

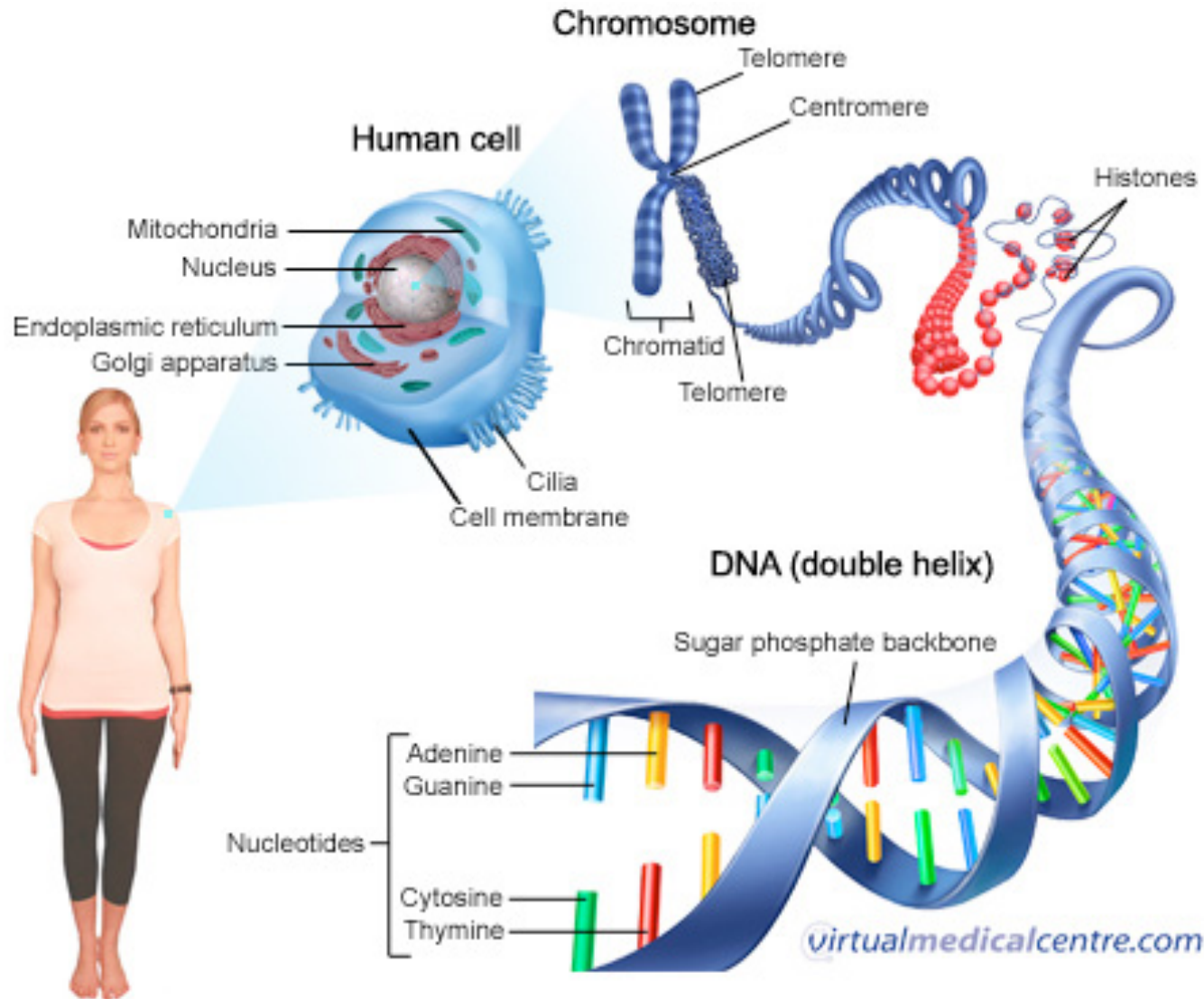
# Machine Learning for Personalized Genomics

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)

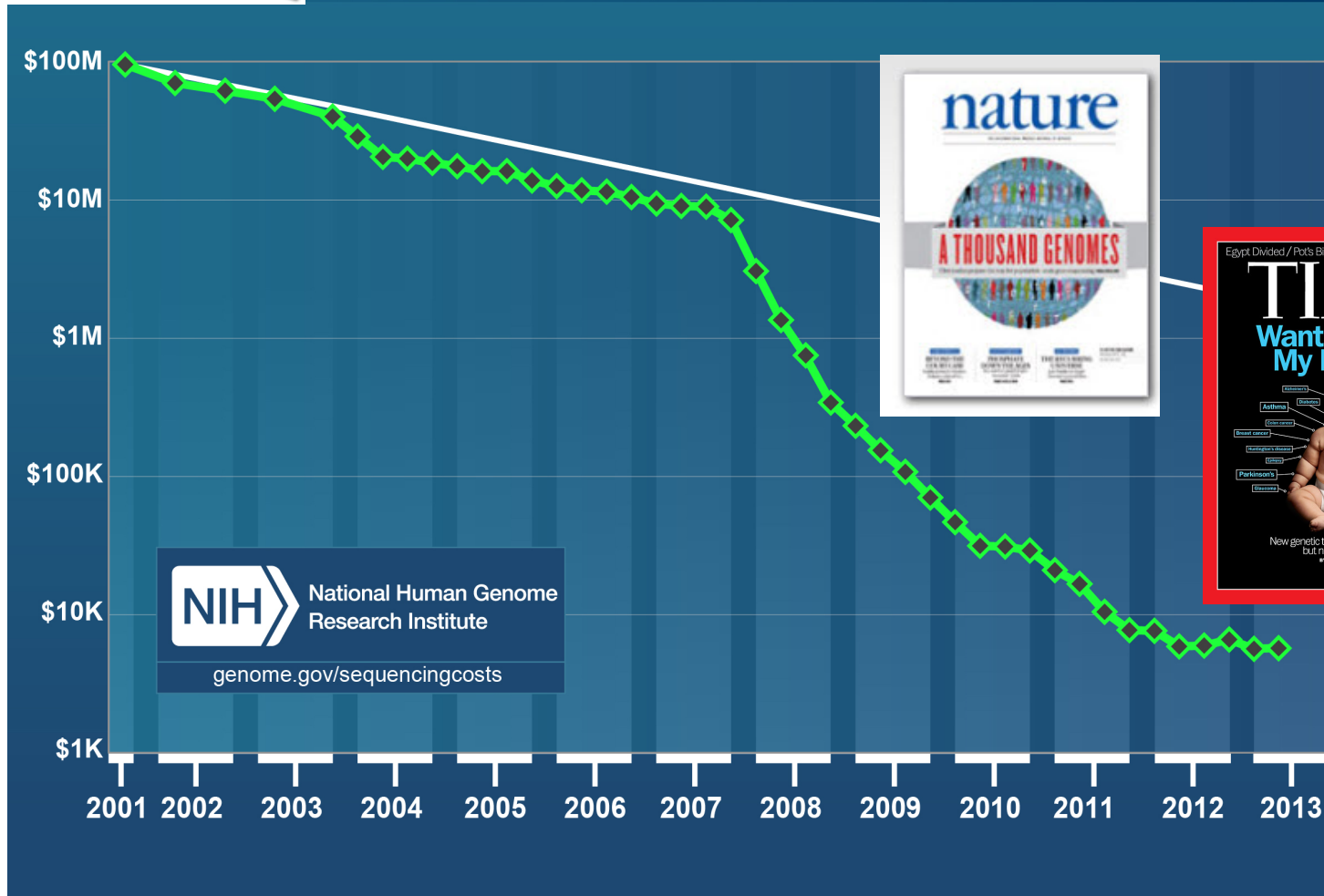
- Goal: sequence the 3,000,000,000 base pairs of the human genome
- Consortium of 20 laboratories, 6 countries
- 13 years, \$3,000,000,000



