Machine Learning for Personalized Genomics

Jean-Philippe Vert

C3BI Kick-off meeting, Institut Pasteur,
Paris, March 16, 2015

Institut Curie / Inserm U900 / MINES ParisTech partnership

institutCurie MINES PARIS ParisTech Inserm

- A joint lab about ``Cancer computational genomics, bioinformatics, biostatistics and epidemiology''
- Located in Institut Curie, a major hospital and cancer research centre in Europe, and MINES ParisTech

4 teams + 1 platform

Systems Biology (Barillot):

- Modelling, simulating biological systems
- Building an *in silico* atlas of cancer pathways

Clinical Biostatistics (Asselain / Paoletti):

- Clinical trials for targeted therapies
- Predictive biomarkers

Cancer Genetic Epidemiology (Andrieu):

- Genetic and environmental factors in breast cancer

Machine learning (Vert):

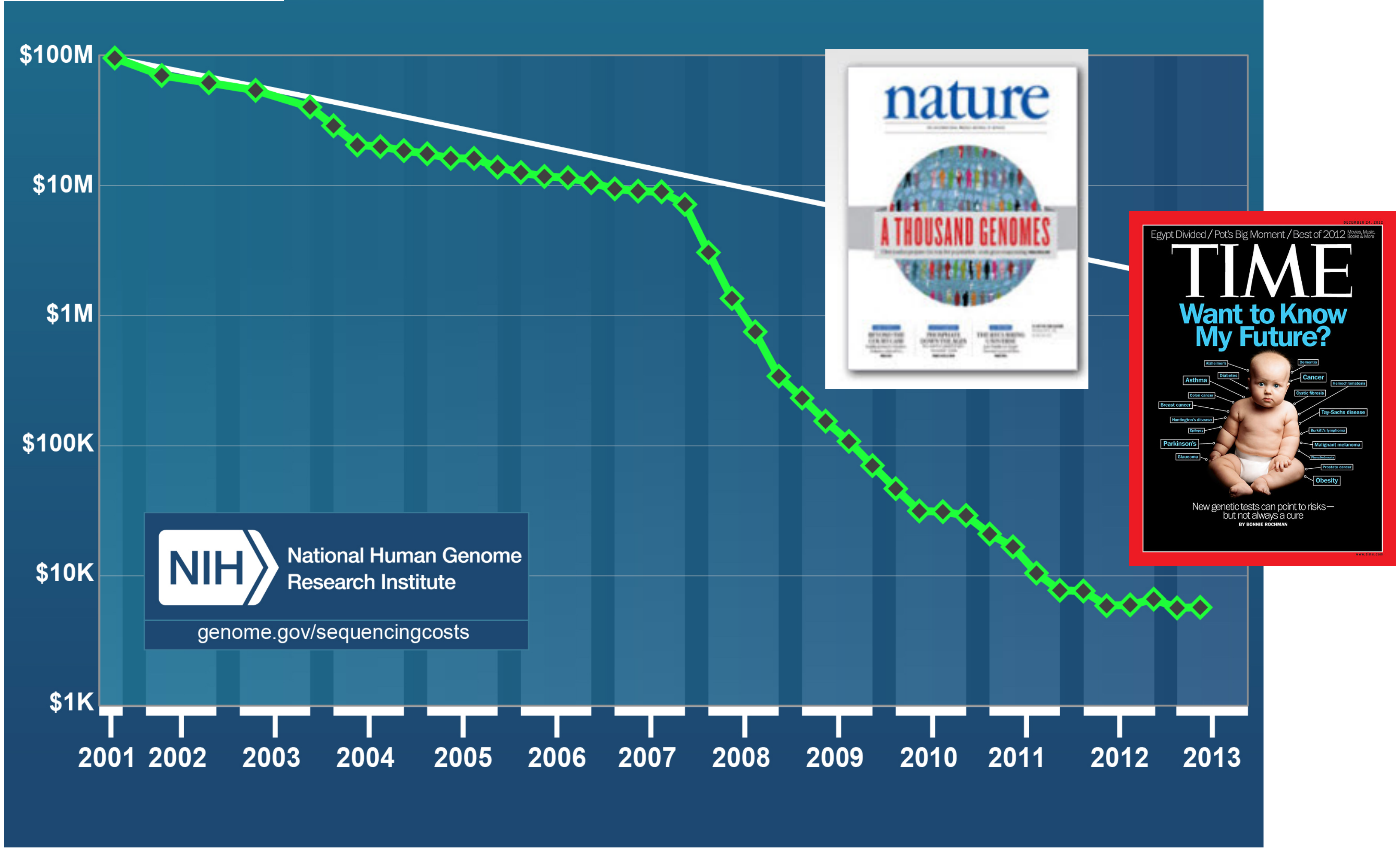
- Learning from « big omics data » for personalized medicine

