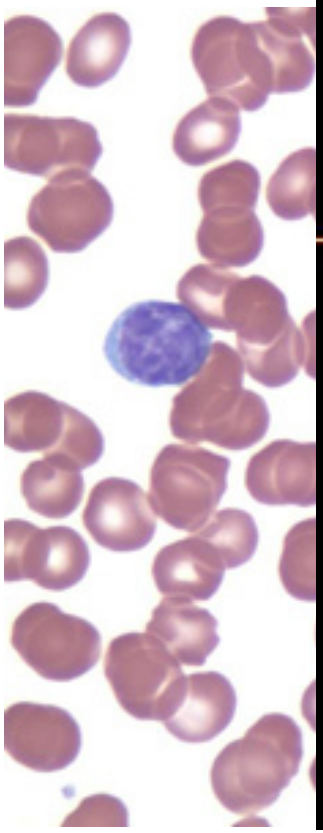


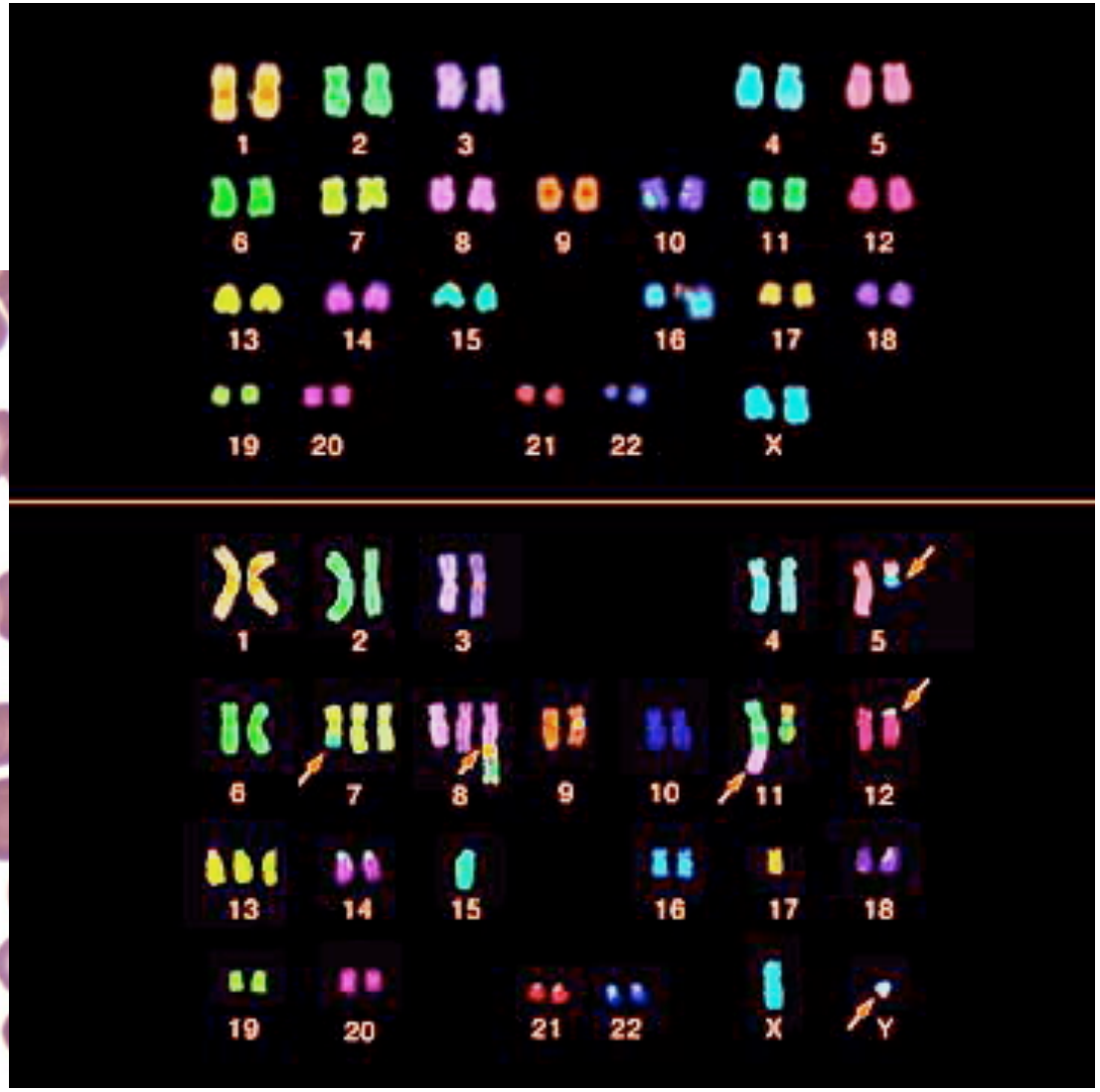
http://rise.duke.edu/seek/pages/page.html?0205

A cancer cell (1900)

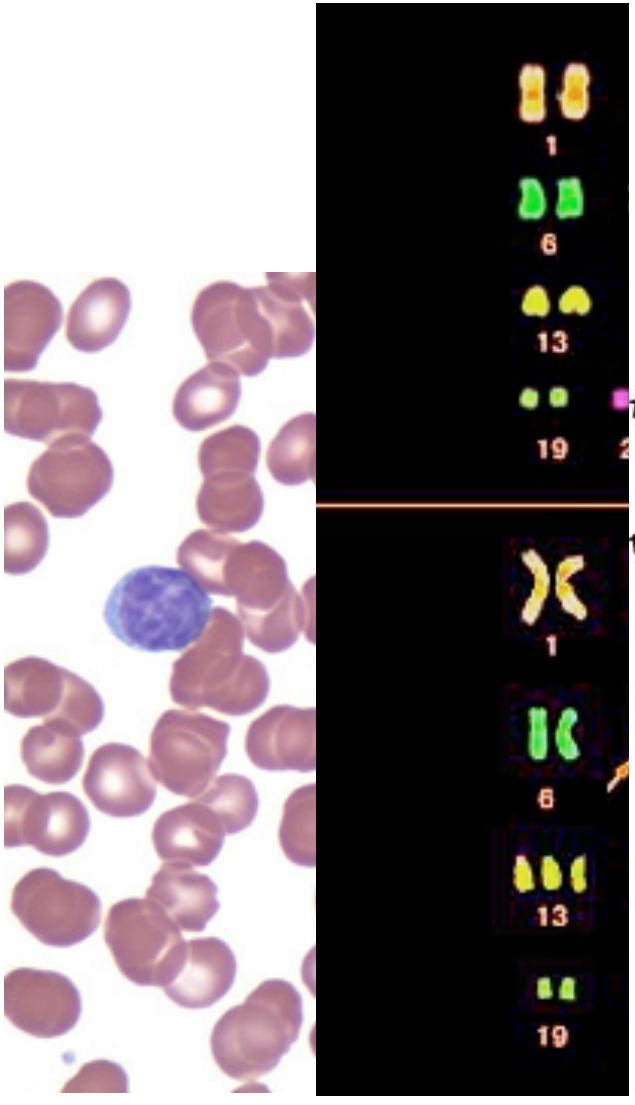


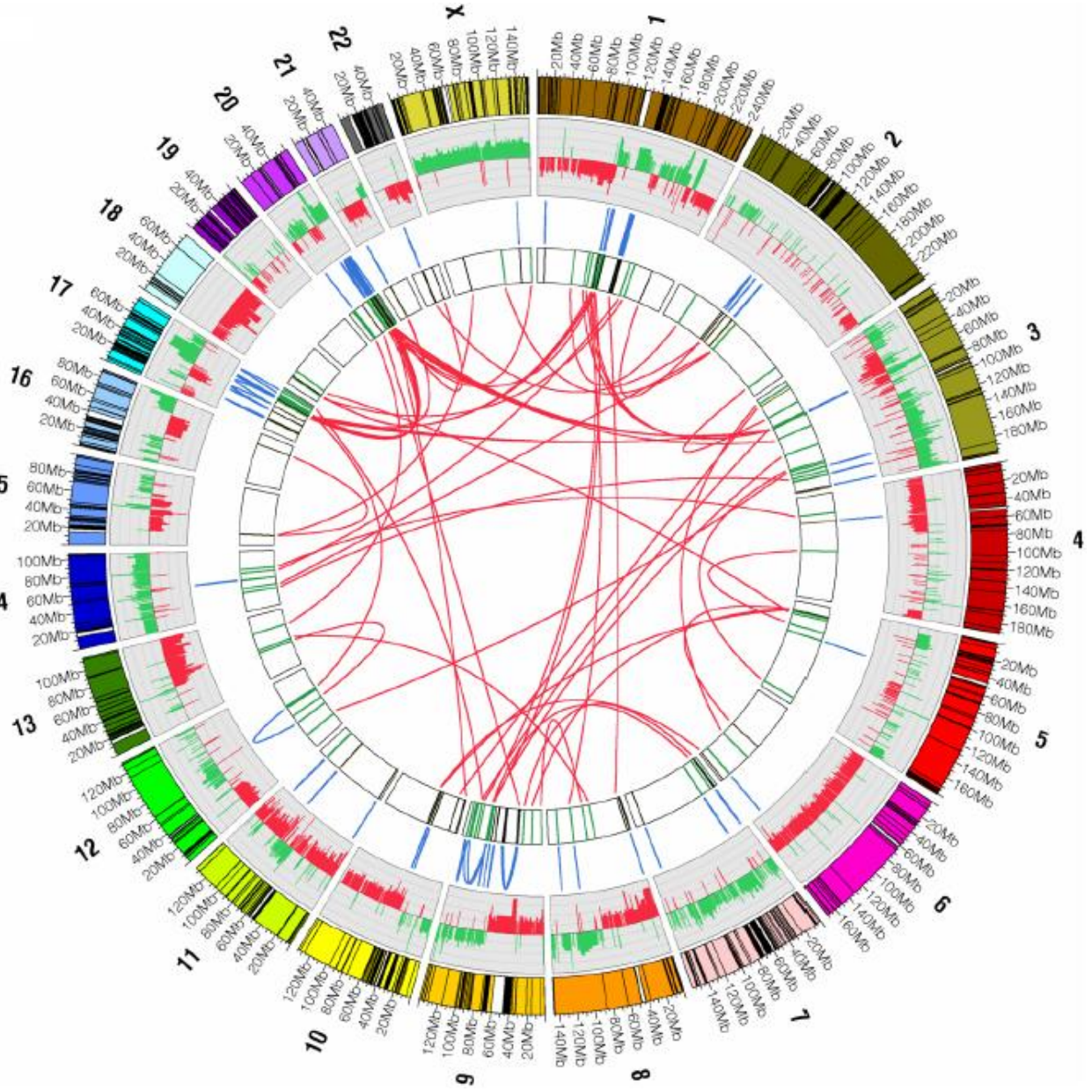
A cancer cell (1960)



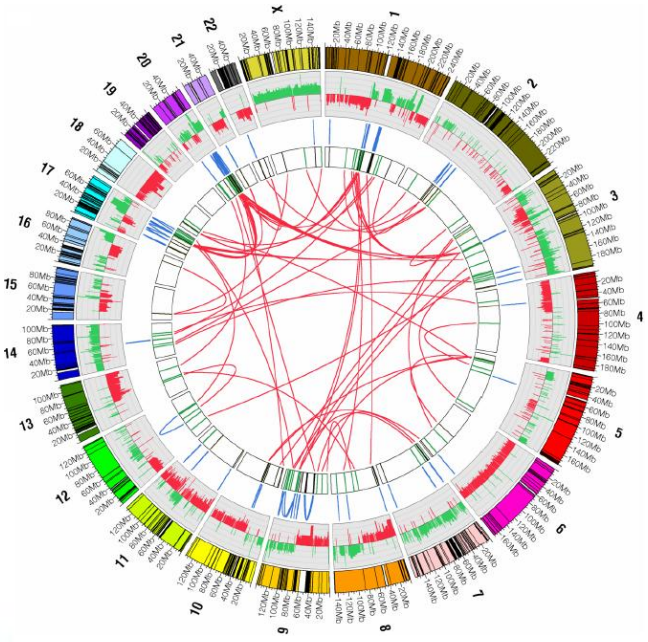


A cancer cell (2010)



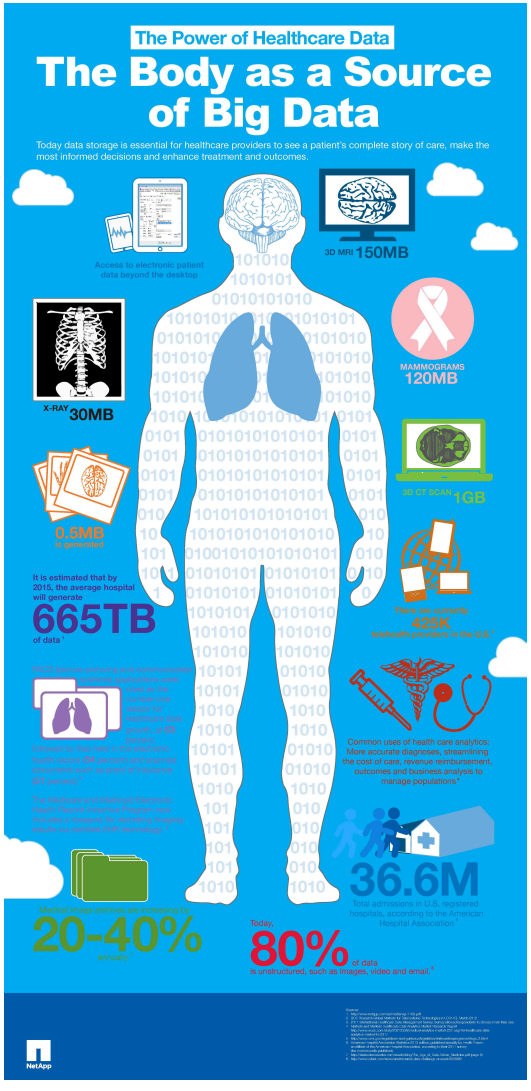


Big data

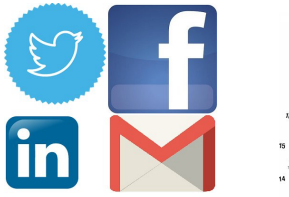


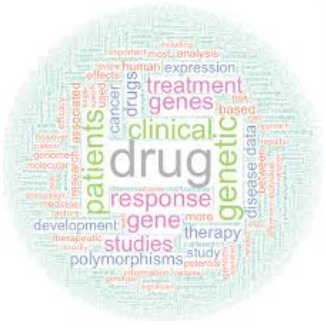

patients
drug
clinical
genetic
response
gene
treatment
genes
expression
human
analysis
drugs
cancer
disease data
development
therapy
studies
polymorphisms
study

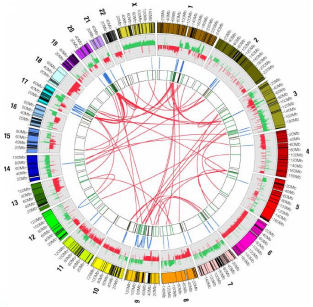

97

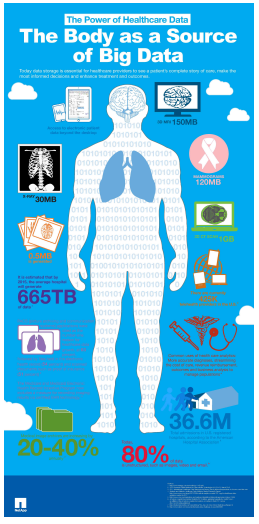

The Power of Healthcare Data
The Body as a Source of Big Data
Today data storage is essential for healthcare providers to see a patient's complete story of care, make the most informed decisions and enhance treatment and outcomes.
Access to electronic patient data beyond the desktop
3D MRI 150MB
X-RAY 30MB
MAMMOGRAMS 120MB
3D CT SCAN 1GB
0.5MB
It is estimated that by 2015, the average hospital will generate
665TB
of data
Common uses of health care analytics: More accurate diagnoses, streamlining the cost of care, revenue reimbursement, outcomes and business analysis to manage populations
36.6M
Medical image archives are increasing by
20-40%
annually
Today,
80%
of data is unstructured, such as images, video and email.
NetApp