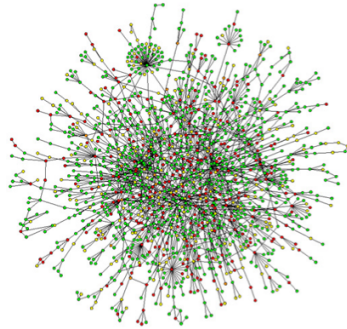


# The *second* revolution

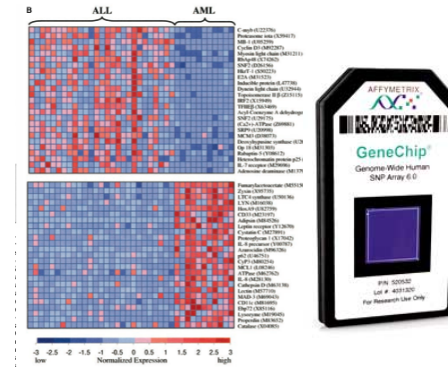
## Cost per Genome



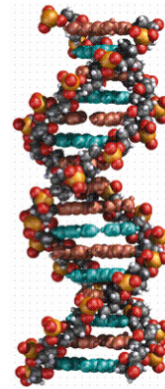
# A flood of *omics* data



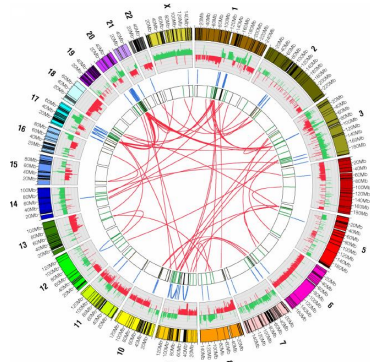
Interactome



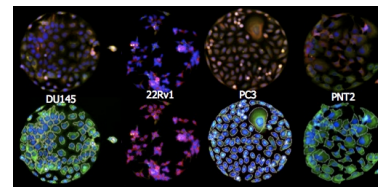
Transcriptome



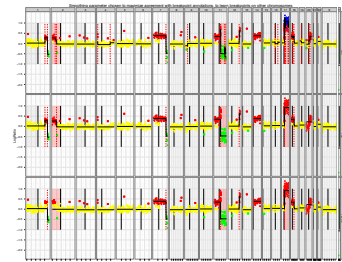
Genome



Mutations  
Structural variations

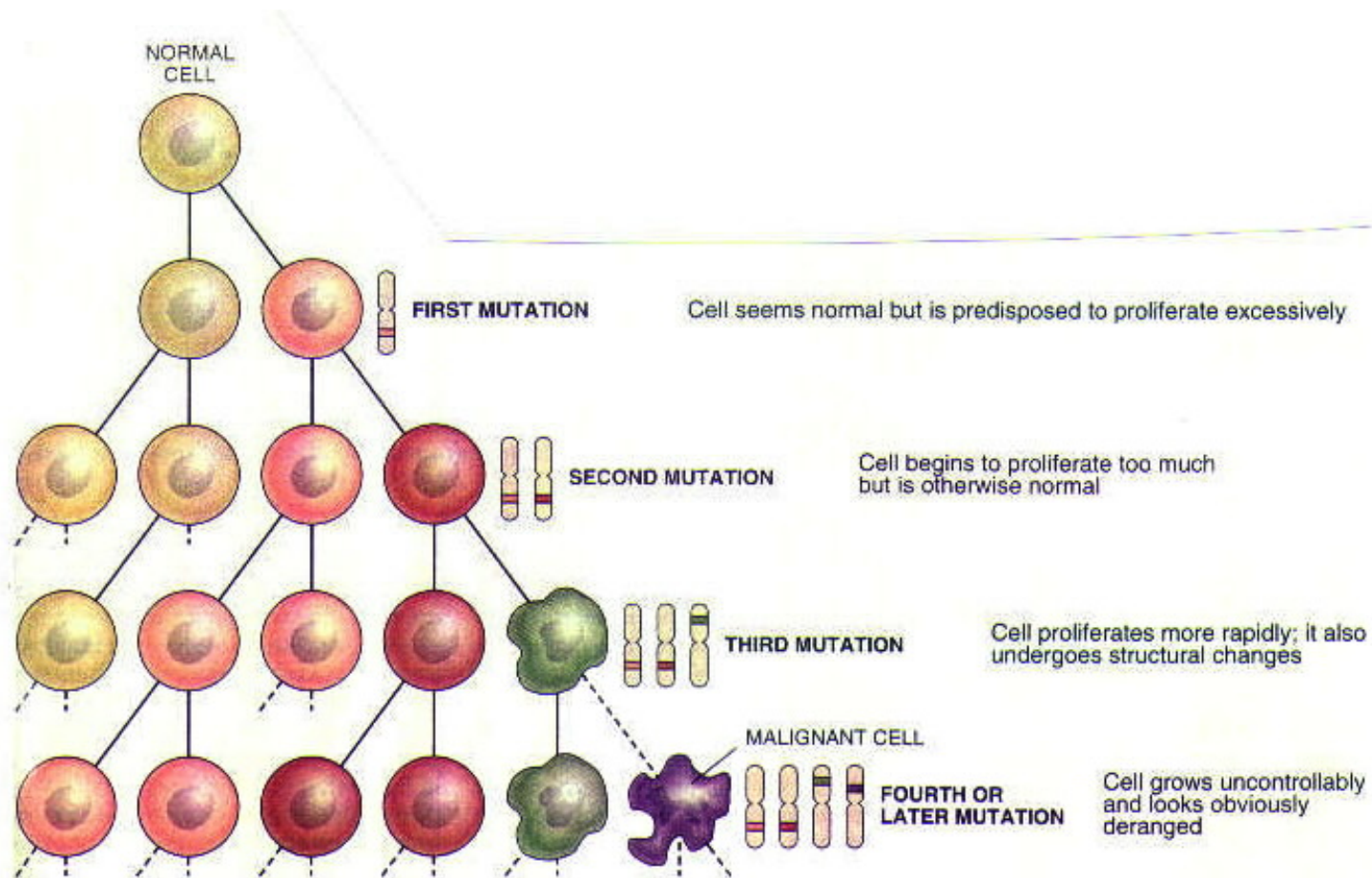


Phenome

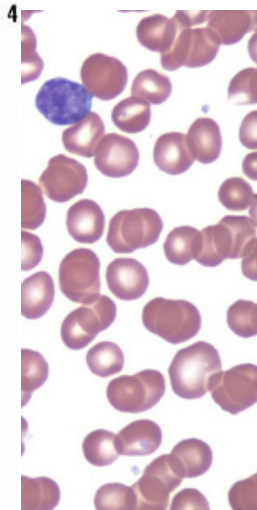
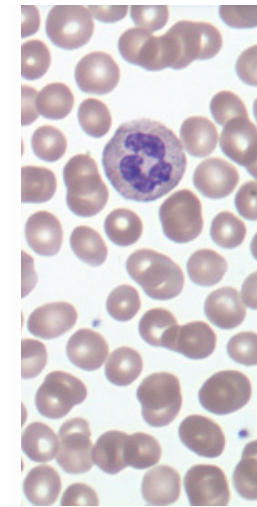
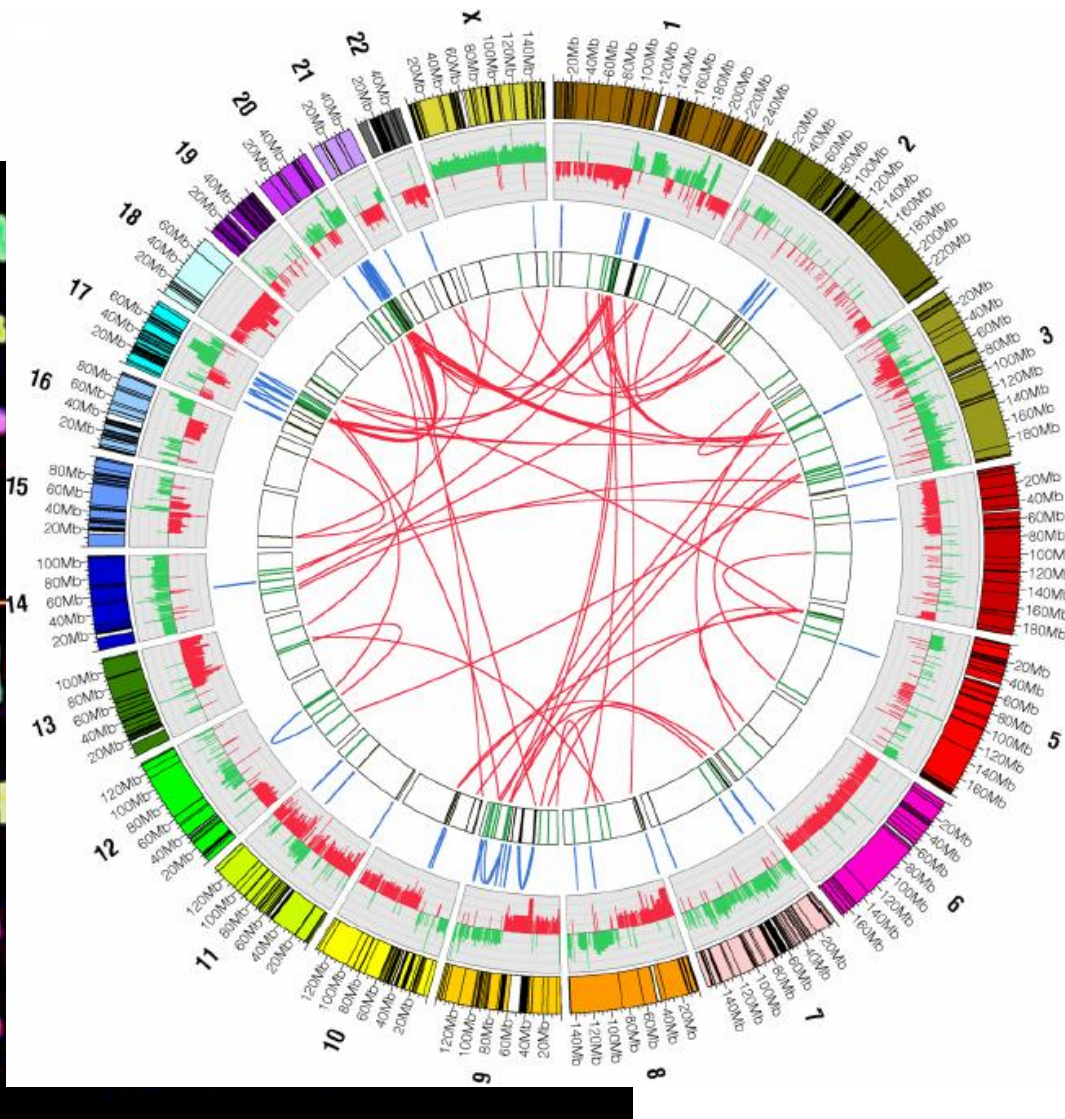