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

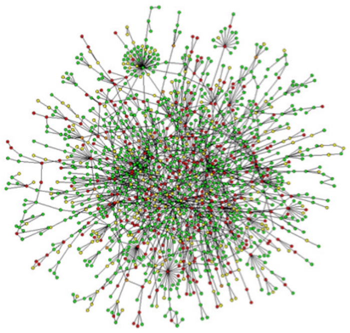
Cost per Genome



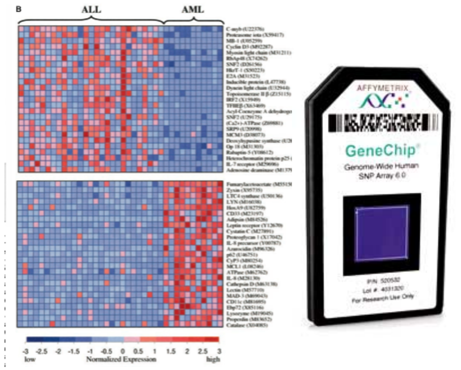
A flood of ***omics*** data

PubMed

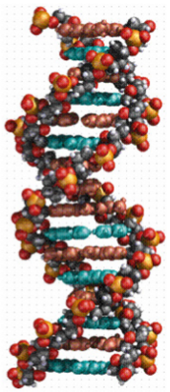
Publications



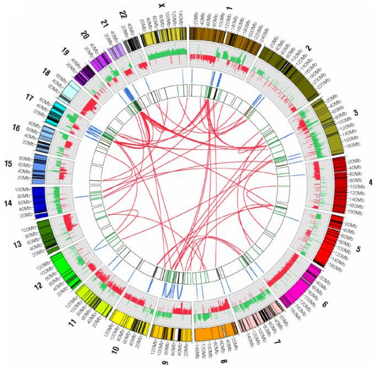
Interactome



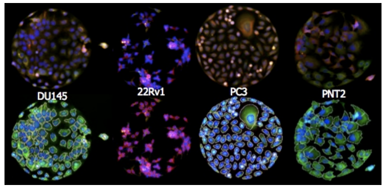
Transcriptome



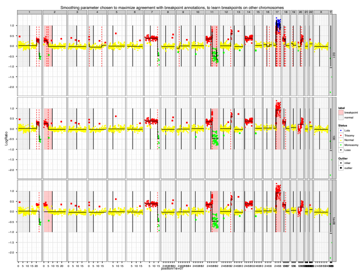
Genome



Mutations
Structural variations

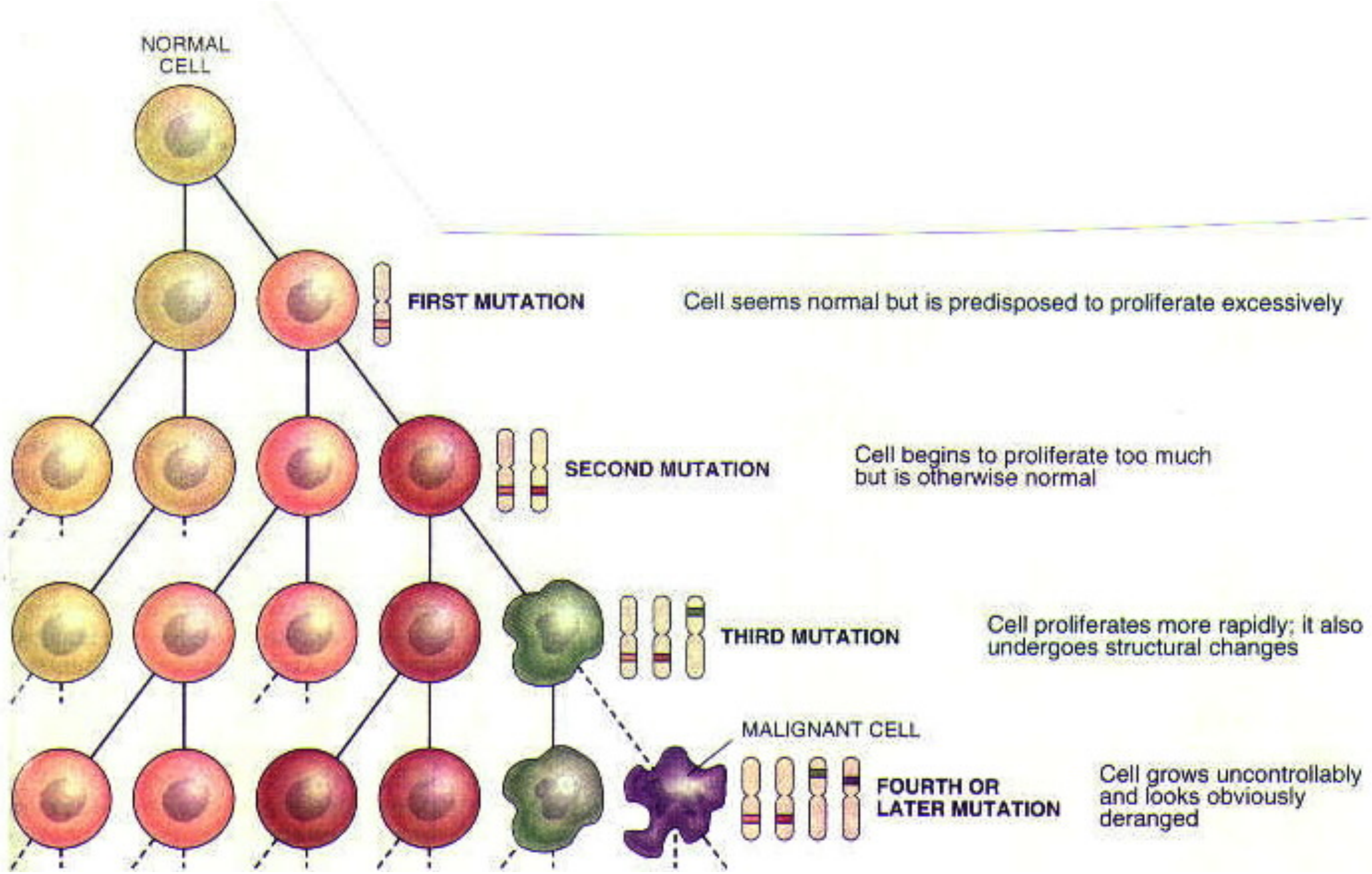


Phenome

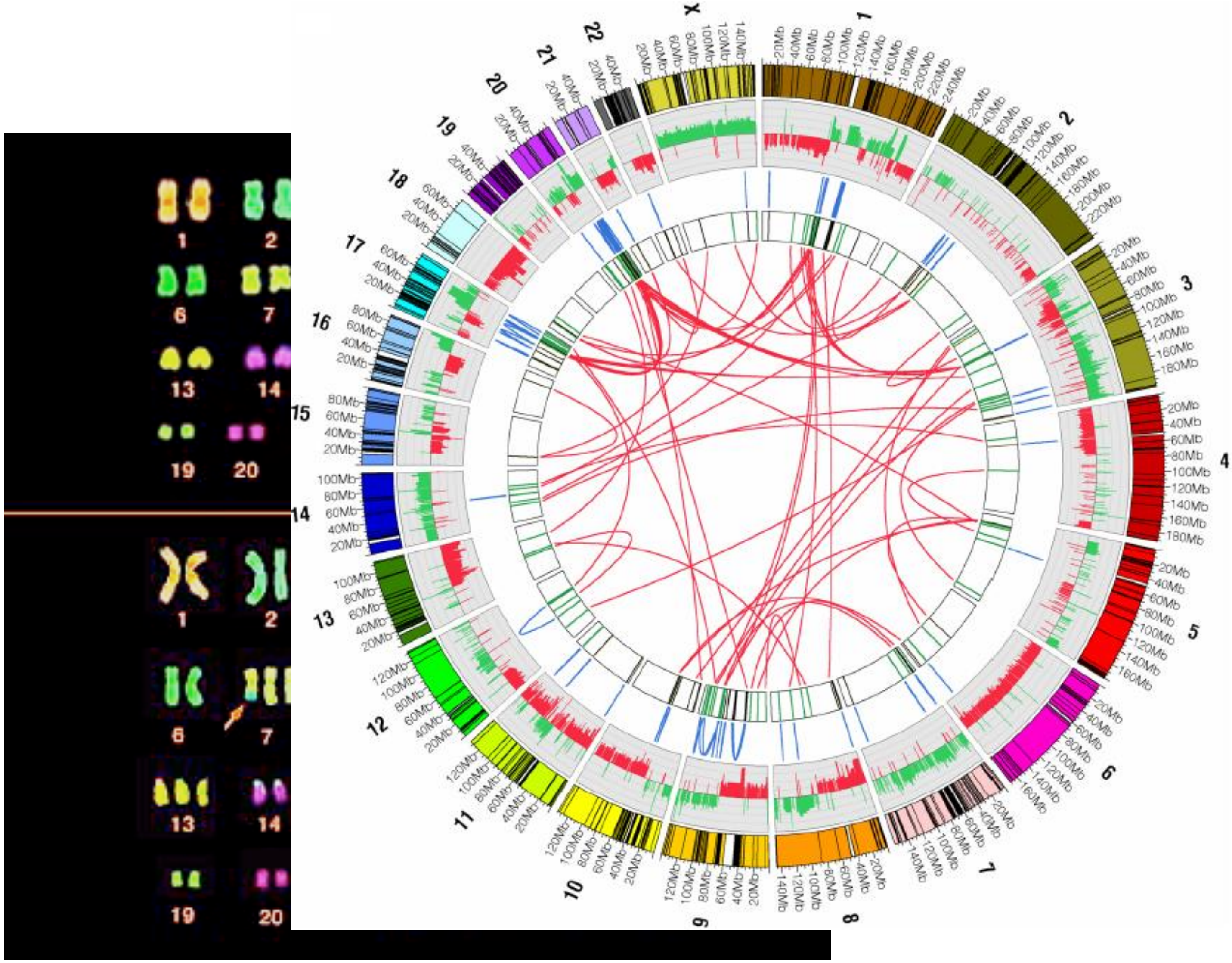


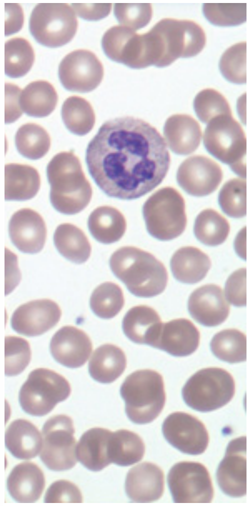
Epigenome

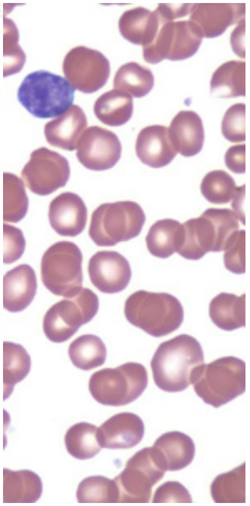
All cancers are different


NORMAL CELL
FIRST MUTATION
Cell seems normal but is predisposed to proliferate excessively
SECOND MUTATION
Cell begins to proliferate too much but is otherwise normal
THIRD MUTATION
Cell proliferates more rapidly; it also undergoes structural changes
MALIGNANT CELL
FOURTH OR LATER MUTATION
Cell grows uncontrollably and looks obviously deranged

Cancer: different views

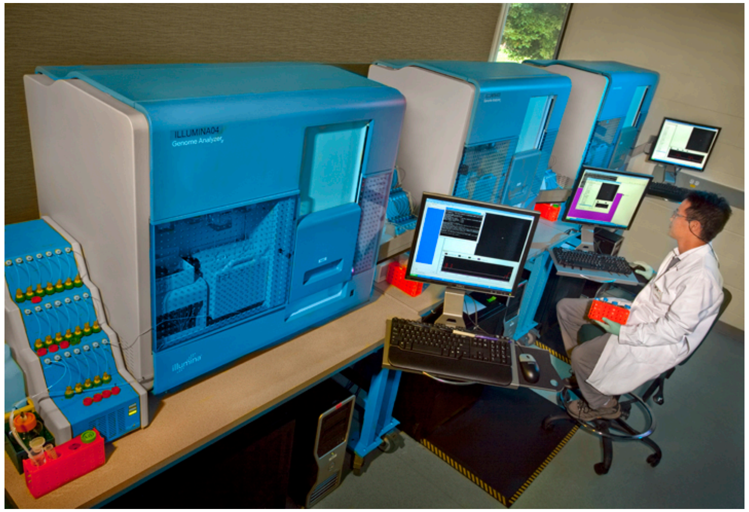


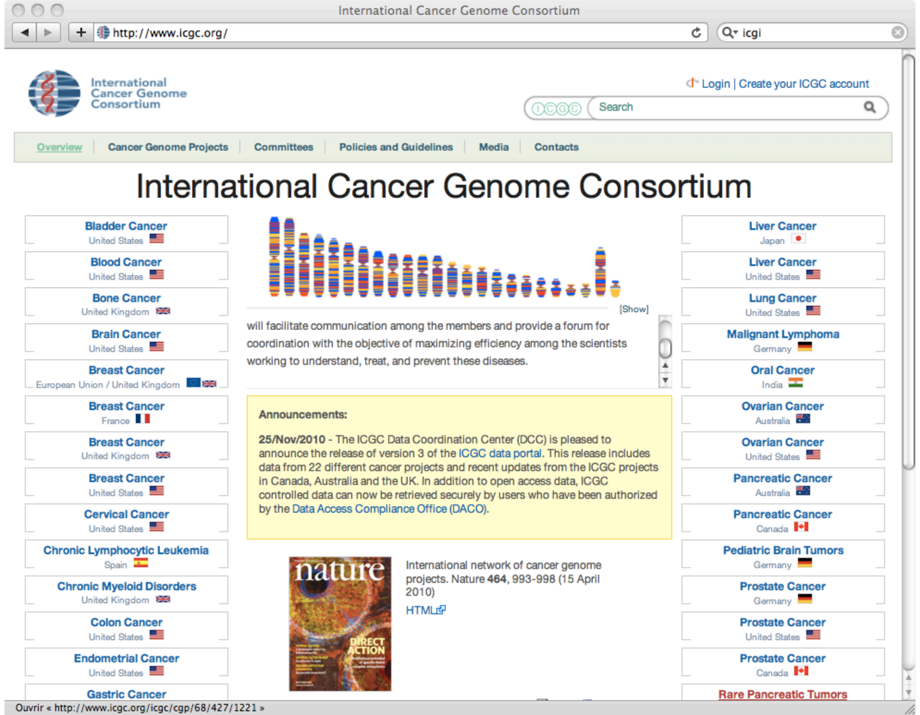




Big data!

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





International Cancer Genome Consortium

Bladder Cancer
United States

Blood Cancer
United States

Bone Cancer
United Kingdom

Brain Cancer
United States

Breast Cancer
European Union / United Kingdom

Breast Cancer
France

Breast Cancer
United Kingdom

Breast Cancer
United States

Cervical Cancer
United States

Chronic Lymphocytic Leukemia
Spain

Chronic Myeloid Disorders
United Kingdom

Colon Cancer
United States

Endometrial Cancer
United States

Gastric Cancer

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.

Announcements:

25/Nov/2010 - The ICGC Data Coordination Center (DCC) is pleased to announce the release of version 3 of the ICGC data portal. This release includes data from 22 different cancer projects and recent updates from the ICGC projects in Canada, Australia and the UK. In addition to open access data, ICGC controlled data can now be retrieved securely by users who have been authorized by the Data Access Compliance Office (DACO).

