A Cautionary Tale about Genomic Biomarkers in Cancer

Benjamin Haibe-Kains

Principal Investigator, Bioinformatics and Computational Genomics Laboratory
Assistant Professor, Medical Biophysics, University of Toronto

The Princess Margaret Cancer Centre
University Health Network

(Random) Prognostic Biomarkers

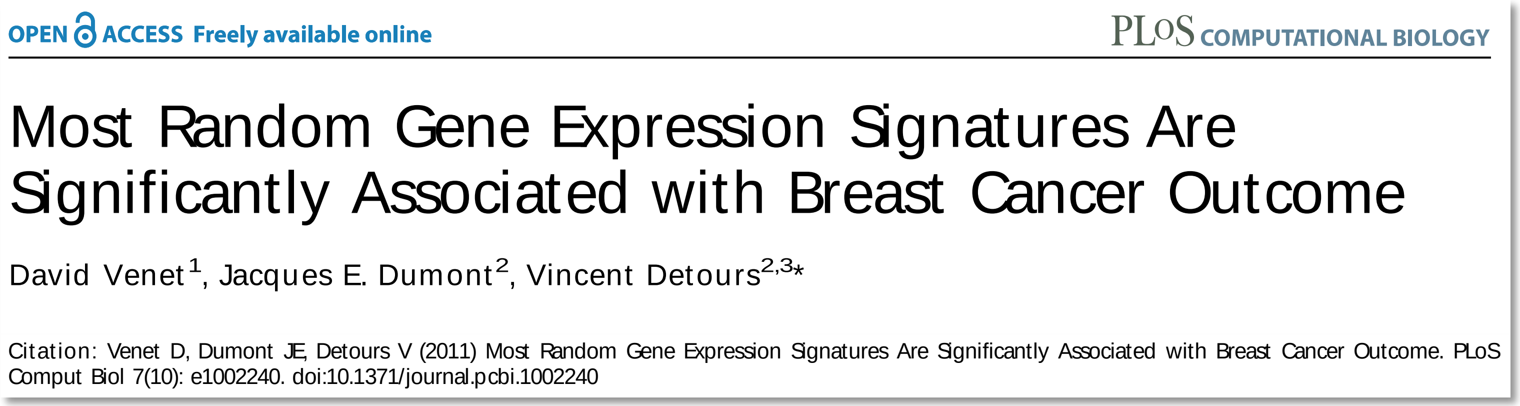
Prognostic gene signatures (aka *genesets*)

- Thanks to high-throughput technologies, the number of publications reporting biomarkers in cancer literally **exploded**
- >3500 gene expression *signatures* have been published so far (MSigDB, GeneSigDB)
- Roughly 300 signatures per year, almost one new signature published every day …

Prognostic value of genesets

- Common practice to claim **biological relevance** for a signature/geneset yielding significant prognostic value

- Venet et al. showed that most **random** genesets can be used to significantly discriminate between low and high-risk breast cancer patients



OPEN ACCESS Freely available online

PLoS COMPUTATIONAL BIOLOGY

Most Random Gene Expression Signatures Are Significantly Associated with Breast Cancer Outcome

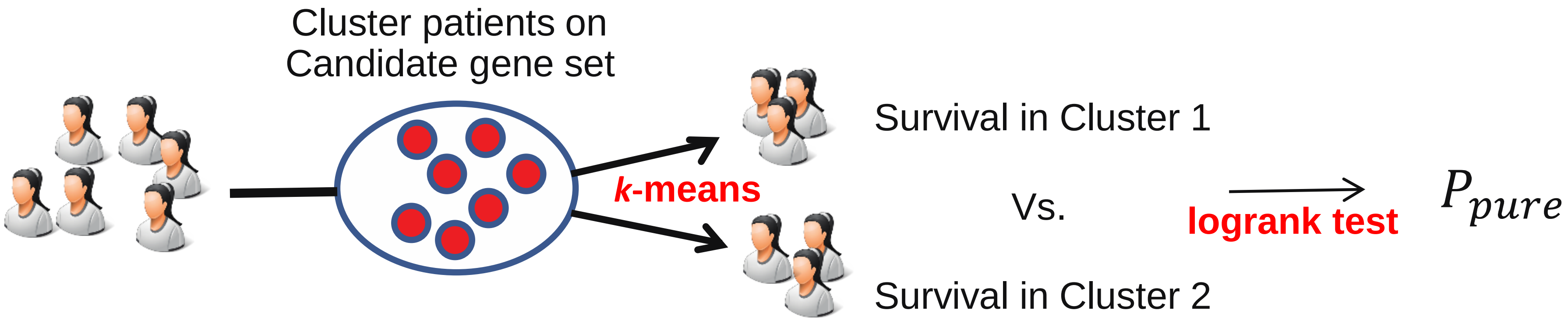
David Venet[1], Jacques E. Dumont[2], Vincent Detours[2,3]*

Citation: Venet D, Dumont JE, Detours V (2011) Most Random Gene Expression Signatures Are Significantly Associated with Breast Cancer Outcome. PLoS Comput Biol 7(10): e1002240. doi:10.1371/journal.pcbi.1002240

Generalization of Venet’s results

- We first checked whether these results still hold when analyzing **(much) more** datasets
- We used a compendium of 36 breast cancer microarray datasets (~4000 patients with survival information)
- Prognostic

These collection of curated datasets will be available soon in ***InSilicoDB***



Cluster patients on Candidate gene set

***k*-means**

Survival in Cluster 1

Vs.

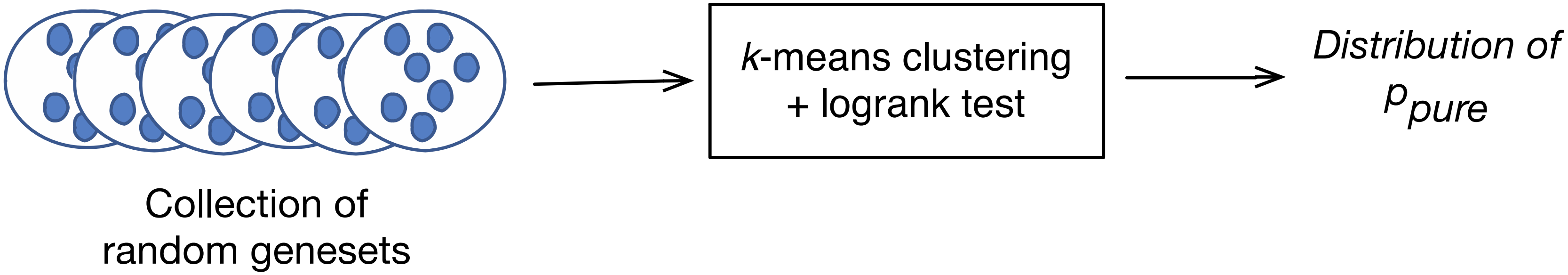
logrank test

P_{pure}

Survival in Cluster 2

Are random genesets prognostic?

- We generated 1000 random genesets for each size and tested their prognostic value