Epigenome

# All cancers are different

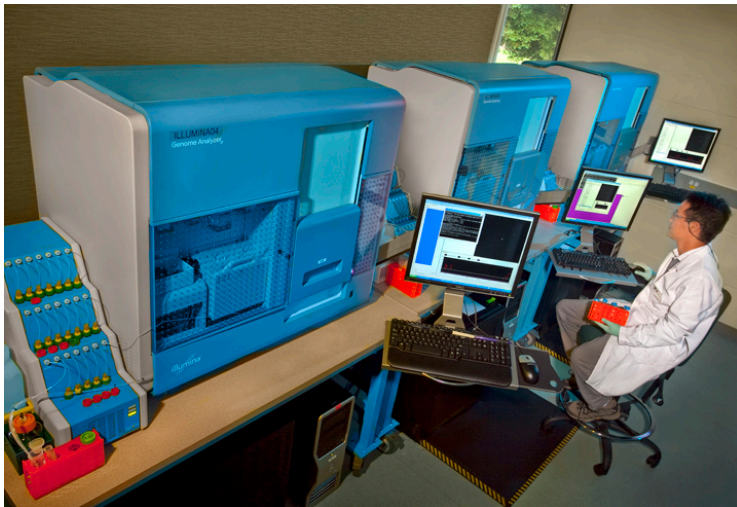


# Cancer: different views



# Big data!

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



The screenshot shows the International Cancer Genome Consortium (ICGC) website. The browser address bar displays 'http://www.icgc.org/'. The website features a navigation menu with 'Overview', 'Cancer Genome Projects', 'Committees', 'Policies and Guidelines', 'Media', and 'Contacts'. The main heading is 'International Cancer Genome Consortium'. Below this, there is a central graphic of a chromosome and a text box stating: 'will facilitate communication among the members and provide a forum for coordination with the objective of maximizing efficiency among the scientists working to understand, treat, and prevent these diseases.' An 'Announcements' section highlights a release on 25/Nov/2010 regarding version 3 of the ICGC data portal. A 'nature' journal cover is also visible, with the text 'International network of cancer genome projects. Nature 464, 993-998 (15 April 2010) HTML'. The website lists various cancer types and their participating countries, including Bladder Cancer (United States), Blood Cancer (United States), Bone Cancer (United Kingdom), Brain Cancer (United States), Breast Cancer (European Union / United Kingdom, France, United Kingdom), Cervical Cancer (United States), Chronic Lymphocytic Leukemia (Spain), Chronic Myeloid Disorders (United Kingdom), Colon Cancer (United States), Endometrial Cancer (United States), Gastric Cancer, Liver Cancer (Japan), Lung Cancer (United States), Malignant Lymphoma (Germany), Oral Cancer (India), Ovarian Cancer (Australia, United States), Pancreatic Cancer (Australia, Canada), Pediatric Brain Tumors (Germany), Prostate Cancer (Germany, United States), and Rare Pancreatic Tumors (Canada).



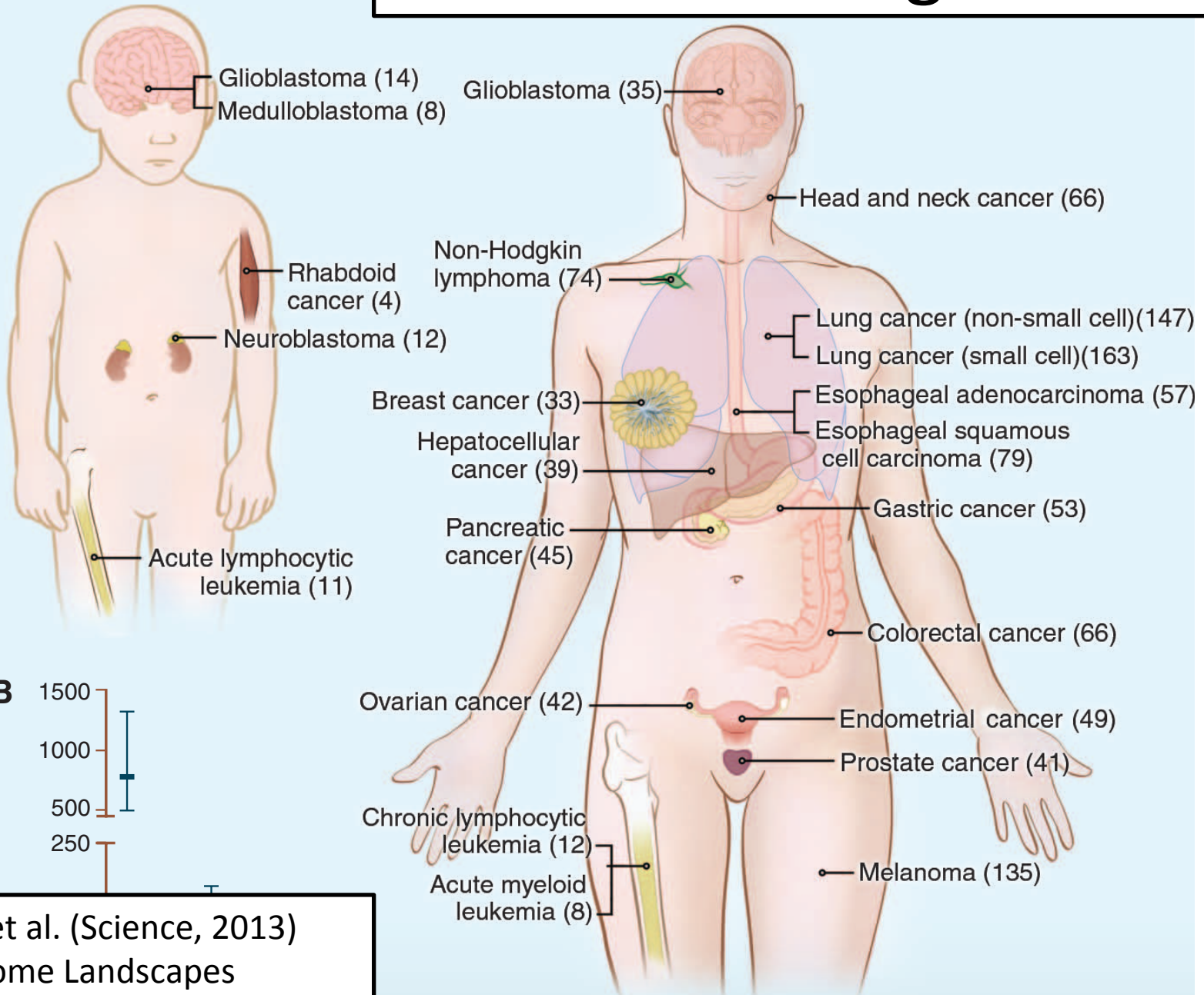


# Opportunities

- **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...

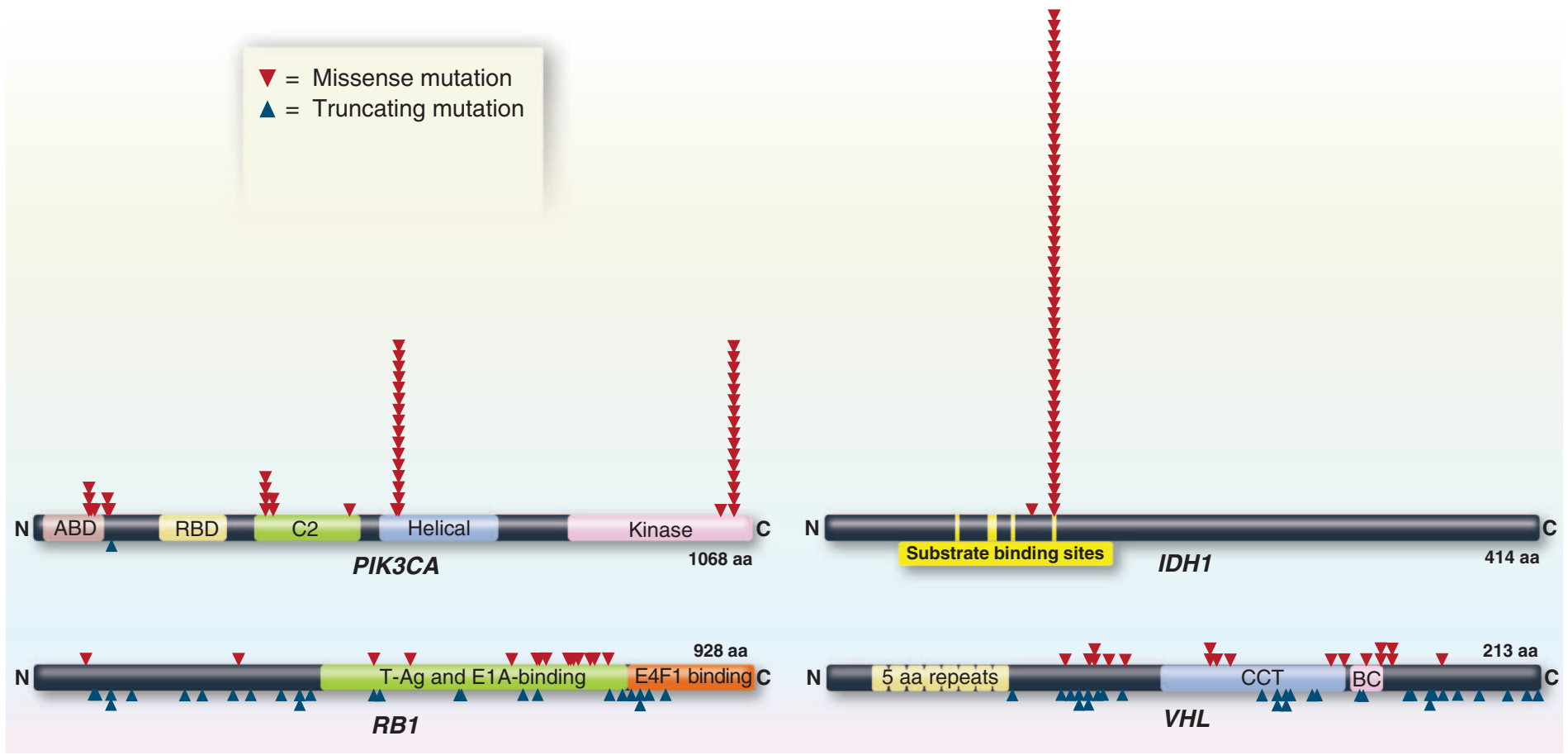
# « Understanding cancer »