International network of cancer genome projects. Nature **464**, 993-998 (15 April 2010)

Liver Cancer
Japan

Liver Cancer
United States

Lung Cancer
United States

Malignant Lymphoma
Germany

Oral Cancer
India

Ovarian Cancer
Australia

Ovarian Cancer
United States

Pancreatic Cancer
Australia

Pancreatic Cancer
Canada

Pediatric Brain Tumors
Germany

Prostate Cancer
Germany

Prostate Cancer
United States

Prostate Cancer
Canada

Rare Pancreatic Tumors

Ouvrir « http://www.icgc.org/icgc/cgp/68/427/1221 »

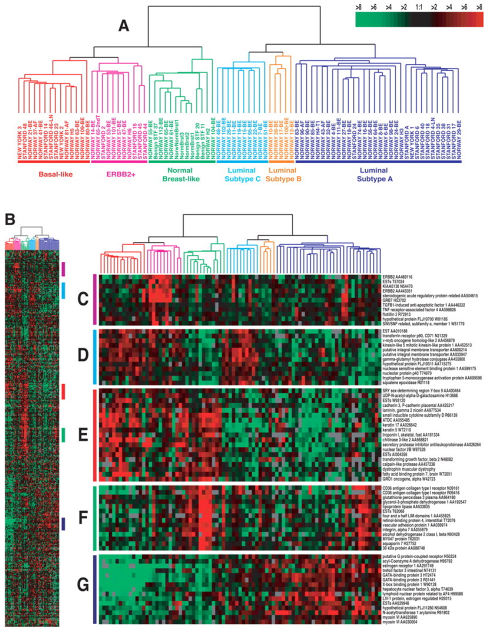
23andMe

Predictive

P4 Medicine

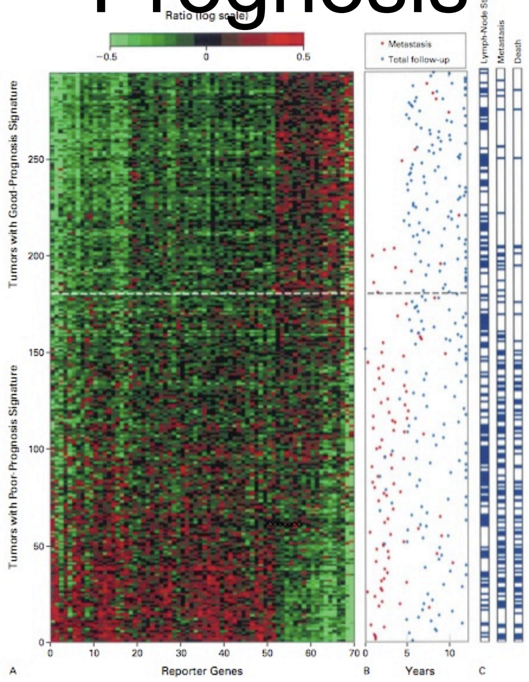
● PREDICT ● PREVENT ● PERSONALIZE ● PARTICIPATE

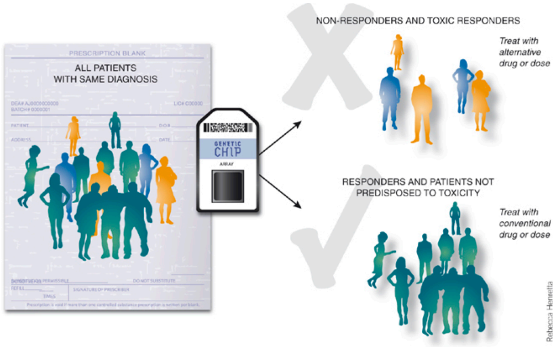
Opportunities



Diagnosis

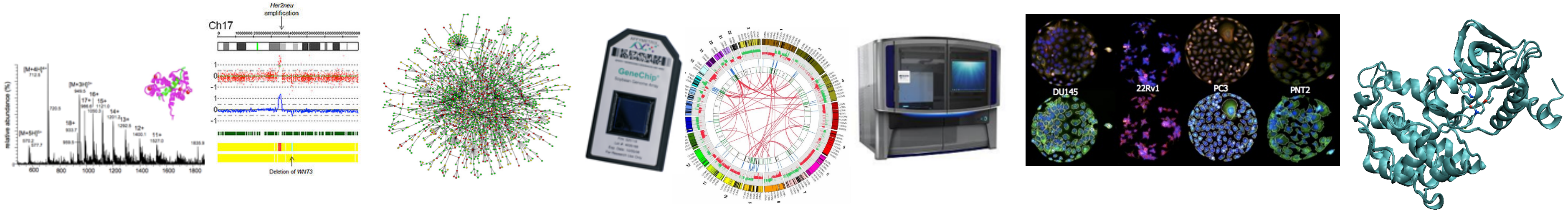
Prognosis

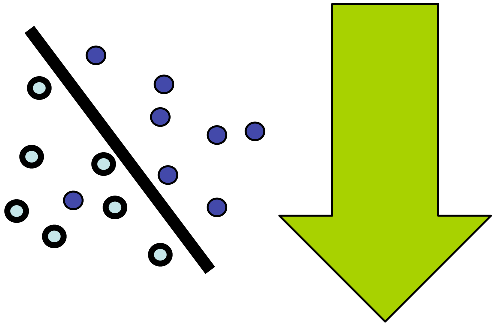



NON-RESPONDERS AND TOXIC RESPONDERS
Treat with alternative drug or dose
ALL PATIENTS WITH SAME DIAGNOSIS
GENETIC CHIP
RESPONDERS AND PATIENTS NOT PREDISPOSED TO TOXICITY
Treat with conventional drug or dose
Rebecca Henretta

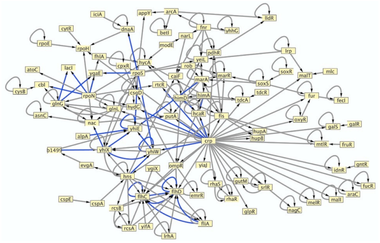
Response to drugs

Rationale of my team

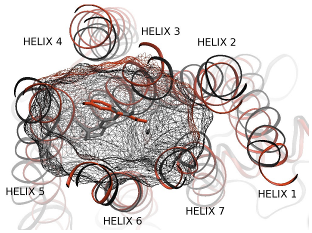




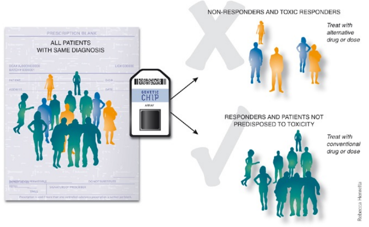
Machine learning



Mecanisms, drug targets

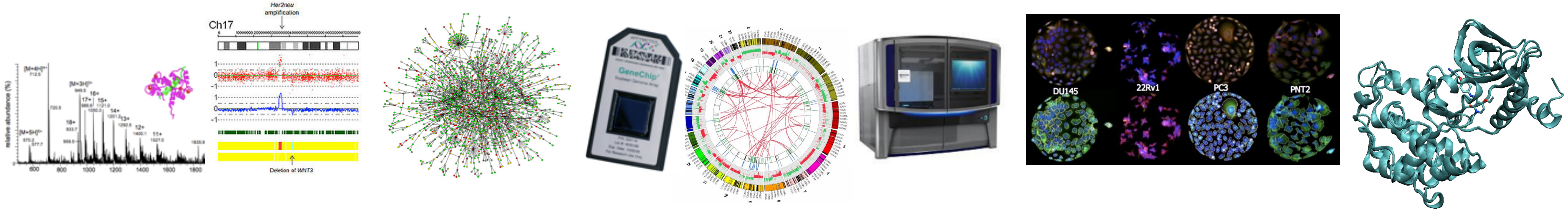


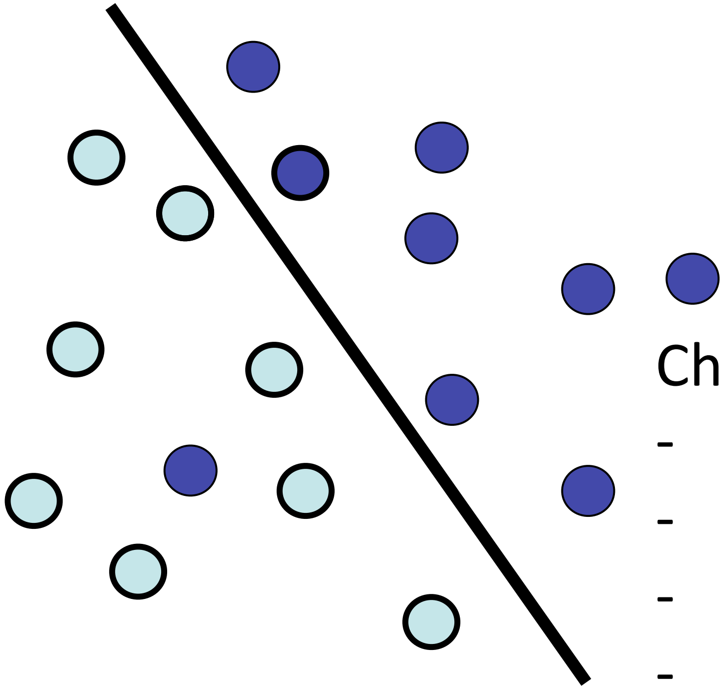
Drug design


NON-RESPONDERS AND TOXIC RESPONDERS
Treat with alternative drug or dose
ALL PATIENTS WITH SAME DIAGNOSIS
GENETIC CHIP
RESPONDERS AND PATIENTS NOT PREDISPOSED TO TOXICITY
Treat with conventional drug or dose

Personalized medicine

Machine Learning?