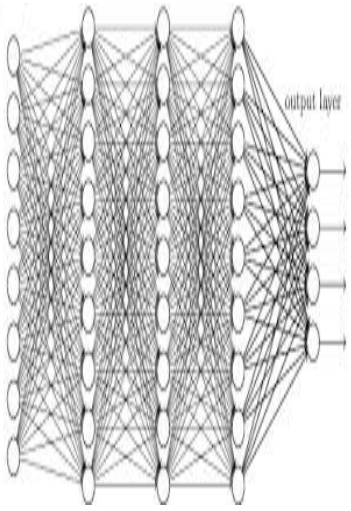
Opportunities





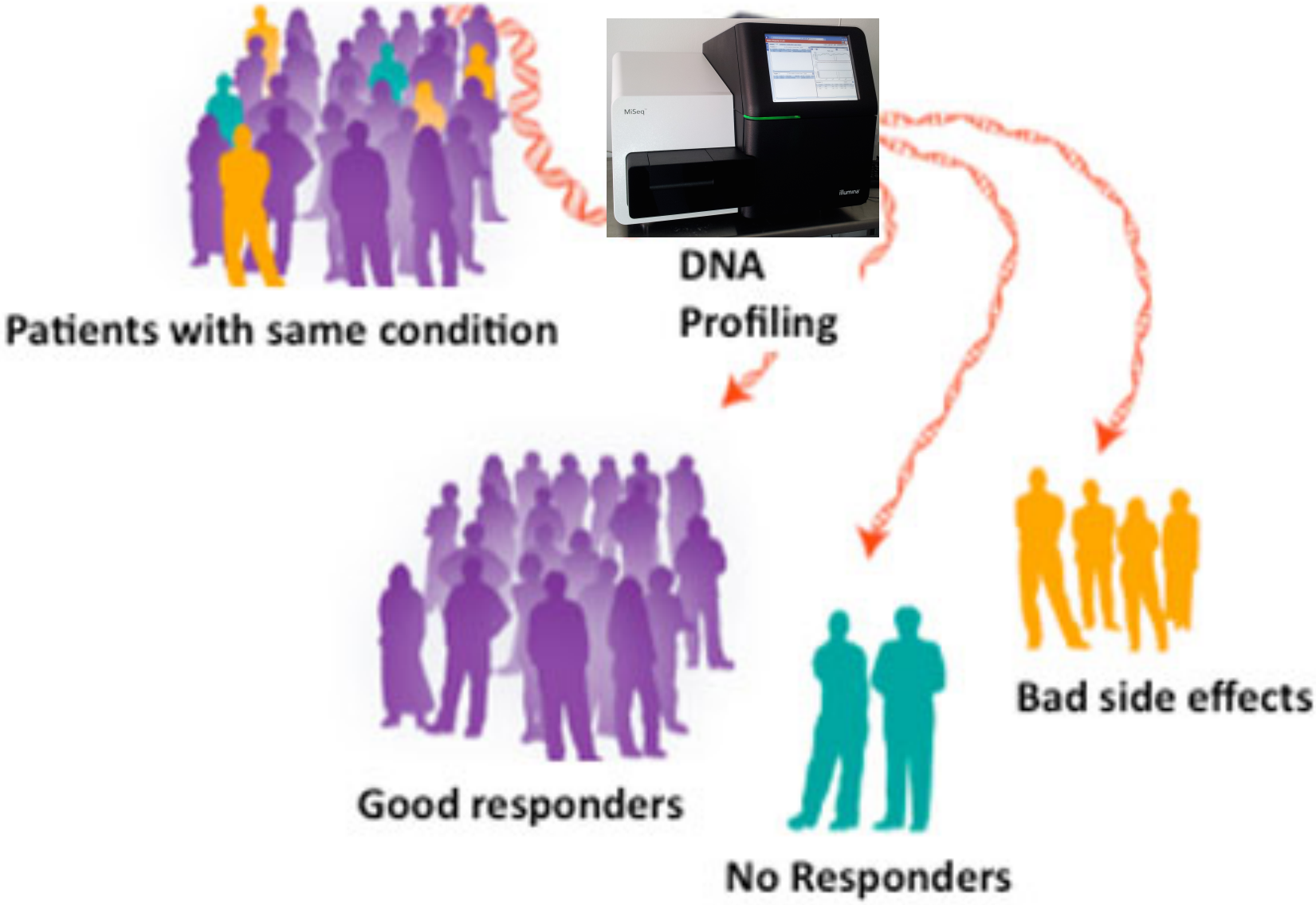






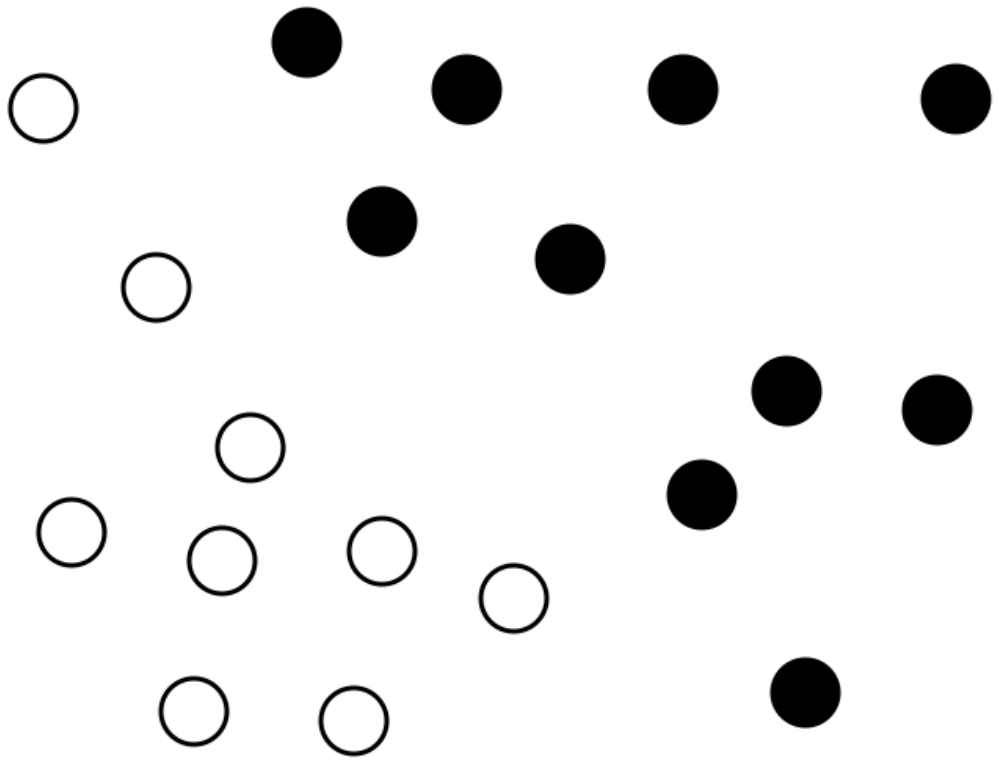
- What is your risk of developing a cancer? (*prevention*)
- Once detected, what precisely is your cancer? (*diagnosis*)
- After treatment, are you cured? (*prognosis*)
- What is the best way to treat your cancer? (*precision medicine*)

Example: precision medicine



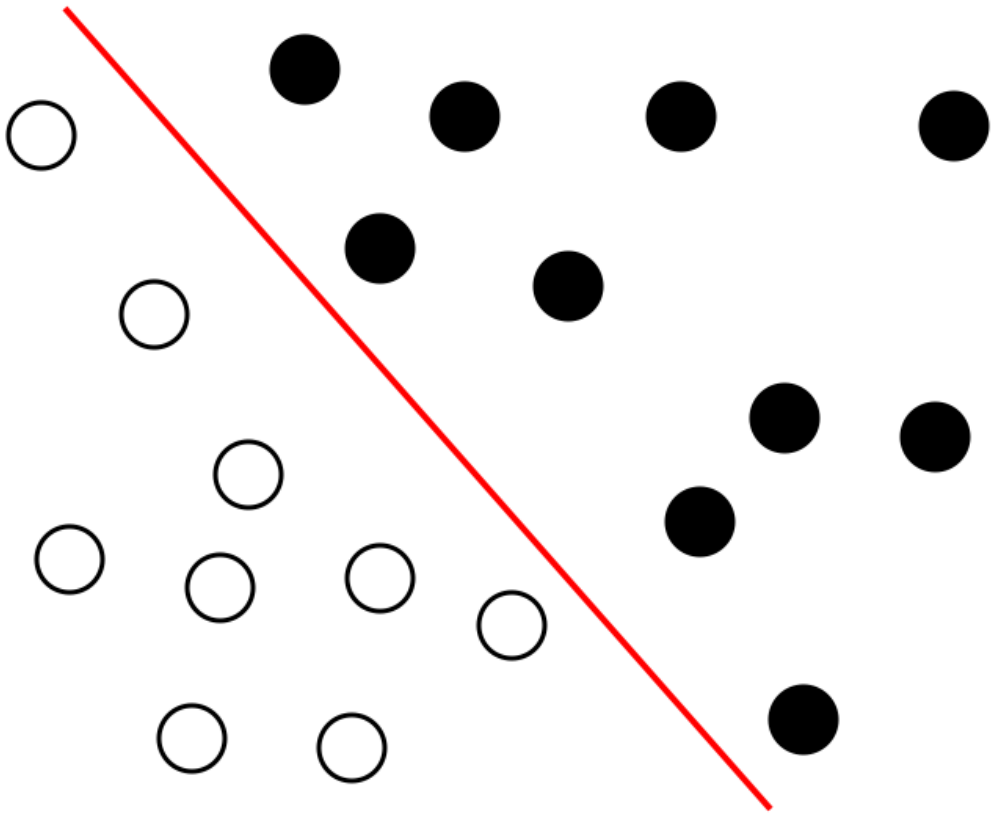
Learning from data (EASY case)

- Good vs Bad responders
- $n(= 19)$ patients $>>$ $p(= 2)$ genes



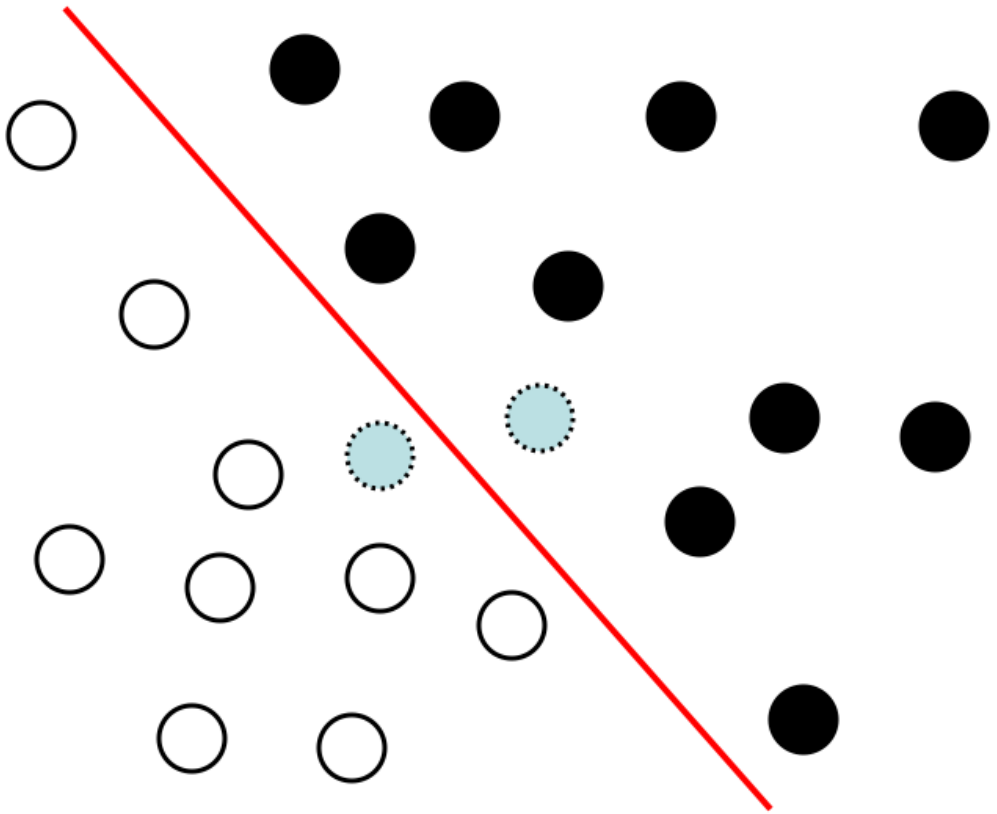
Learning from data (EASY case)

- Good vs Bad responders
- $n(= 19)$ patients $>>$ $p(= 2)$ genes



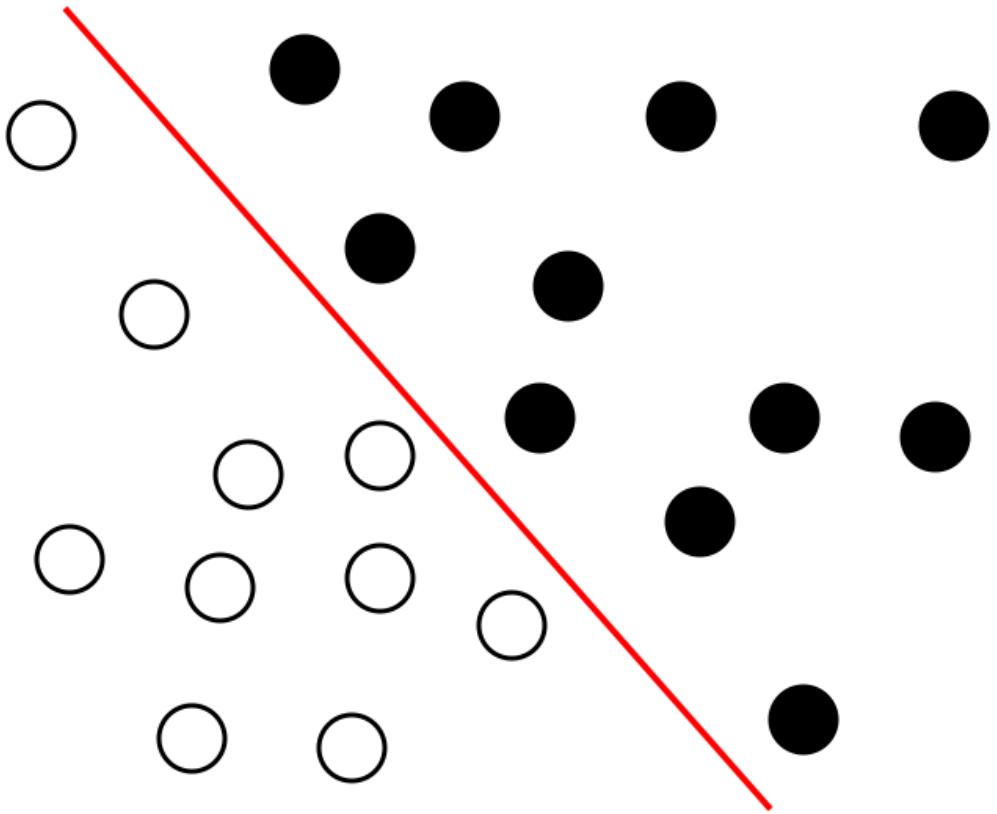
Learning from data (EASY case)

