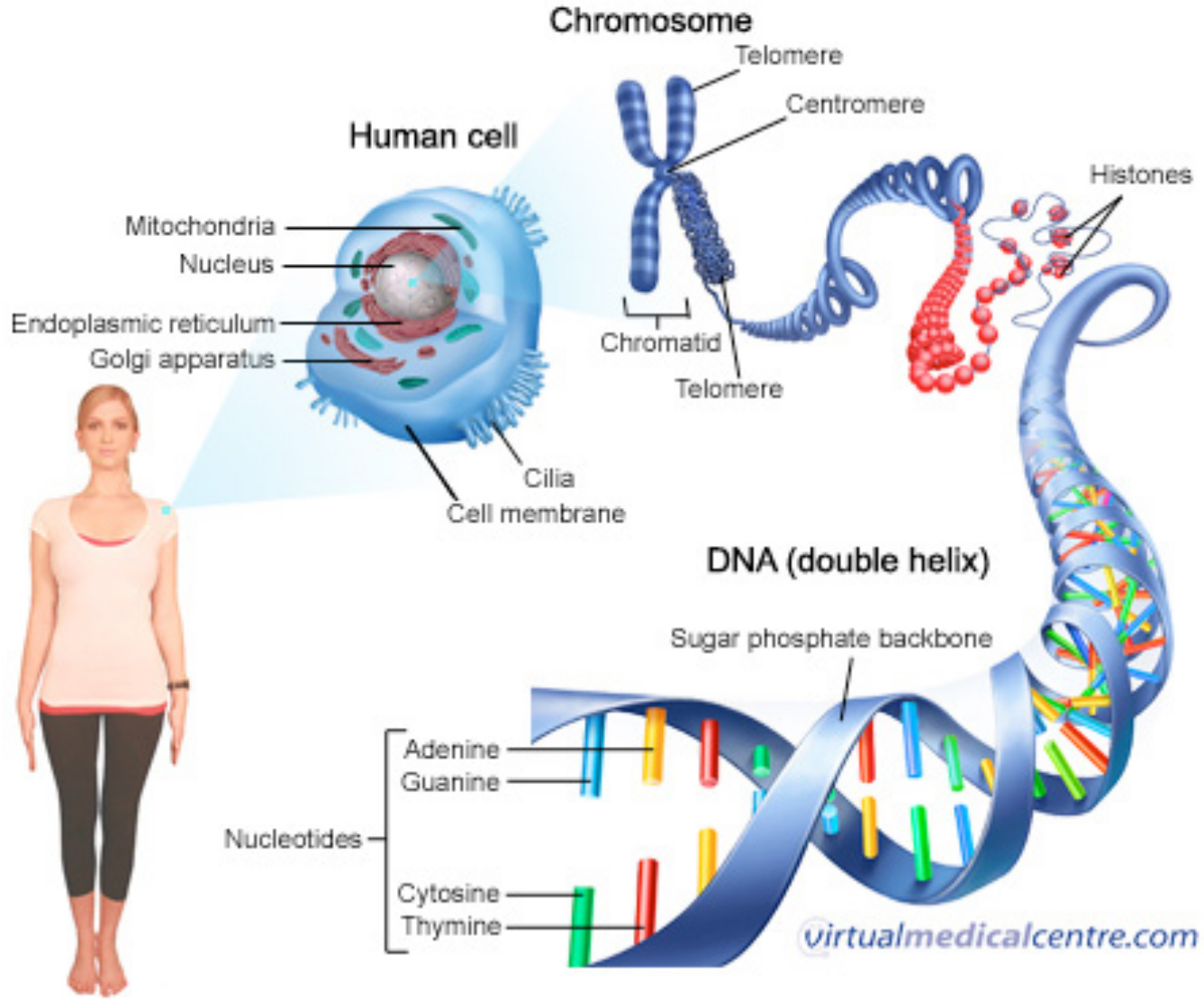
Machine Learning for Personalized Genomics

Jean-Philippe Vert

Paris-Saclay Center for Data Science kick-off meeting
LAL, Saclay, June 30, 2014

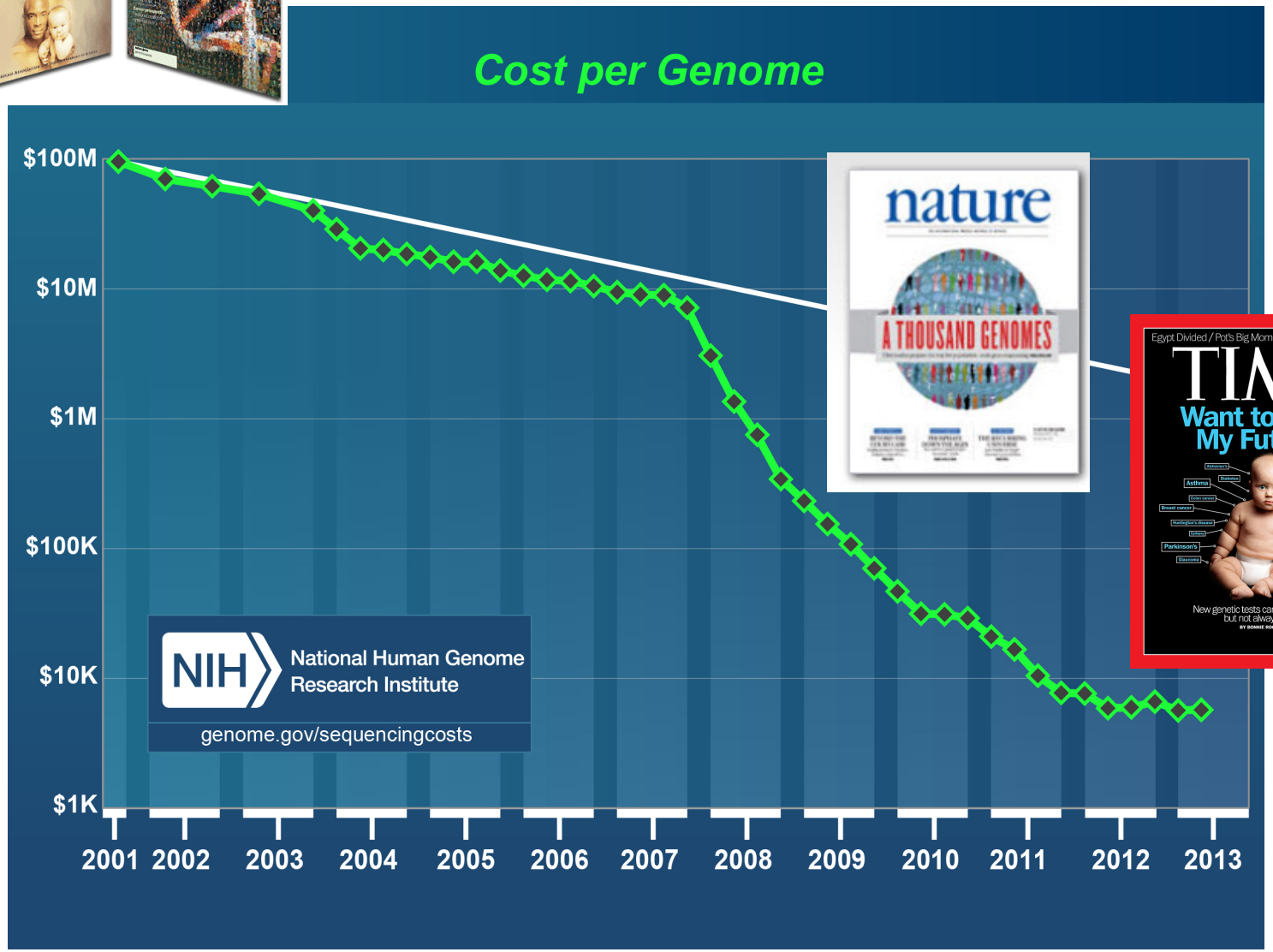
1 body = 100 trillions cells
1 cell = 6 billions ACGT in DNA



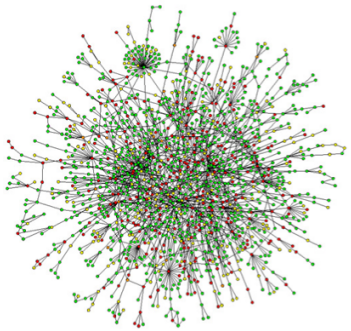
Human genome project (1990-2003)

- Goal: sequence the 3,000,000,000 base pairs of the human genome
- Consortium of 20 laboratories, 6 countries
- 13 years, $3,000,000,000