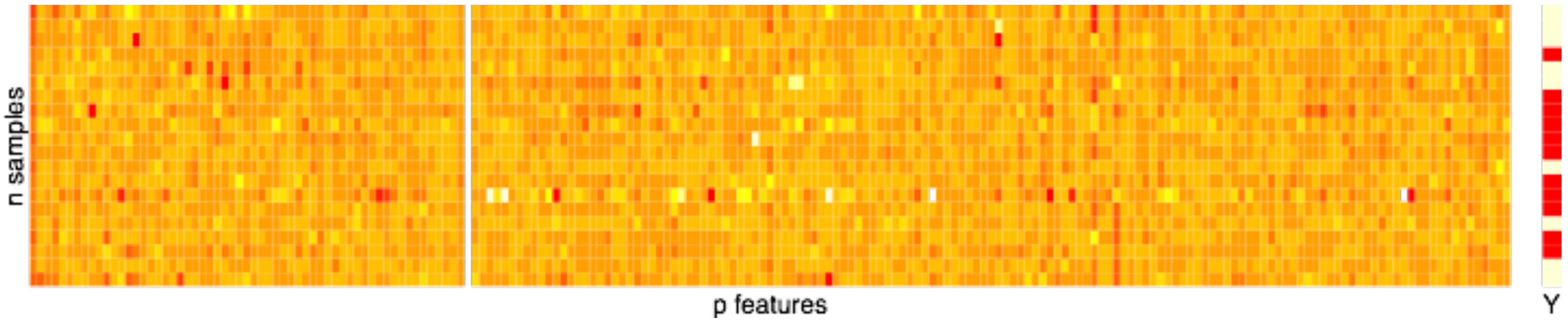


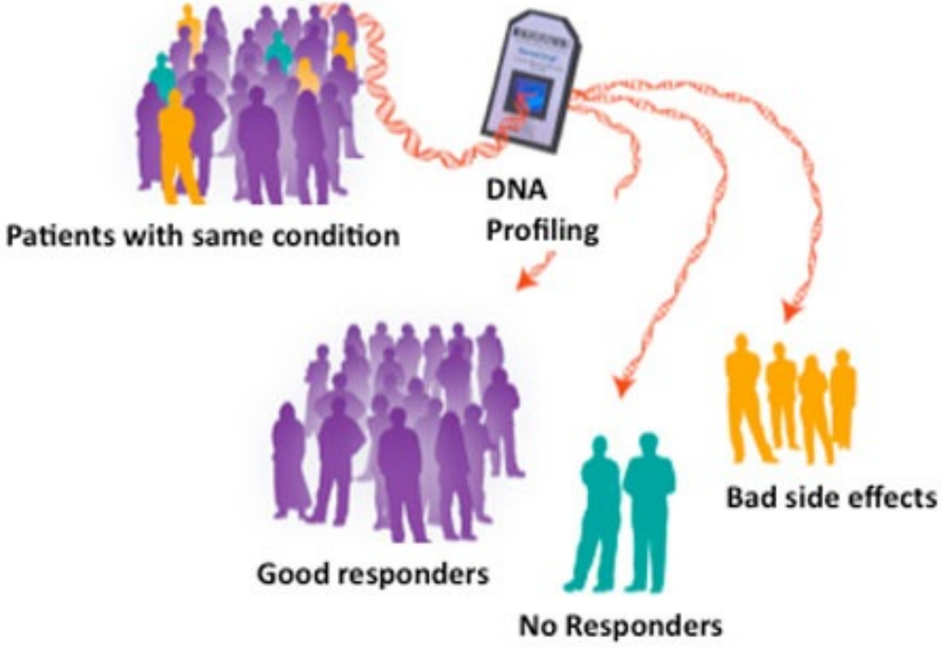


Challenges

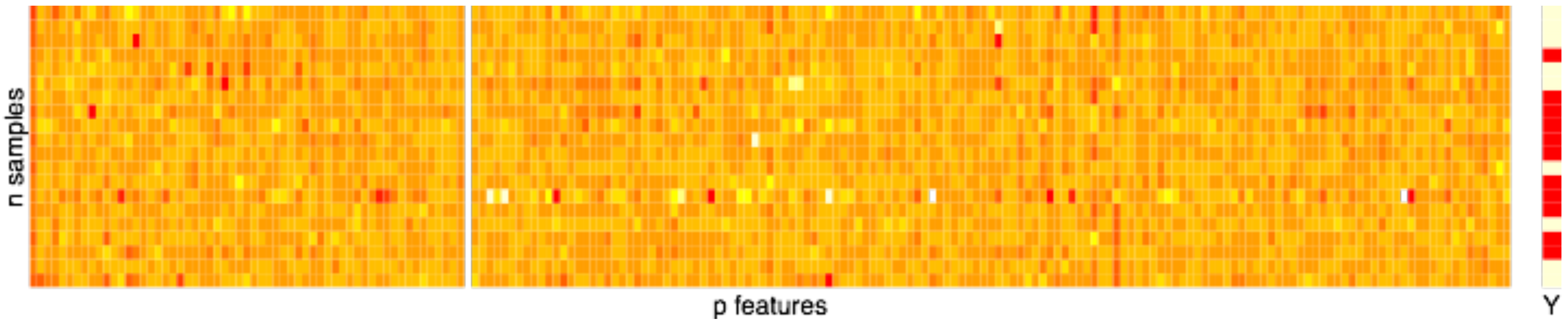
- **High dimension**
- **Few examples**
- Structured data
- Efficient algorithms
- Interpretability

Example: Patient stratification





Problem : n << p

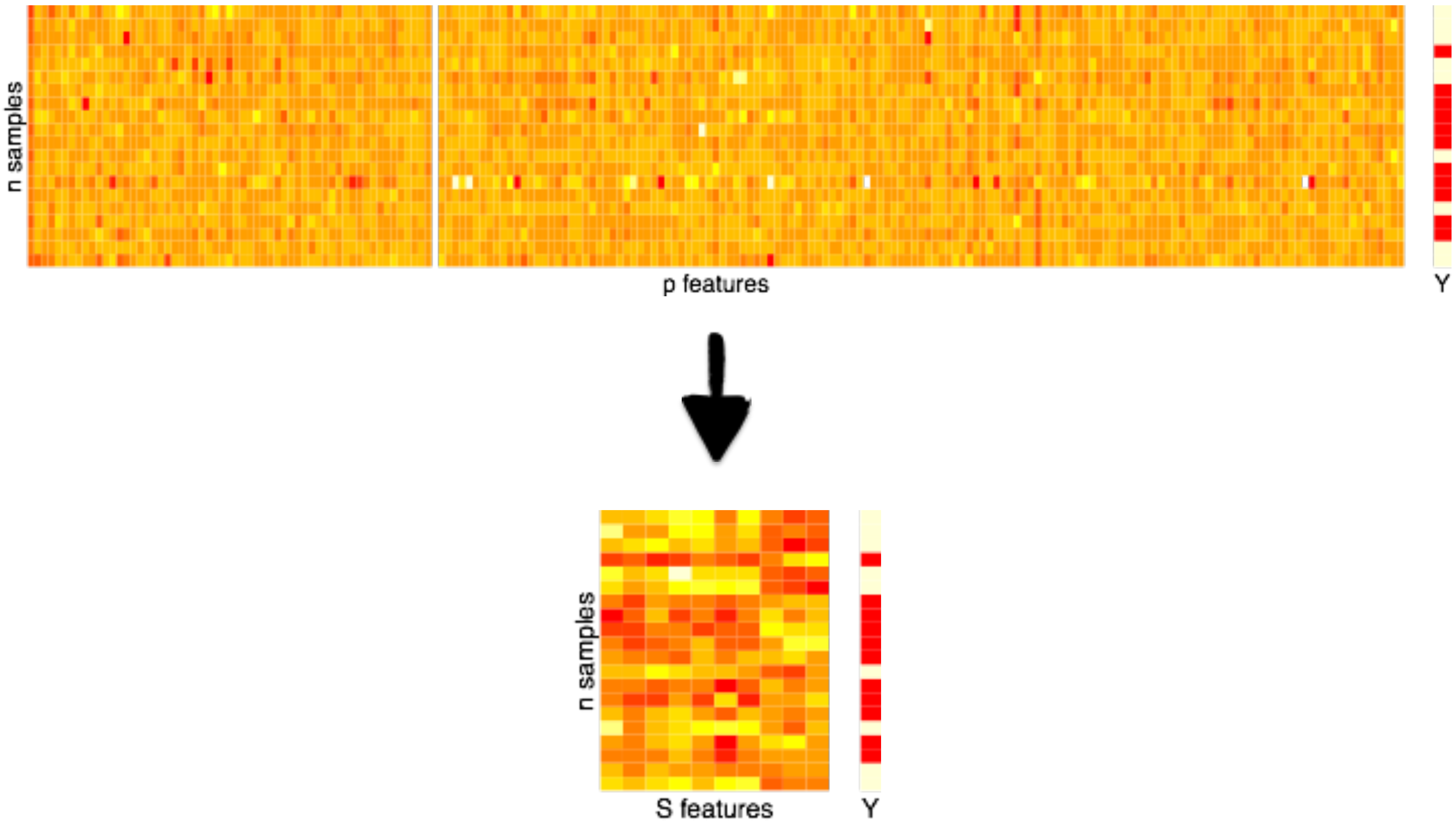


Patients with same condition
DNA
Profiling
Good responders
No Responders
Bad side effects

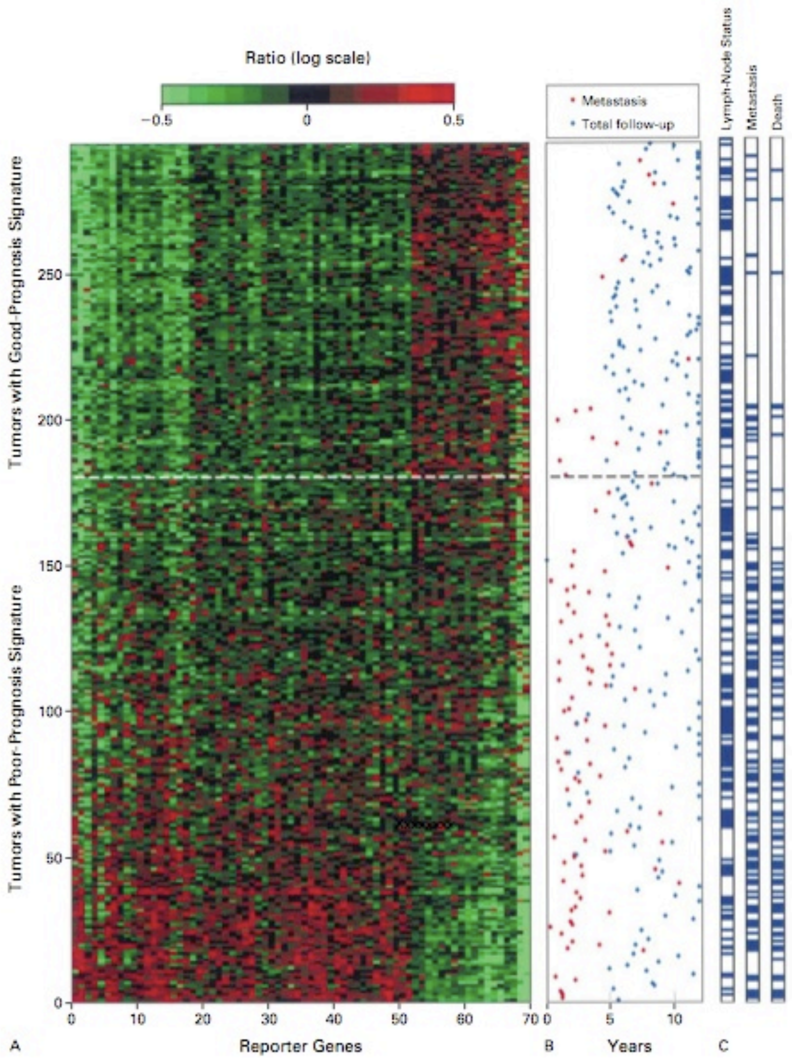
n = 1E2 ~ 1E4
(patients)

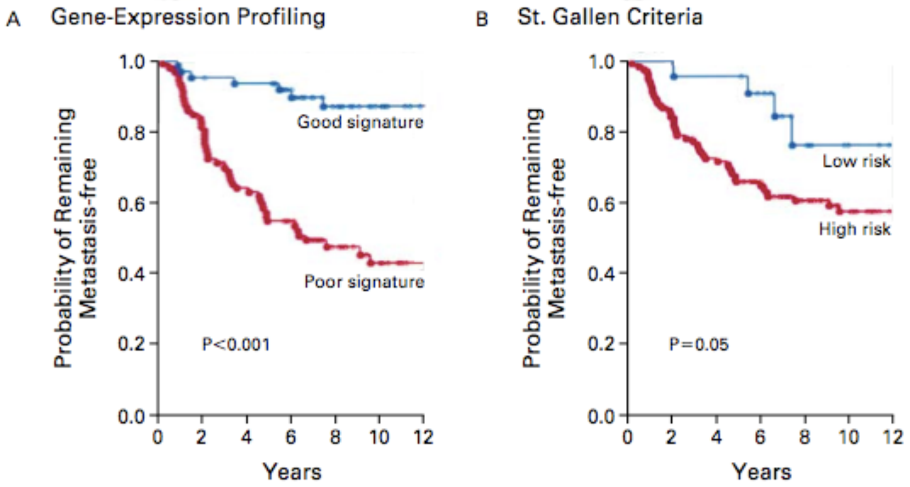
p = 1E4 ~ 1E7
(genes, mutations,
copy numbers, …)