**A**



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

# Finding « cancer genes »



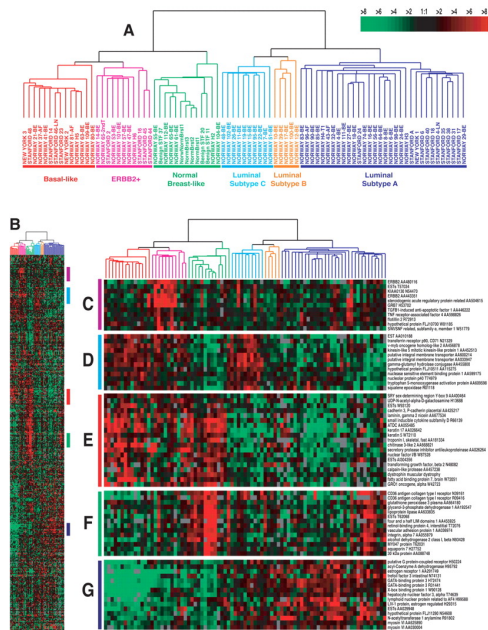
Vogelstein et al. (Science, 2013)  
Cancer genome Landscapes



**P4 Medicine**  
 ● PREDICT ● PREVENT ● PERSONALIZE ● PARTICIPATE

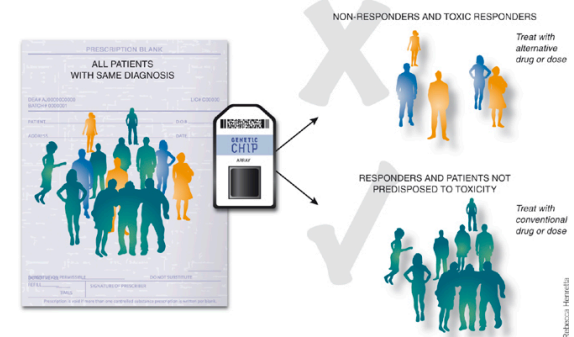
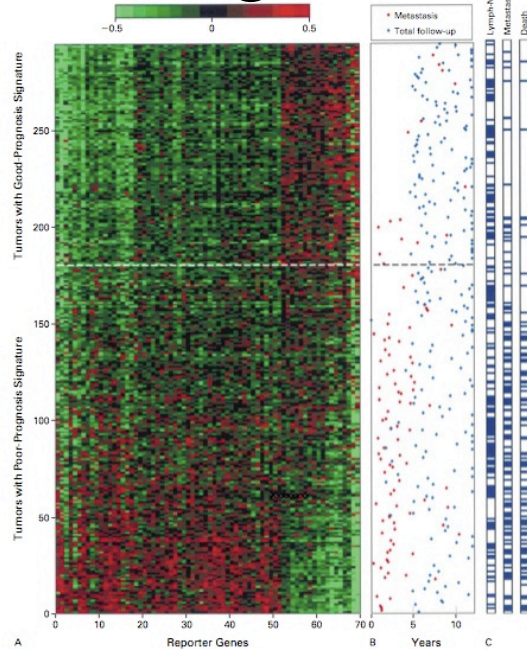