The ***second*** revolution



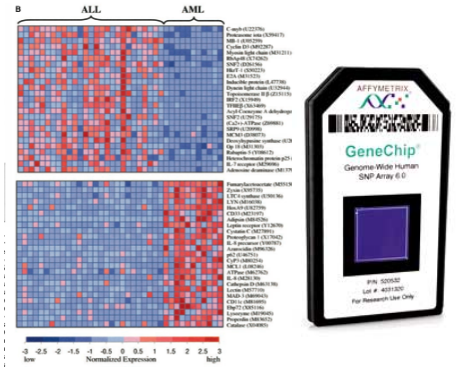
A flood of ***omics*** data



Interactome

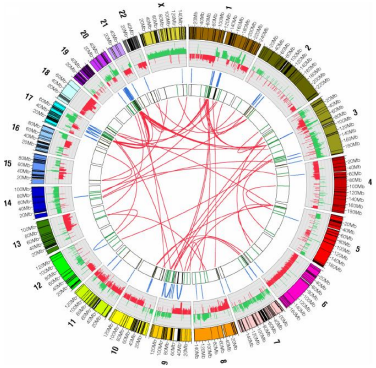
PubMed

Publications



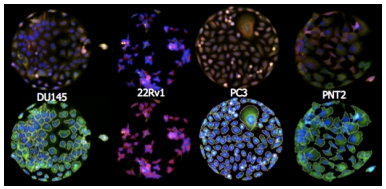
Transcriptome

Genome

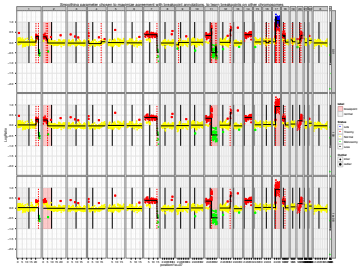


Mutations

Structural variations

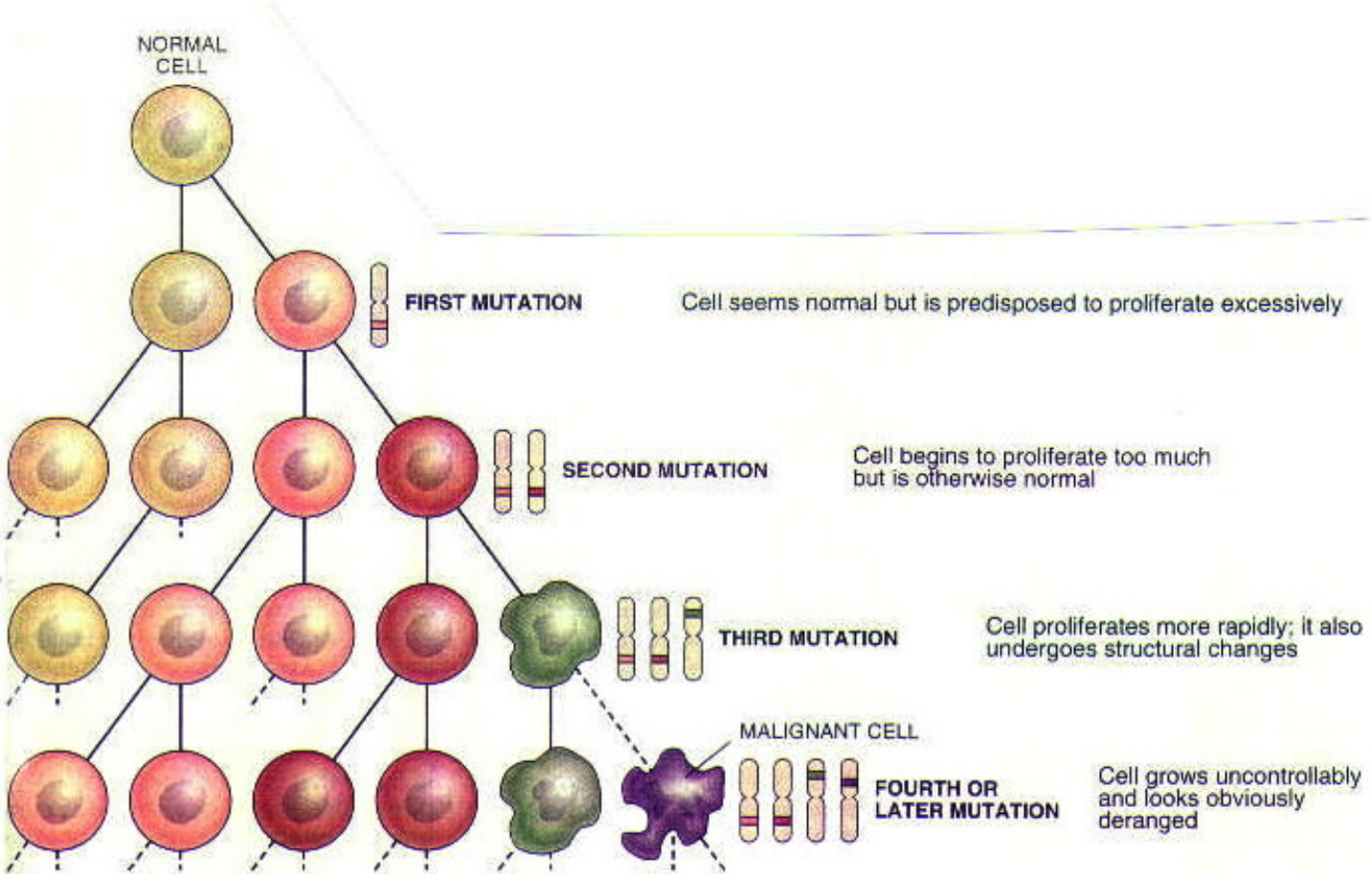


Phenome

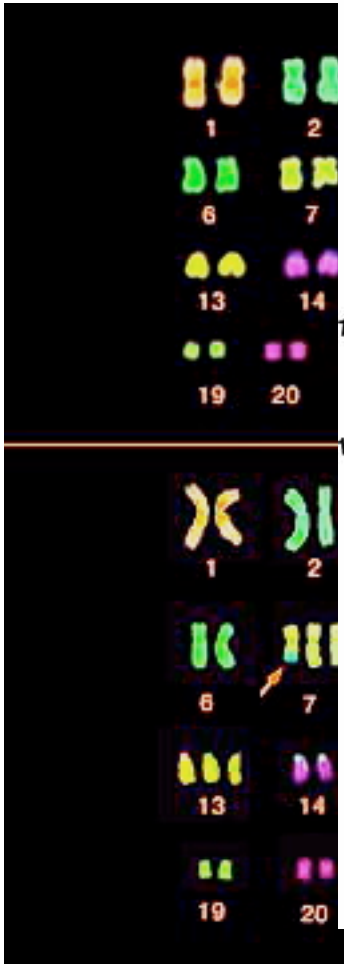


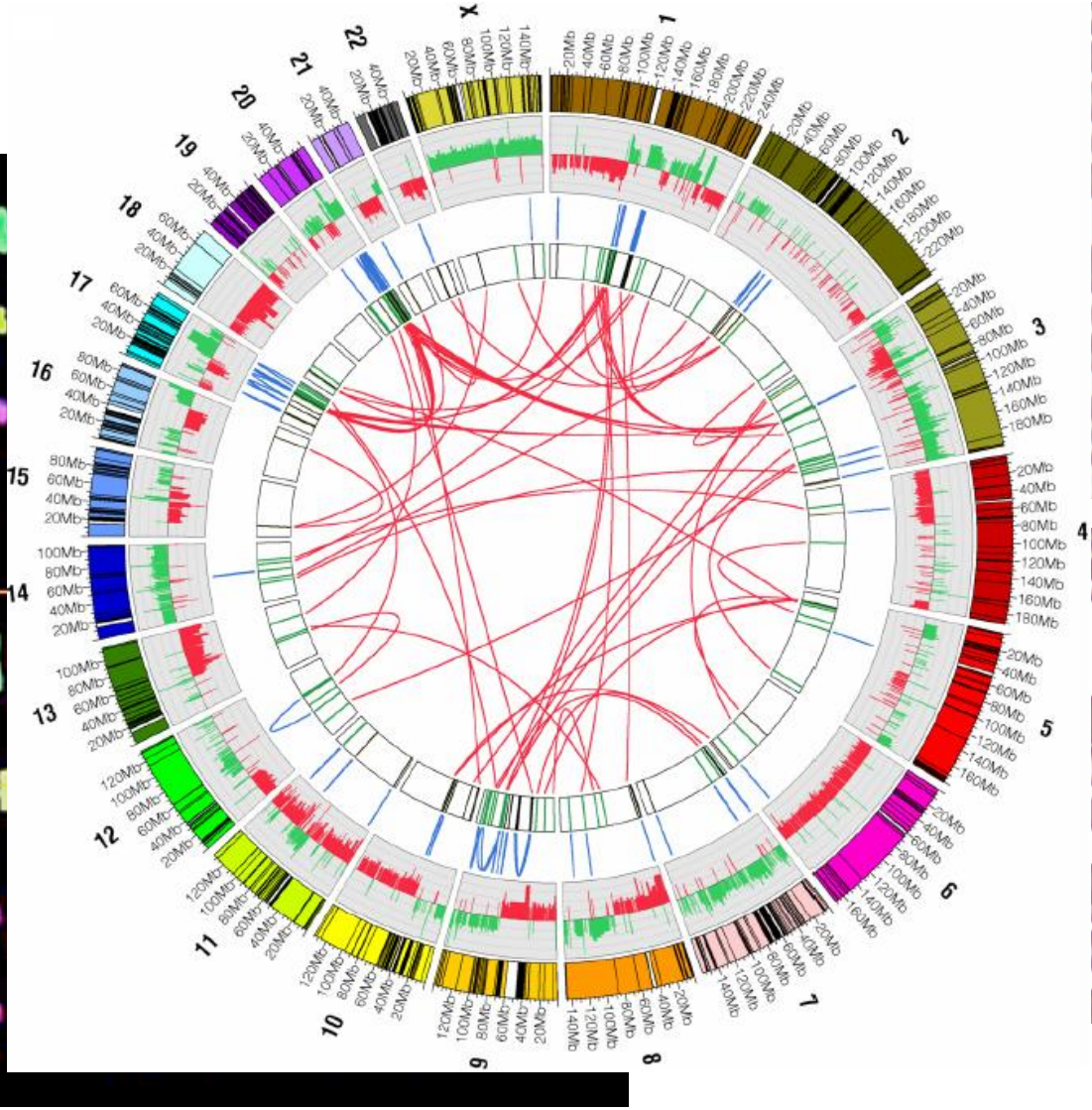
Epigenome

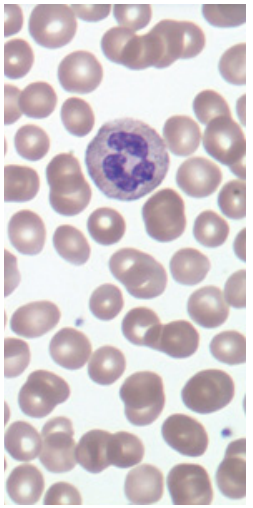
All cancers are different