- Good vs Bad responders
- $n(= 19)$ patients $>>$ $p(= 2)$ genes

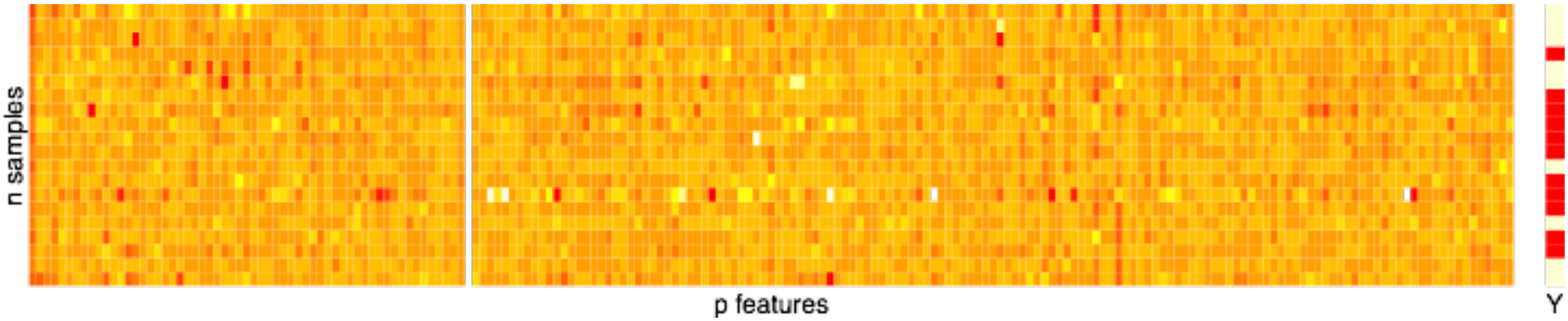


Learning from data (EASY case)

- Good vs Bad responders
- $n(= 19)$ 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

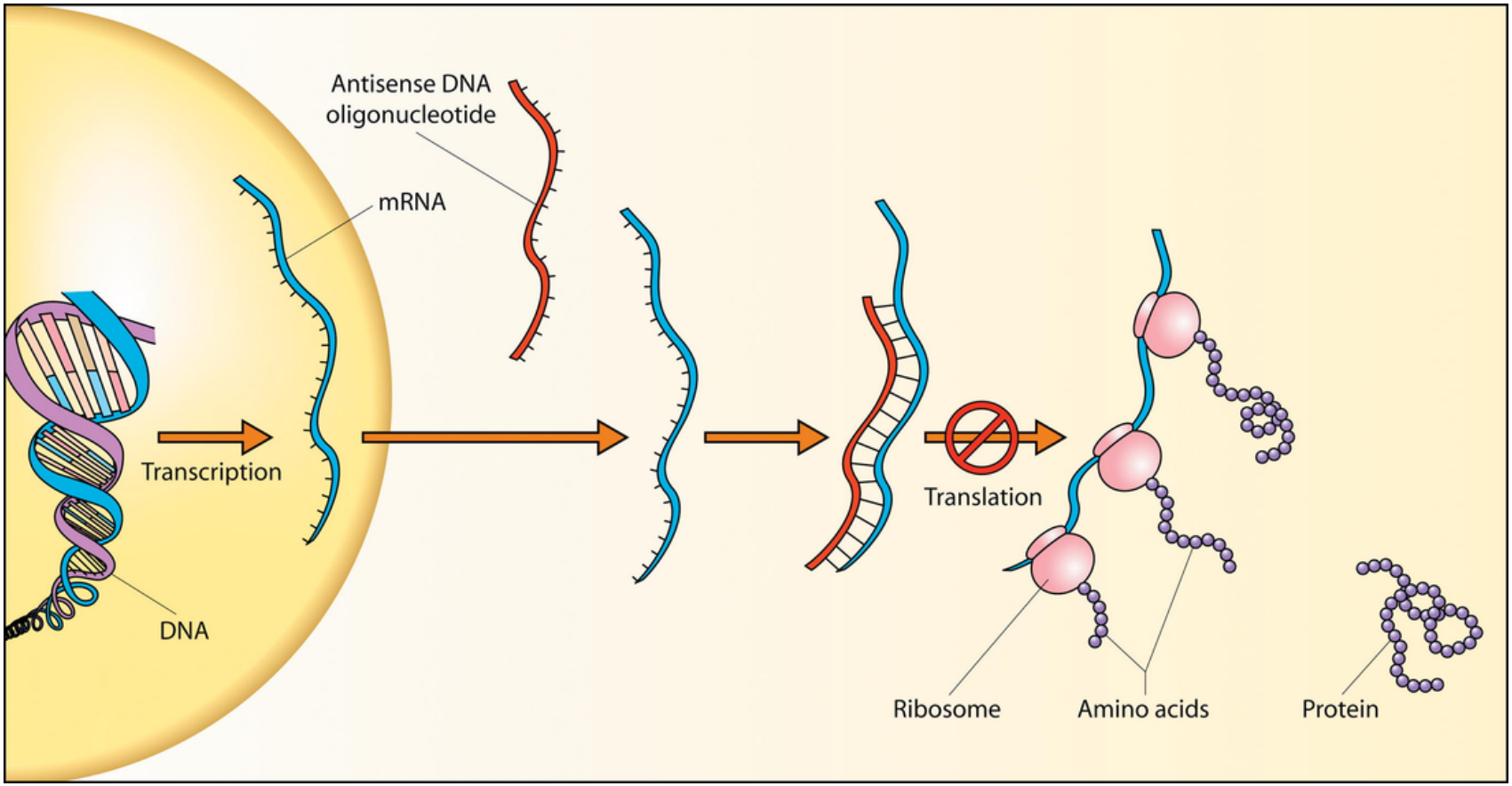
Outline

Outline

Joint work with...

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

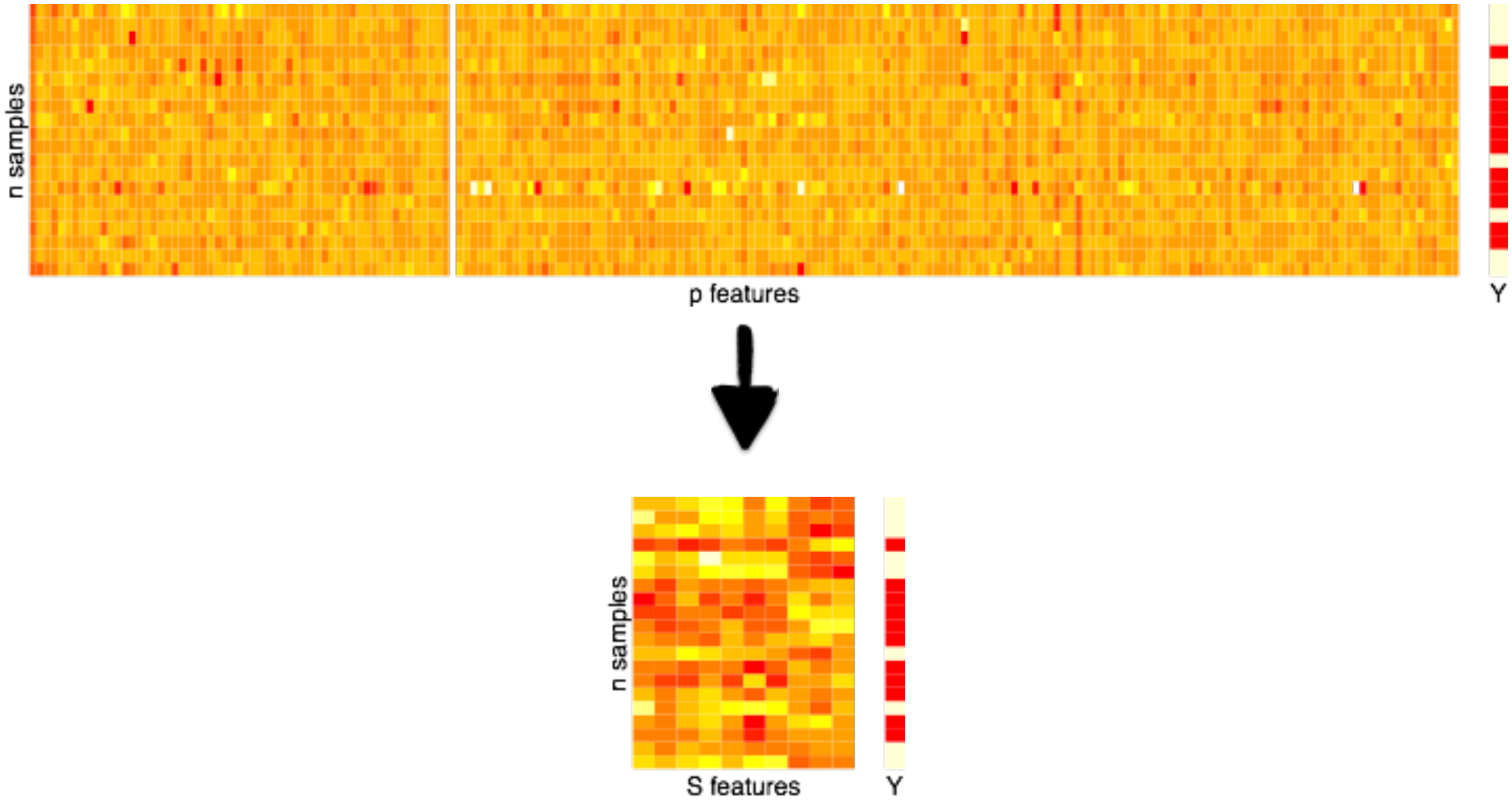
Gene expression



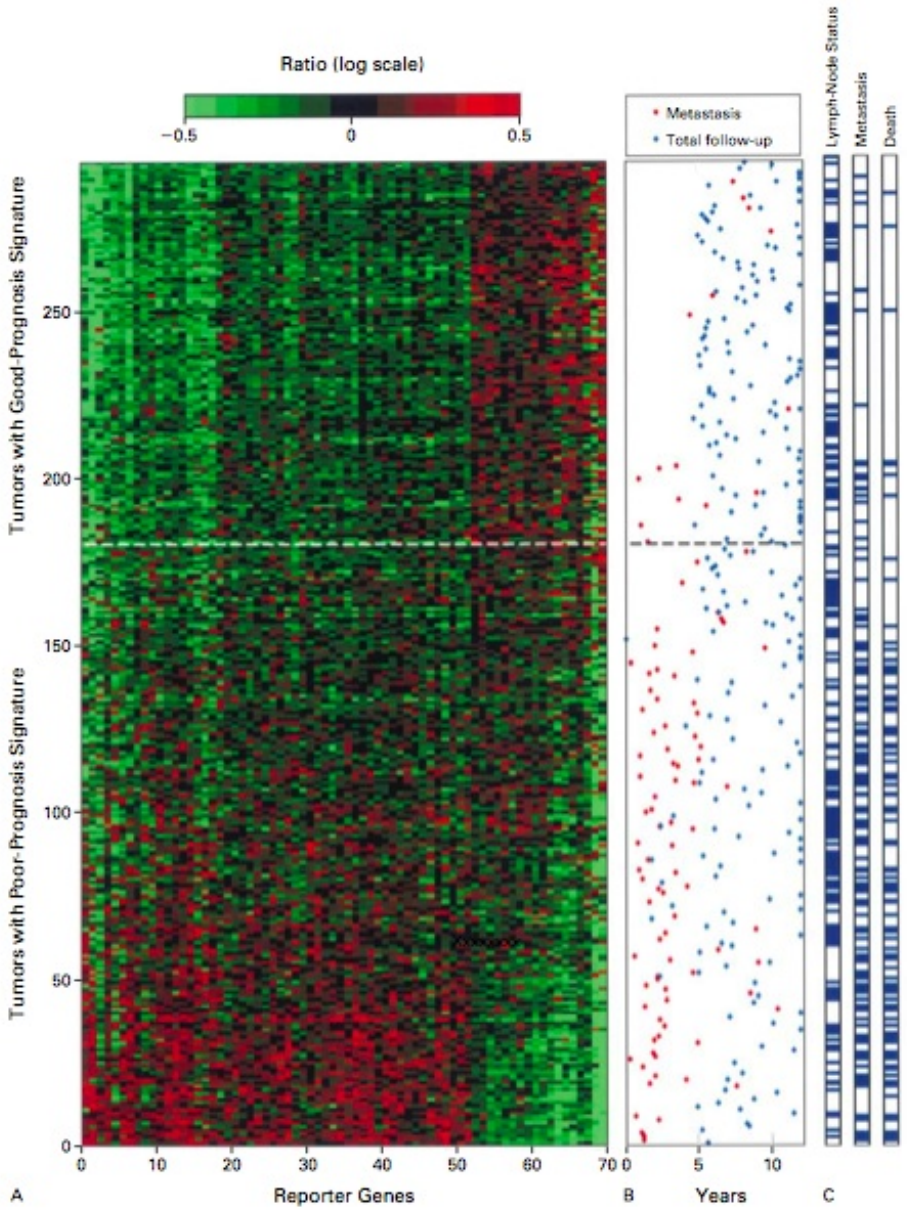
http://mrsbabbkv.weebly.com/rna--protein.html

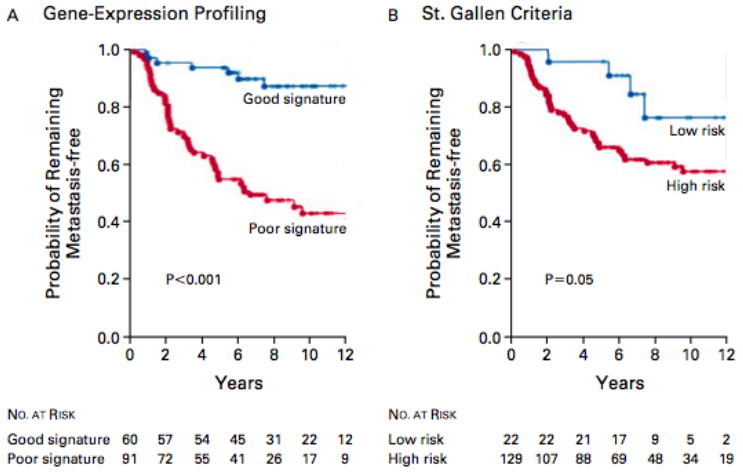
- About 22,000 genes encoded in DNA (same for all cells)
- Expression of each gene (= RNA synthesis) varies between cells
- Can be measured for all genes simultaneously with sequencing

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*

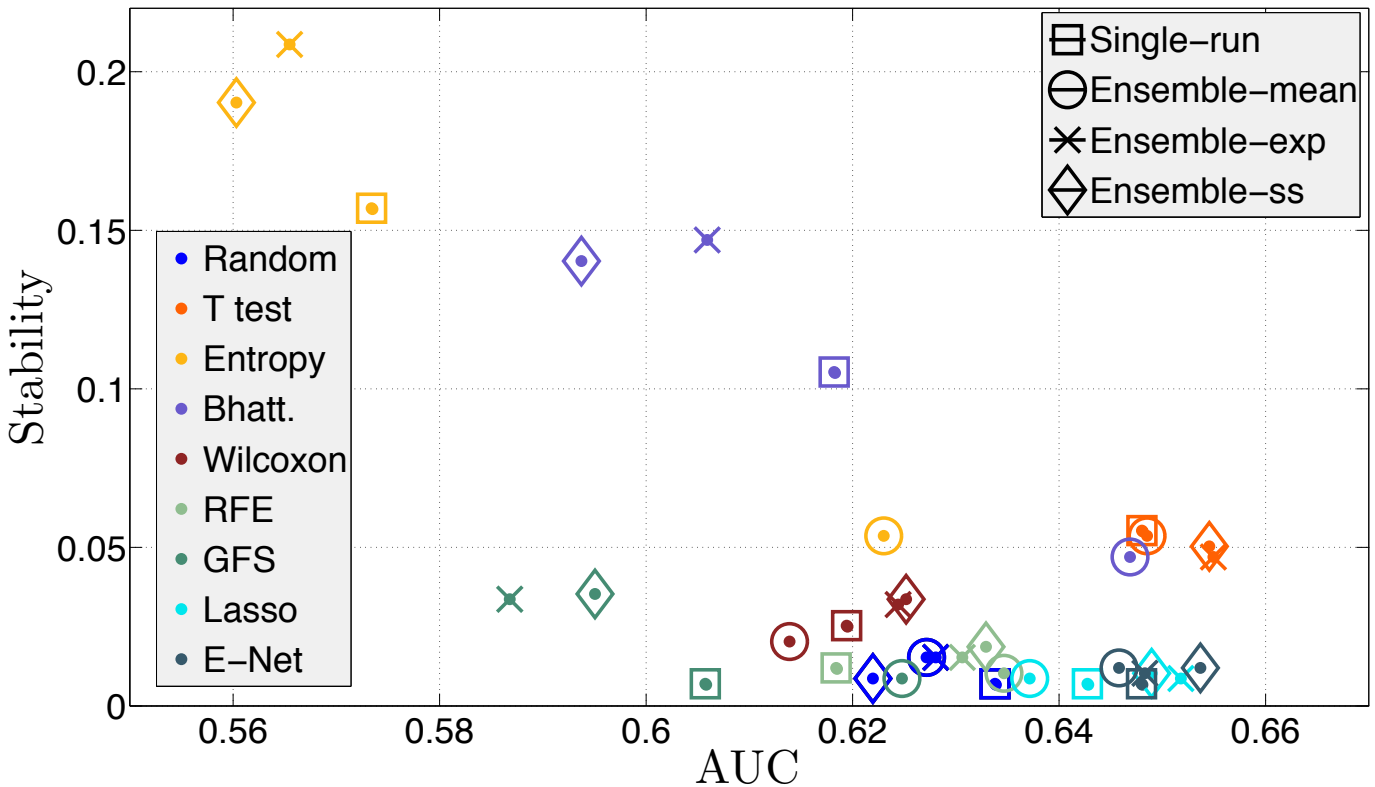
OPEN ACCESS Freely available online

PLoS one

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

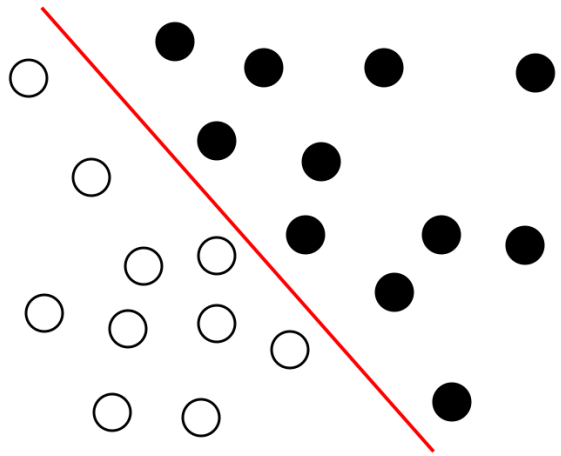
Anne-Claire Haury[1,2,3]*, Pierre Gestraud[1,2,3], Jean-Philippe Vert[1,2,3]