**Opportunities**



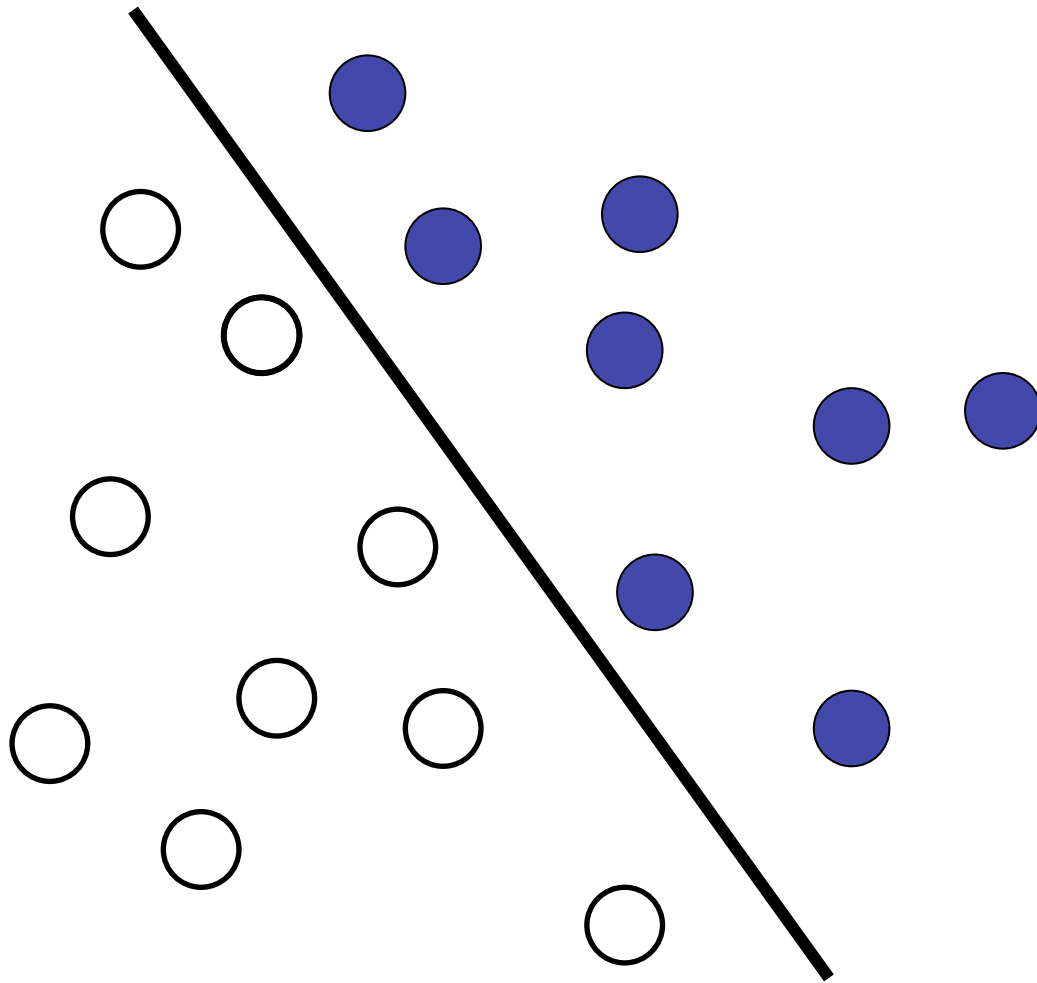
**Diagnosis**

**Prognosis**



**Response to drugs**

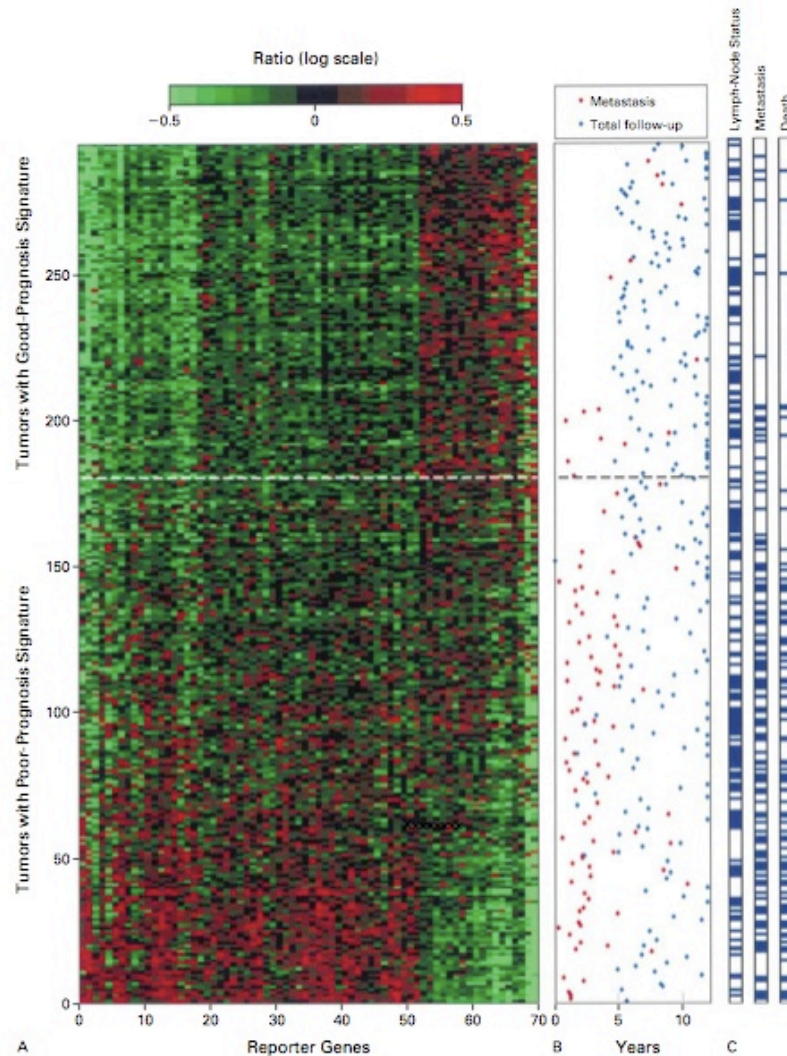
# Supervised machine learning



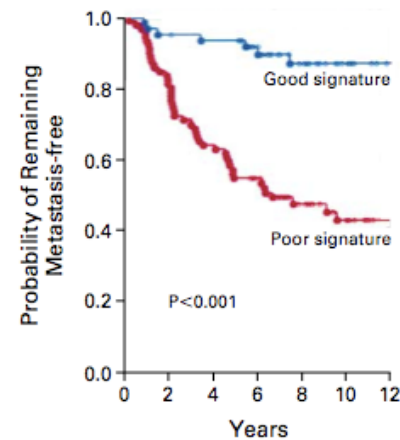
## Challenges

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

# Example: Breast cancer prognostic signature



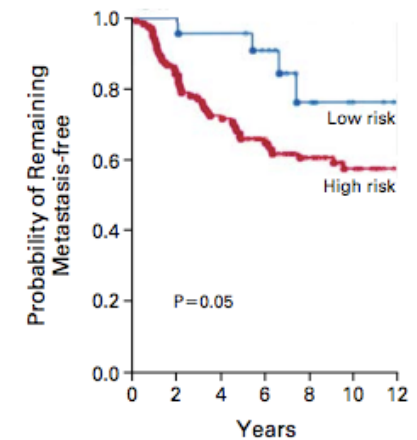
A Gene-Expression Profiling



NO. AT RISK

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

B St. Gallen Criteria



NO. AT RISK

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

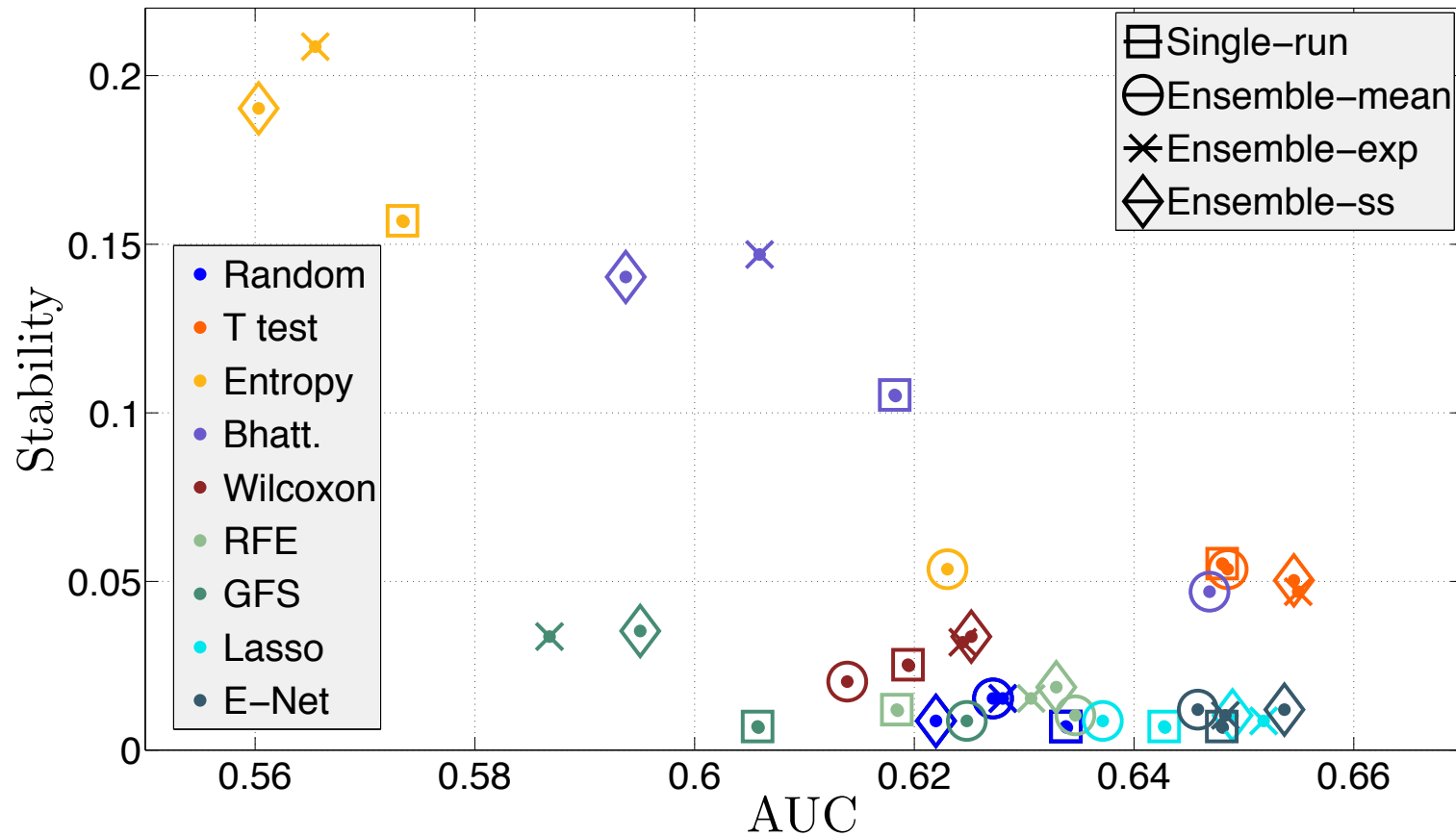
(Van de Vijver et al 2002)

# Two signatures have less than 5% genes in common...

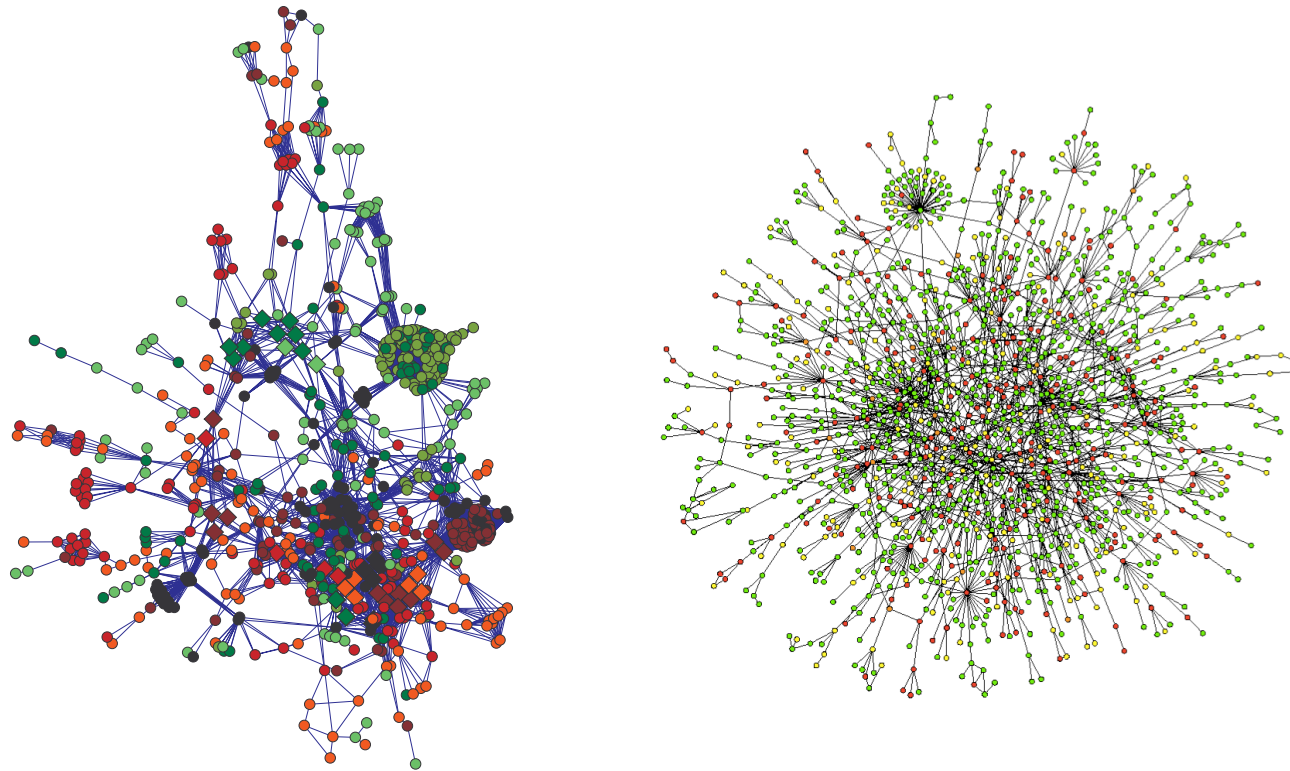
**Gene expression profiling predicts clinical outcome of breast cancer**