Feature Selection



Example:
Breast cancer prognostic signature





No. at Risk

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

No. at Risk

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

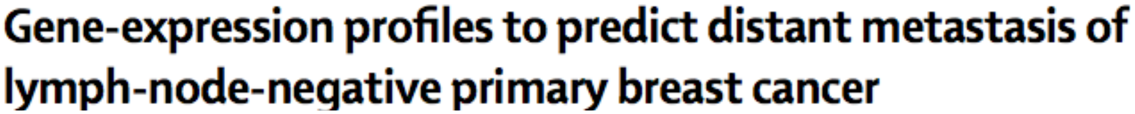
(Van de Vijver et al 2002)

But…

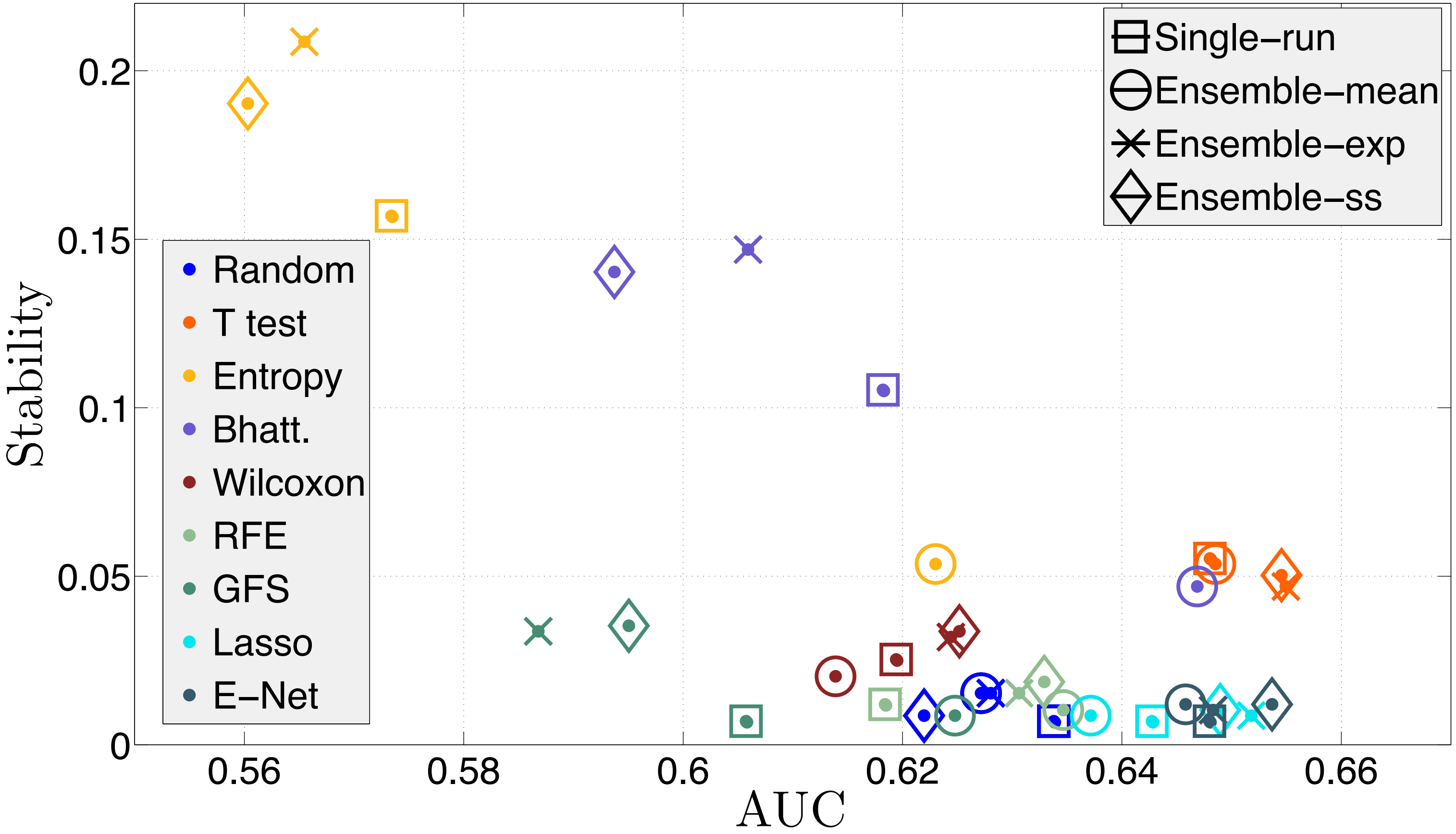
Gene expression profiling predicts clinical outcome of breast cancer

Laura J
Yudong
Karin va
George
Peter S.

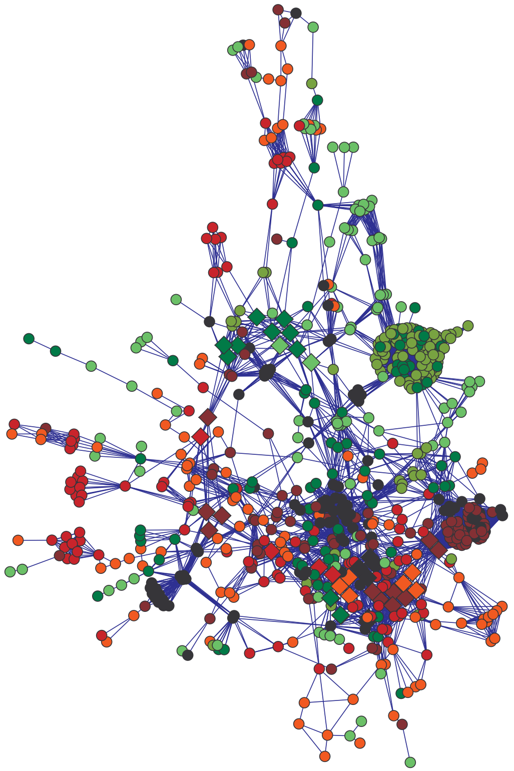
* Divisio
and Cen
121 Ples
‡ Rosetta

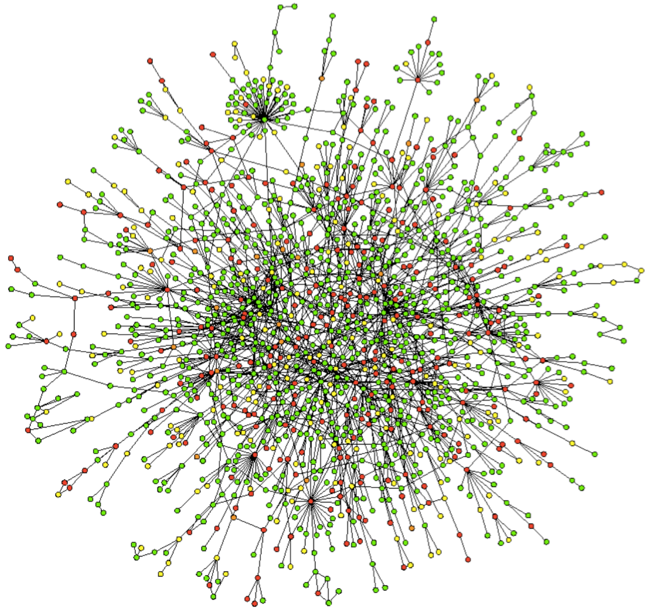


Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer



Prior knowledge: gene network

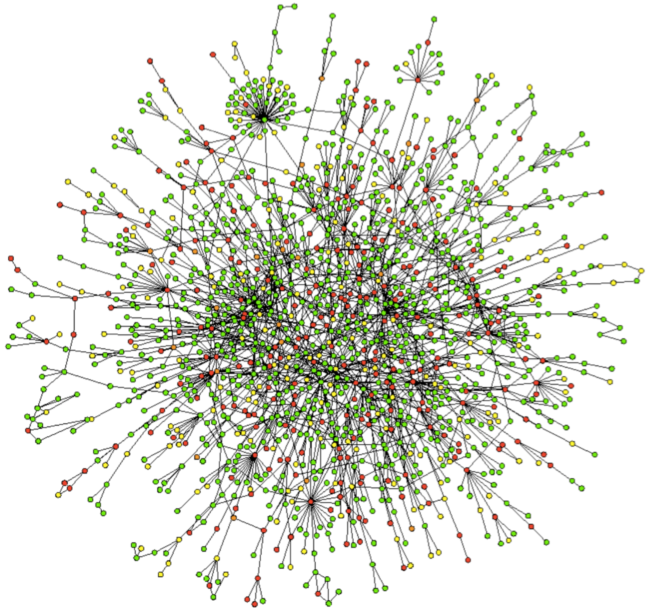


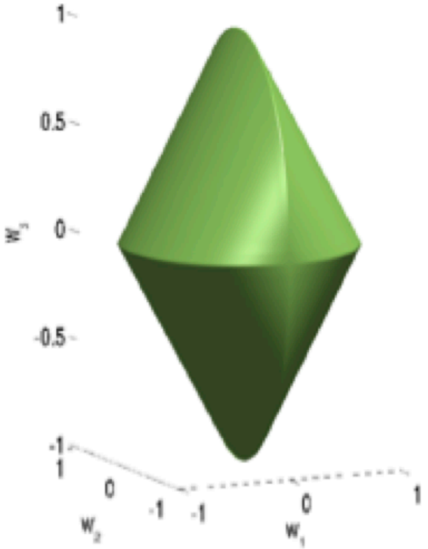


Can we « force » the signature to be « coherent » with a known network?