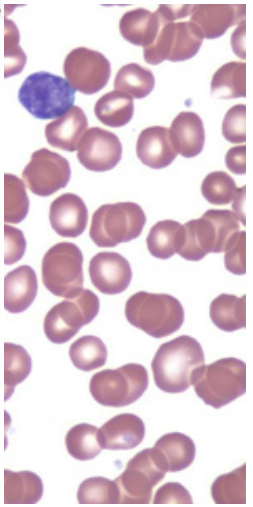


Cancer: different views



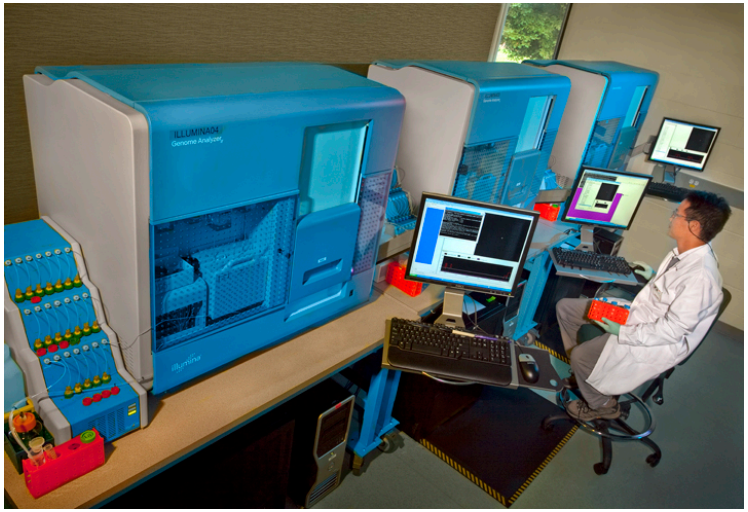


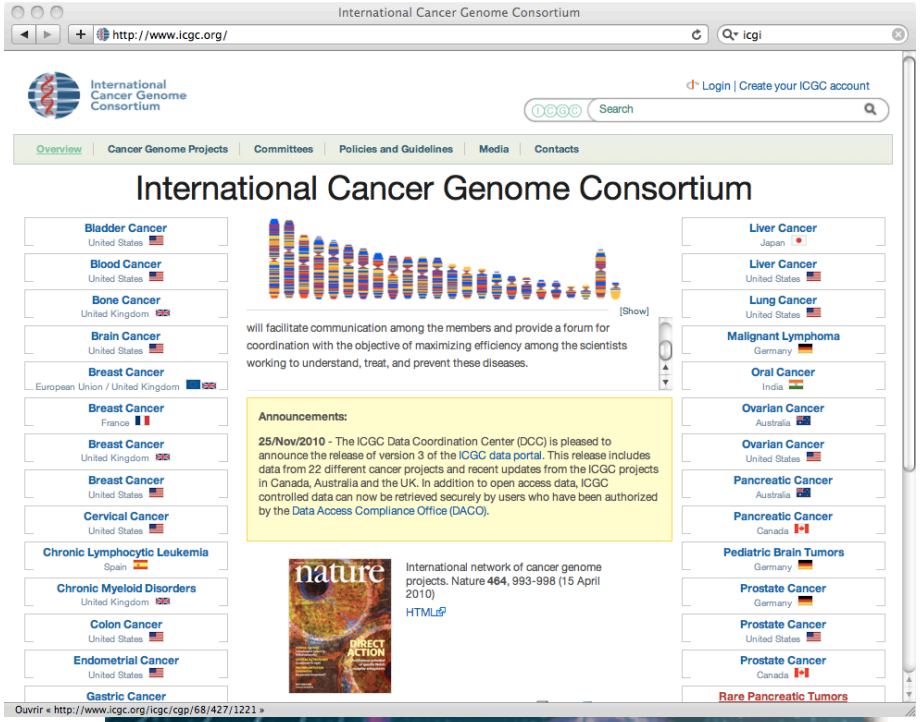




Big data!

- http://aws.amazon.com/1000genomes/

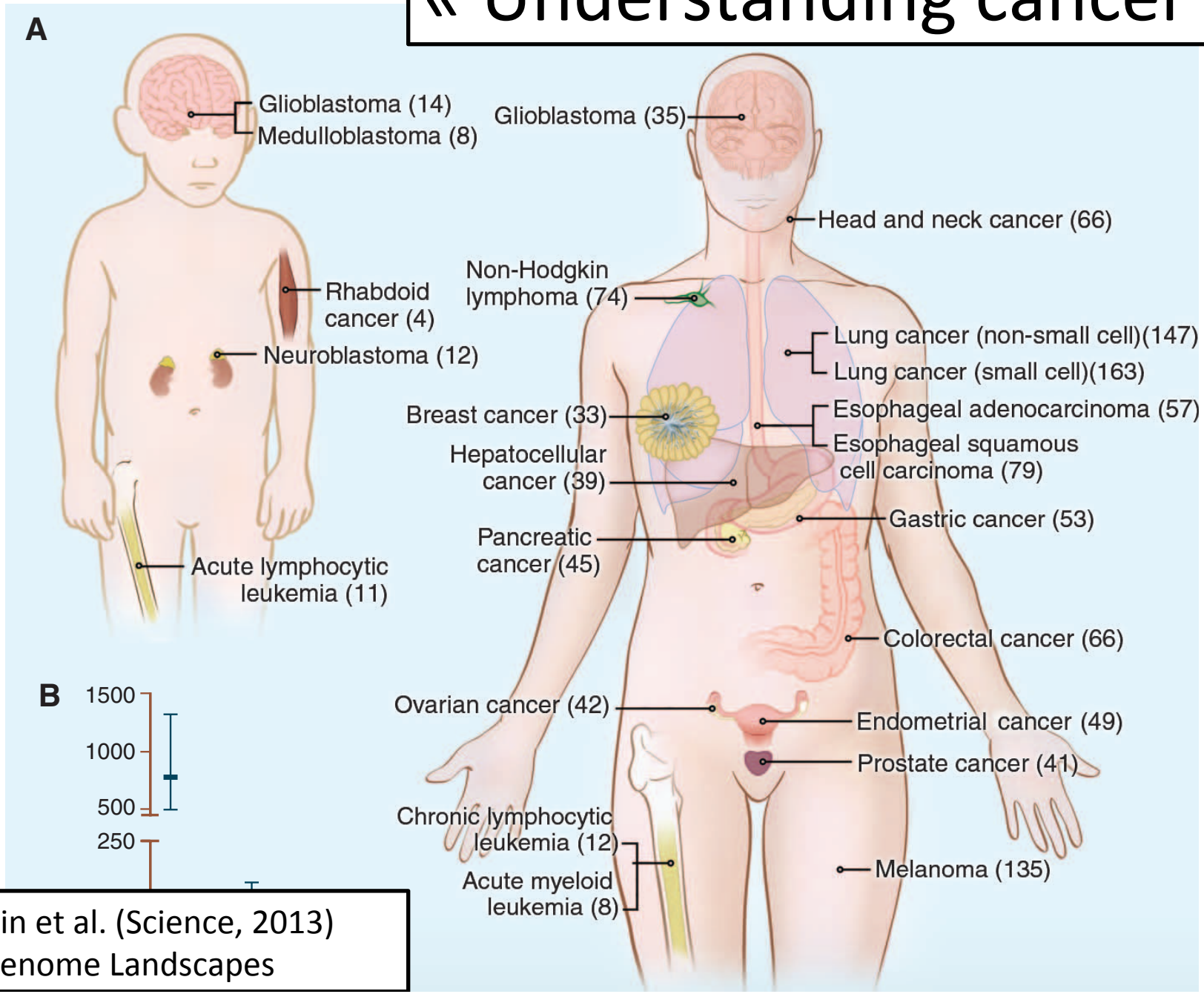




Opportunities

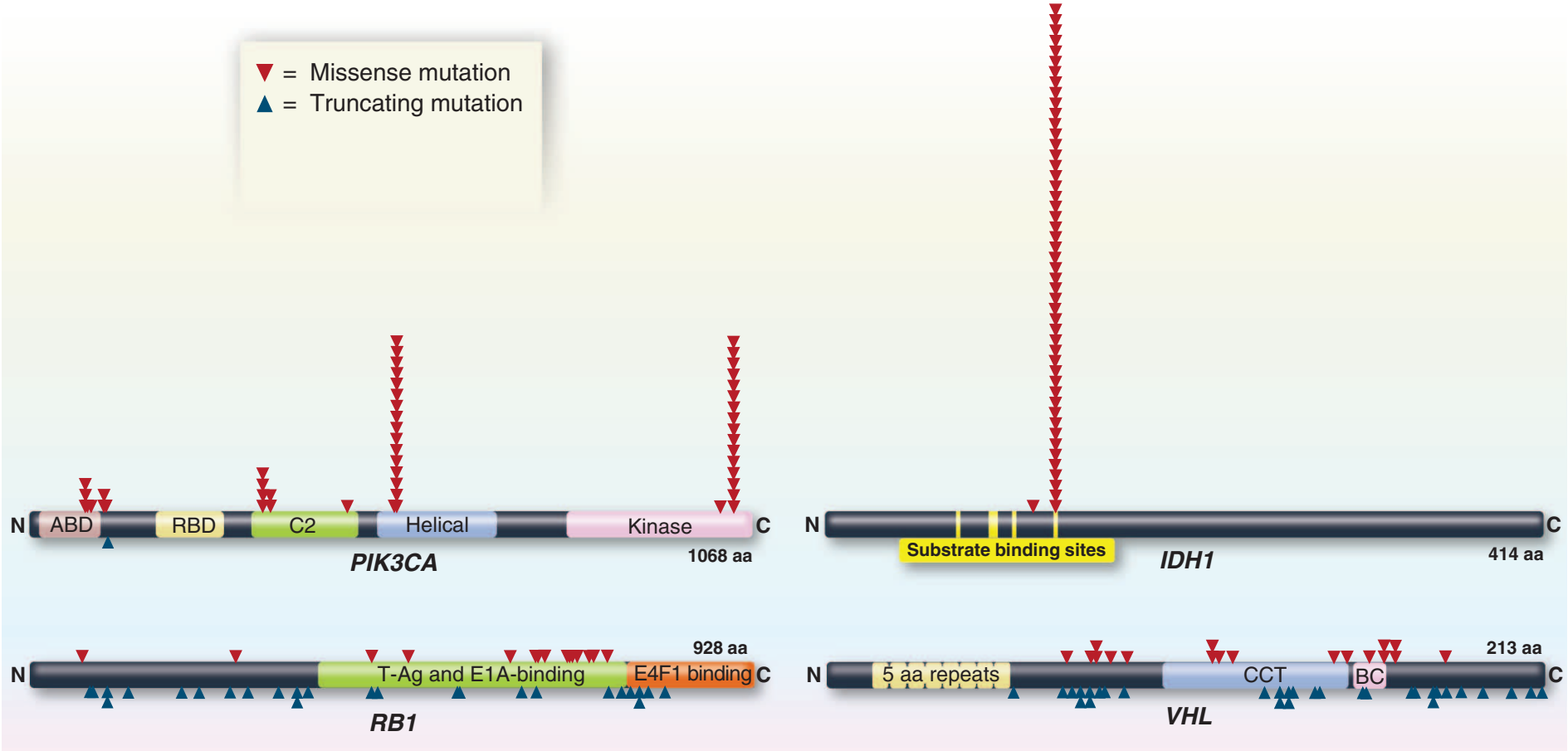
- **New drug targets and therapies**
  - By analyzing specificities of cancer cells at the molecular level
- **Precision medicine**
  - By developing predictive models for diagnosis, prognosis, response to drugs…