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

Lai  
Yur  
Kar  
Ger  
Pet  
  
\*D  
and  
121  
‡R



# Prior knowledge: gene network

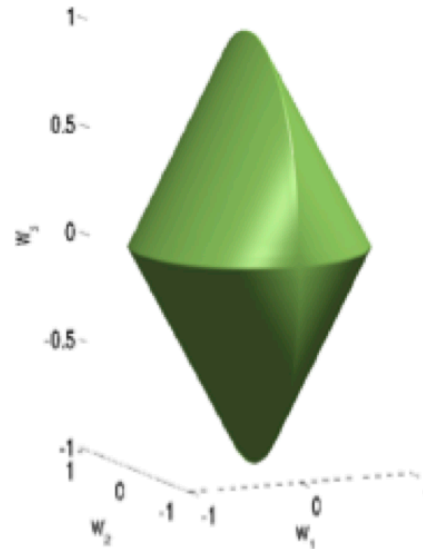
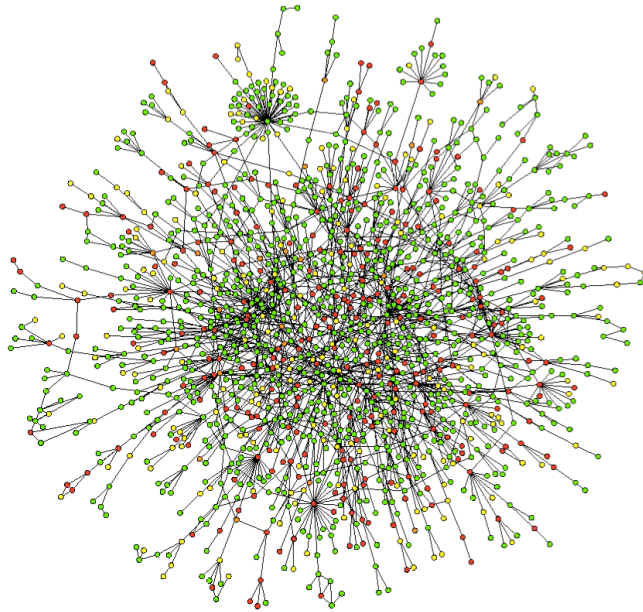


*Can we « force » the signature to be « coherent »  
with a known network?*



# Example: the graph lasso

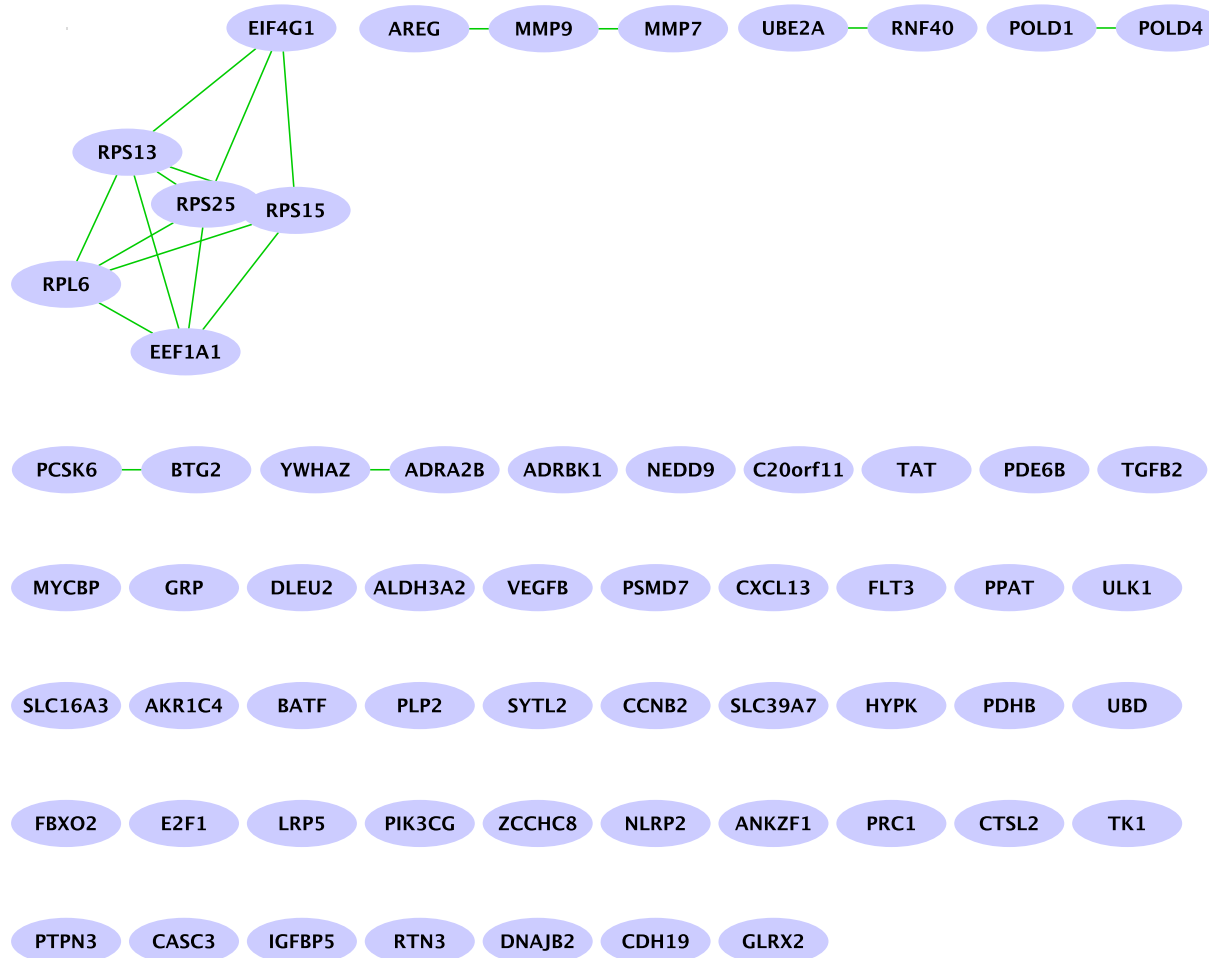
- Step 1: Using the network, define a subset of « candidate » signatures



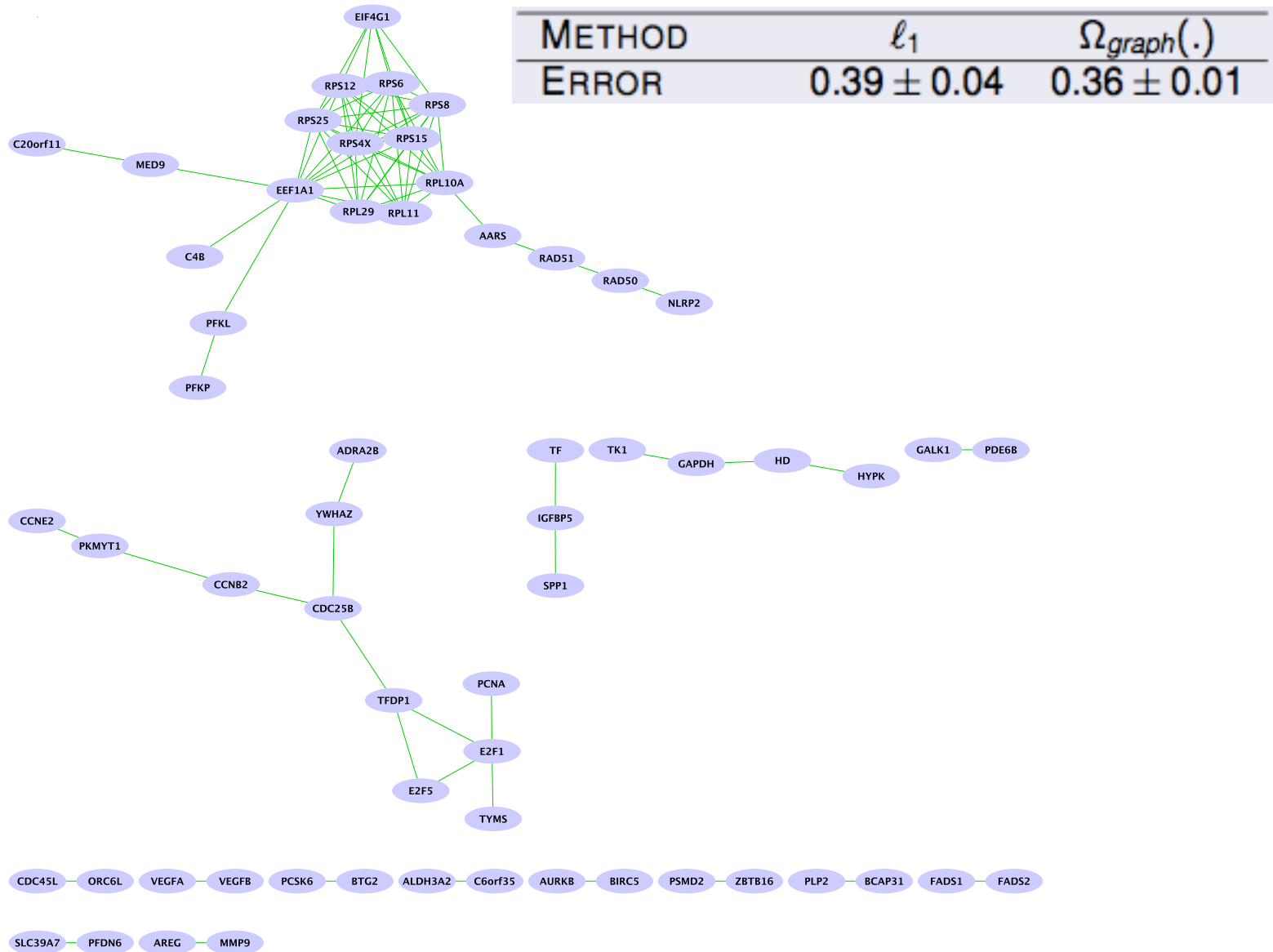
- Step 2: Among the candidates, find the best signature to explain the data