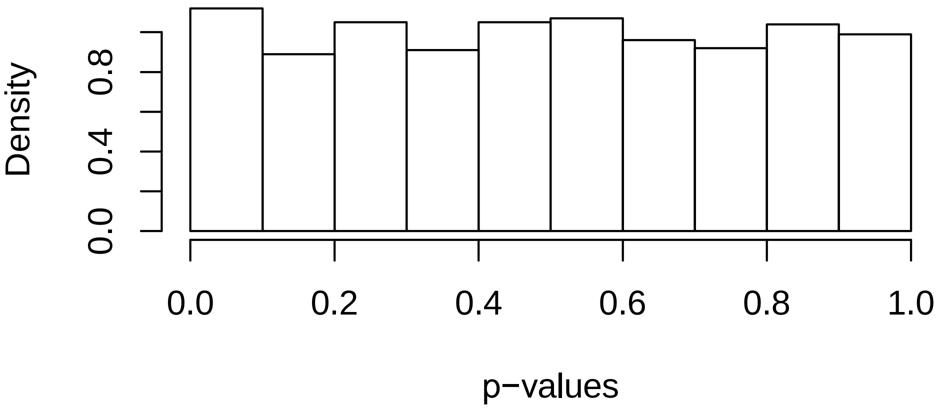


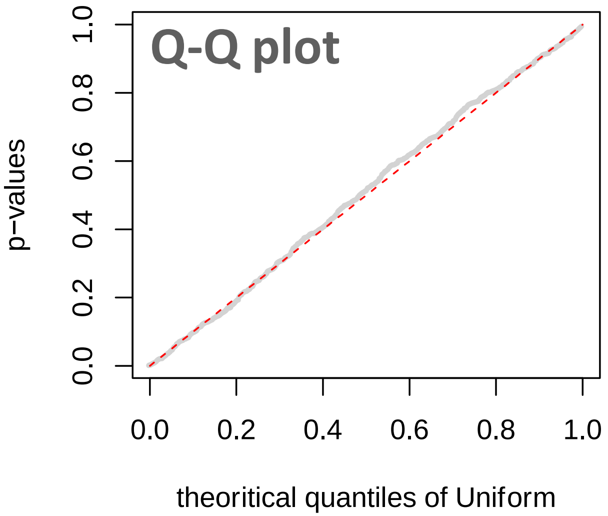
Collection of random genesets

k-means clustering + logrank test

Distribution of p_{pure}

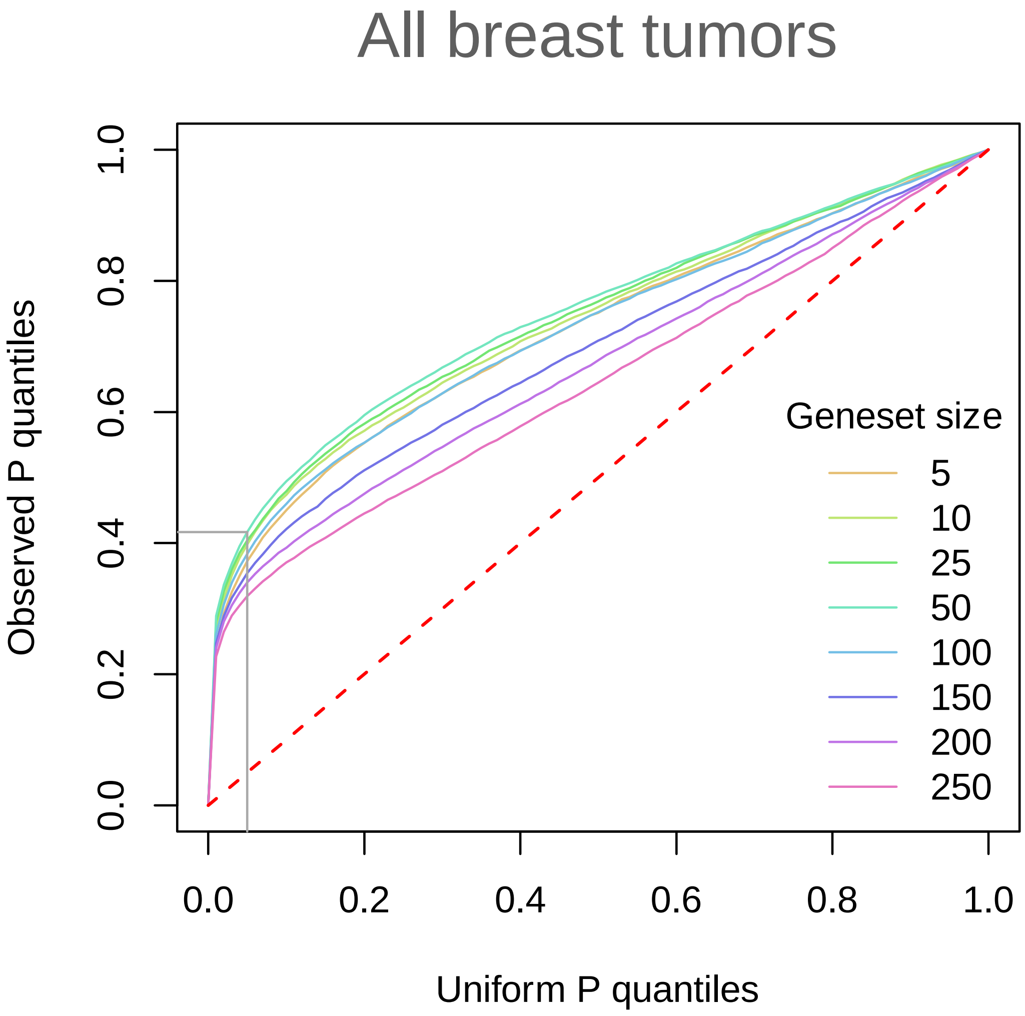
- If the assumptions of the log-rank test are met, p-values for random genesets should be approx. uniformly distributed





Many **random** genesets are prognostic

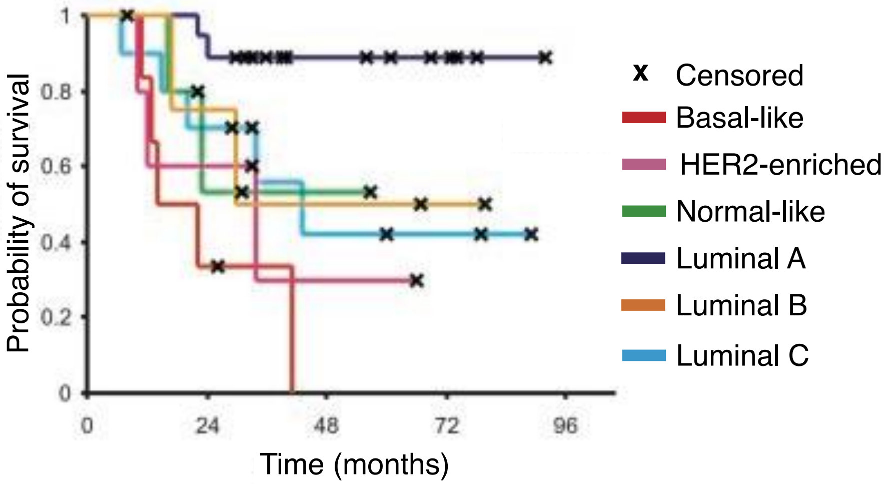
Global population of breast cancer patients

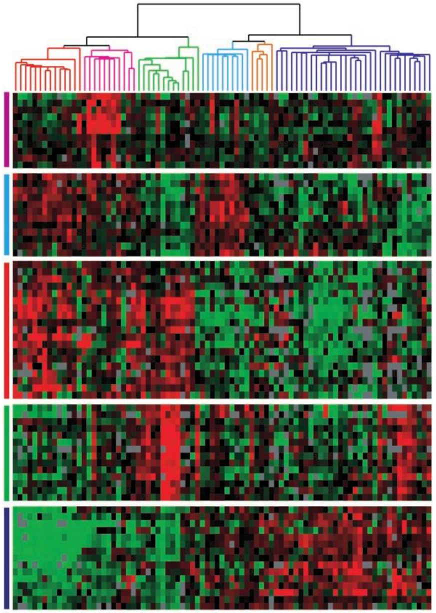


WARNING Breast cancers are heterogeneous

- Breast cancers are clinically diverse
    - Tumors with identical clinical parameters may lead to different outcomes
- And **molecularly** heterogeneous
    - Identification of subtypes based on gene expressions

- Perou et al. identified 4-6 subtypes exhibiting different clinical outcome

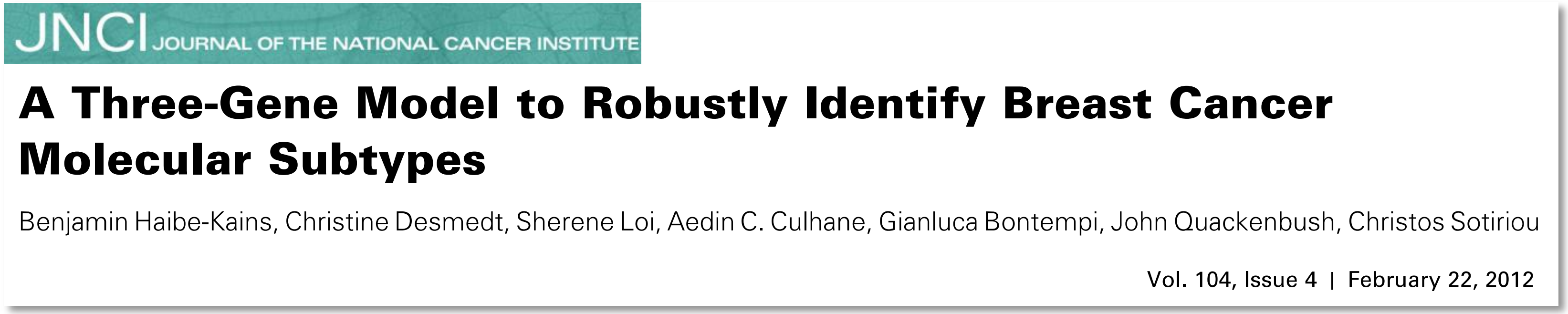




Stratification by molecular subtypes

- Are the results confounded by the presence of molecular subtypes?

- Subtyping using the robust **SCMGENE** classification model



JNCI JOURNAL OF THE NATIONAL CANCER INSTITUTE

A Three-Gene Model to Robustly Identify Breast Cancer Molecular Subtypes

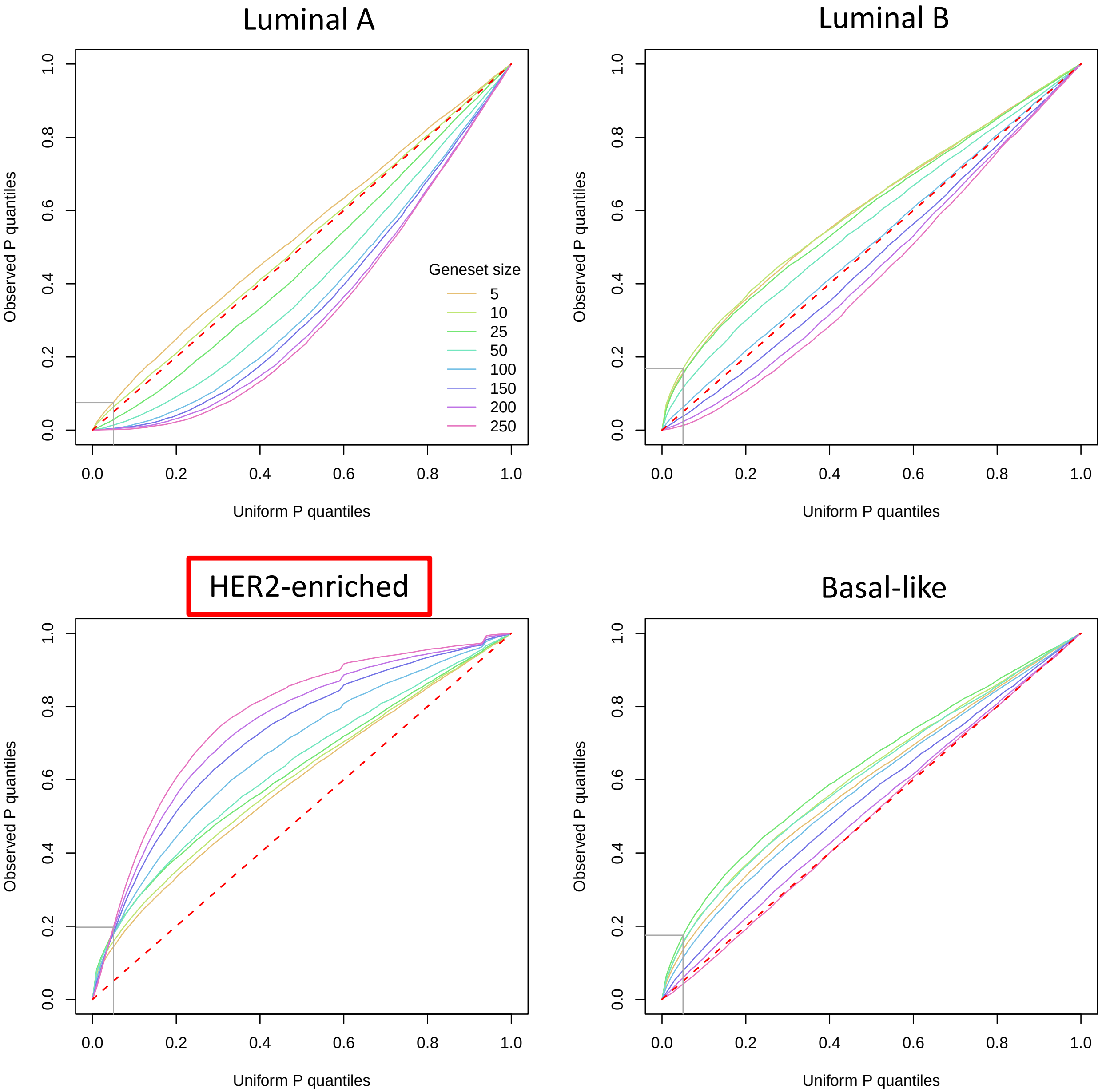
Benjamin Haibe-Kains, Christine Desmedt, Sherene Loi, Aedin C. Culhane, Gianluca Bontempi, John Quackenbush, Christos Sotiriou

Vol. 104, Issue 4 | February 22, 2012

- Four main subtypes:
  - ER+/HER2- Low Proliferation (**Luminal A**)
  - ER+/HER2- High Proliferation (**Luminal B**)
  - HER2+ (**HER2-enriched**)
  - ER-/HER2- (**Basal-like**)