Example: the graph lasso

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

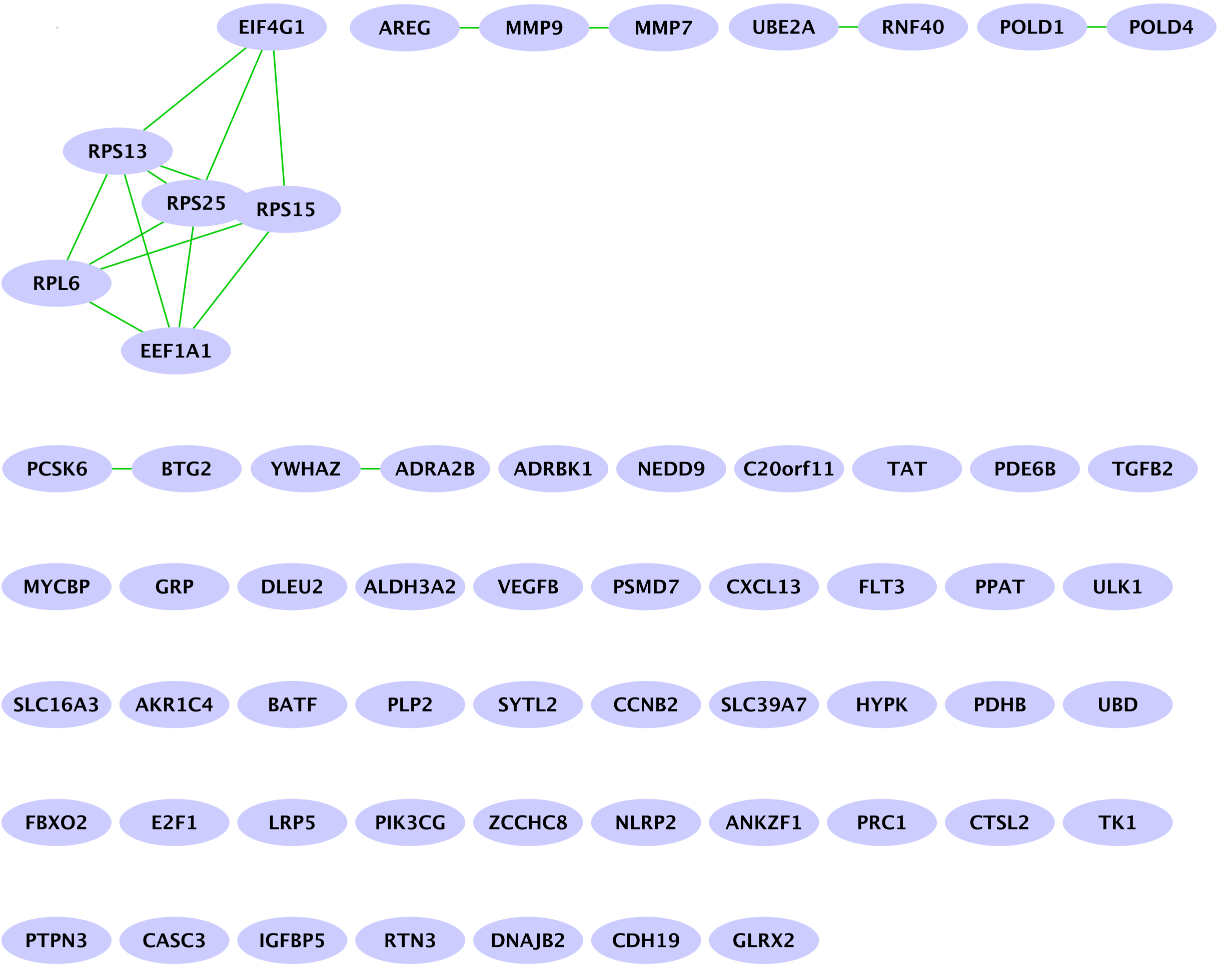




- Step 2: Among the candidates, find the best signature to explain the data

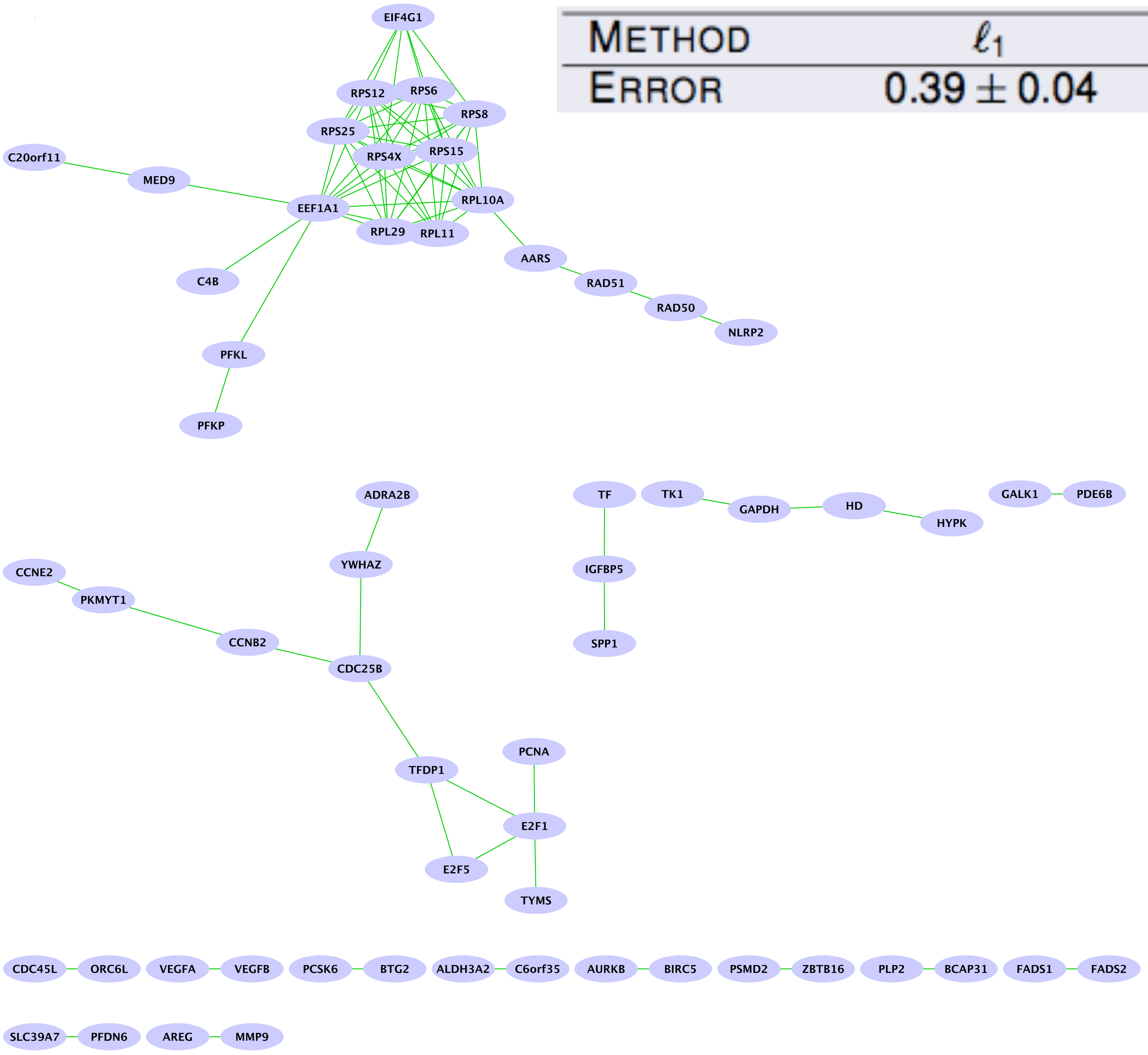
(Jacob et al 2009)