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)

Learning with regularization



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

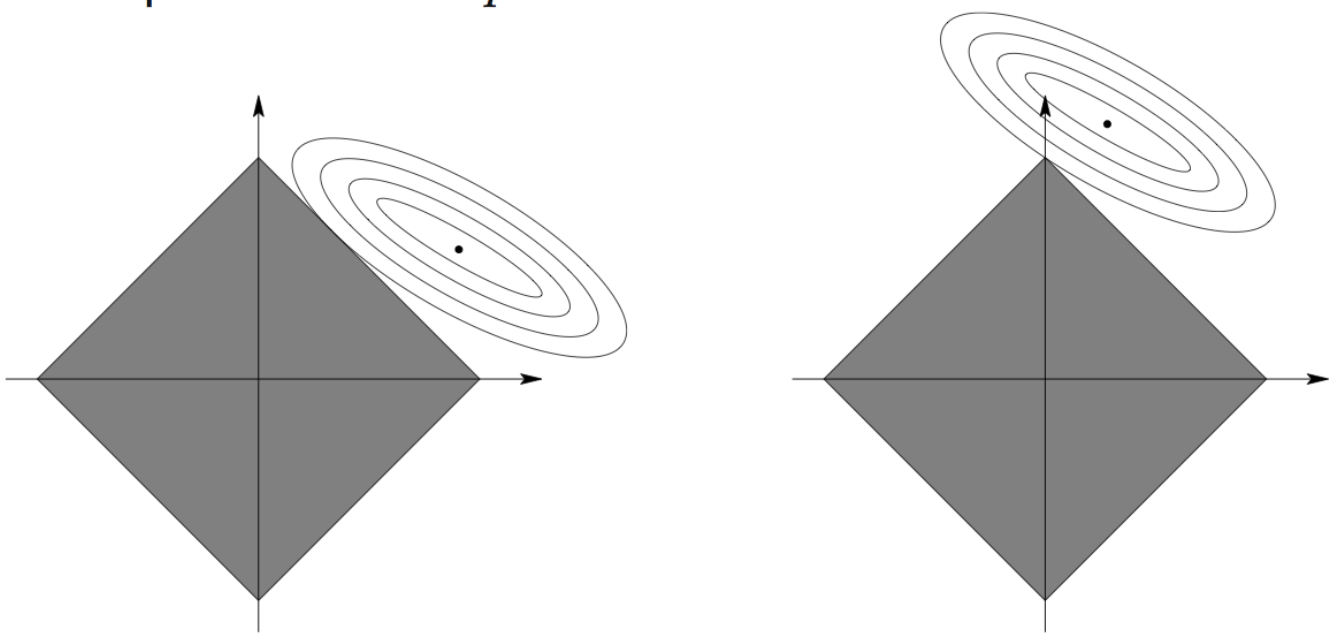
$$f_\beta(x) = \beta^\top x \qquad \min_{\beta \in \mathbb{R}^p} R(f_\beta) + \lambda \Omega(\beta)$$

- $R(f_\beta)$ 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,...)

Sparsity with ℓ_1 regularization

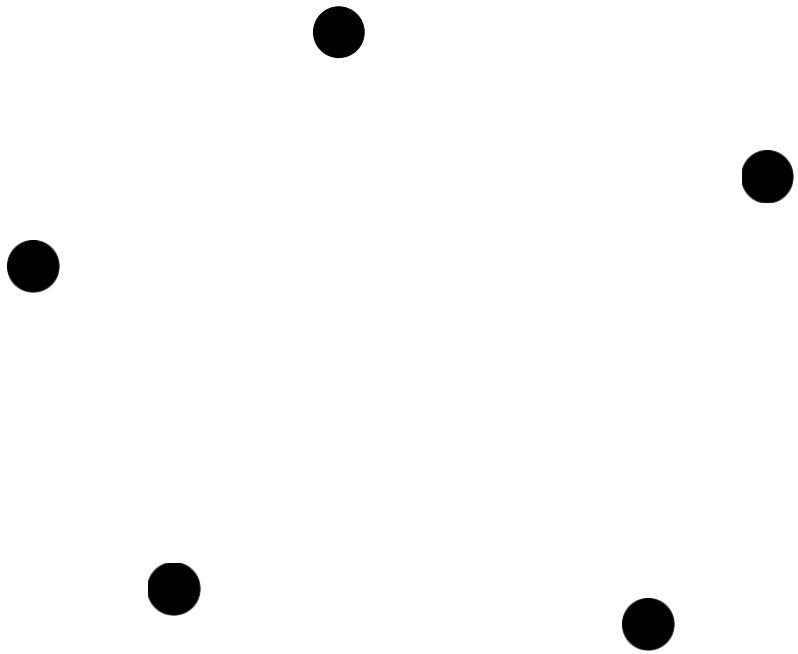
$$\min_{\beta} R(f_\beta) + \lambda \sum_{i=1}^{p} |\beta_i| \quad \Leftrightarrow \quad \min_{\beta} R(f_\beta) \text{ such that } \sum_{i=1}^{p} |\beta_i| \leq C$$

Geometric interpretation with $p = 2$



Leads to sparse models (feature selection)

Atomic Norm (Chandrasekaran et al., 2012)



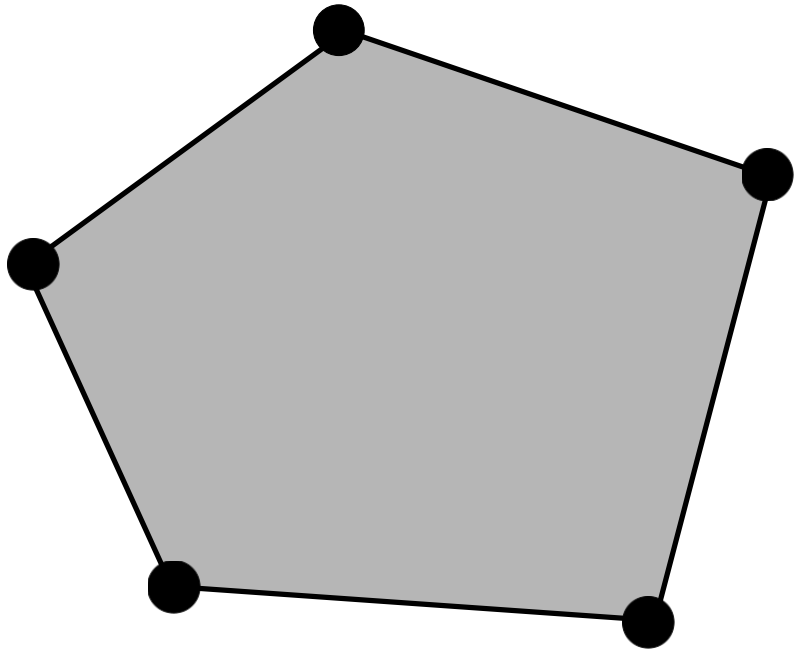
Definition

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

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

$\mathcal{A}$ should be centrally symmetric and span $\mathbb{R}^p$

Atomic Norm (Chandrasekaran et al., 2012)



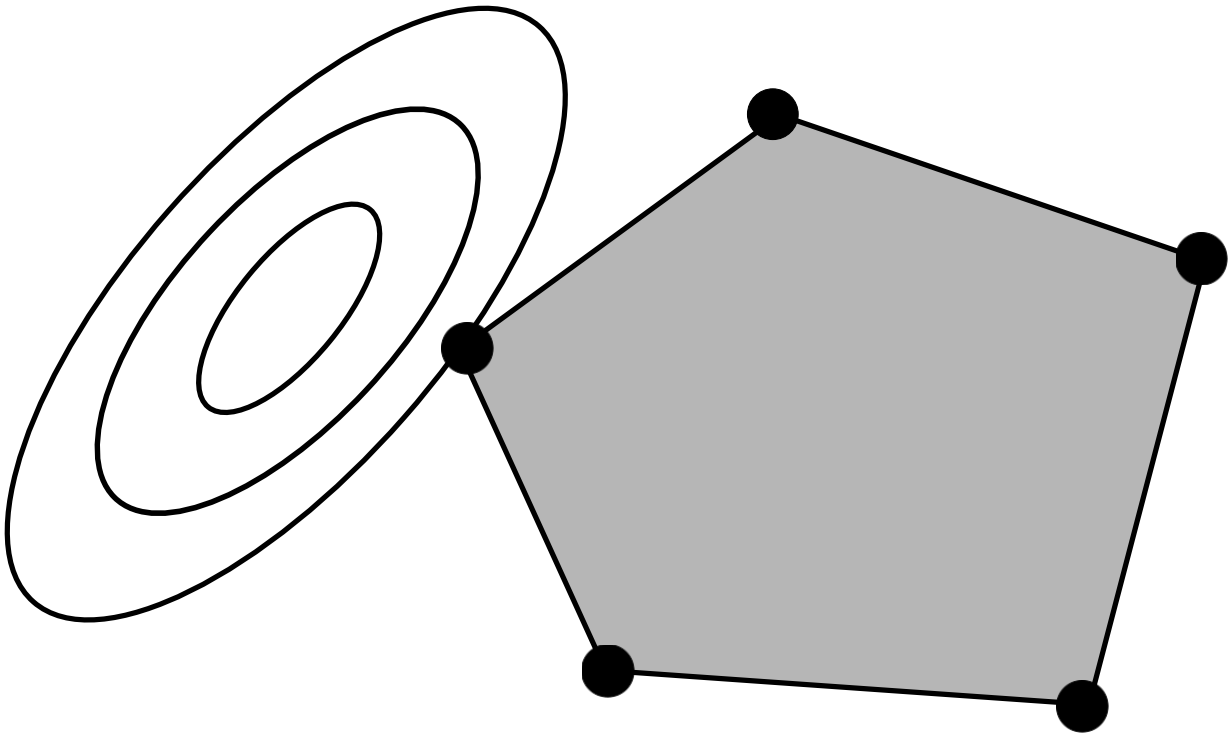
Definition

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

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

$\mathcal{A}$ should be centrally symmetric and span $\mathbb{R}^p$

Atomic Norm (Chandrasekaran et al., 2012)