Prognostic value depends on the subtypes

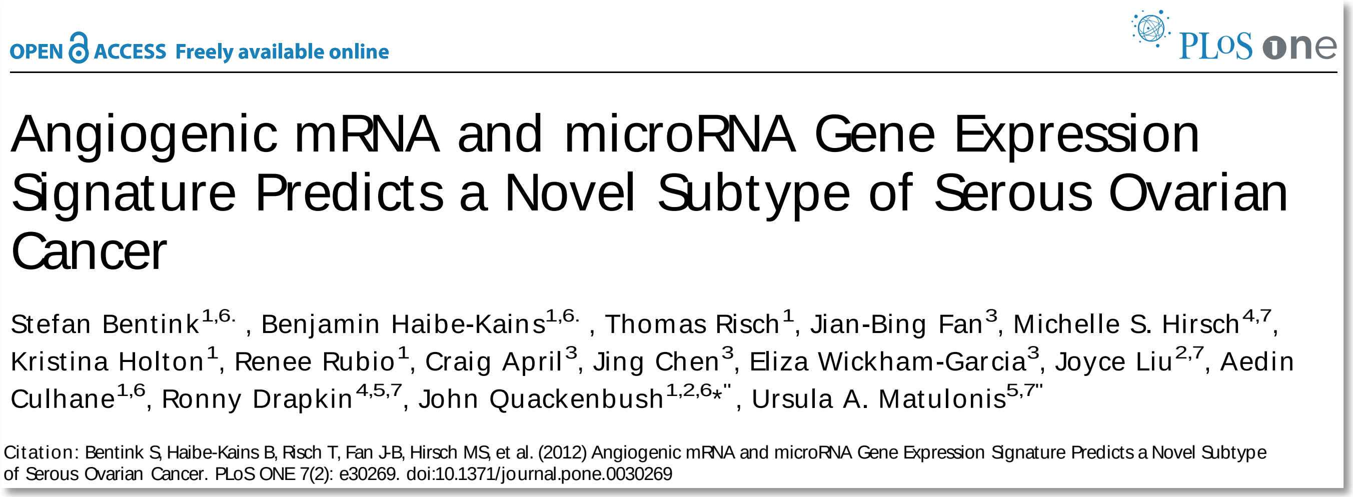
Breast cancer molecular subtypes



Analysis of ovarian cancer

- Collection of 11 datasets including ~1700 high-grade serous ovarian tumors

- Subtyping using the **AngioS** classification model



OPEN ACCESS Freely available online

PLoS one

Angiogenic mRNA and microRNA Gene Expression Signature Predicts a Novel Subtype of Serous Ovarian Cancer

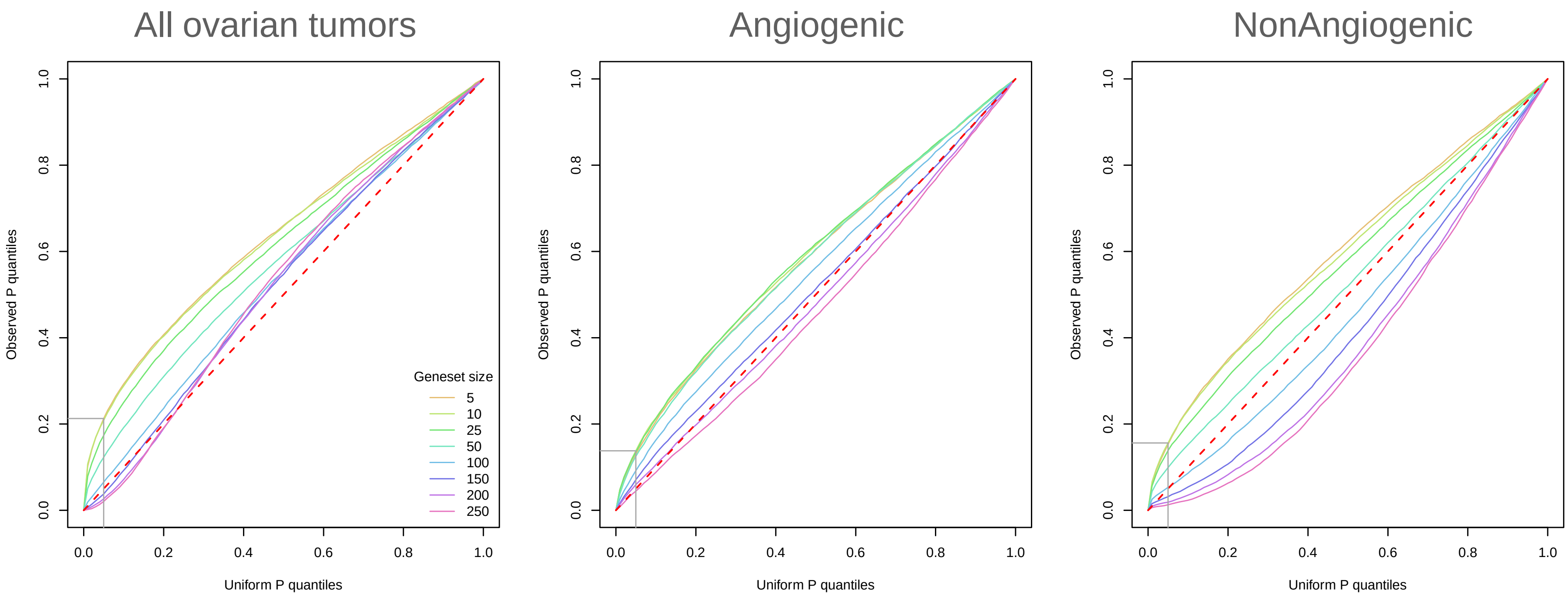
Stefan Bentink[1,6.], Benjamin Haibe-Kains[1,6.], Thomas Risch[1], Jian-Bing Fan[3], Michelle S. Hirsch[4,7], Kristina Holton[1], Renee Rubio[1], Craig April[3], Jing Chen[3], Eliza Wickham-Garcia[3], Joyce Liu[2,7], Aedin Culhane[1,6], Ronny Drapkin[4,5,7], John Quackenbush[1,2,6*"], Ursula A. Matulonis[5,7"]

Citation: Bentink S, Haibe-Kains B, Risch T, Fan J-B, Hirsch MS, et al. (2012) Angiogenic mRNA and microRNA Gene Expression Signature Predicts a Novel Subtype of Serous Ovarian Cancer. PLoS ONE 7(2): e30269. doi:10.1371/journal.pone.0030269

- Two main subtypes:
  - Angiogenic
  - NonAngiogenic

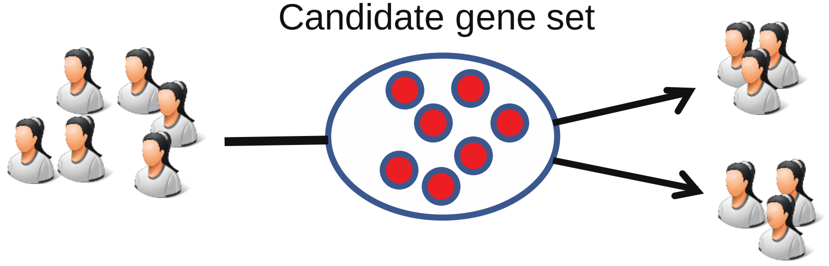
Prognostic value depends on the disease

Global population of ovarian cancer patients and molecular subtypes



Significance analysis of prognostic signatures (**SAPS**)

Cluster patients on
Candidate gene set

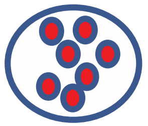


Survival in Cluster 1

Vs.

Survival in Cluster 2

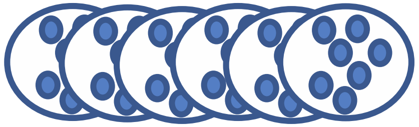
$\longrightarrow$ P_{pure}



Candidate
genesets

P_{pure}

Vs.

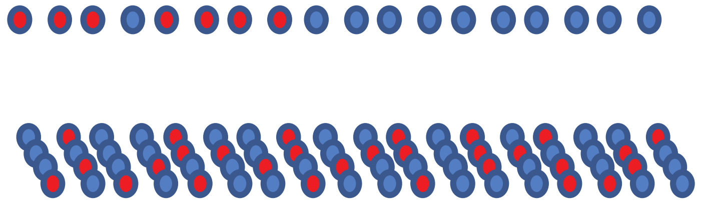


Collection of
random genesets

P_{pure}

$\longrightarrow$ P_{random}

Genes ranked based on prognostic association



Candidate geneset
Enrichment Score

Vs.

Collection of
random geneset
Enrichment Scores

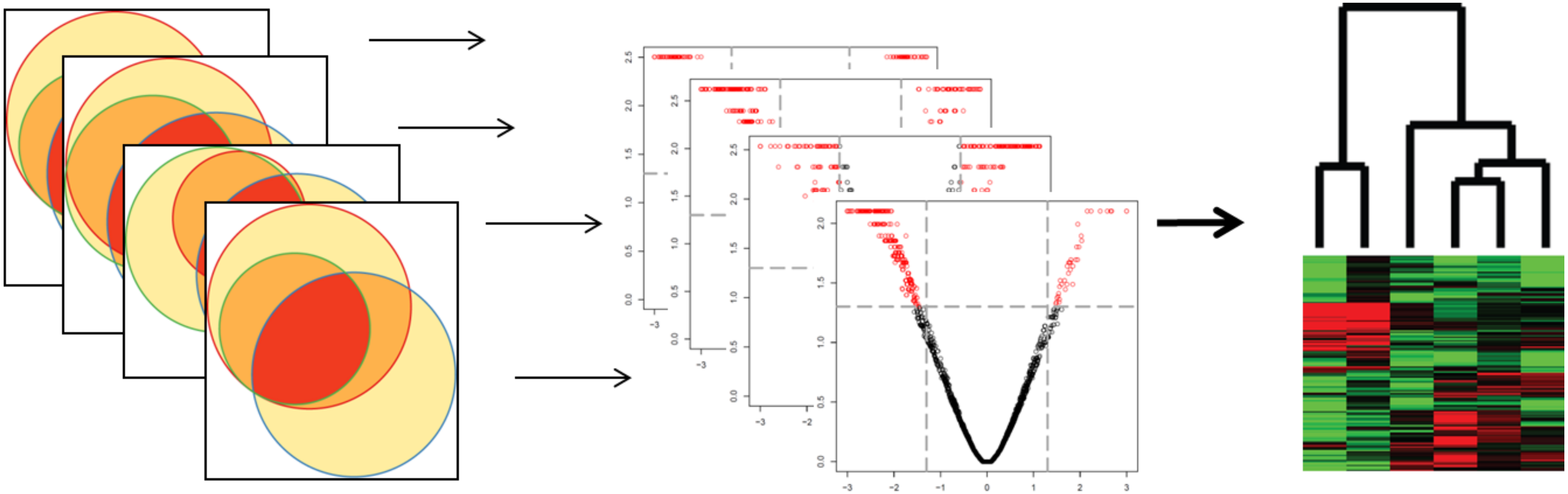
$\longrightarrow$ $P_{enrichment}$

$$\Rightarrow \quad SAPS\ score = -\log_{10}\max(P_{pure}, P_{random}, P_{enrichment}) \times Direction$$