« Understanding cancer »



Vogelstein et al. (Science, 2013)
Cancer genome Landscapes

Finding « cancer genes »



Vogelstein et al. (Science, 2013)
Cancer genome Landscapes

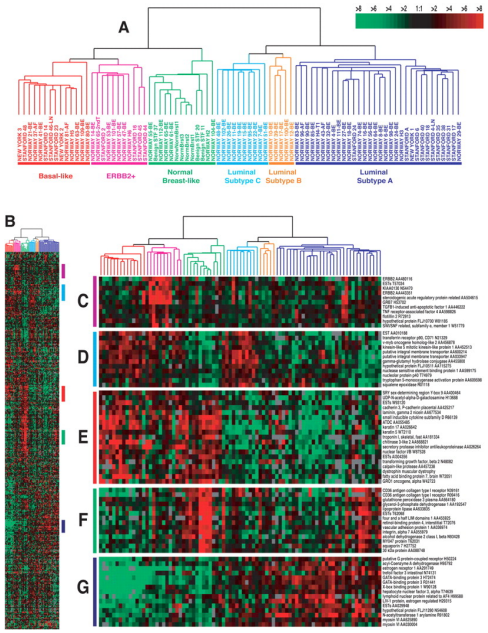
23andMe

Predictive

P4 Medicine

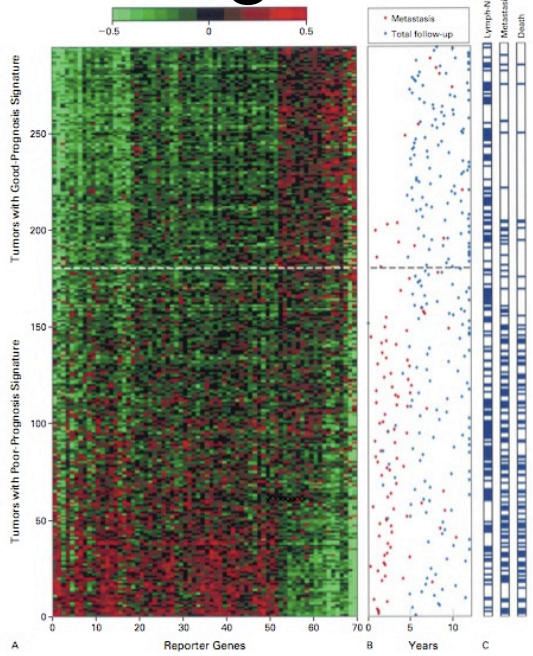
PREDICT • PREVENT • PERSONALIZE • PARTICIPATE

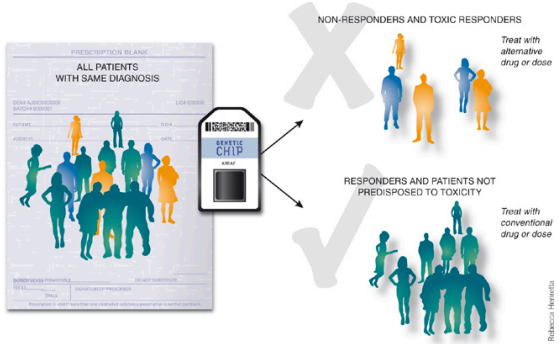
Opportunities



Diagnosis

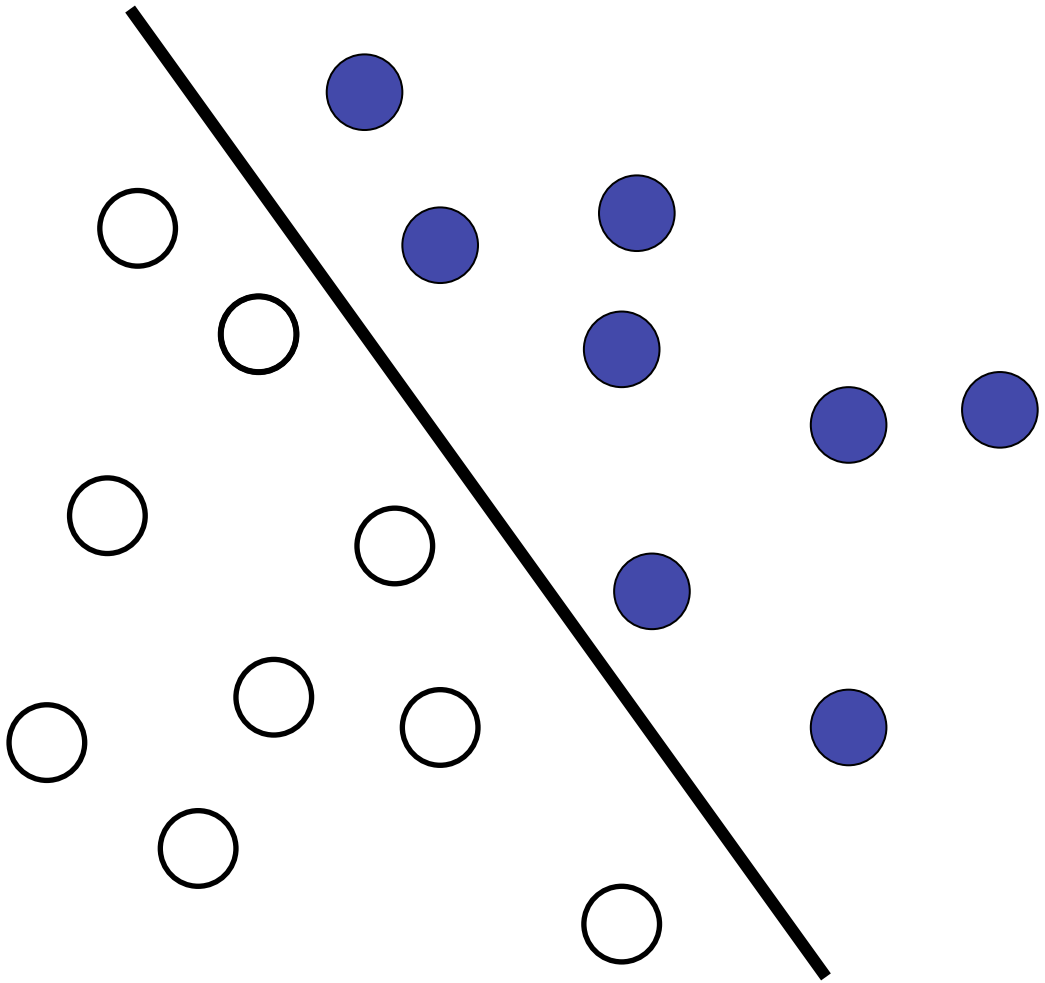
Prognosis





Response to drugs

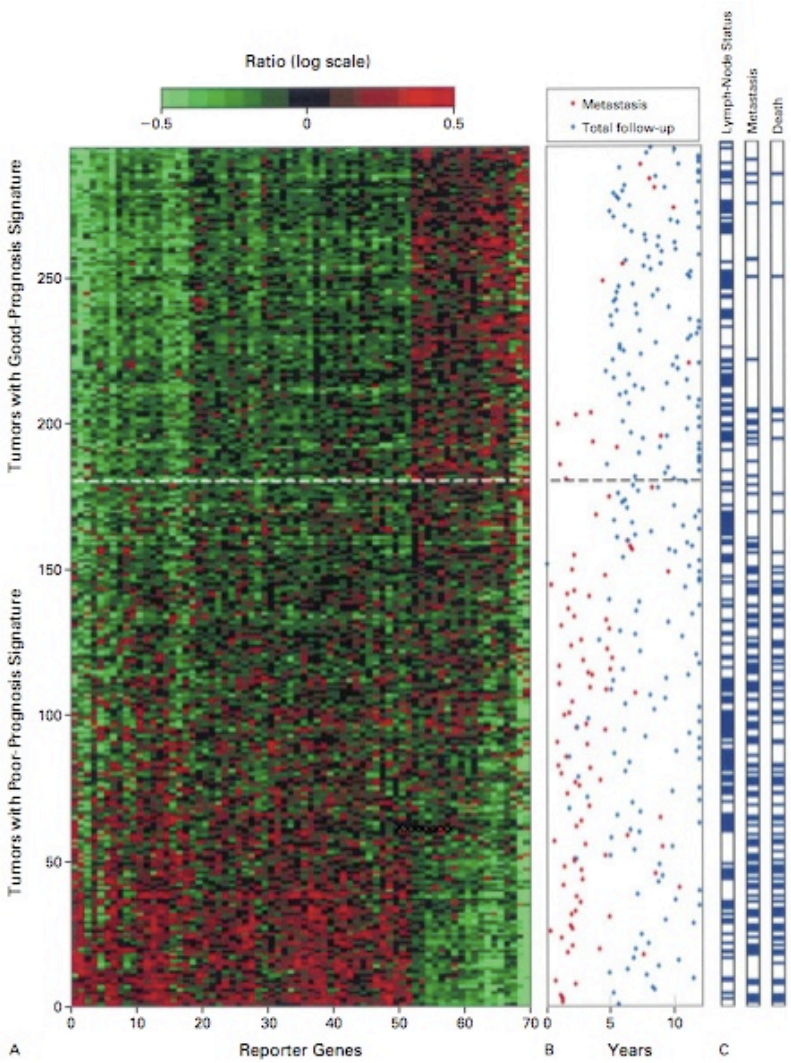
Supervised machine learning

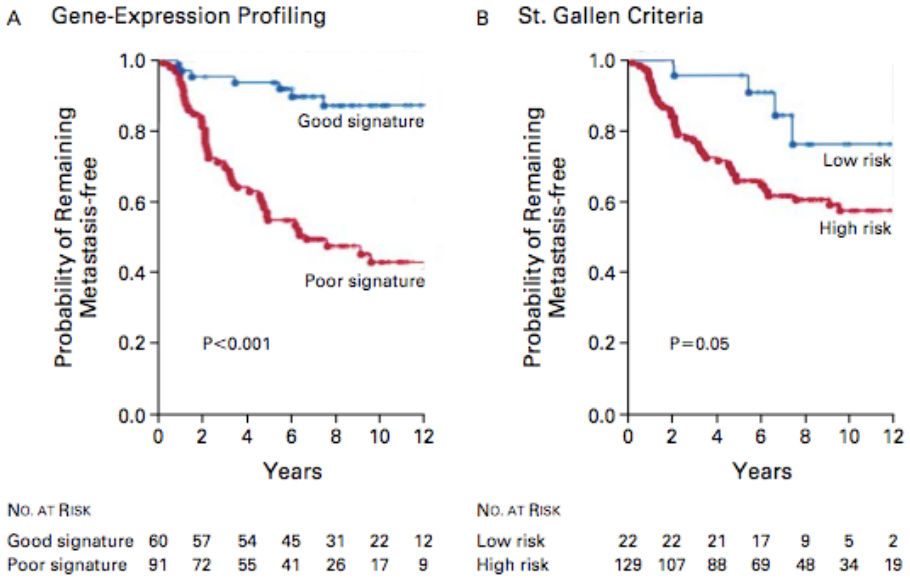


Challenges

- **High dimension**
- **Few examples**
- Structured data
- Efficient algorithms
- Interpretability