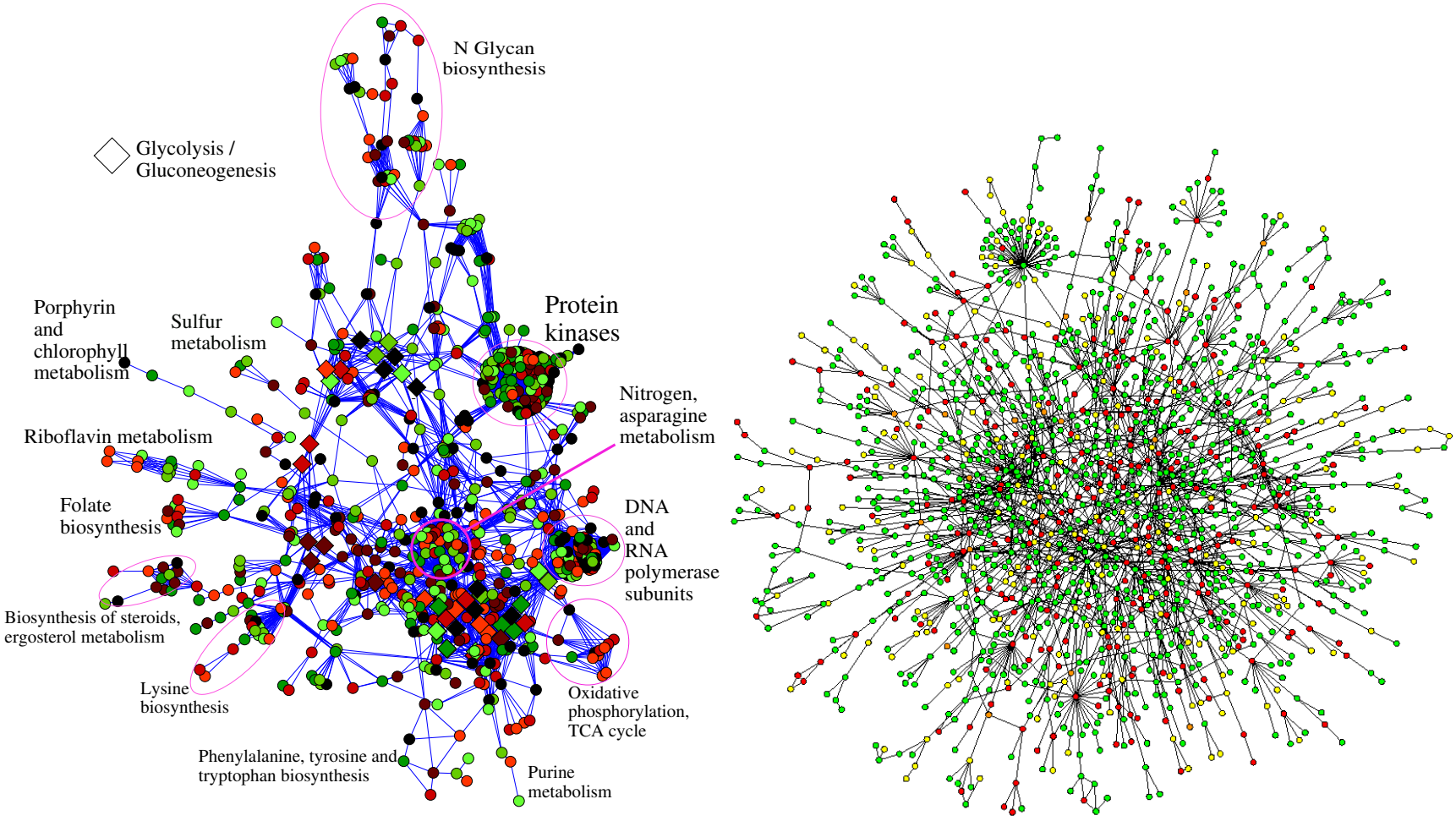
Definition

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

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

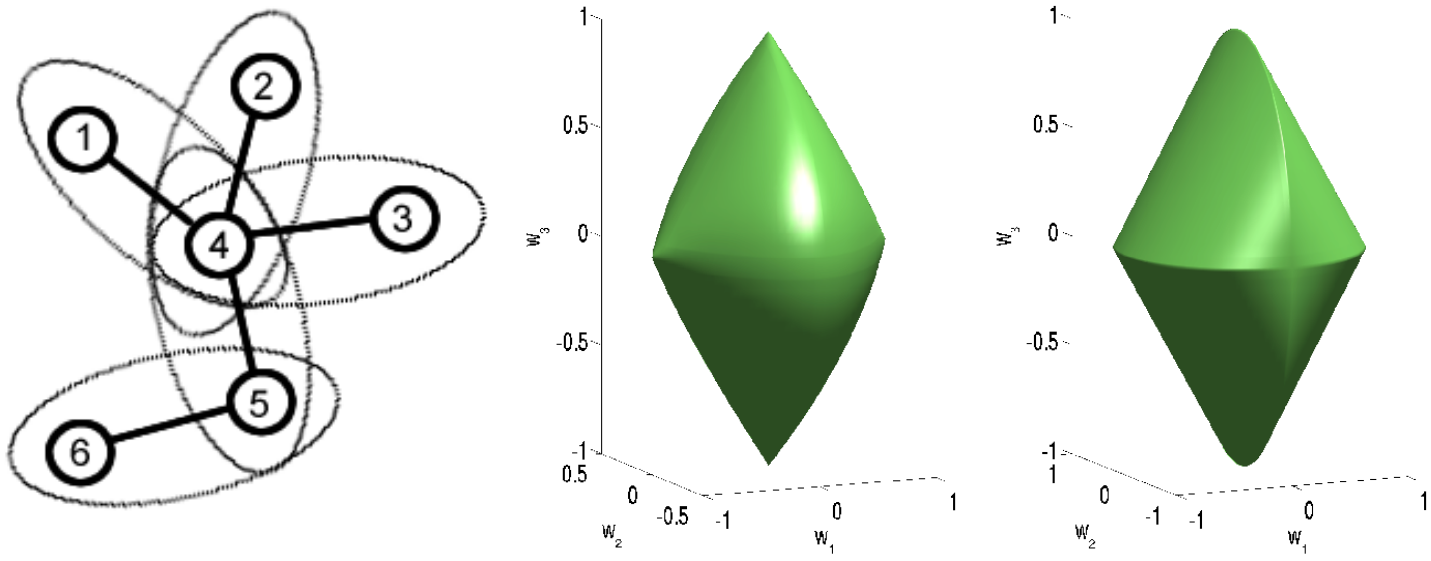
$\mathcal{A}$ should be centrally symmetric and span $\mathbb{R}^p$

Gene networks as prior knowledge



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

Graph-based structured feature selection

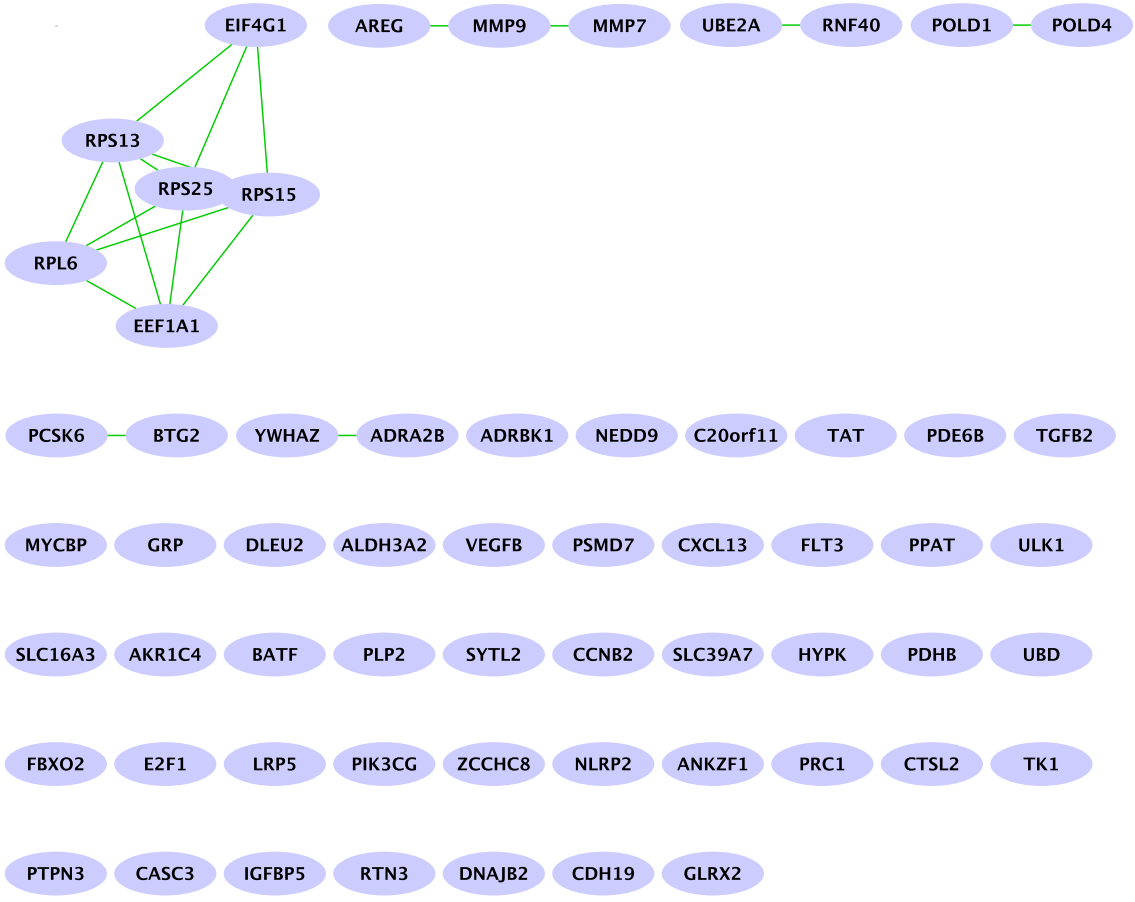


Graph lasso(s)

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

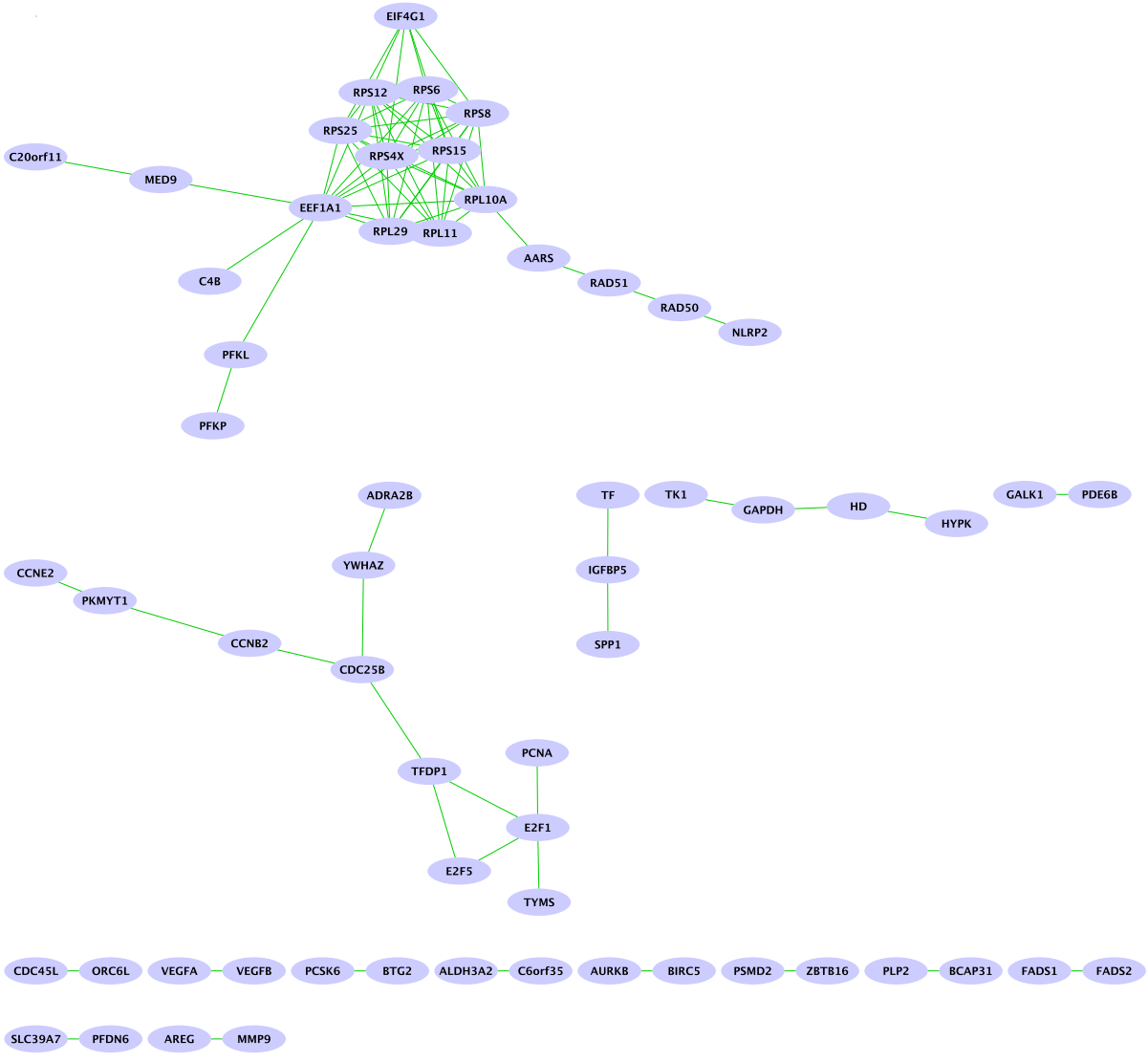
$$\Omega_2(\beta) = \sup_{\alpha \in \mathbb{R}^p : \forall i \sim j, \|\alpha_i^2 + \alpha_j^2\| \leq 1} \alpha^\top \beta \qquad \text{(Jacob et al., 2009)}$$

Lasso signature (accuracy 0.61)



Breast cancer prognosis, Jacob et al. (2009)

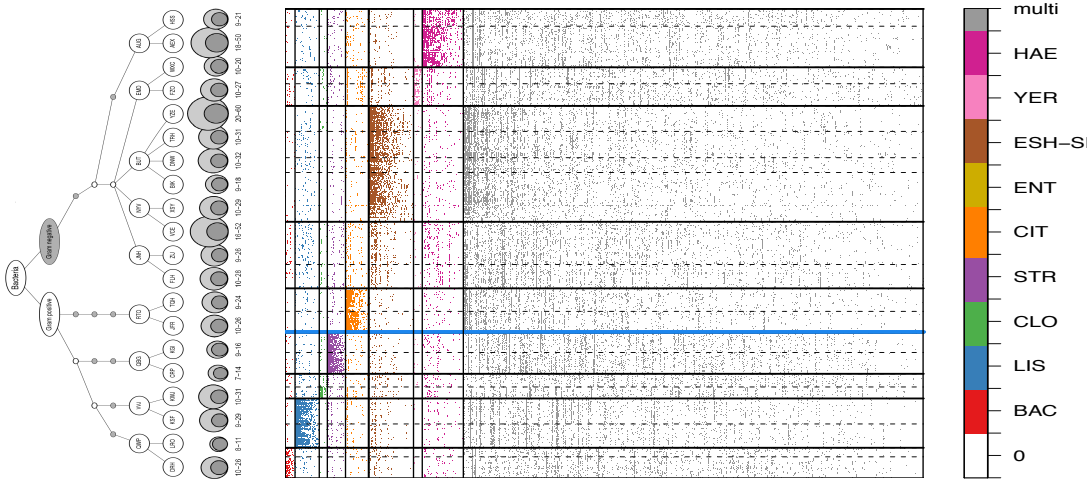
Graph Lasso signature (accuracy 0.64)



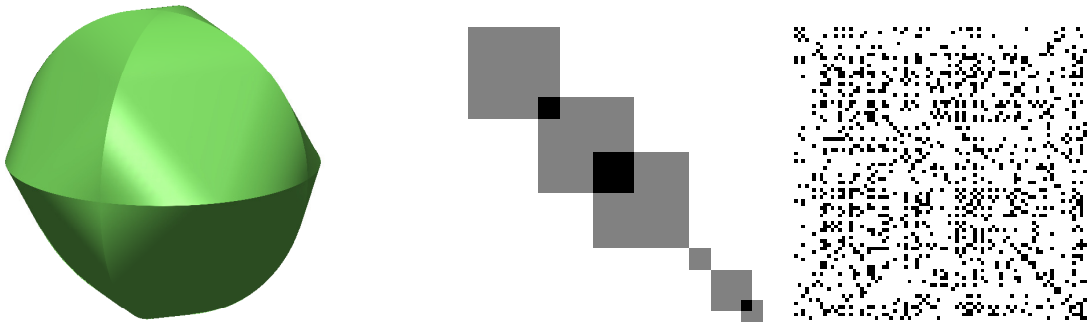
Breast cancer prognosis, Jacob et al. (2009)

Other atomic norms

- Disjoint feature selection for hierarchical classification (Vervier et al., 2014)

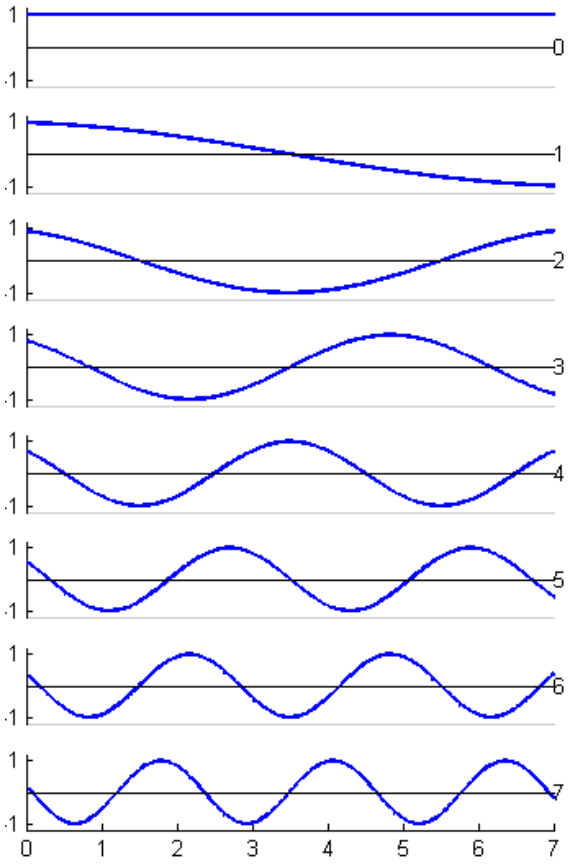


- Sparse low-rank matrices for sparse PCA and regression (Richard et al., 2014)