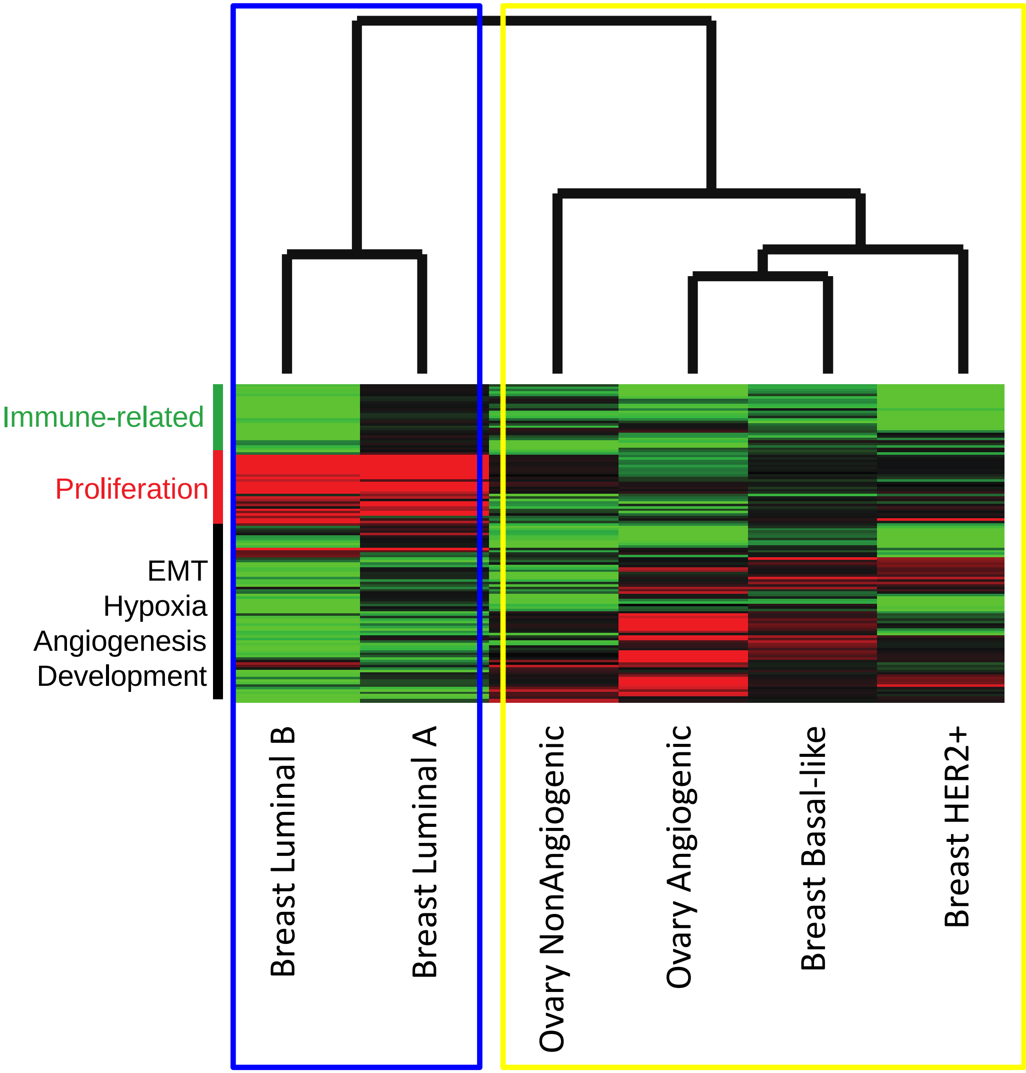
Genesets identified by **SAPS**

- We identified 1300 genesets (out of 5320, MSigDB) which yielded significant SAPS scores in at least one cancer subtype

- We clustered genesets and disease subtypes using hierarchical clustering



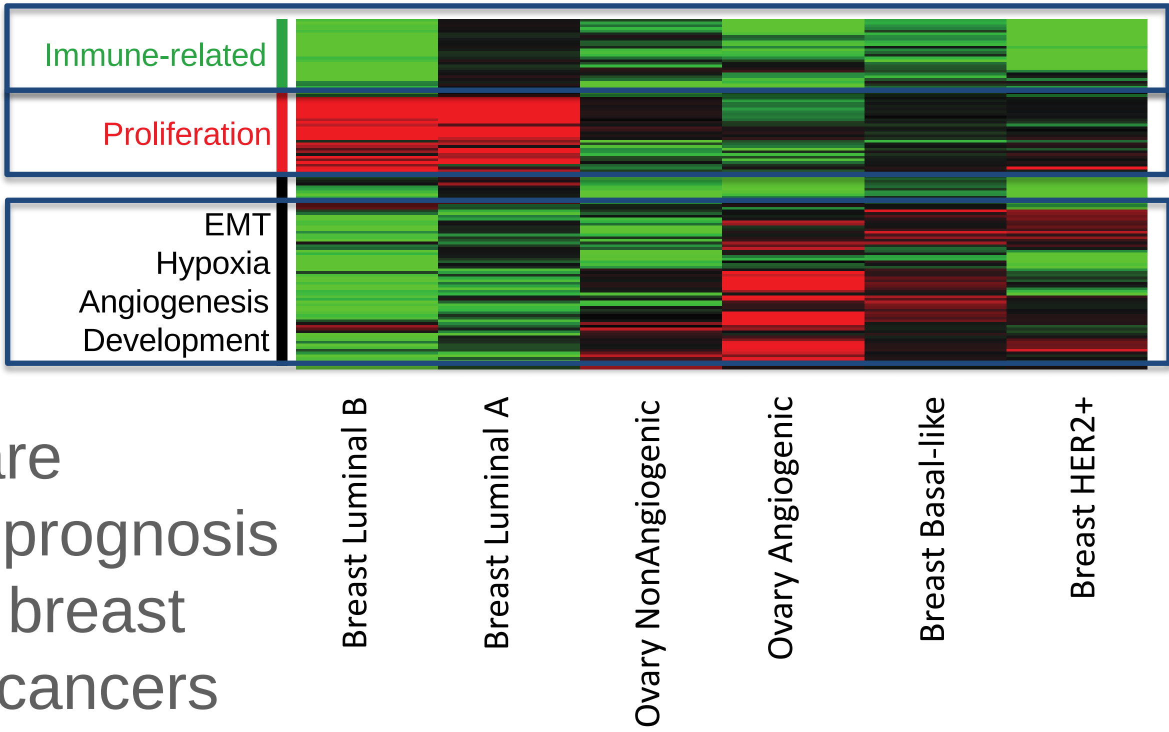
Prognostic genesets in ovarian and breast cancers



- 2 main clusters of subtypes, **not based on cancer type**

Prognostic genesets in ovarian and breast cancers

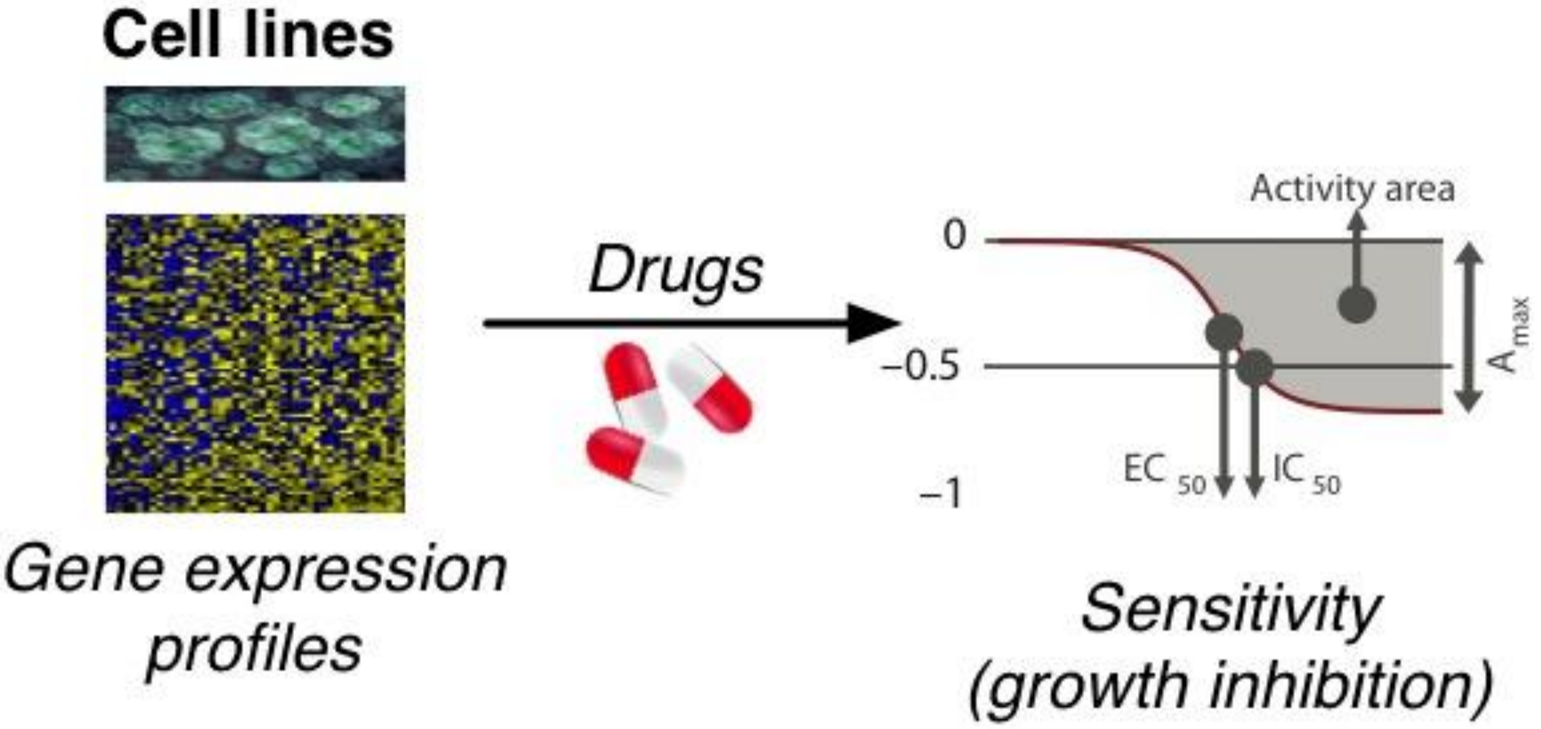
- Proliferation-related genesets are highly prognostic in luminal breast cancers
- Immune-related genesets are associated with good prognosis in all subtypes but Lumina A breast cancers



- The other genesets are associated with poor prognosis in Basal-like, HER2+ breast cancers and ovarian cancers

(Random) Predictive Biomarkers

Pharmacogenomic data



Resistant vs. sensitive cell lines

Large pharmacogenomic datasets

- Large-scale studies have been published recently in ***Nature***

  - The Cancer Genome Project (**CGP**) initiated by the Sanger Institute
    - **131** drugs (IC_{50})
    - **727** cancer cell lines