Example:
Breast cancer prognostic signature



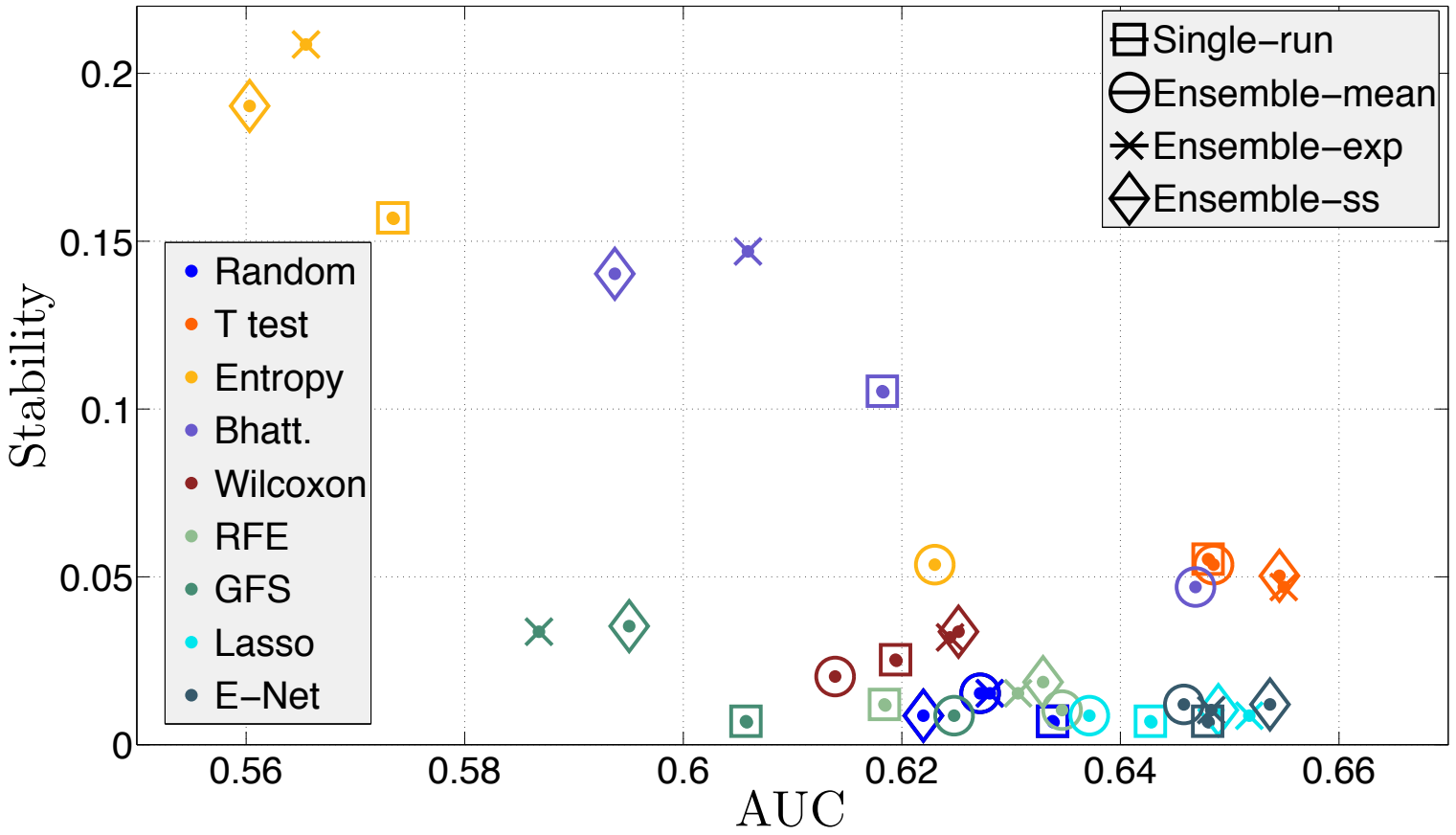


(Van de Vijver et al 2002)

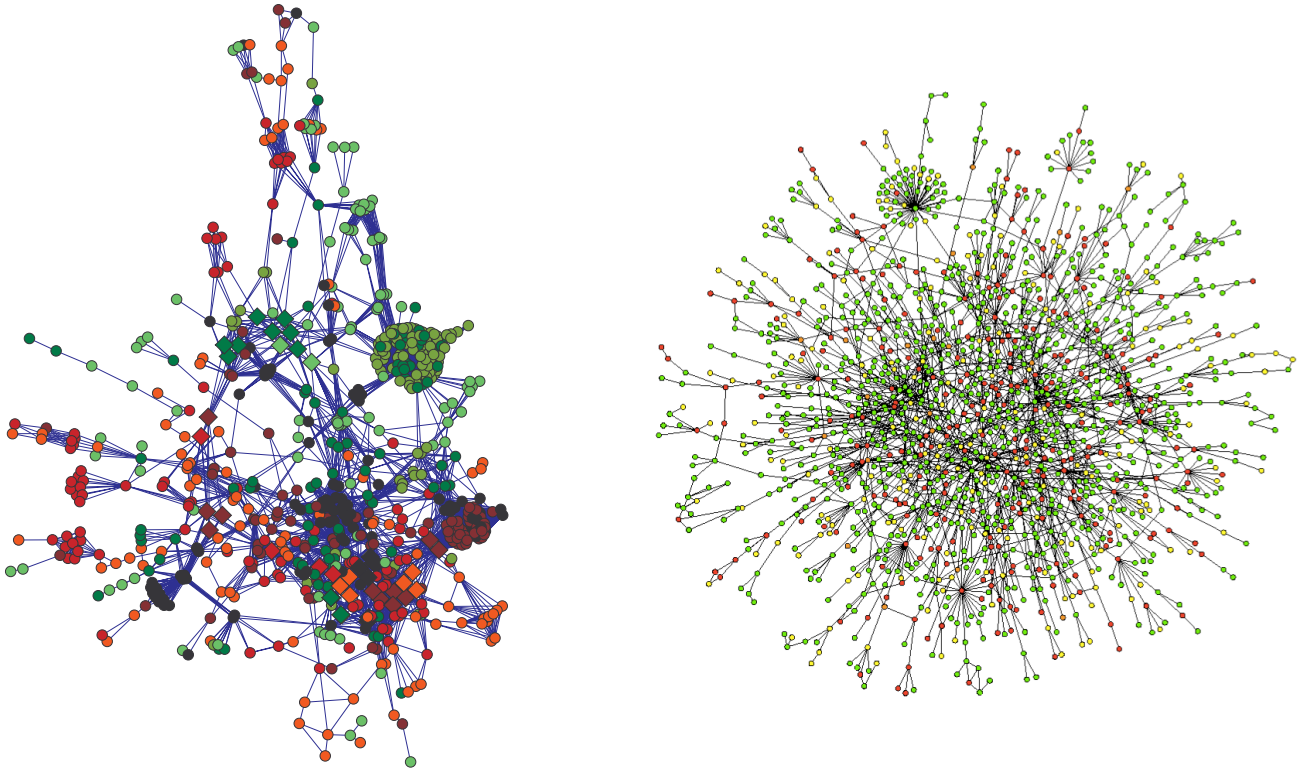
Two signatures have less than 5% genes in common…

Gene expression profiling predicts clinical outcome of breast cancer

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



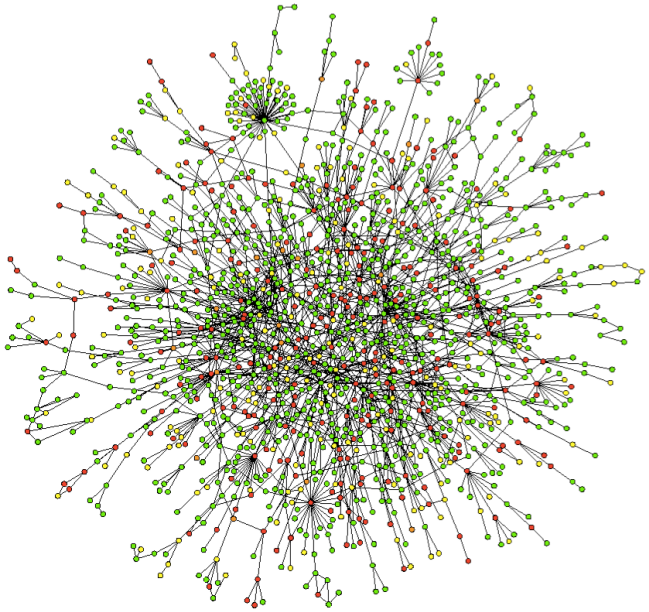
Prior knowledge: gene network

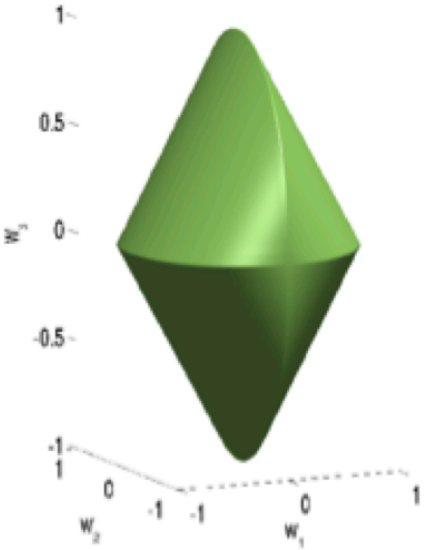


Can we « force » the signature to be « coherent » with a known network?