*(Jacob et al 2009)*

# Classical signature



# The graph lasso signature



# Example: Pharmacogenomics / Toxicogenomics

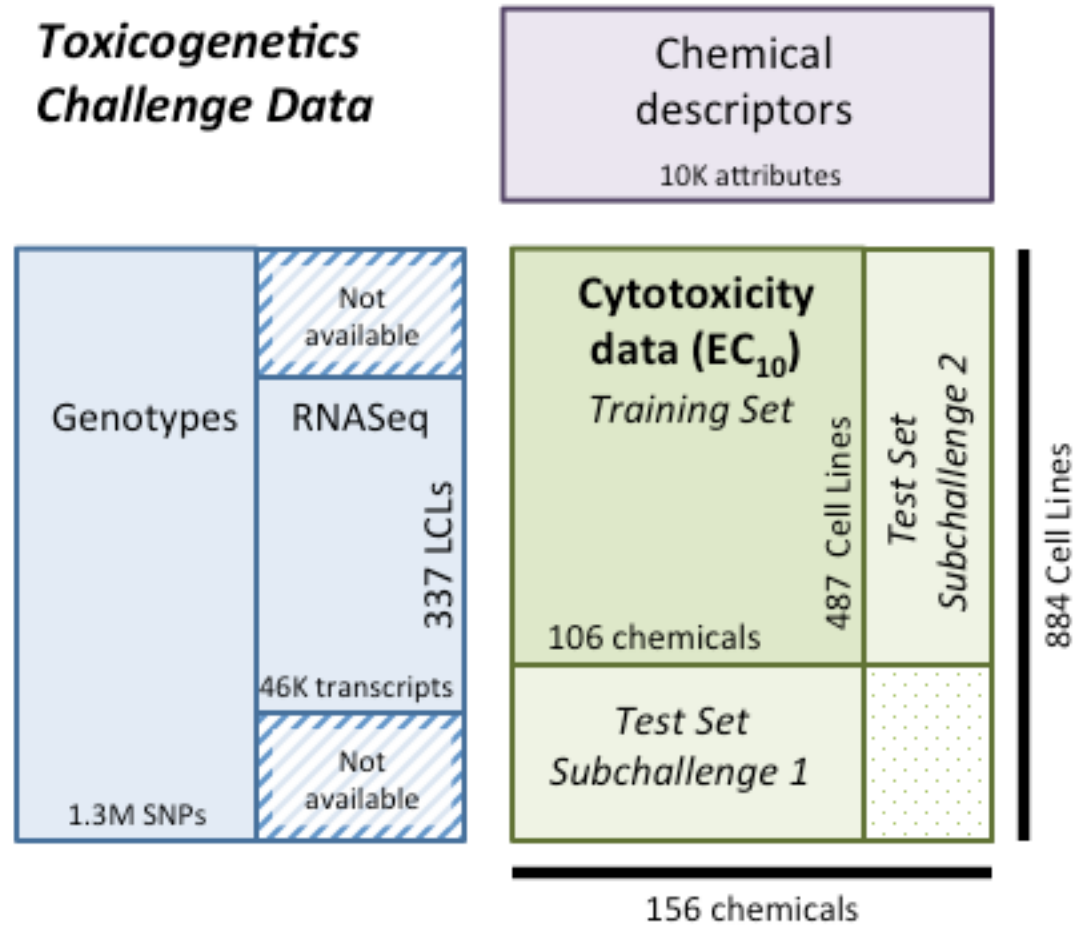


# Crowd-sourcing initiatives

The screenshot shows a web browser window with the following elements:

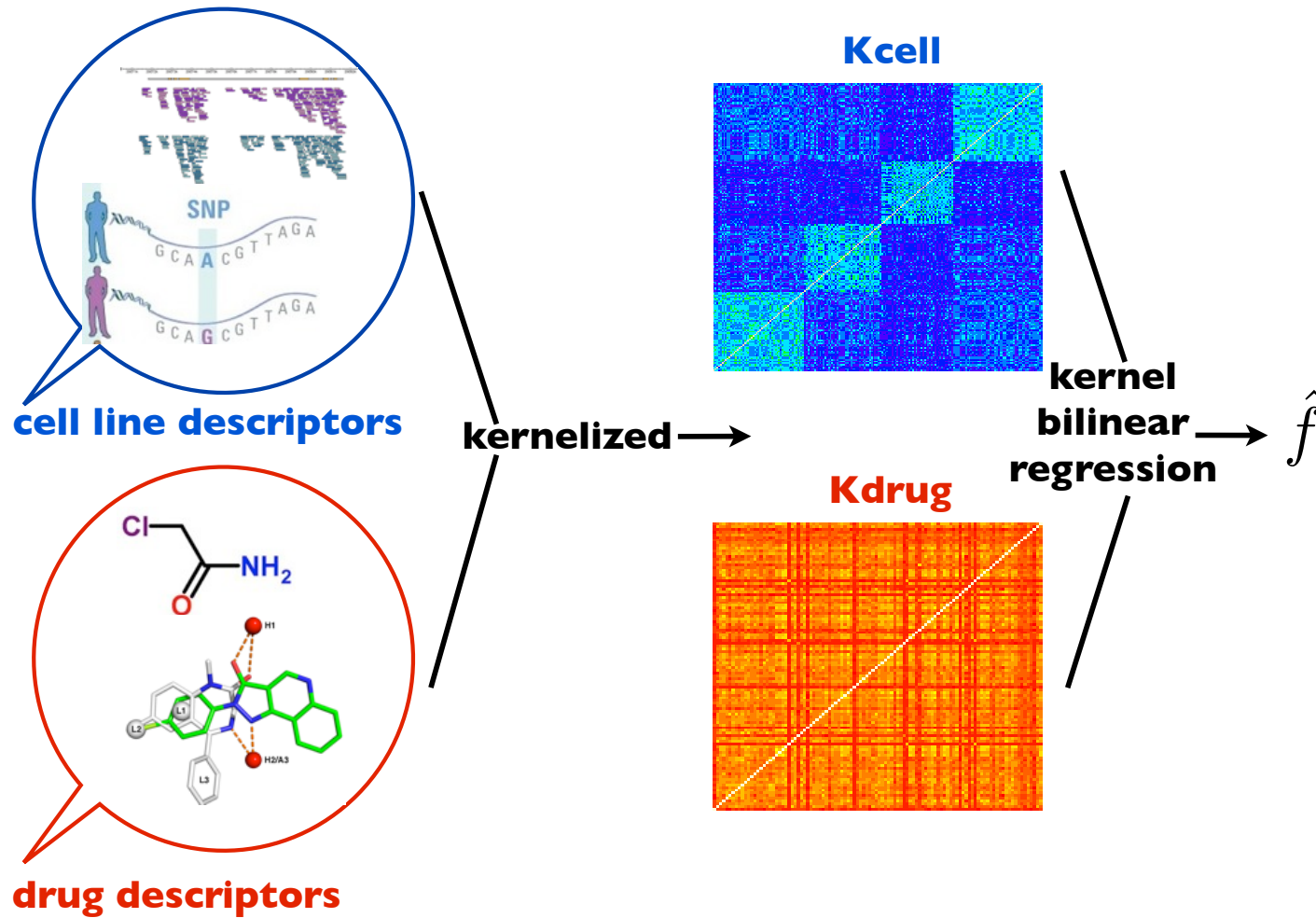
- Browser Tab:** NIEHS-NCATS-UNC DREAM Toxicogenetics Challenge - syn1761567
- Address Bar:** <https://www.synapse.org/#!/Synapse:syn1761567>
- Page Header:** Sage Synapse: Contribute to the Cure | NIEHS-NCATS-UNC DREAM Toxicogenetics Challenge - syn1761567
- Navigation:** Synapse logo, "CONTRIBUTE to the CURE", Search bar, Forum, Register, Login
- Page Title:** NIEHS-NCATS-UNC DREAM Toxicogenetics Challenge ★
- Metadata:** Synapse ID: syn1761567, DOI: (doi:10.7303/syn1761567)
- Navigation Tabs:** Wiki (selected), Files
- Wiki Subpages:**
  - ▲ NIEHS-NCATS-UNC DREAM Toxicogenetics Challenge (Current Page)
    - Data Description
    - Data File Description
  - ▲ Subchallenge 1
    - Subchallenge 1 Final Scoring
    - Subchallenge 1 Leaderboard
  - ▲ Subchallenge 2
    - ▲ Subchallenge 2 Final Scoring
      - Additional metrics
    - Updates to Challenge Information

# DREAM8 challenge (jun-sep 2013)



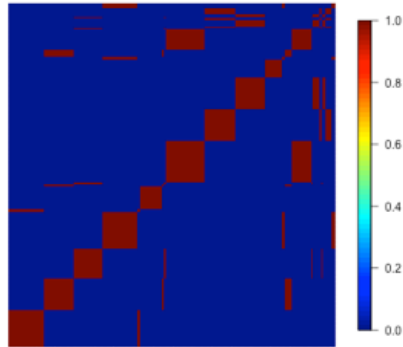
*Genotypes from the 1000 genome project; RNASeq from the Geuvadis project*