  - The Cancer Cell Line Encyclopedia (**CCLE**) initiated by Novartis/Broad Institute
    - **24** drugs (IC_{50})
    - **1036** cancer cell lines

CGP ∩ CCLE

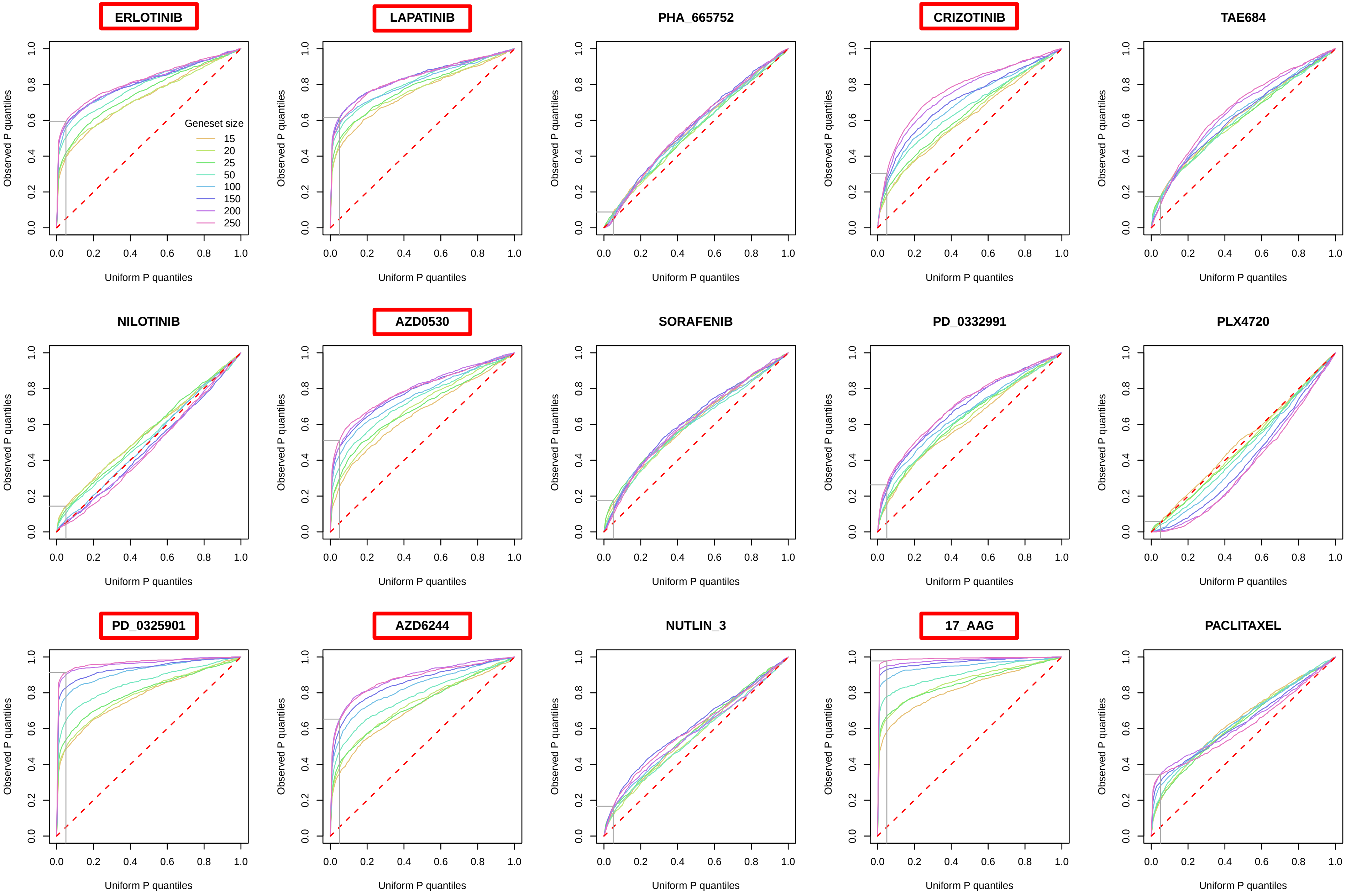
- Drugs: **15** drugs have been investigated both in CGP and CCLE

Paclitaxel	Microtubules depolymerization inhibitor
PD-0325901, AZD6244	Mitogen-activated protein kinase kinase (MEK) inhibitor
AZD0530 (Saracatinib)	Proto-oncogene tyrosine-protein Src inhibitor
Nutlin-3	Ubiquitin-protein ligase MDM2 inhibitor
Nilotinib	BCR-ABL fusion protein inhibitor
17-AAG (Tanespamycin)	Heat shock protein (Hsp90) inhibitor
PD-0332991	CDK4/6-Cyclin D inhibitor
PLX4720, Sorafenib	RAF kinase inhibitors
Crizotinib, TAE684	ALK kinase inhibitors
Erlotinib, Lapatinib	EGFR/HER2 kinase inhibitors
PHA-665752	Proto-oncogene c-MET kinase inhibitor

Significant Analysis of **Predictive** Signatures

- Joint analysis of CGP and CCLE in a meta-analysis framework
- Genesets are summarized by their first principal component
- Significance is computed using a linear regression model controlled for tissue type
- We generated 1000 random genesets for each size and tested the significance of their predictive value

Are **random** genesets predictive of drug sensitivity?



Wrap-up

Take home messages

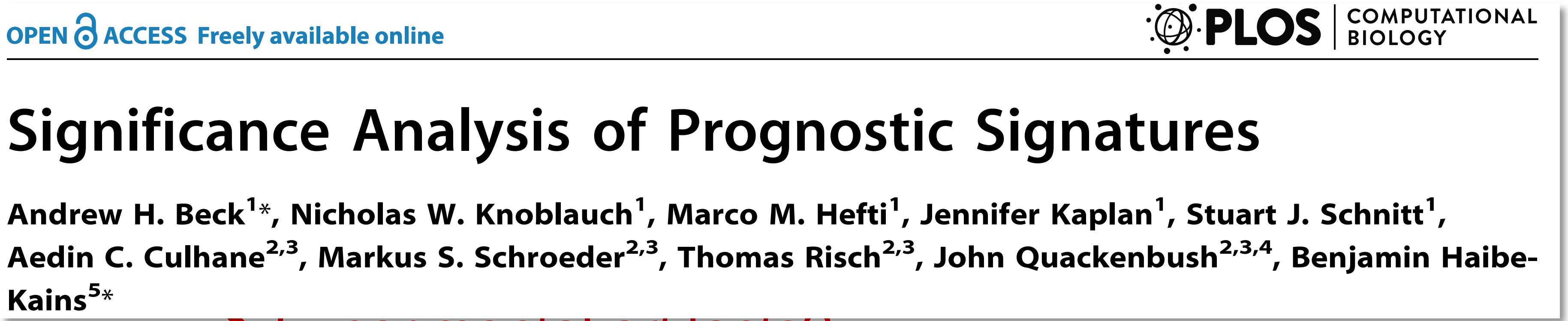
Prognostic biomarkers

- Cancers are molecularly heterogeneous

  ➔ subtypes should be taken into account

- Many, many genes might be prognostic

  ➔ Prognostic value of genesets should be tested against random sets of genes



OPEN ACCESS Freely available online

PLOS | COMPUTATIONAL BIOLOGY

Significance Analysis of Prognostic Signatures

Andrew H. Beck[1]*, Nicholas W. Knoblauch[1], Marco M. Hefti[1], Jennifer Kaplan[1], Stuart J. Schnitt[1], Aedin C. Culhane[2,3], Markus S. Schroeder[2,3], Thomas Risch[2,3], John Quackenbush[2,3,4], Benjamin Haibe-Kains[5]*

➔ Experimental artifacts?

Acknowledgements

IRCM

- Nehme Hachem
- Pierre-Olivier Bachant-Winner
- Simon Papillon-Cavanagh
- Nicolas De Jay

INSILICO

- Alain Coletta
- David Weiss
- David Steenhoff
- Robin Duque

DANA-FARBER
CANCER INSTITUTE

- Hugo Aerts
- John Quackenbush

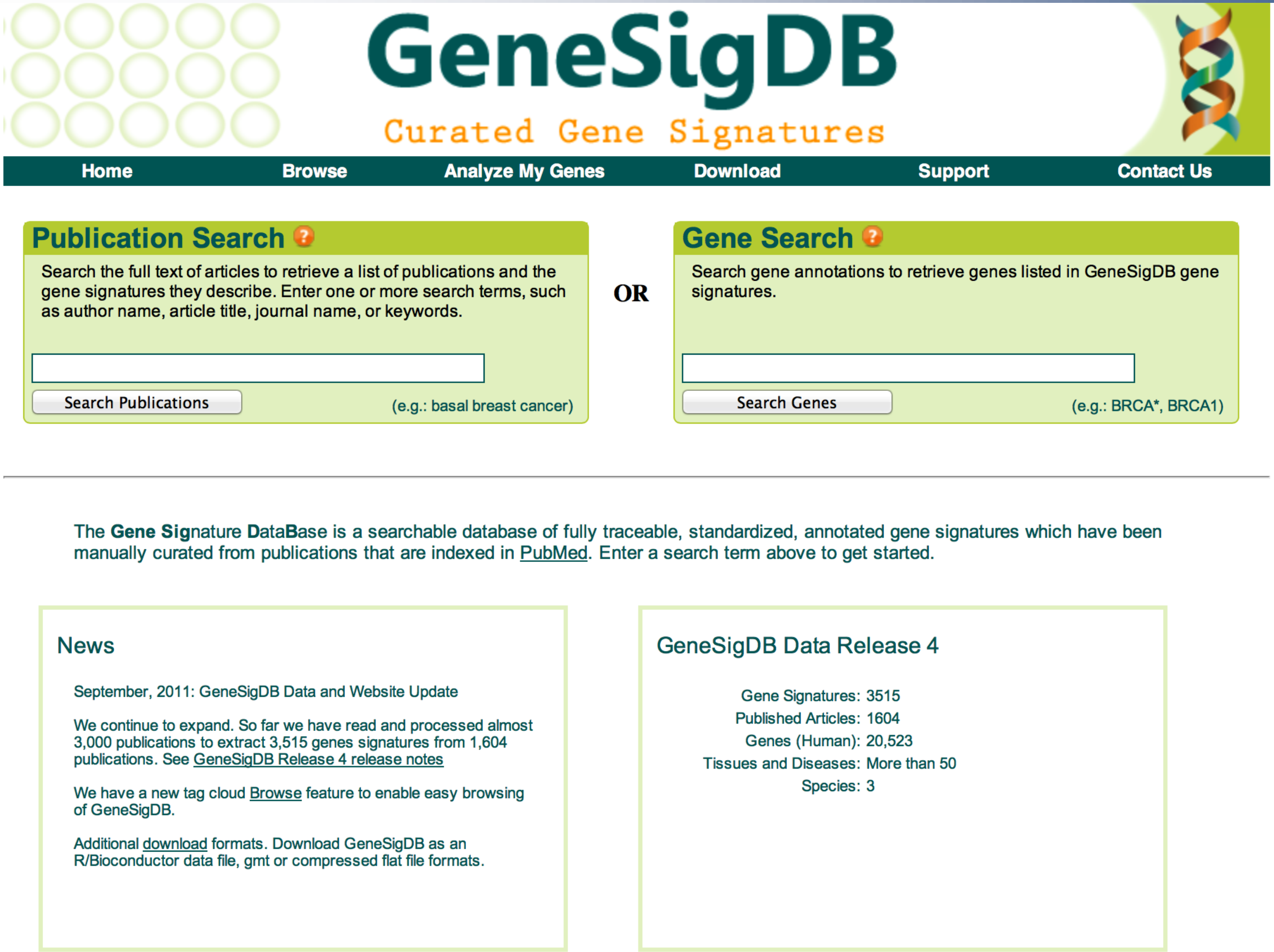
Beth Israel Deaconess
Medical Center
A TEACHING HOSPITAL OF HARVARD MEDICAL SCHOOL

- Andrew Beck
- Pier Paolo Pandolfi
- Nina Seitzer

Thank you for your attention!