Example: the graph lasso

- Step 1: Using the network, define a subset of « candidate » signatures

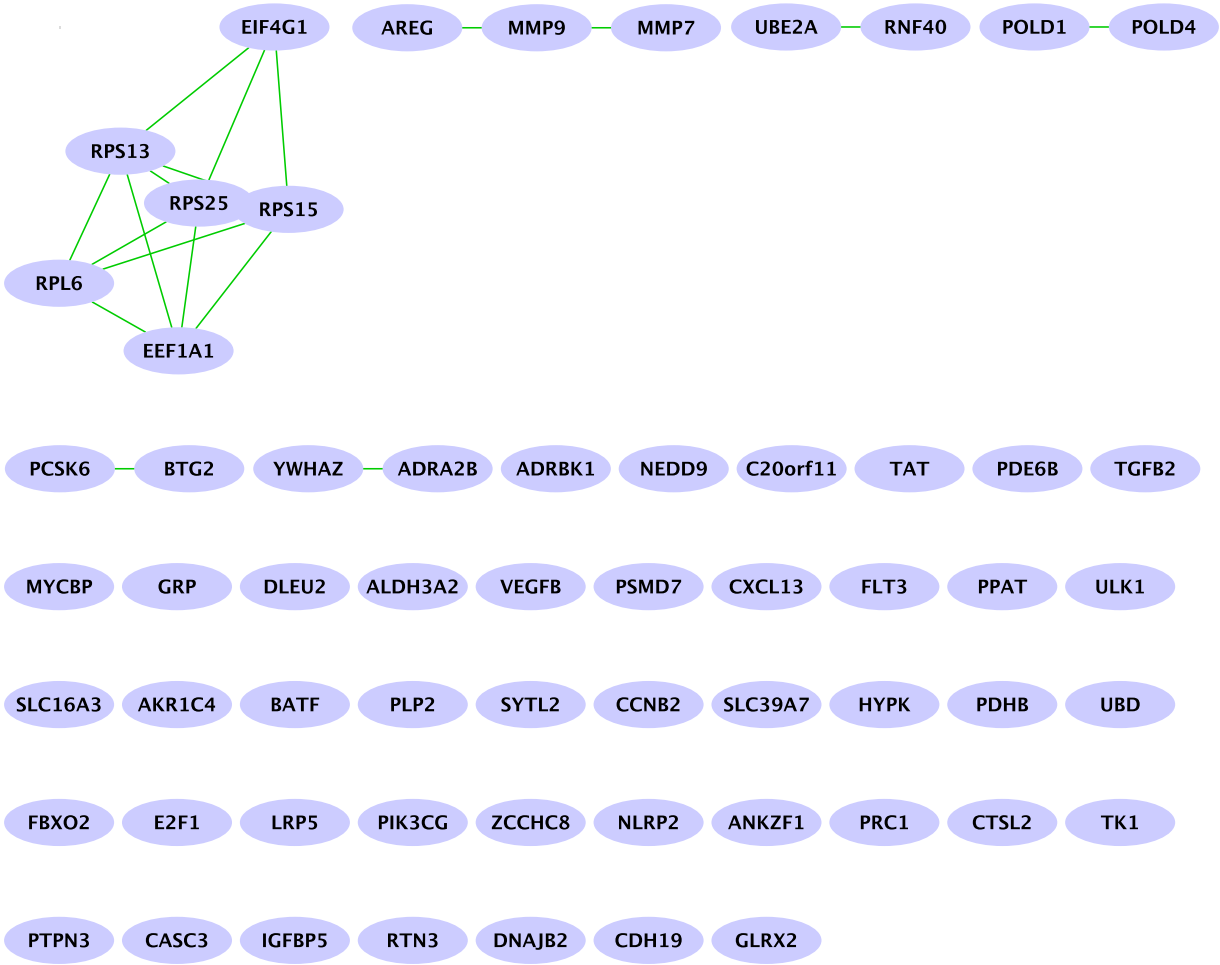




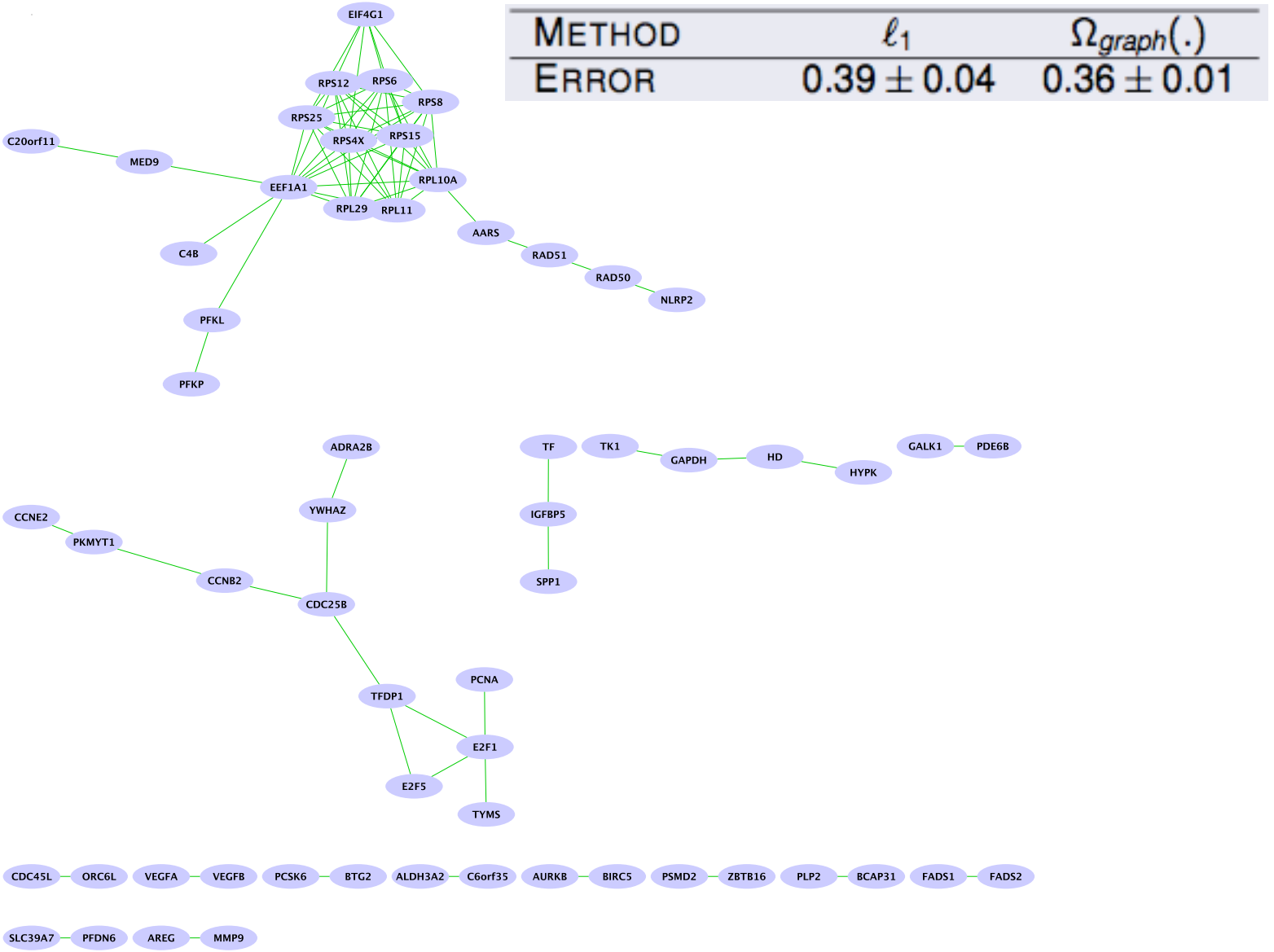
- Step 2: Among the candidates, find the best signature to explain the data

(Jacob et al 2009)