Classical signature

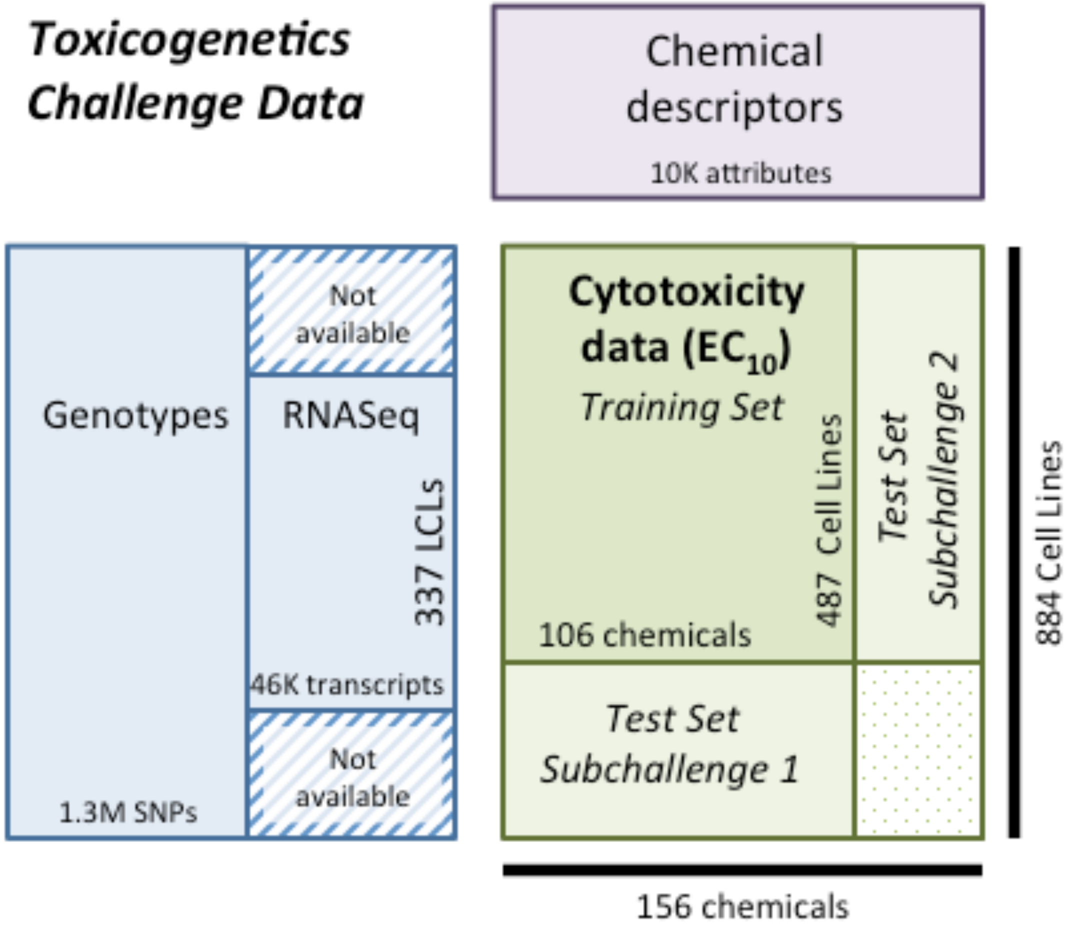


The graph lasso signature

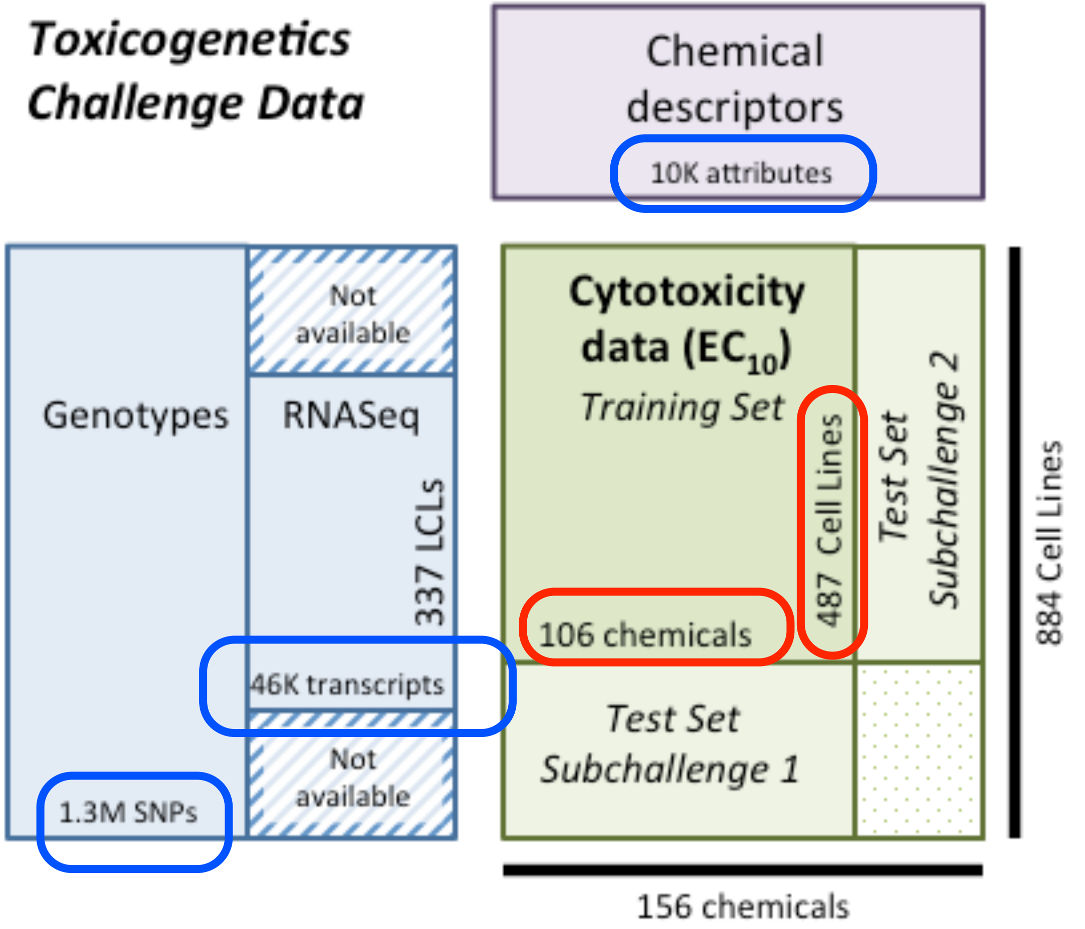
METHOD	ℓ_1	$\Omega_{graph}(.)$
ERROR	0.39 ± 0.04	0.36 ± 0.01



Example: Toxicogenetics / Pharmacogenomics



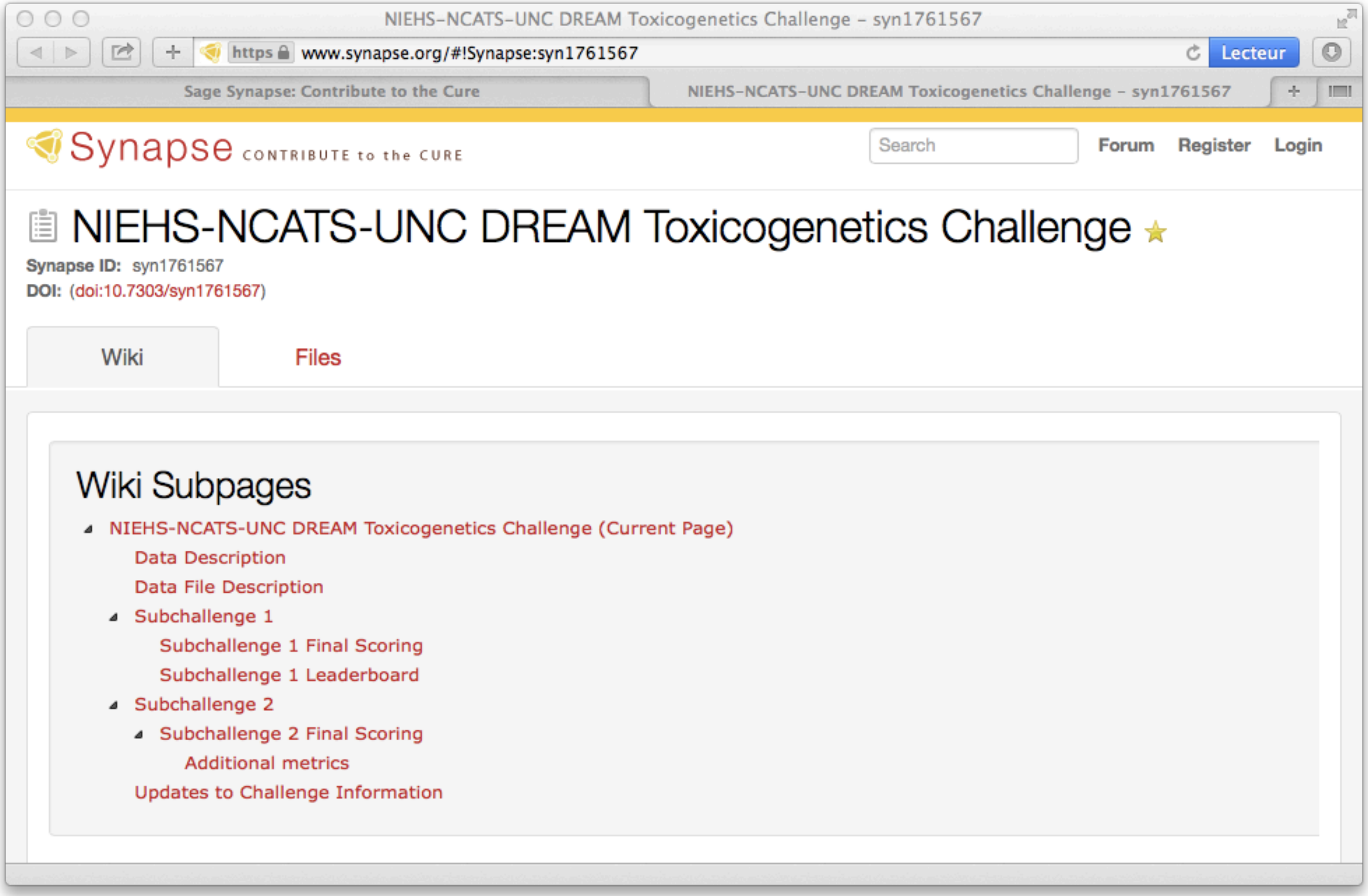
Problem: n << p



n = 5E4

p = 1E10

Crowd-sourcing initiatives



NIEHS-NCATS-UNC DREAM Toxicogenetics Challenge - syn1761567
https www.synapse.org/#!Synapse:syn1761567
Lecteur
Sage Synapse: Contribute to the Cure
NIEHS-NCATS-UNC DREAM Toxicogenetics Challenge - syn1761567
Synapse CONTRIBUTE to the CURE
Search
Forum
Register
Login

NIEHS-NCATS-UNC DREAM Toxicogenetics Challenge ★

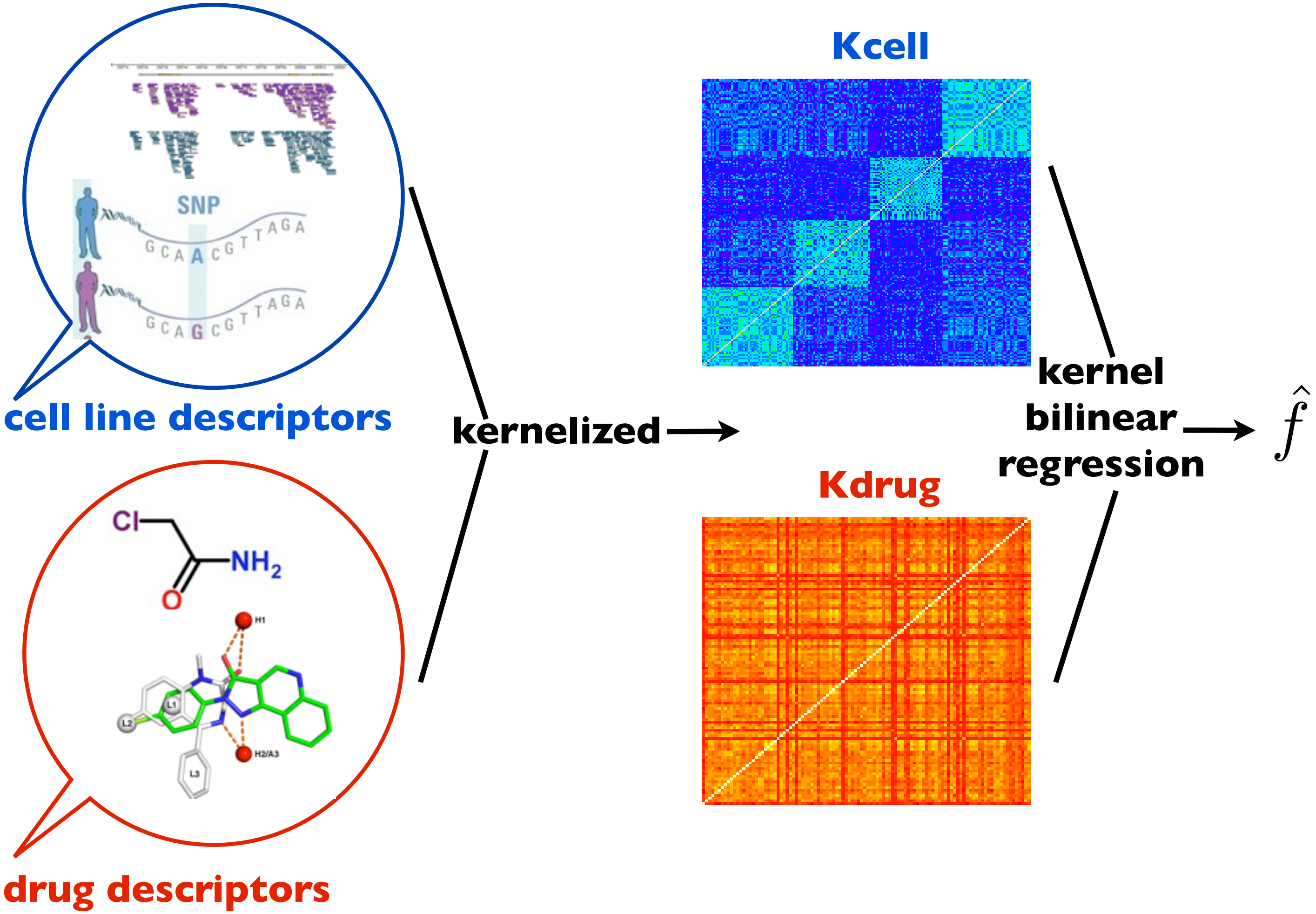
Synapse ID: syn1761567

DOI: (doi:10.7303/syn1761567)

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

Our approach



cell line descriptors

drug descriptors

kernelized →

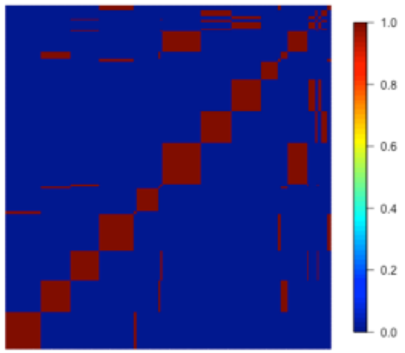
Kcell

Kdrug

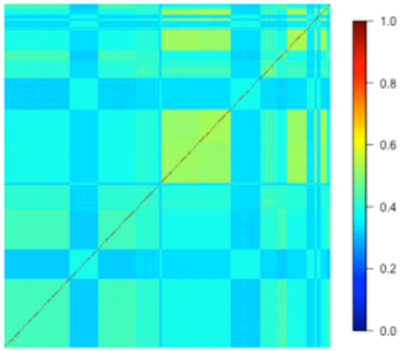
kernel bilinear regression → $\hat{f}$

Cell line descriptors (30 kernels)