Appendix

Published gene signatures



Publication Search

Search the full text of articles to retrieve a list of publications and the gene signatures they describe. Enter one or more search terms, such as author name, article title, journal name, or keywords.

Gene Search

Search gene annotations to retrieve genes listed in GeneSigDB gene signatures.

The **Gene Sig**nature **DataB**ase is a searchable database of fully traceable, standardized, annotated gene signatures which have been manually curated from publications that are indexed in PubMed. Enter a search term above to get started.

News

September, 2011: GeneSigDB Data and Website Update

We continue to expand. So far we have read and processed almost 3,000 publications to extract 3,515 genes signatures from 1,604 publications. See GeneSigDB Release 4 release notes

We have a new tag cloud Browse feature to enable easy browsing of GeneSigDB.

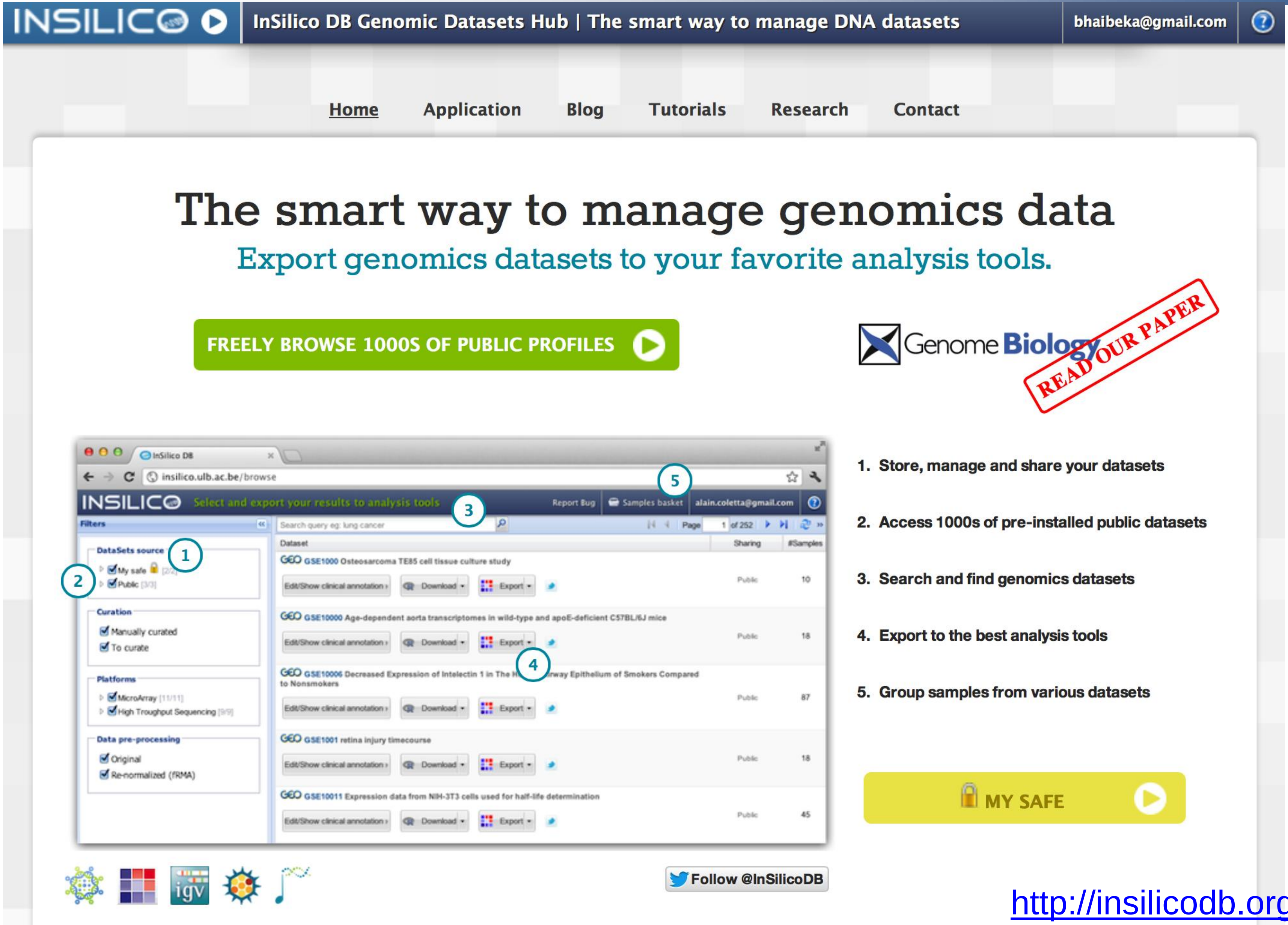
Additional download formats. Download GeneSigDB as an R/Bioconductor data file, gmt or compressed flat file formats.

GeneSigDB Data Release 4

Gene Signatures:	3515
Published Articles:	1604
Genes (Human):	20,523
Tissues and Diseases:	More than 50
Species:	3

http://compbio.dfci.harvard.edu/genesigdb/

Compendium of datasets



INSILICO | InSilico DB Genomic Datasets Hub | The smart way to manage DNA datasets | bhaibeka@gmail.com

Home | Application | Blog | Tutorials | Research | Contact

The smart way to manage genomics data

Export genomics datasets to your favorite analysis tools.

FREELY BROWSE 1000S OF PUBLIC PROFILES

Genome Biology — READ OUR PAPER

InSilico DB

insilico.ulb.ac.be/browse

INSILICO Select and export your results to analysis tools | Report Bug | Samples basket | alain.coletta@gmail.com

Filters

Search query eg: lung cancer | Page 1 of 252

DataSets source
- ☑ My safe [2/2]
- ☑ Public [3/3]

Curation
- ☑ Manually curated
- ☑ To curate

Platforms
- ☑ MicroArray [11/11]
- ☑ High Troughput Sequencing [9/9]

Data pre-processing
- ☑ Original
- ☑ Re-normalized (fRMA)

Dataset	Sharing	#Samples
GEO GSE1000 Osteosarcoma TE85 cell tissue culture study	Public	10
GEO GSE10000 Age-dependent aorta transcriptomes in wild-type and apoE-deficient C57BL/6J mice	Public	18
GEO GSE10006 Decreased Expression of Intelectin 1 in The H[illegible]rway Epithelium of Smokers Compared to Nonsmokers	Public	87
GEO GSE1001 retina injury timecourse	Public	18
GEO GSE10011 Expression data from NIH-3T3 cells used for half-life determination	Public	45

1. Store, manage and share your datasets
2. Access 1000s of pre-installed public datasets
3. Search and find genomics datasets
4. Export to the best analysis tools
5. Group samples from various datasets

MY SAFE

Follow @InSilicoDB

http://insilicodb.org

Advantages of InSilicoDB

- One of the main issues in meta-analysis is data curation
- InSilicoDB allows you to store and access your **own** curation

- You can download R workspaces directly from the web interface

- Even better, you can programmatically download and access the curated genomic and clinical data

Advantages of InSilicoDB

Example of code:

```
> library(inSilicoDb2)
> InSilicoLogin(login="bhaibeka@gmail.com",
password="747779bec8a754b91076d6cc1f700831")
> platf <- inSilicoDb2::getPlatforms(dataset="GSE2034")
> esets <- inSilicoDb2::getDatasets(dataset="GSE2034",
norm="FRMA", curation="22068", features="PROBE")
> InSilicoLogout()
```

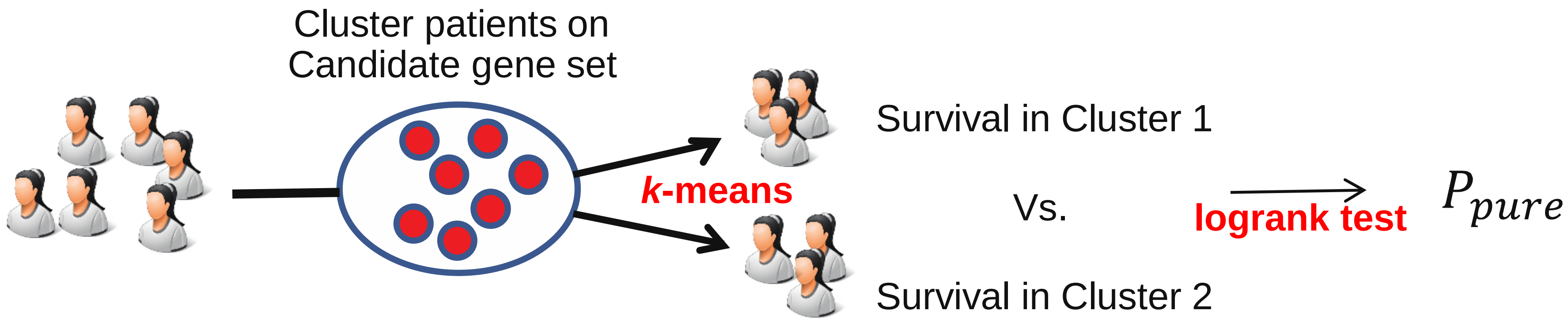
Advantages of InSilicoDB

Output:

```
> print(esets)
ExpressionSet (storageMode: lockedEnvironment)
assayData: 22283 features, 286 samples
  element names: exprs
protocolData: none
phenoData
  Measurements: GSM36777 GSM36778 ... GSM37062 (286 total)
  varLabels: tissue age ... e.dmfs (19 total)
featureNames: 1007_s_at 1053_at ... AFFX-r2-P1-cre-5_at (22283 total)
  fvarLabels: ENTREZID SYMBOL GENENAME
Annotation: hgu133a
```

Generalization of Venet’s results (cont’d)