Graph smoothing penalty

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

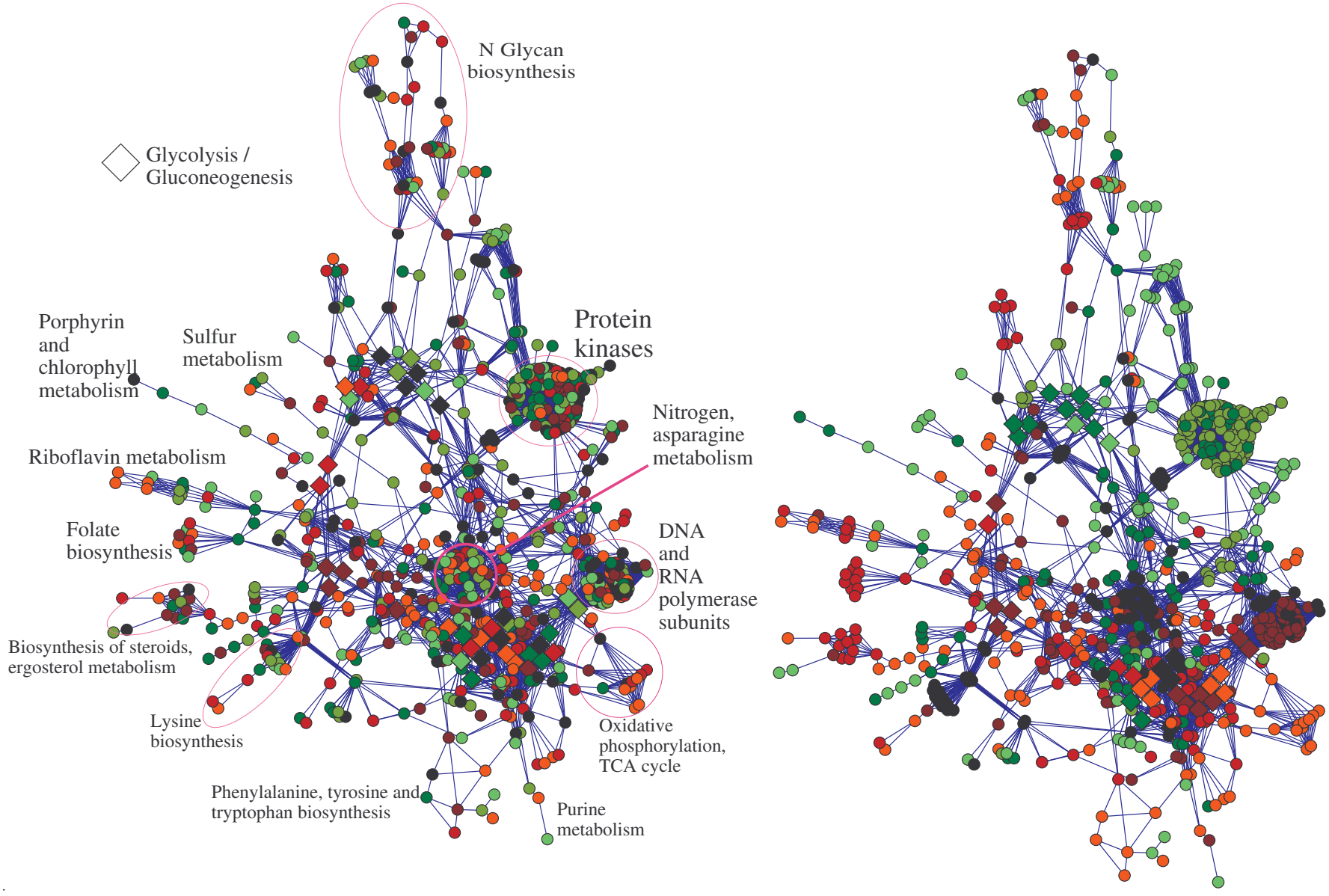


- Let $(e_i, \lambda_i)_{i=1,\ldots,p}$ the Fourier basis of the graph (eigenvectors of the Laplacian)
- Learning with $\Omega_G(\beta)$ on data x is the same as learning with $\Omega(\beta) = \|\beta\|^2$ on the smoothed data $\Phi(x)$:

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

- See (Rapaport et al., 2007) for other variants

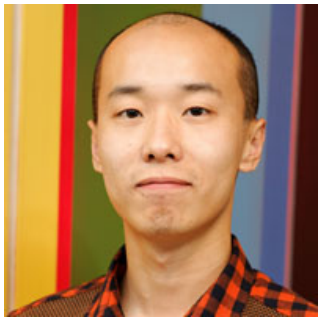
Classifiers



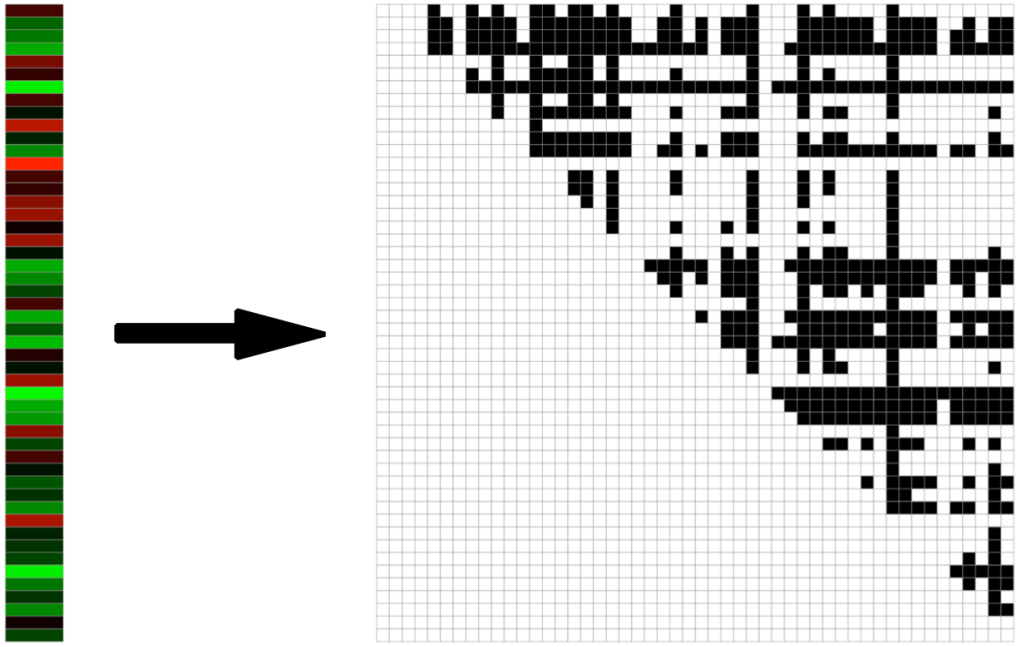
Another representation: ranking

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

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



Yunlong Jiao

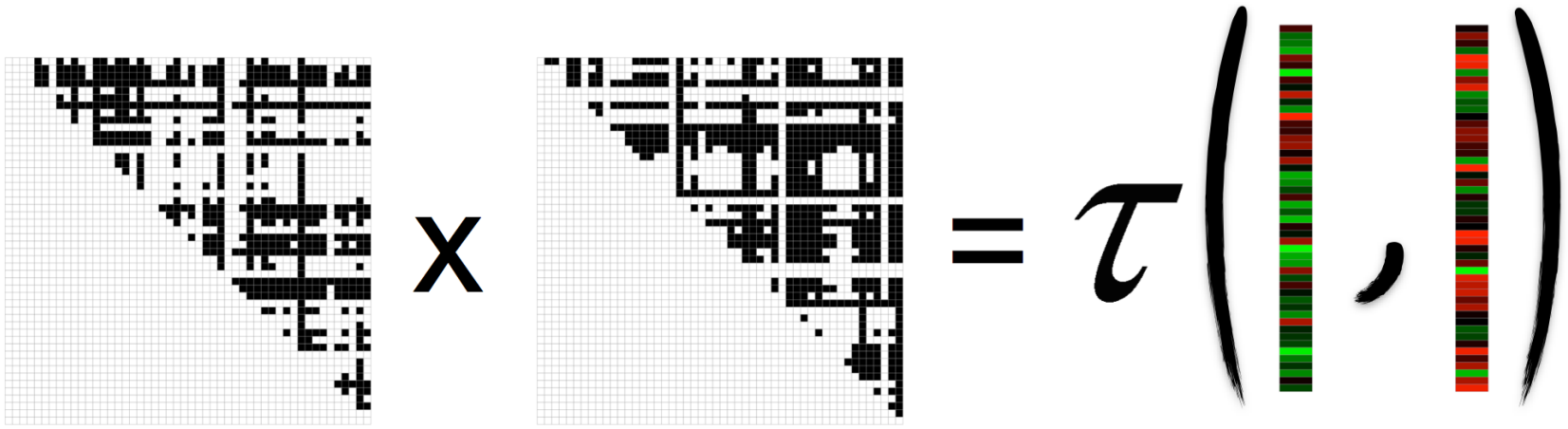


One sample x
p features

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

Kernel trick

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



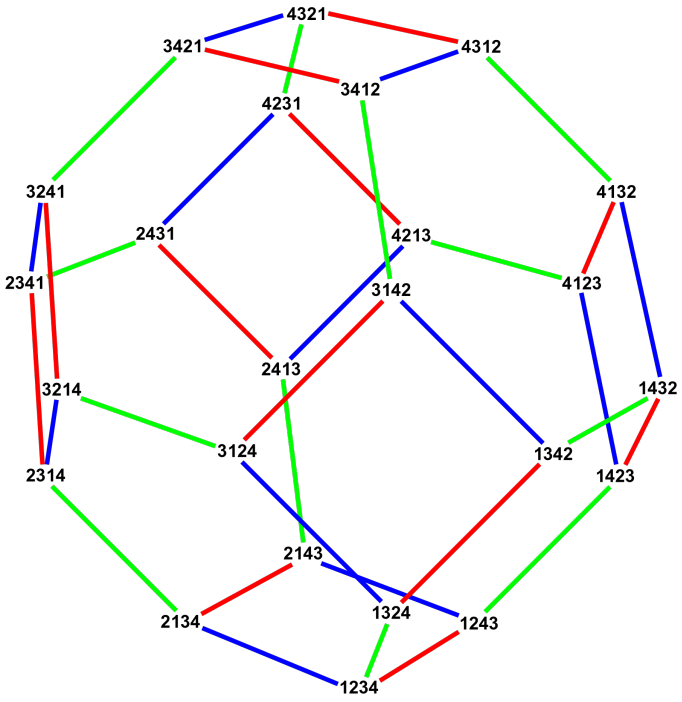
O(p^2)

O(p log(p))

Theorem ((Jiao and Vert, 2015))

The Kendall and Mallows kernels are positive definite and can be evaluated in $O(p \log p)$ time.

Related work



Cayley graph of $\mathbb{S}_4$

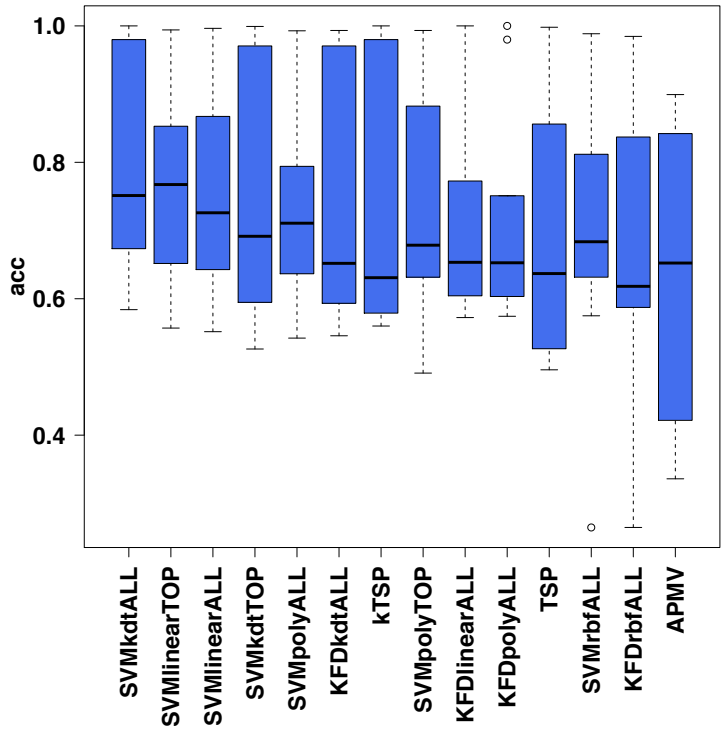
- Kondor and Barbarosa (2010) proposed the diffusion kernel on the Cayley graph of the symmetric group generated by adjacent transpositions.
- Computationally intensive ($O(p^p)$)
- Mallows kernel is written as

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

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

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

Application



Dataset	No. of features	No. of samples (training/test)	
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)

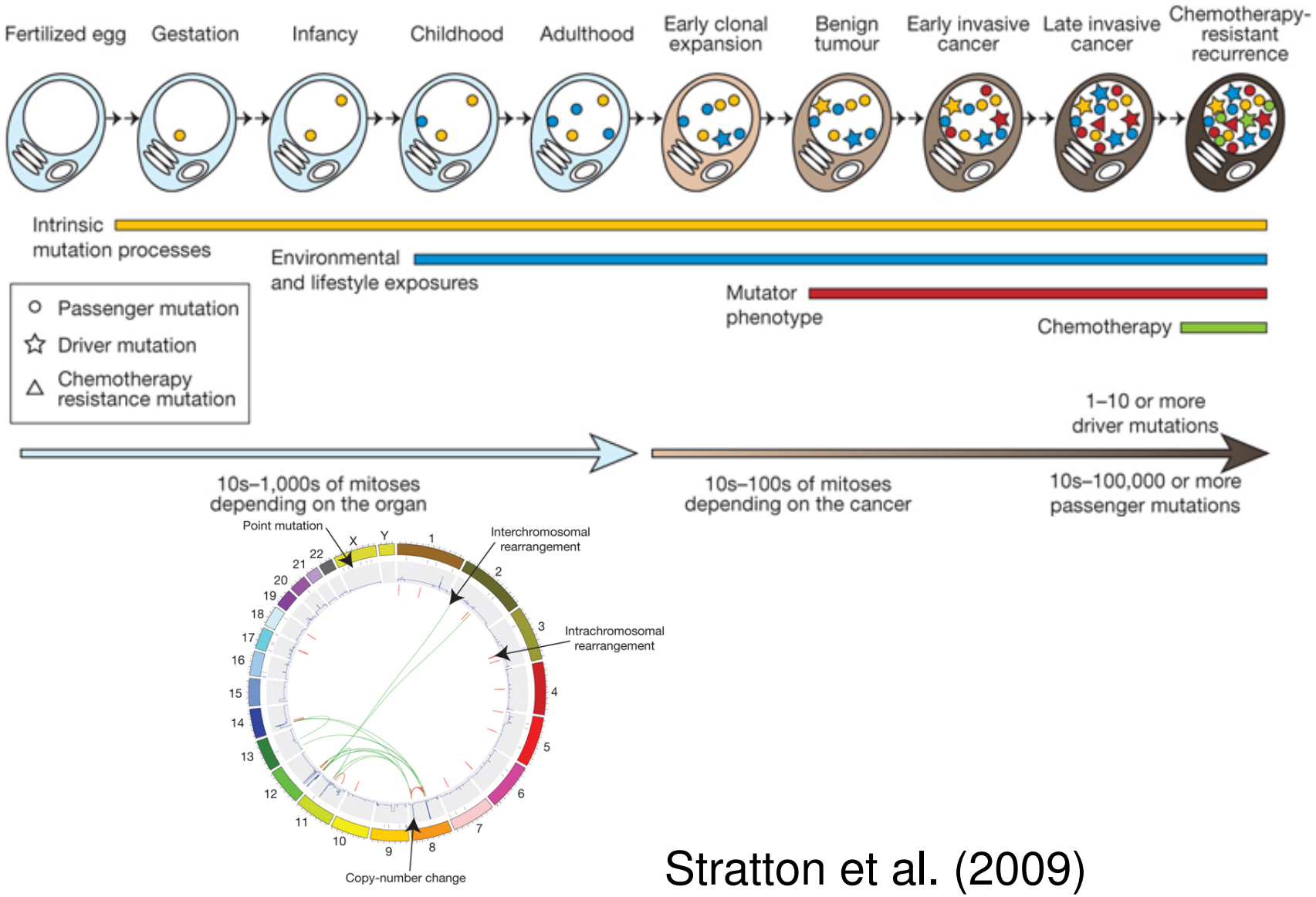
Outline

Joint work with

Marine Le Morvan

Andrei Zinovyev

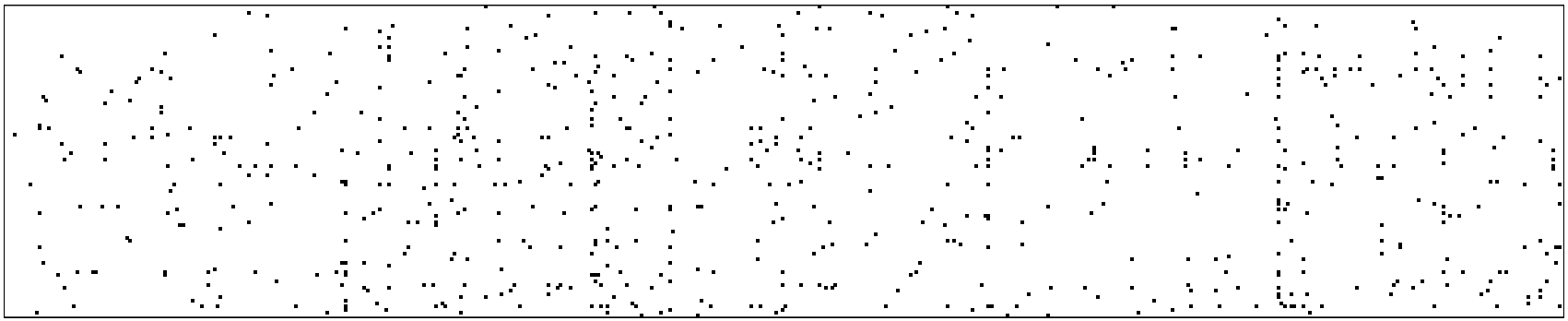
Somatic mutations in cancer



Stratton et al. (2009)

Large-scale efforts to collect somatic mutations

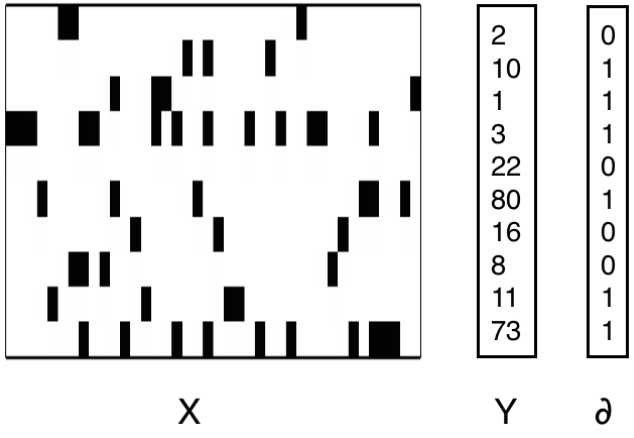
- 3,378 samples with survival information from 8 cancer types
- downloaded from the TCGA / cBioPortal portals.