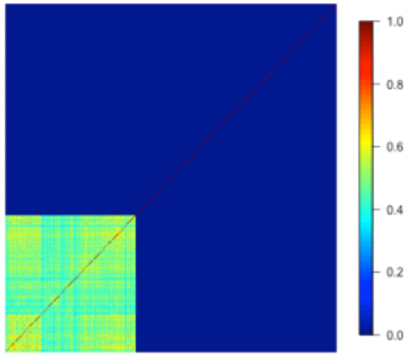
Covariates
. linear kernel



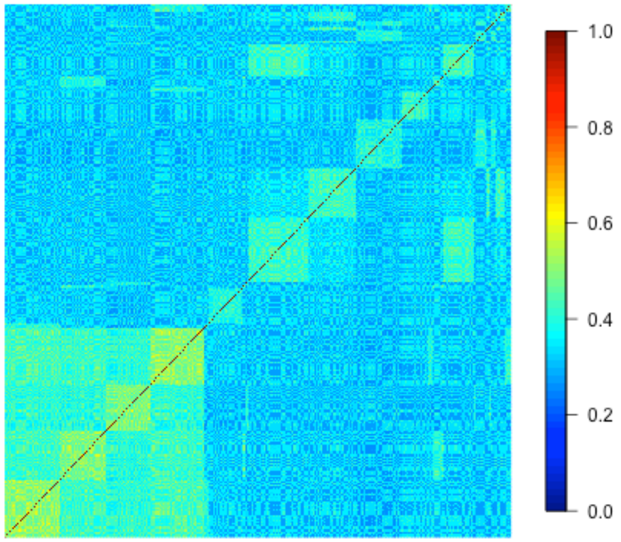
SNPs
. 10 gaussian kernels



RNA-seq
. 10 gaussian kernels

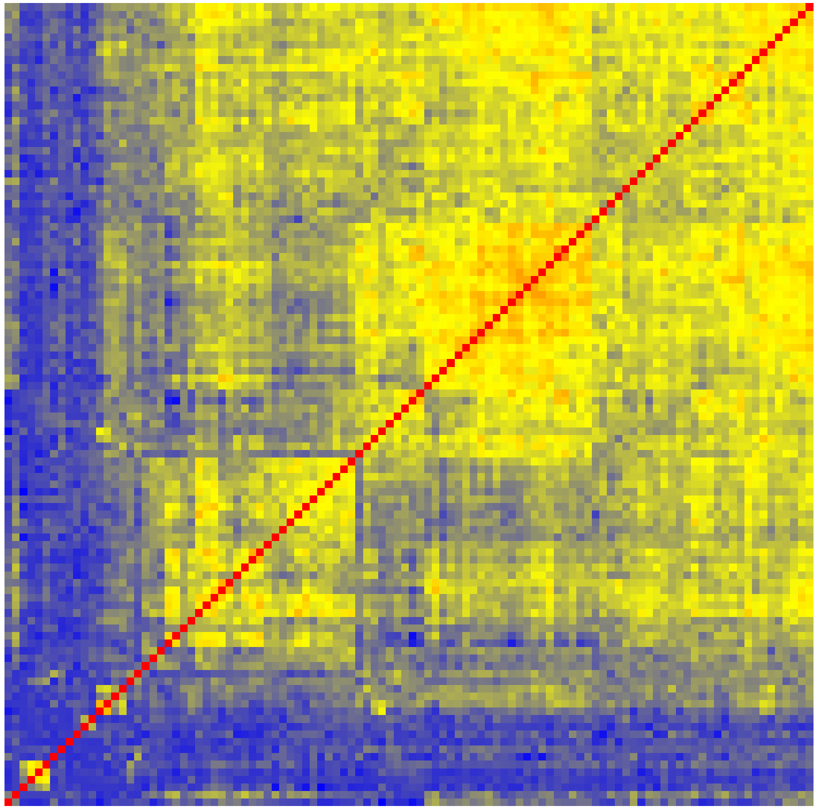


Integrated kernel

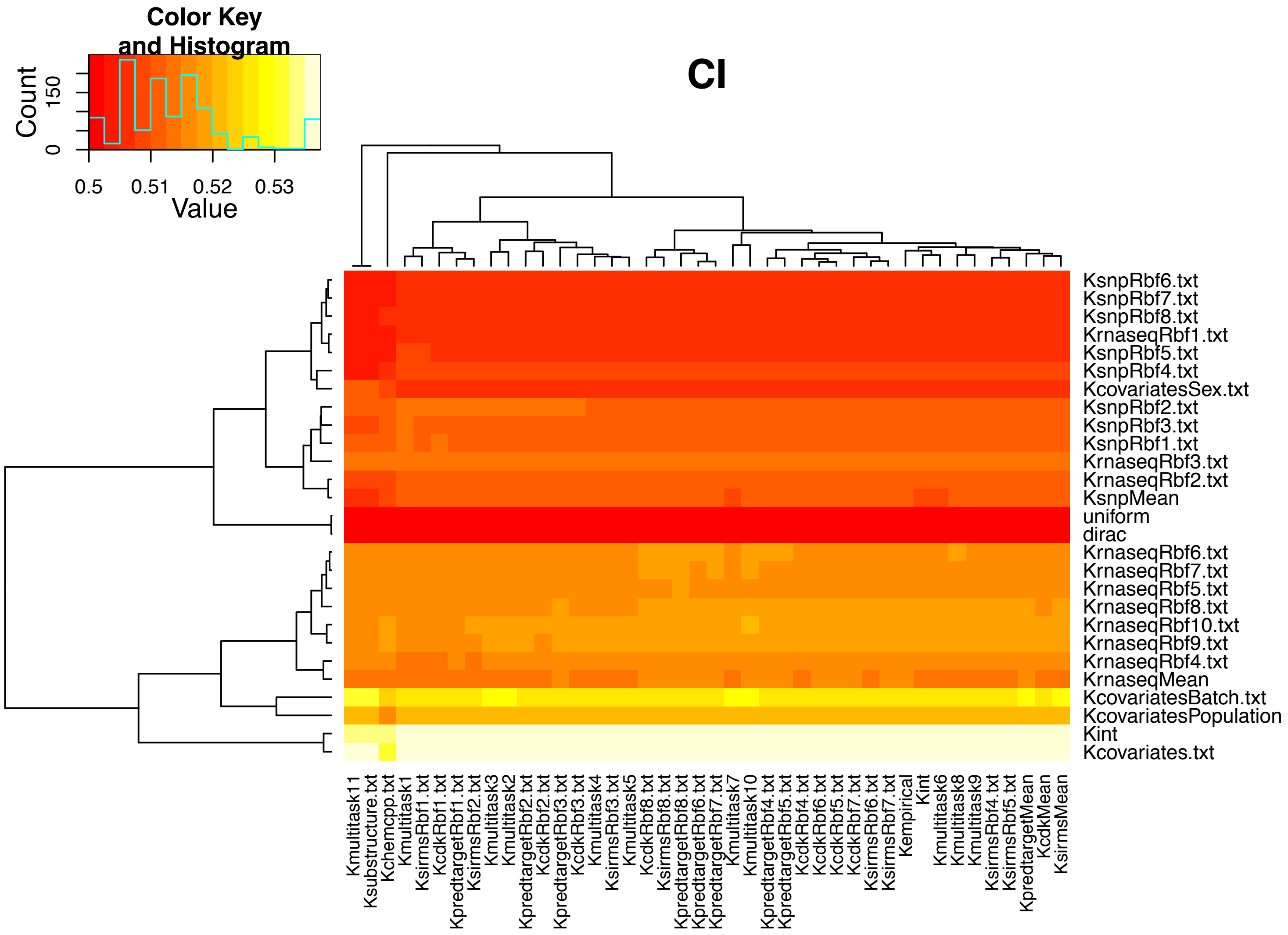


Chemical descriptors (49 kernels)

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

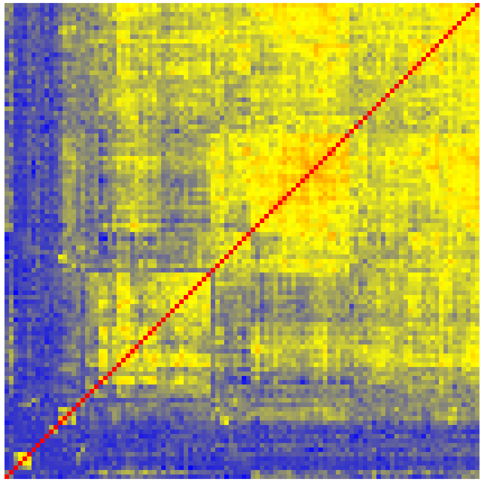


Learning occurs…

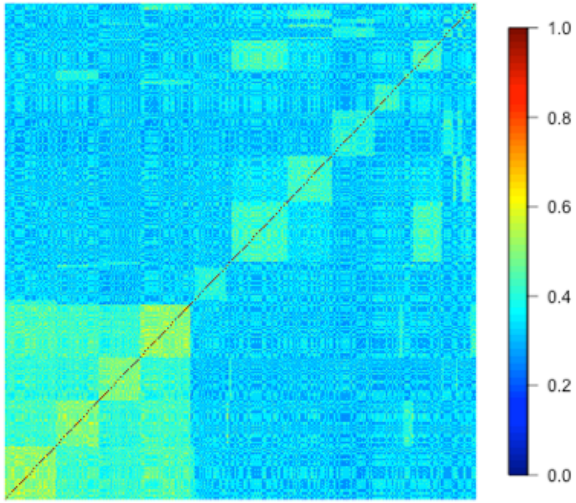


Final submission (ranked 2nd)

Empirical kernel on drugs



Integrated kernel on cell lines



0.54

RECOMB/ISCB Conference on Regulatory and Systems Genomics, with DREAM Challenges 2013

TORONTO, ONTARIO
NOV 8 - 12, 2013

Conclusion

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

I'M WORRIED THAT HEALTH CARE HAS BECOME TOO IMPERSONAL, DOC.
NONSENSE... JUST RELAX AND LIE BACK ON THE BAR CODE SCANNER.
2229901055

Rob Rogers / Pittsburgh Post-Gazette

Thanks!

erc

SEVENTH FRAMEWORK PROGRAMME

INVESTISSEMENTS D'AVENIR

AGENCE NATIONALE DE LA RECHERCHE ANR

BIOMÉRIEUX

MINES ParisTech

institutCurie
Together, let's beat cancer.

Inserm
Institut national
de la santé et de la recherche médicale

PSL RESEARCH UNIVERSITY