Classical signature



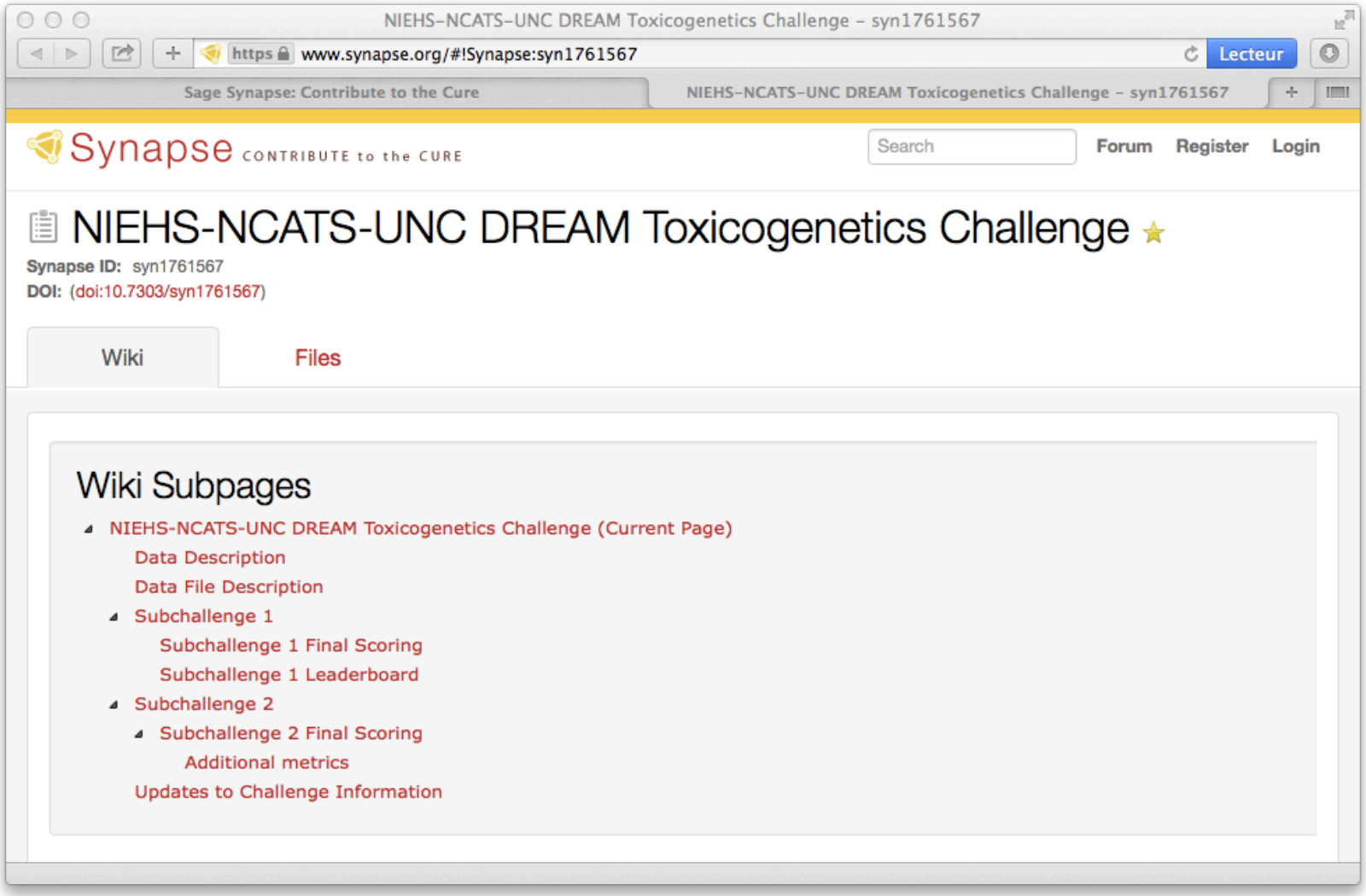
The graph lasso signature



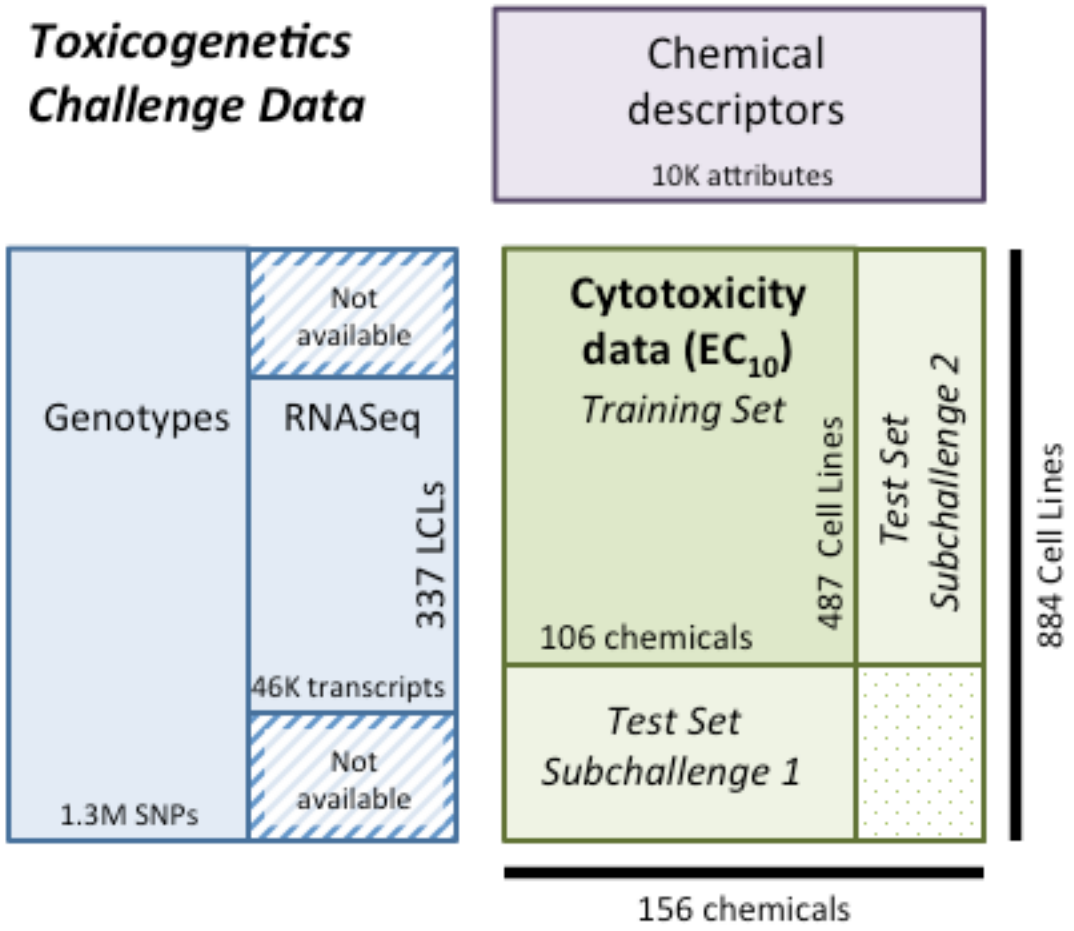
METHOD	ℓ_1	$\Omega_{graph}(.)$
ERROR	0.39 ± 0.04	0.36 ± 0.01

Example:
Pharmacogenomics / Toxicogenomics

Crowd-sourcing initiatives

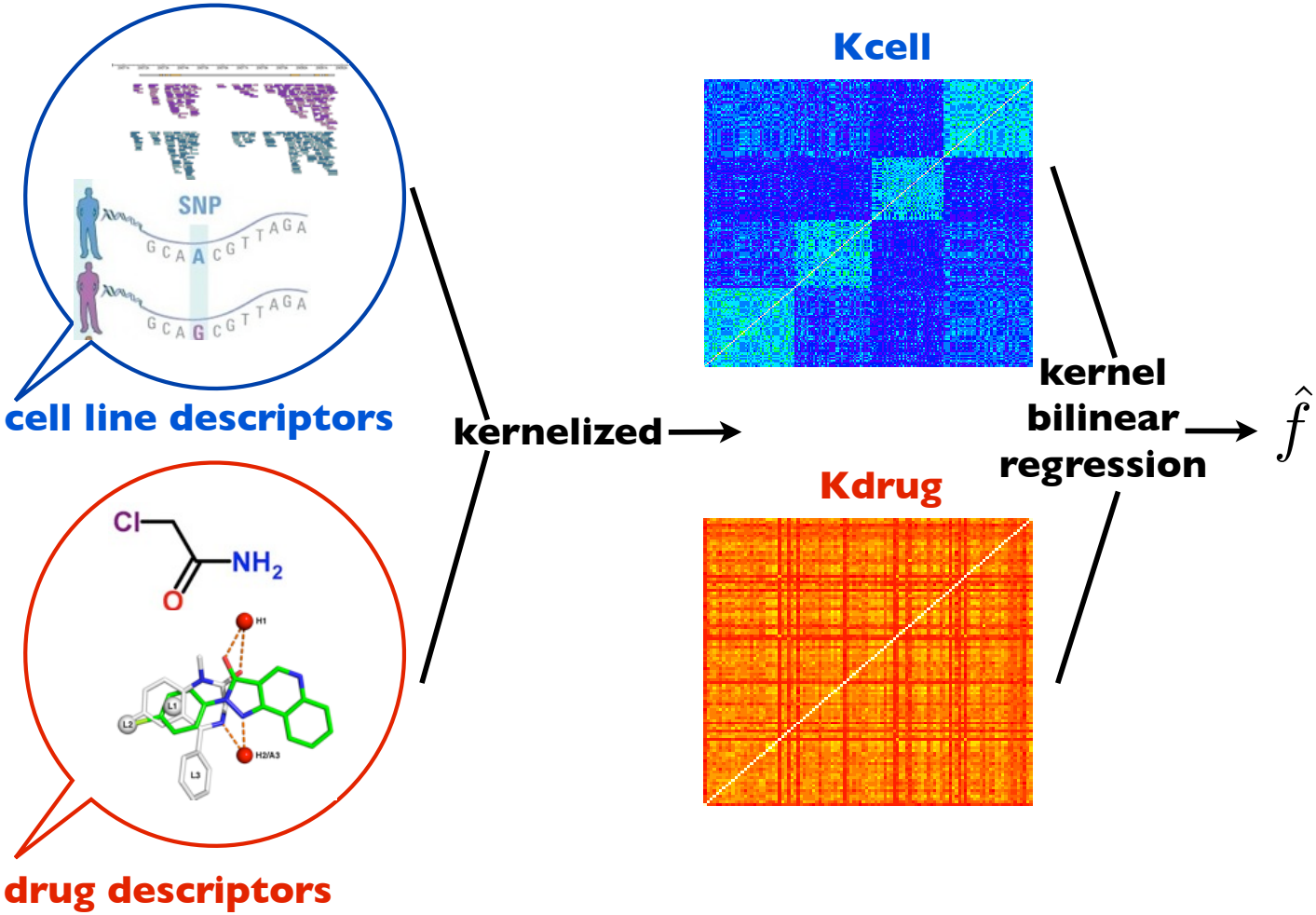


DREAM8 challenge (jun-sep 2013)



Genotypes from the 1000 genome project; RNASeq from the Geuvadis project

Our approach

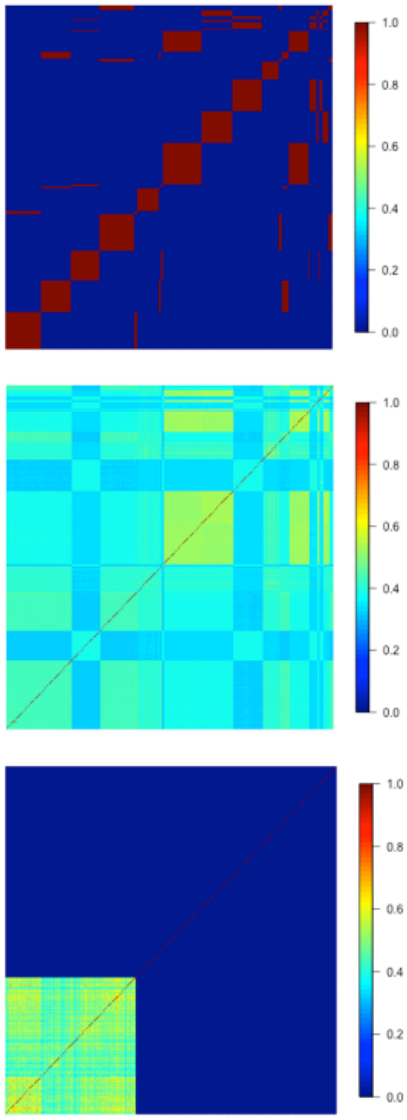


Cell line descriptors (30 kernels)