- We scaled all the datasets to make them “comparable”
  - Z score (mu=0 and sd=1) for each gene
- We used ***k*-means** (*k*=2; unsupervised learning) to classify patients into low- and high-risk group
  - Significance computed using **logrank test**



Cluster patients on Candidate gene set

***k*-means**

Survival in Cluster 1

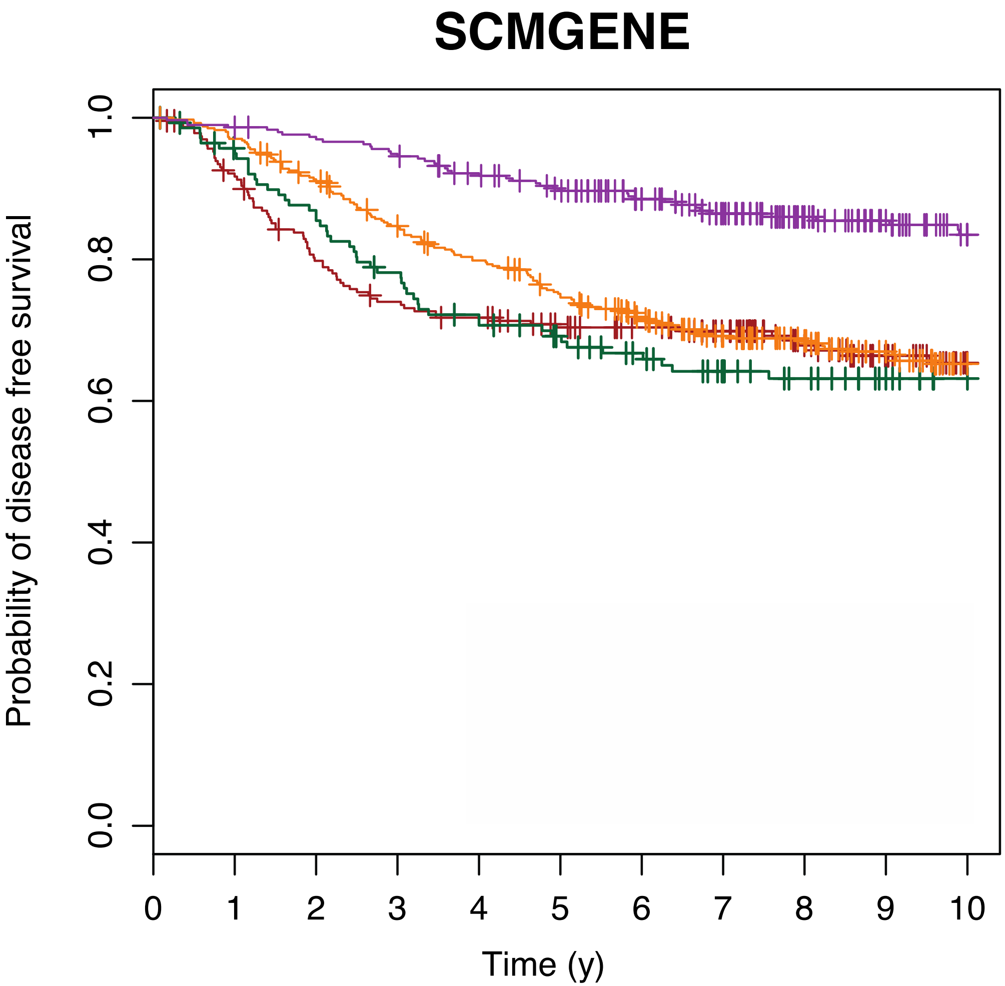
Vs.

Survival in Cluster 2

logrank test

P_{pure}

Subtypes exhibit different clinical outcome



No. at risk											
Basal-like	231	210	181	166	160	149	141	130	98	78	56
HER2-enriched	141	131	119	106	96	89	80	69	61	52	44
Luminal B	405	393	364	333	311	284	249	218	189	161	135
Luminal A	296	292	286	278	264	251	225	202	172	143	117

Predictive biomarkers

- Numerous drug compounds have been designed and many others are under development

- Cancer cell lines can be used as preclinical models to screen thousands of drugs

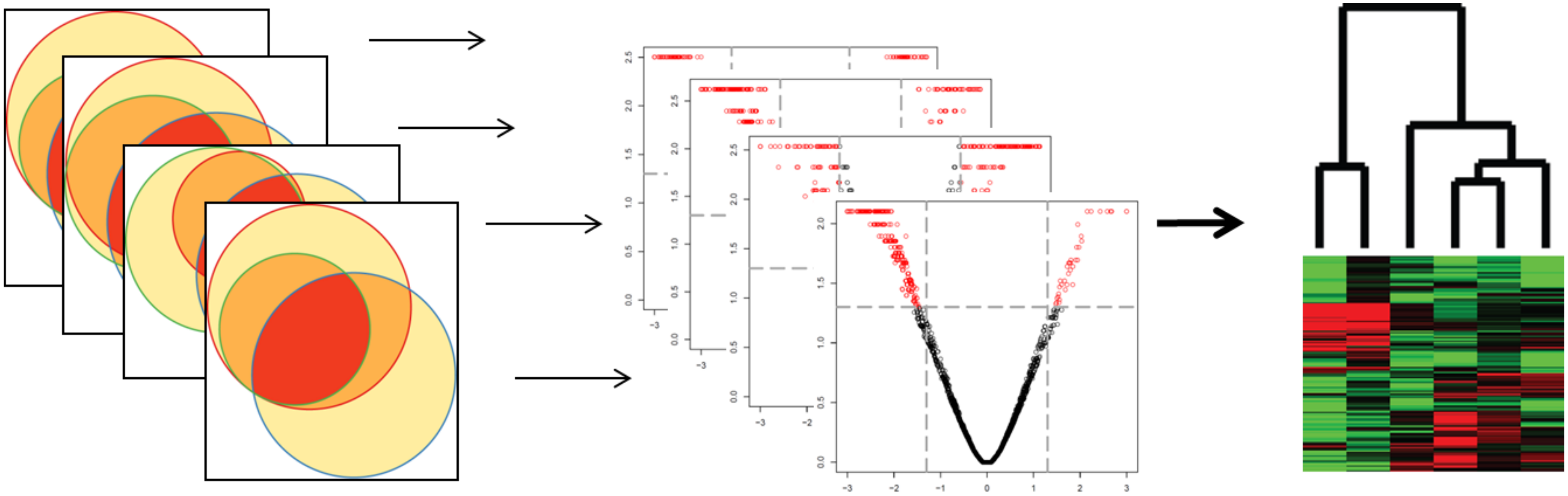
- **Pros**:
  - Cheap and high-throughput
  - Simple models to investigate drugs’ mechanisms of action

- **Cons**:
  - No cell lines are like tumors but they represent well the molecular diversity of cancer

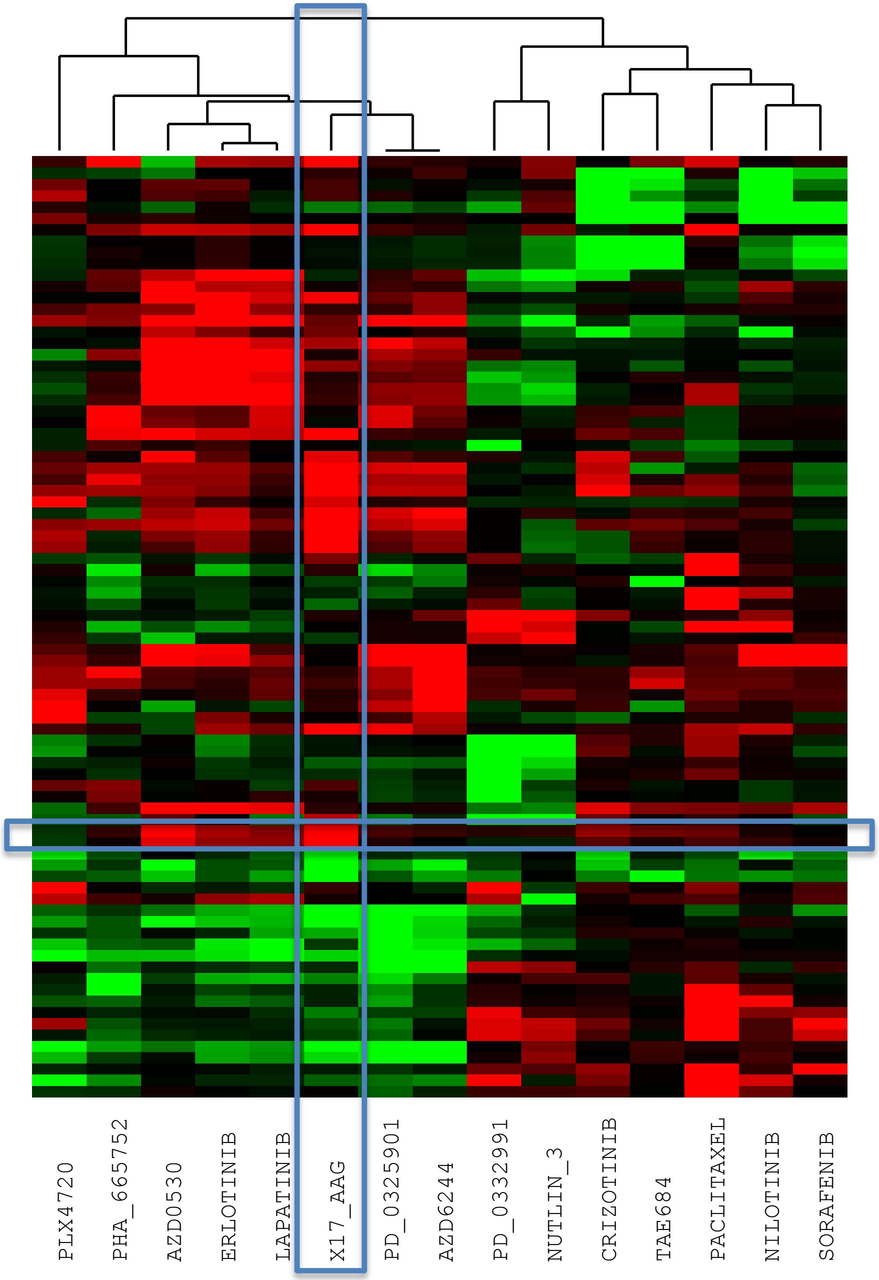
Genesets identified by **SAPS**

- We identified 83 genesets (out of 518 GO biological processes) which yielded significant SAPS scores for at least one drug