# Our approach

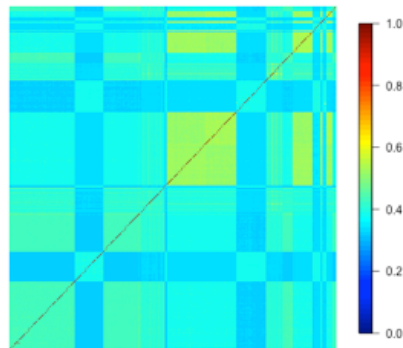


# Cell line descriptors (30 kernels)

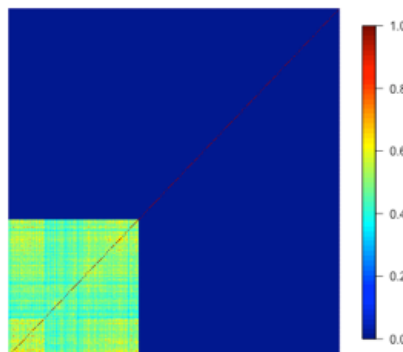
**Covariates**  
. linear kernel



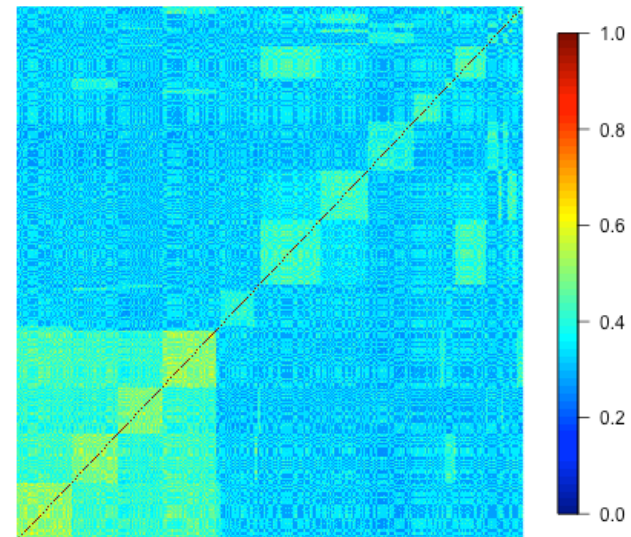
**SNPs**  
. 10 gaussian kernels



**RNA-seq**  
. 10 gaussian kernels



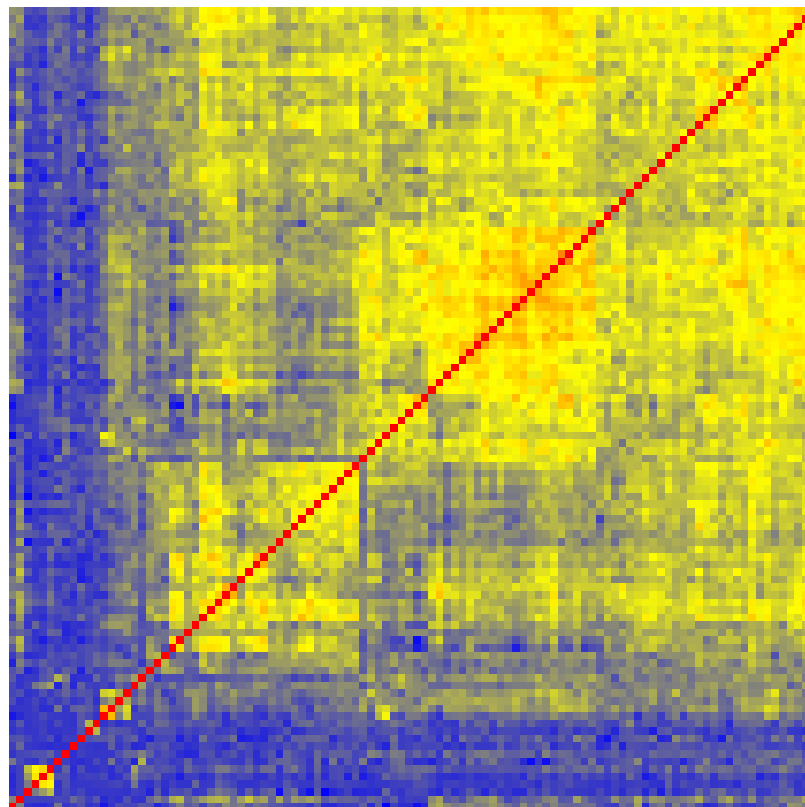
**Integrated kernel**



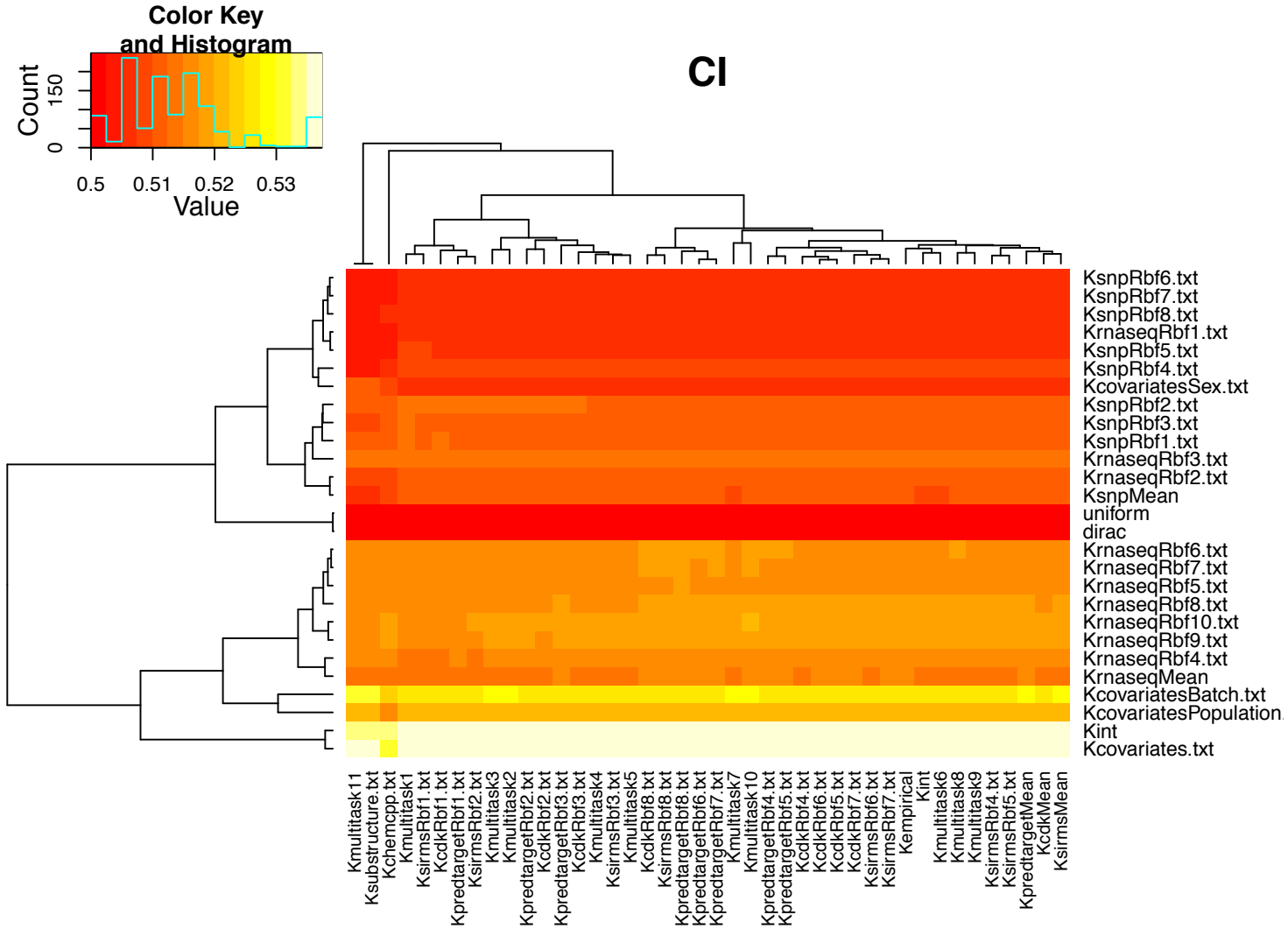


# Chemical descriptors (49 kernels)

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

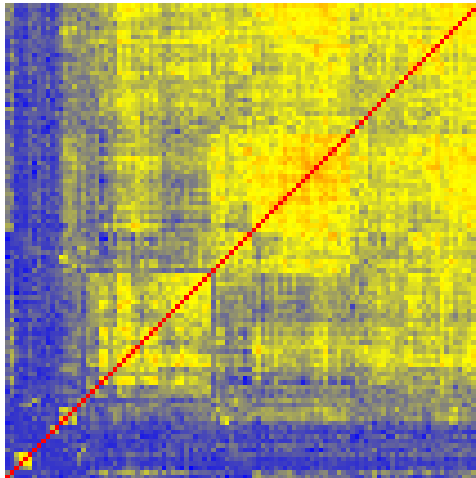


# Learning occurs...

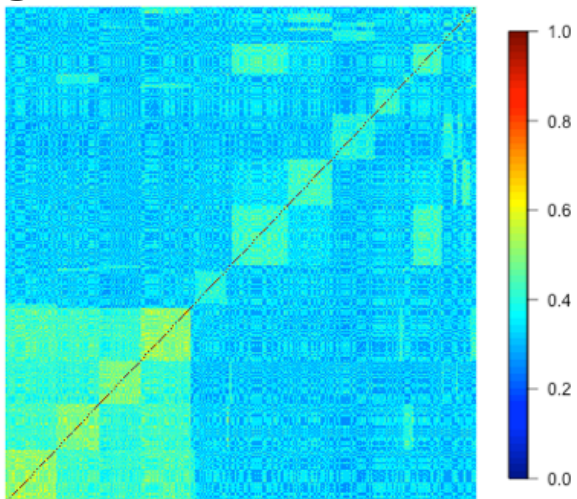


# Final submission (ranked 2<sup>nd</sup>)

**Empirical kernel on drugs**



**Integrated kernel on cell lines**



# Conclusion

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