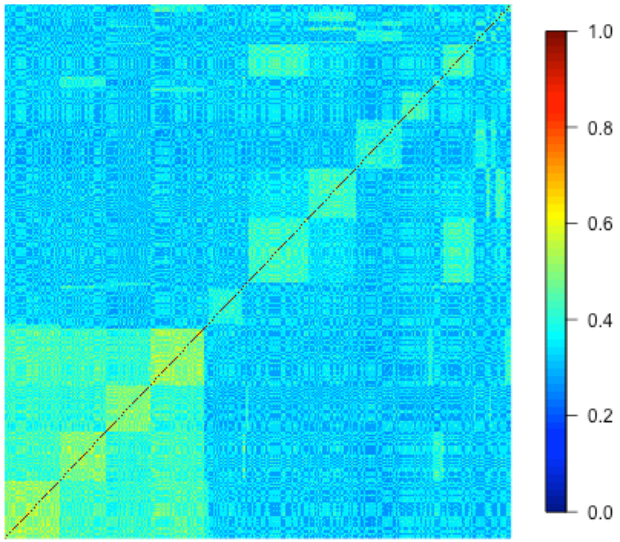
Covariates
. linear kernel

SNPs
. 10 gaussian kernels

RNA-seq
. 10 gaussian kernels

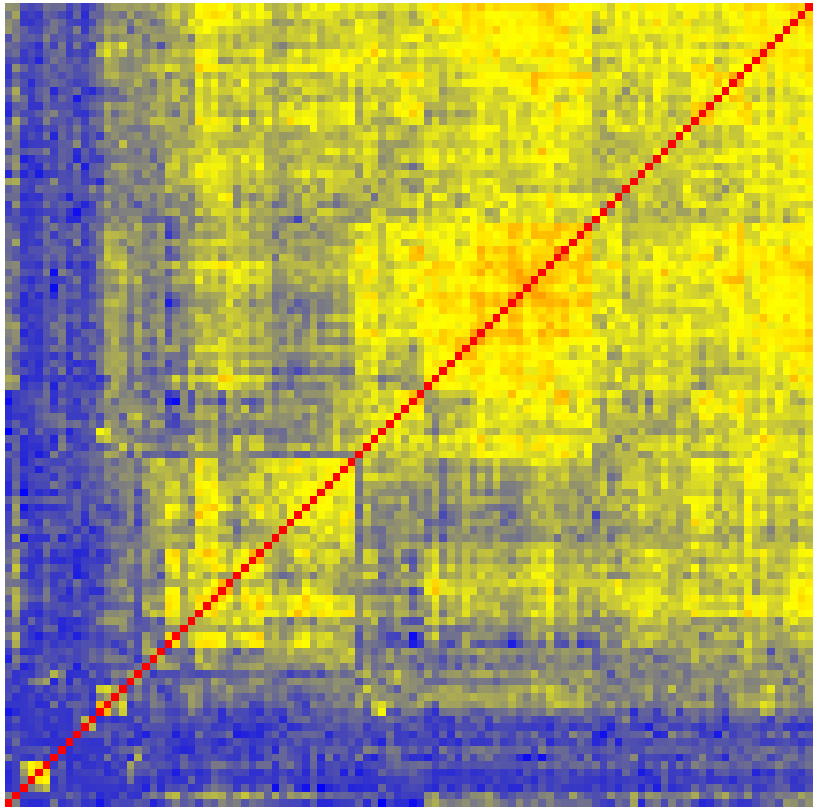


Integrated kernel

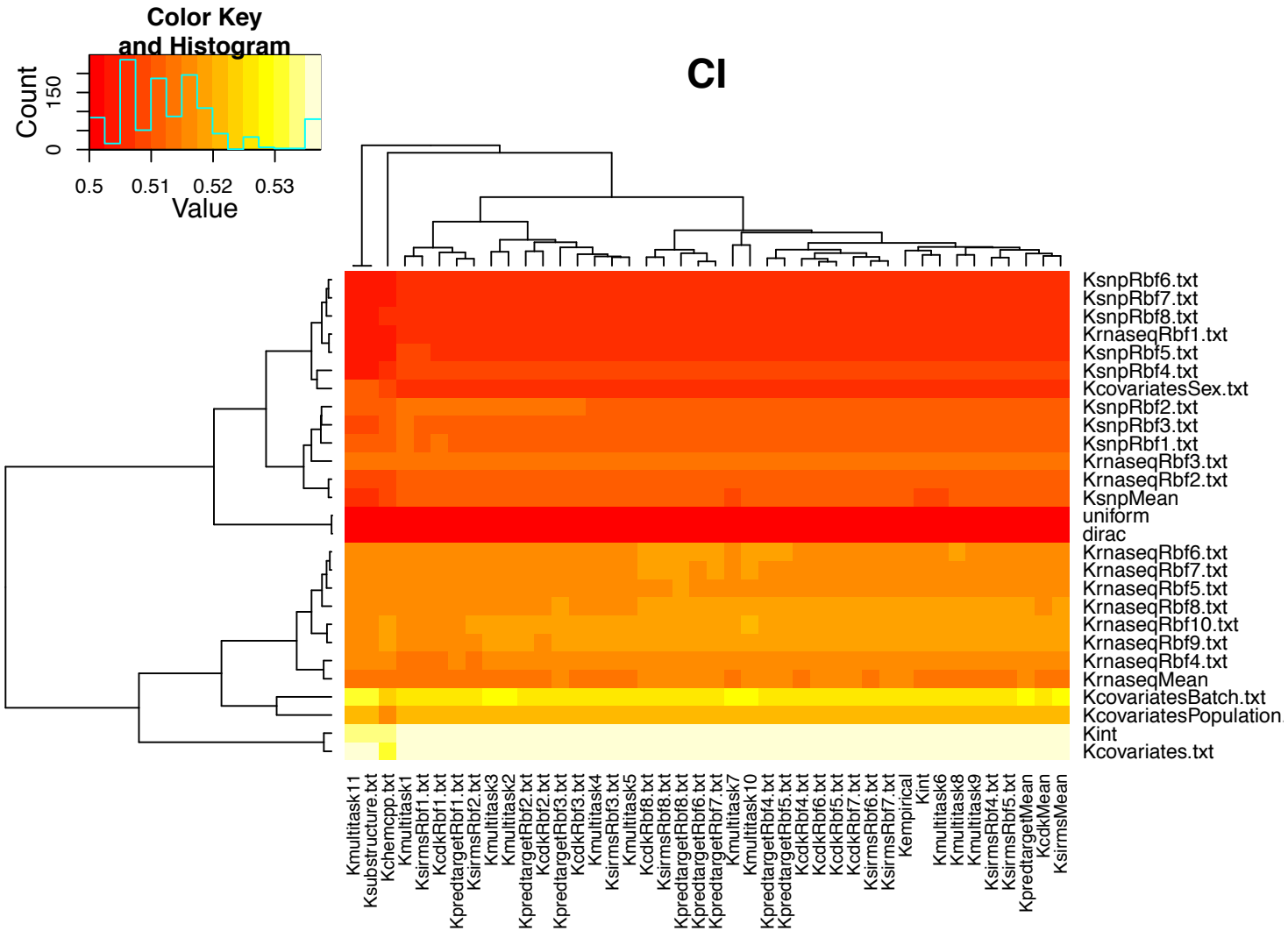


Chemical descriptors (49 kernels)

- Descriptors of chemical structures
- Multitask kernels
- Empirical correlation
- Integrated kernel

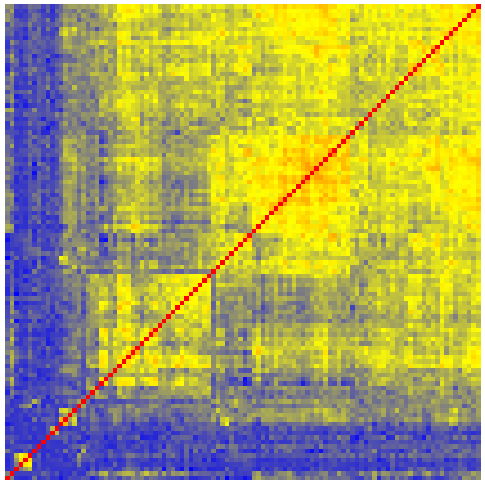


Learning occurs…

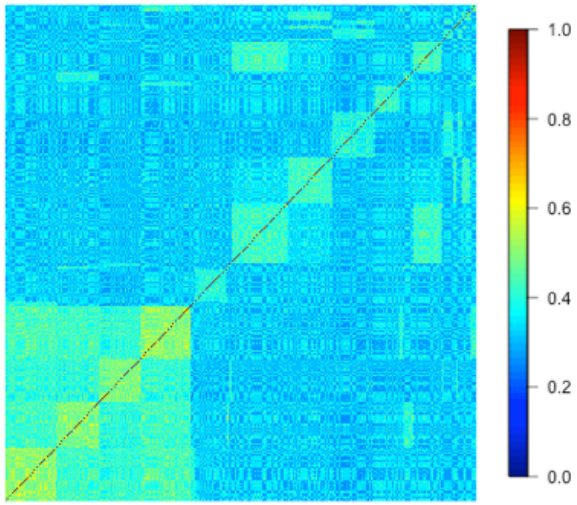


Final submission (ranked 2nd)

Empirical kernel on drugs



Integrated kernel on cell lines



0.54

Conclusion

- Lots of data due to technological progress
- **Opportunities**: precision medicine, quantitative biology
- **Challenges**: « small N », weak signal, complex systems