- We clustered genesets and drugs using hierarchical clustering



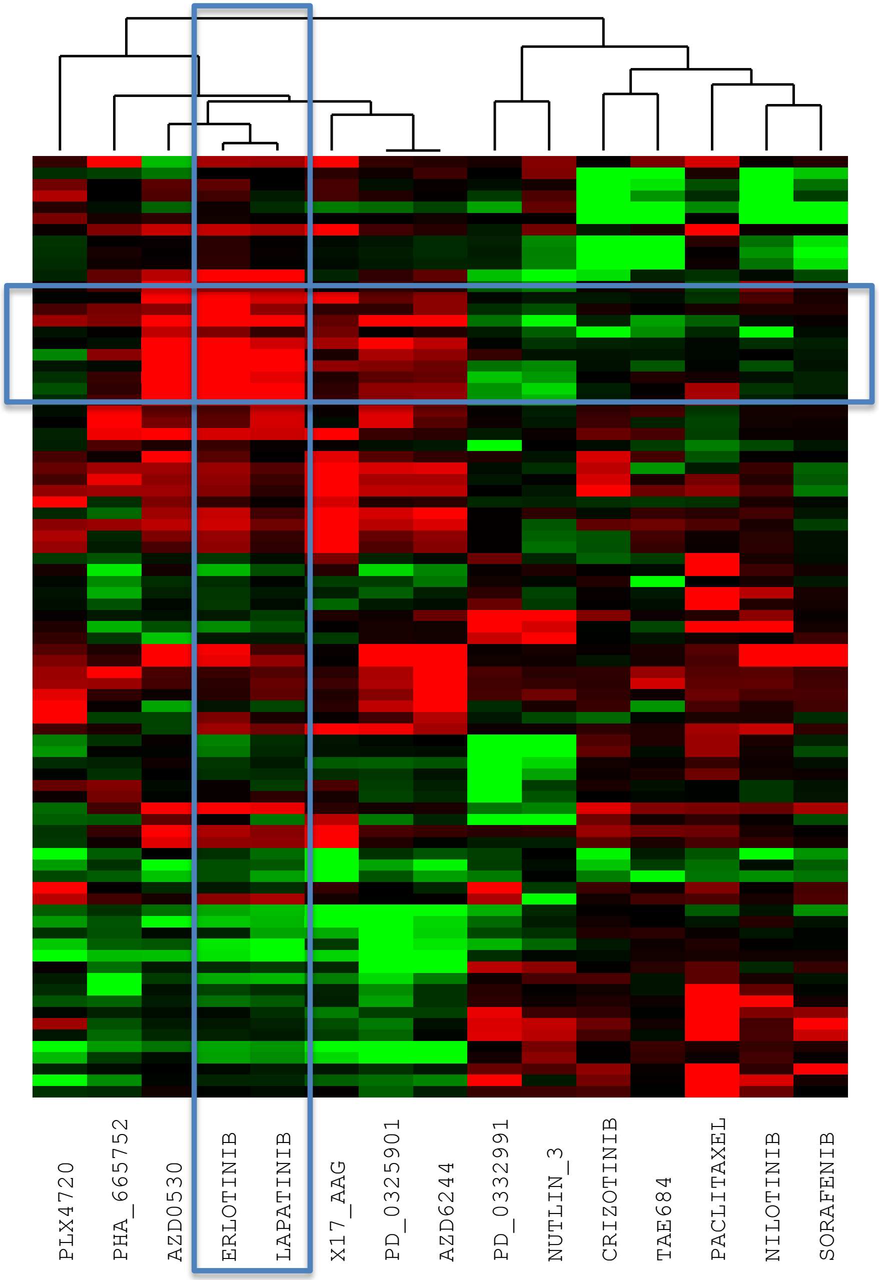
Predictive genesets



NQO1 is associated with sensitivity [FDR < 10^{-54}] as it metabolizes the drug to its active hydroquinone form

HSP90 inhibitor

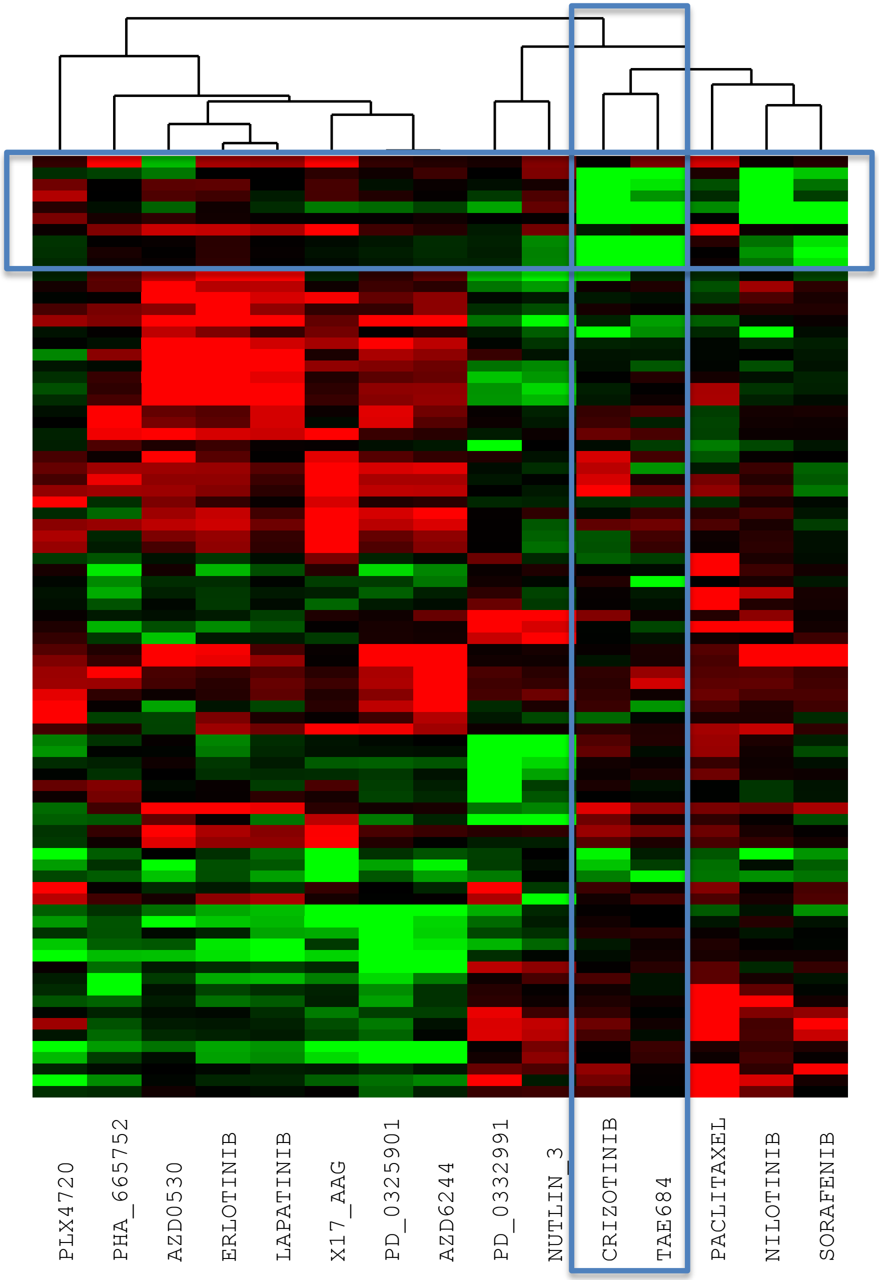
Predictive genesets



EGFR dependent pathways and downstream regulators are associated to sensitivity

EGFR/HER2 kinase inhibitors

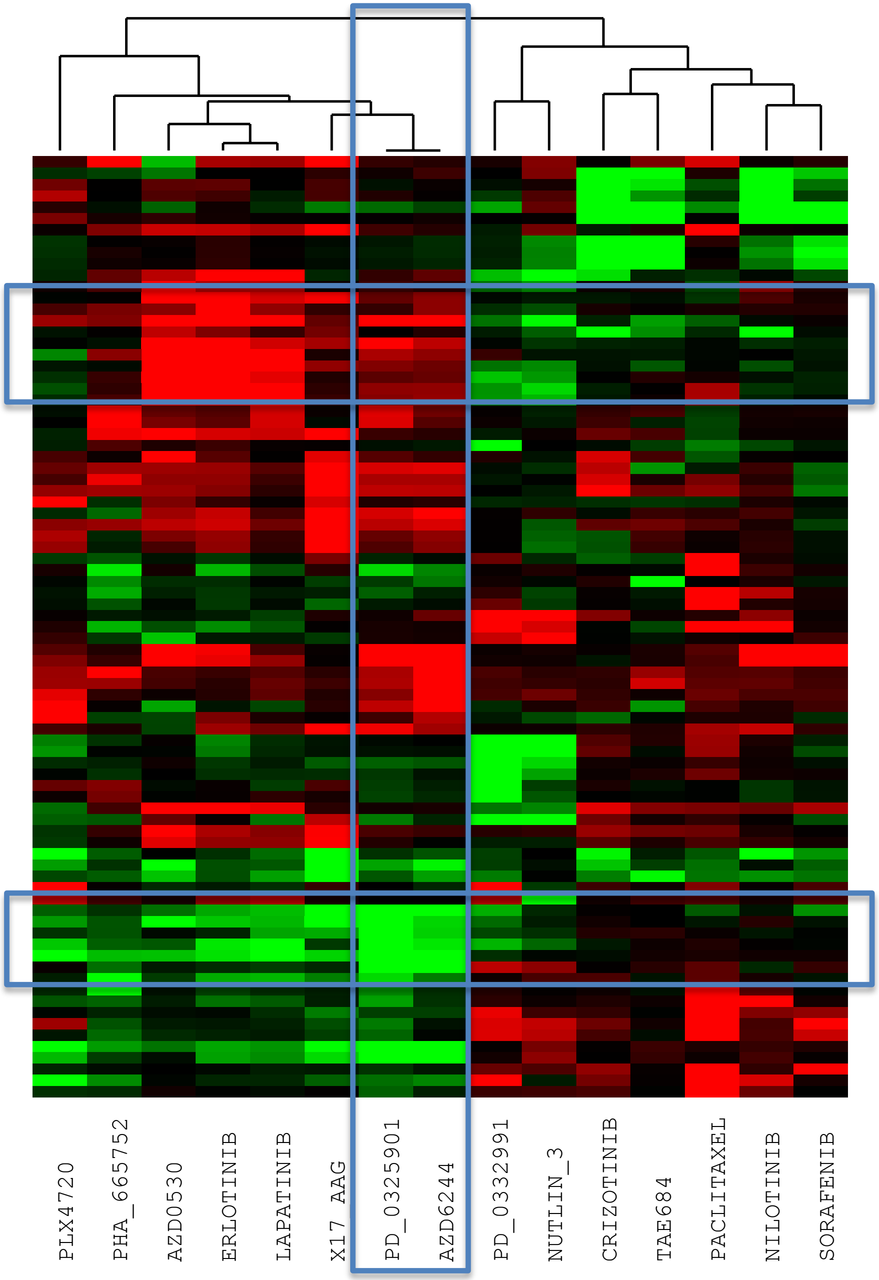
Predictive genesets



Genes involved in immune response are associated to resistance

ALK inhibitors

Predictive genesets



EGFR dependent pathways and downstream regulators are associated to sensitivity

Genes involved in cytoskeleton organization and microtubules are associated to resistance

MEK inhibitors