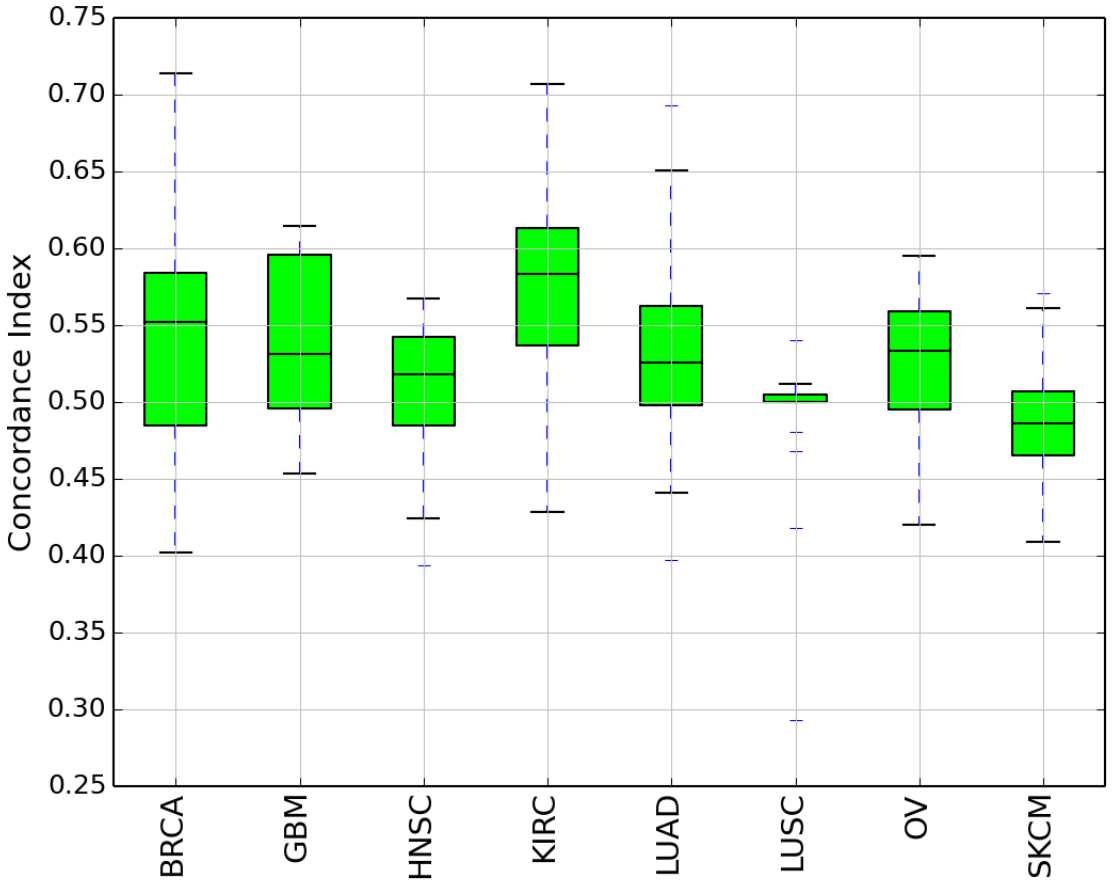


Cancer type	Patients	Genes
LUAD (Lung adenocarcinoma)	430	20 596
SKCM (Skin cutaneous melanoma)	307	17 463
GBM (Glioblastoma multiforme)	265	14 750
BRCA (Breast invasive carcinoma)	945	16 806
KIRC (Kidney renal clear cell carcinoma)	411	10 609
HNSC (Head and Neck squamous cell carcinoma)	388	17 022
LUSC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

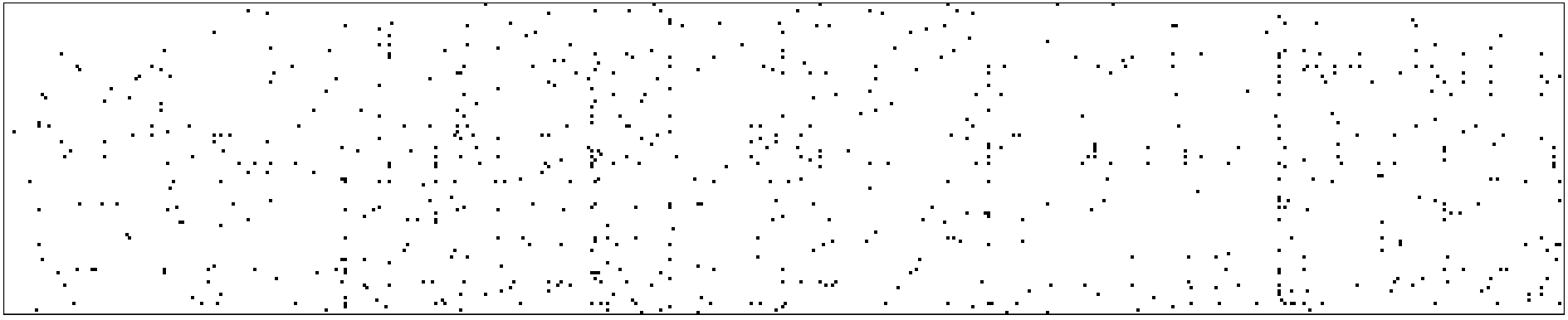
Survival prediction from raw mutation profiles

- 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





Changing the representation?



Can we replace

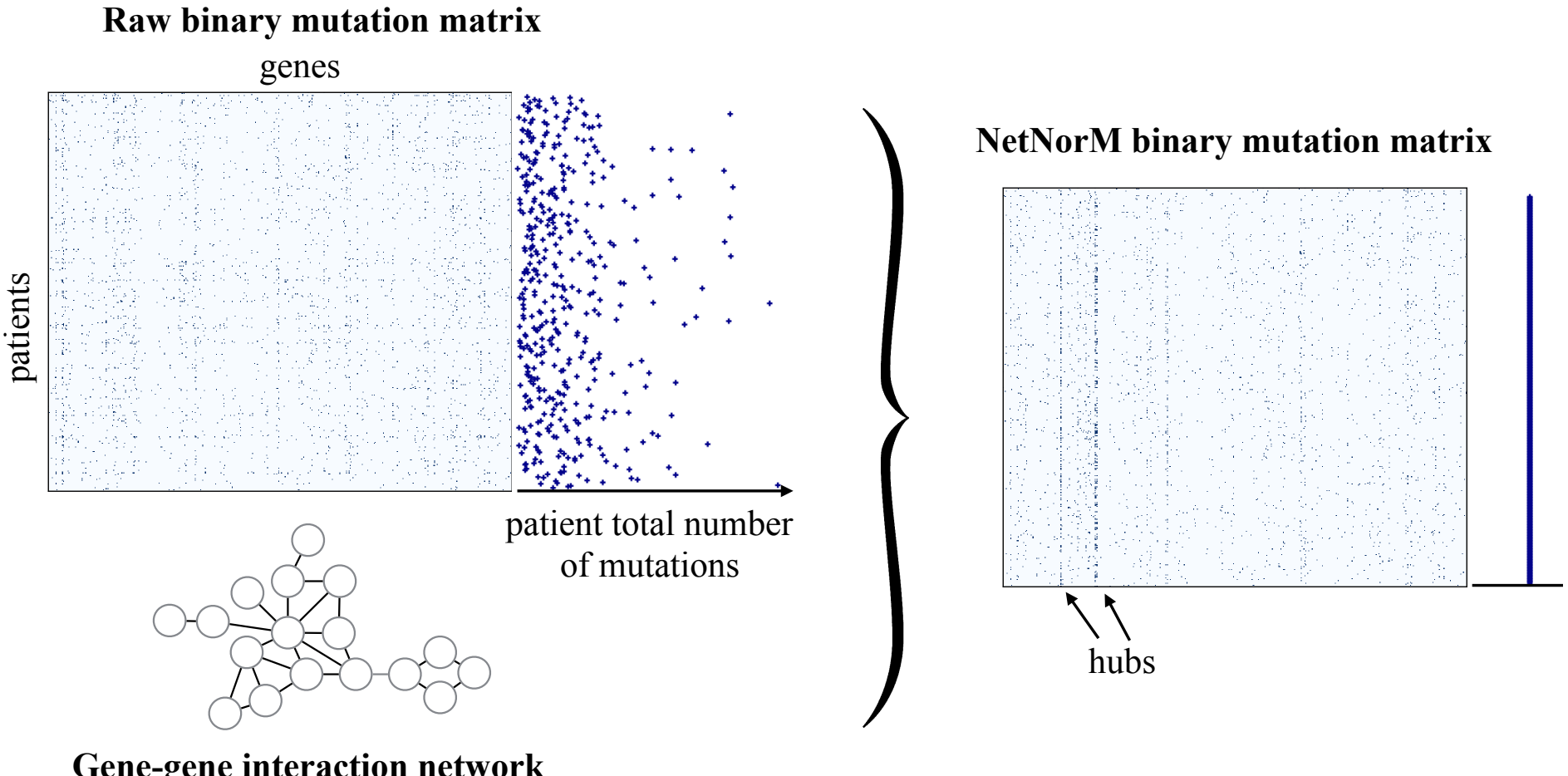
$$x \in \{0,1\}^p \quad \text{with } p \text{ very large, very sparse}$$

by a representation with more information shared between samples

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

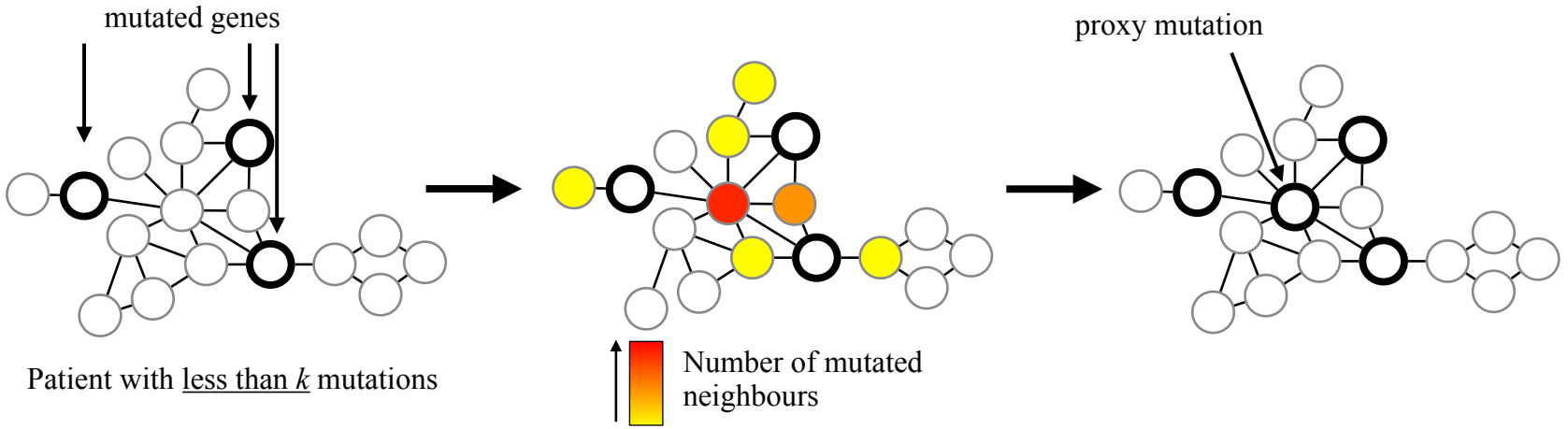
NetNorm Overview (Le Morvan et al., 2016)

- 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

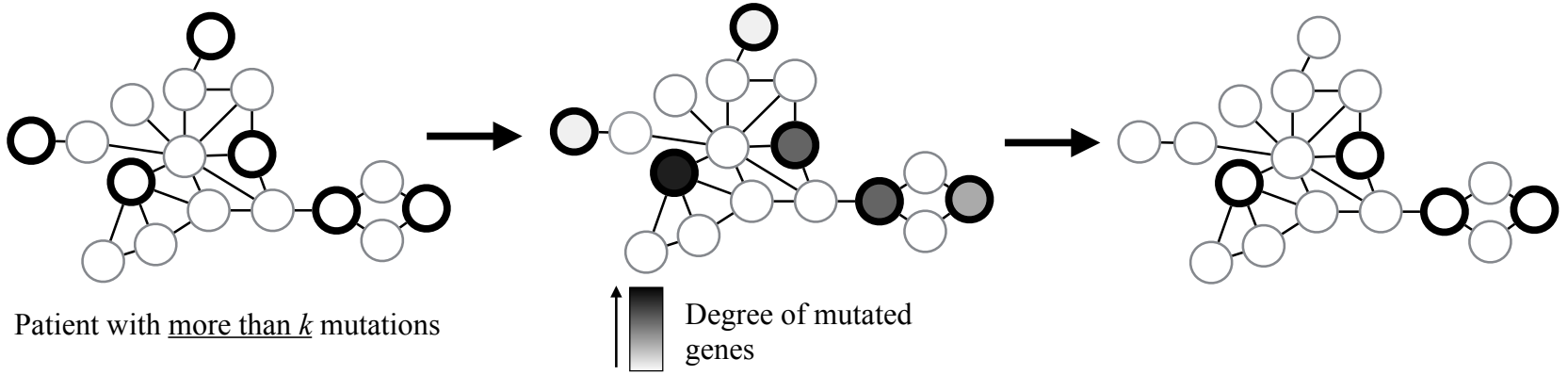


NetNorm detail (k=4)

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



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



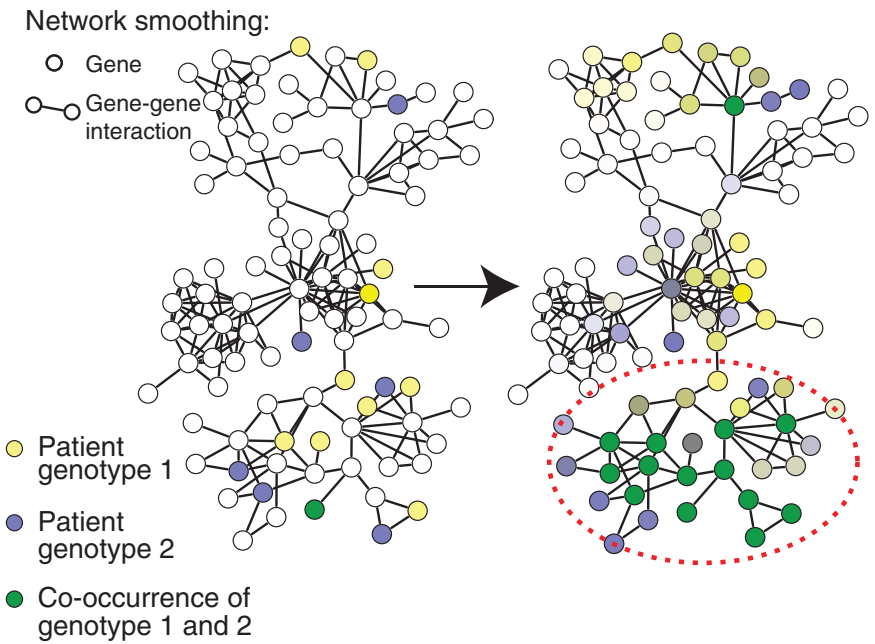
Related work (Hofree et al., 2013)

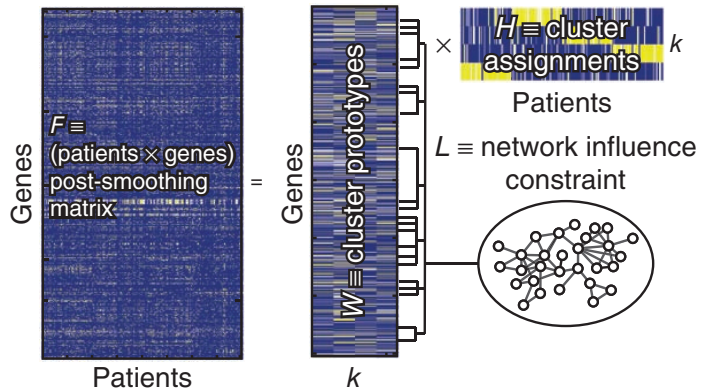
Network-based stratification of tumor mutations

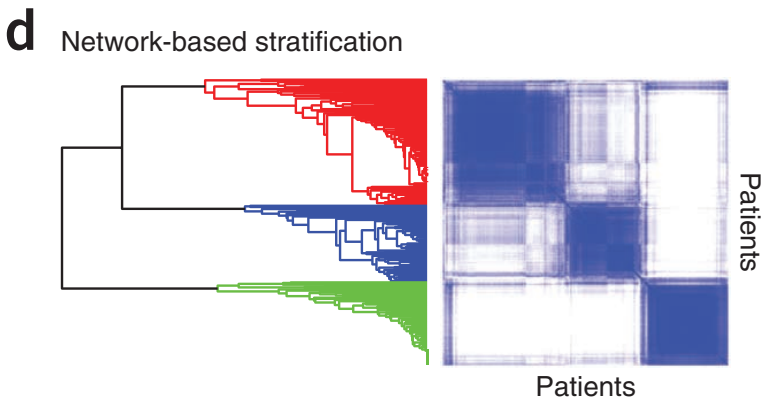
Matan Hofree[1], John P Shen[2], Hannah Carter[2], Andrew Gross[3] & Trey Ideker[1–3]

[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).

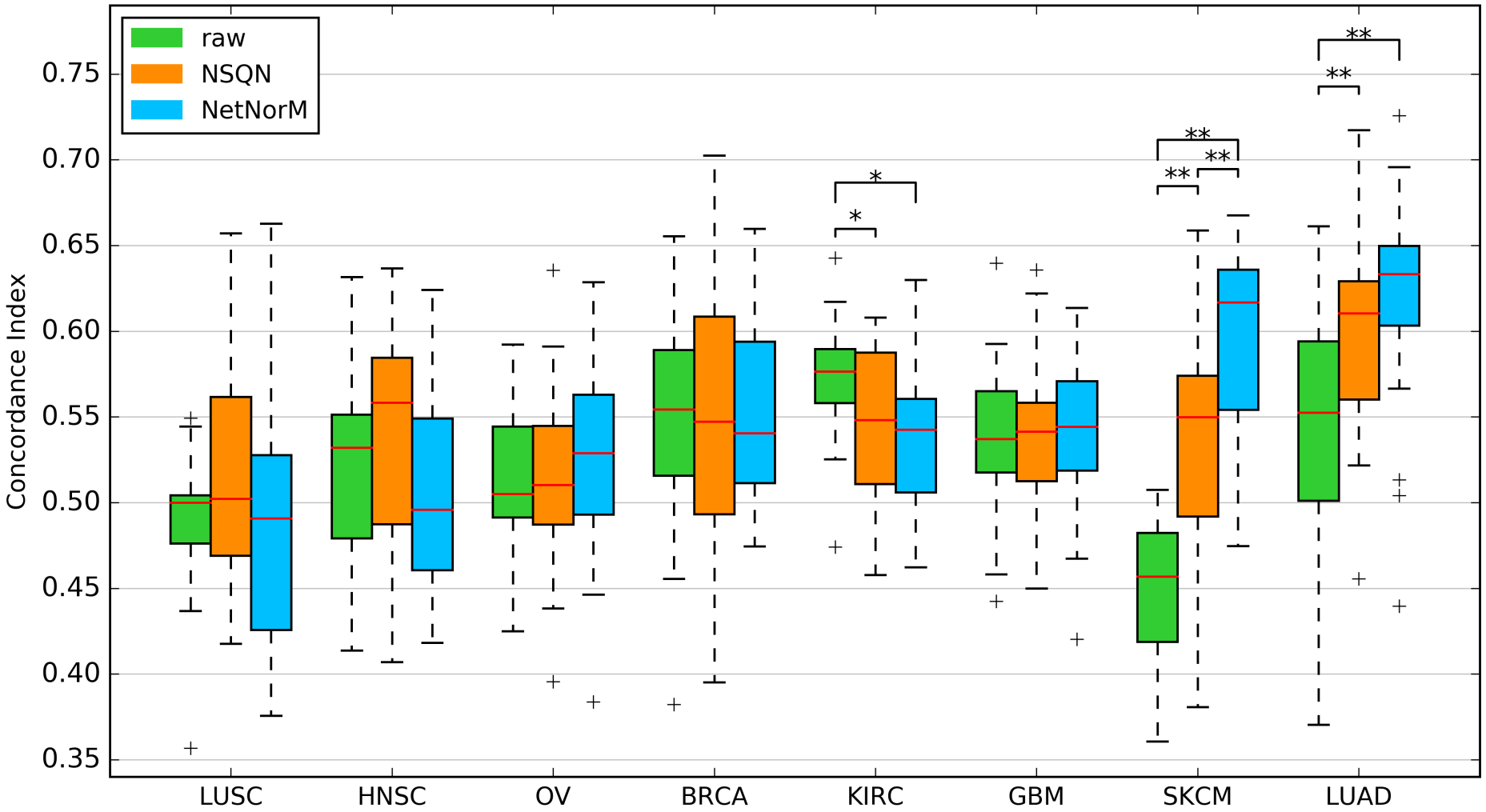
RECEIVED 14 FEBRUARY; ACCEPTED 12 AUGUST; PUBLISHED ONLINE 15 SEPTEMBER 2013; DOI:10.1038/NMETH.2651







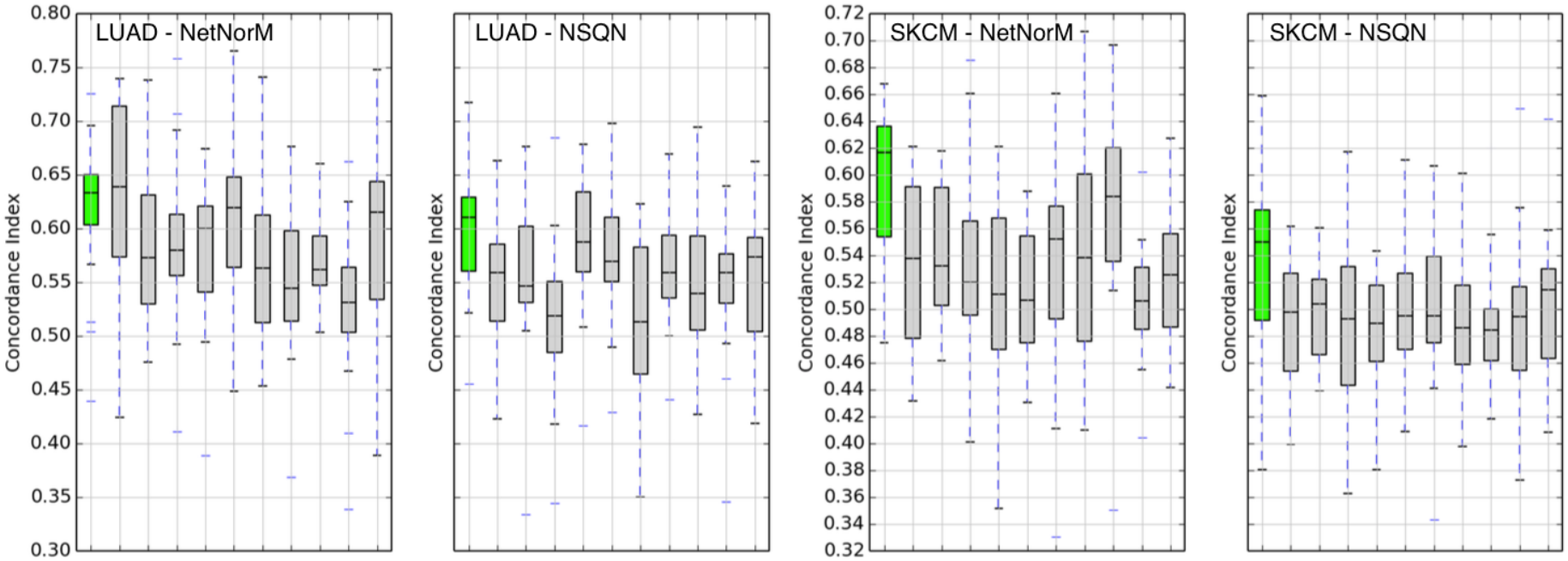
Performance on survival prediction



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):

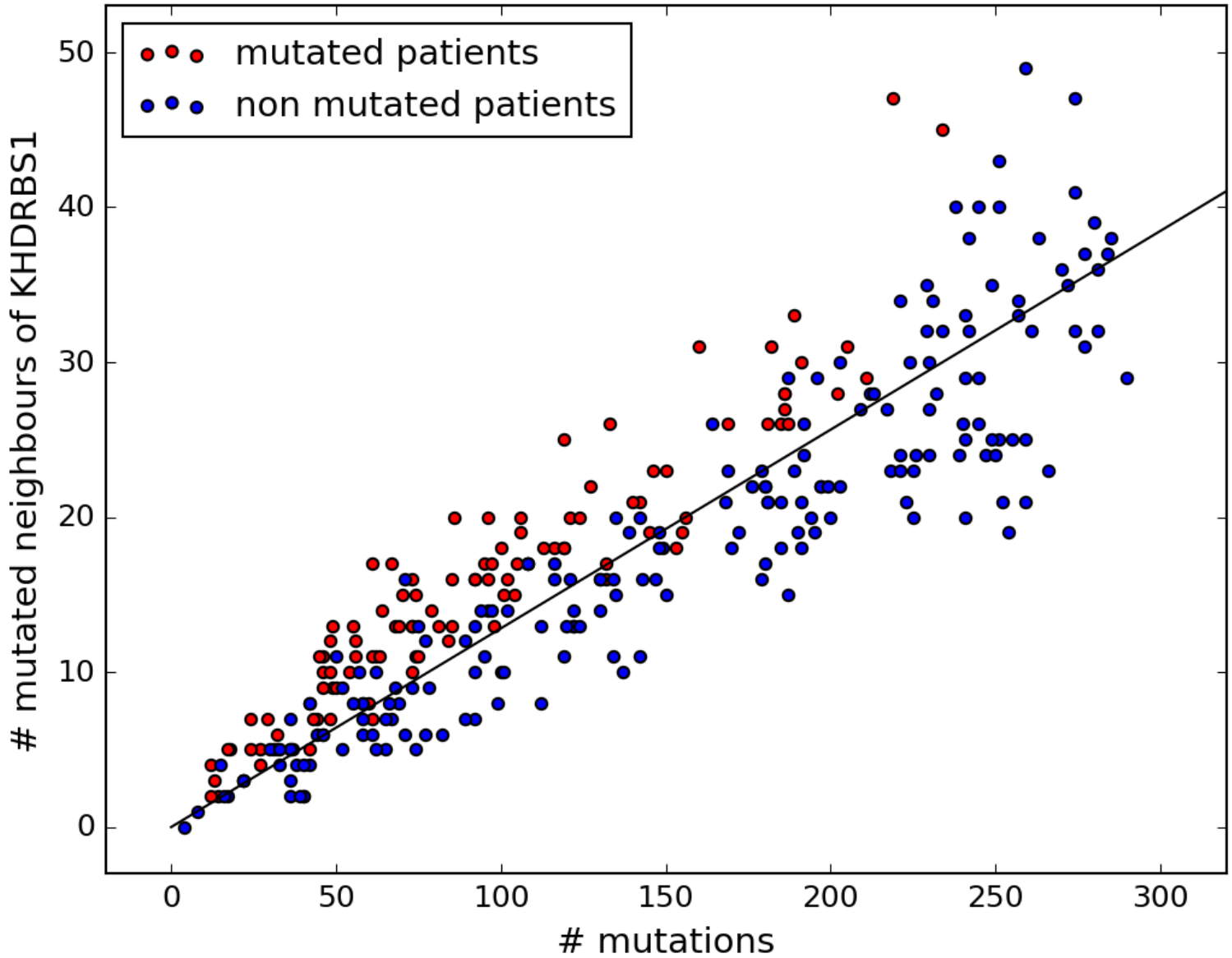
	NSQN	NetNorM
LUAD	2×10^{-3}	3.5×10^{-2}
SKCM	1.2×10^{-2}	1×10^{-4}

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

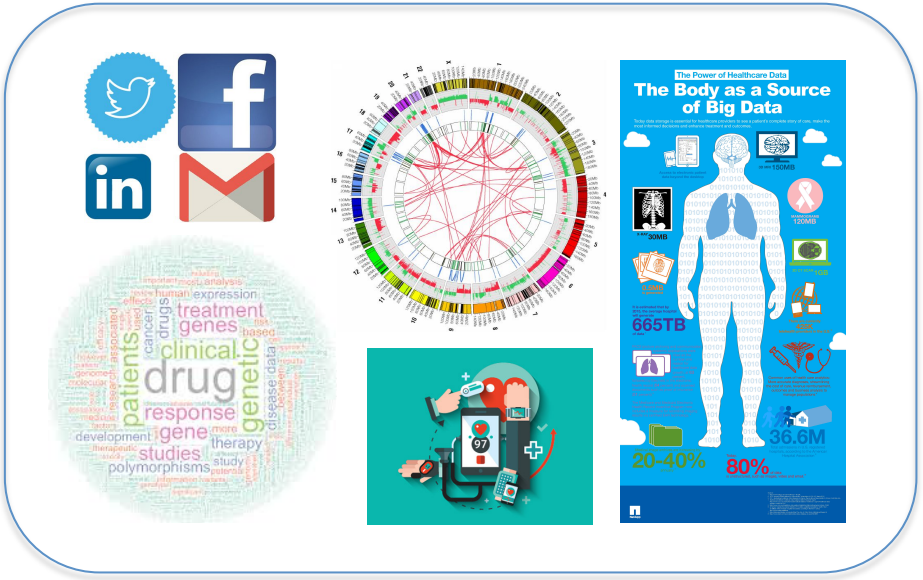
Proxy mutations encode local mutational burden



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

Outline

Conclusion



- Many new exciting 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, ...)
- Some problems seem inherently complicated
- Big data analytics will help, but is certainly not a magic bullet

Thanks

References

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.

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.

References (cont.)

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.

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.

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.

E. Richard, G. Obozinski, and J.-P. Vert. Tight convex relaxations for sparse matrix factorization. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Adv. Neural. Inform. Process Syst.*, volume 27, pages 3284–3292. Curran Associates, Inc., 2014. URL https://papers.nips.cc/paper/5408-tight-convex-relaxations-for-sparse-matrix-factorization.

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.

References (cont.)

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.

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.