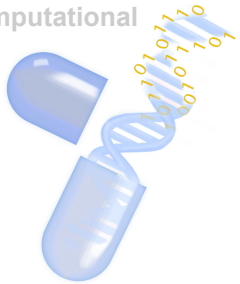
Bioinformatics and Computational
Genomics Laboratory



PharmacoGx: Data Sharing and Research Reproducibility in Pharmacogenomics

Benjamin Haibe-Kains

Princess Margaret Cancer Centre
University Health Network
University of Toronto
Ontario Institute of Cancer Research

2 open postdoc positions:
Re radiomics and single-cell RNA-seq

June 25, 2016

Reproducibility crisis

- Reproducibility in biomedical sciences has attracted a lot of attention in the last 10 years

Signal in Noise:
Reported Repro
Serum Proteomi
Ovarian Cancer

Keith A. Baggerly,
S. Morris, Sarah R. Edmonson,

Believe it or not: how much can we rely on published data on potential drug targets?

Florian Prinz, Thomas Schlange and Khusru Asadullah

Research Findings

Raise standards for preclinical cancer research

C. Glenn Begley and **Lee M. Ellis** propose how methods, publications and incentives must change if patients are to benefit.

SITIVITY FROM CELL LINES:
IATICS AND REPRODUCIBLE
I-THROUGHPUT BIOLOGY

Y[1] AND KEVIN R. COOMBES[2]

Why data and code sharing?

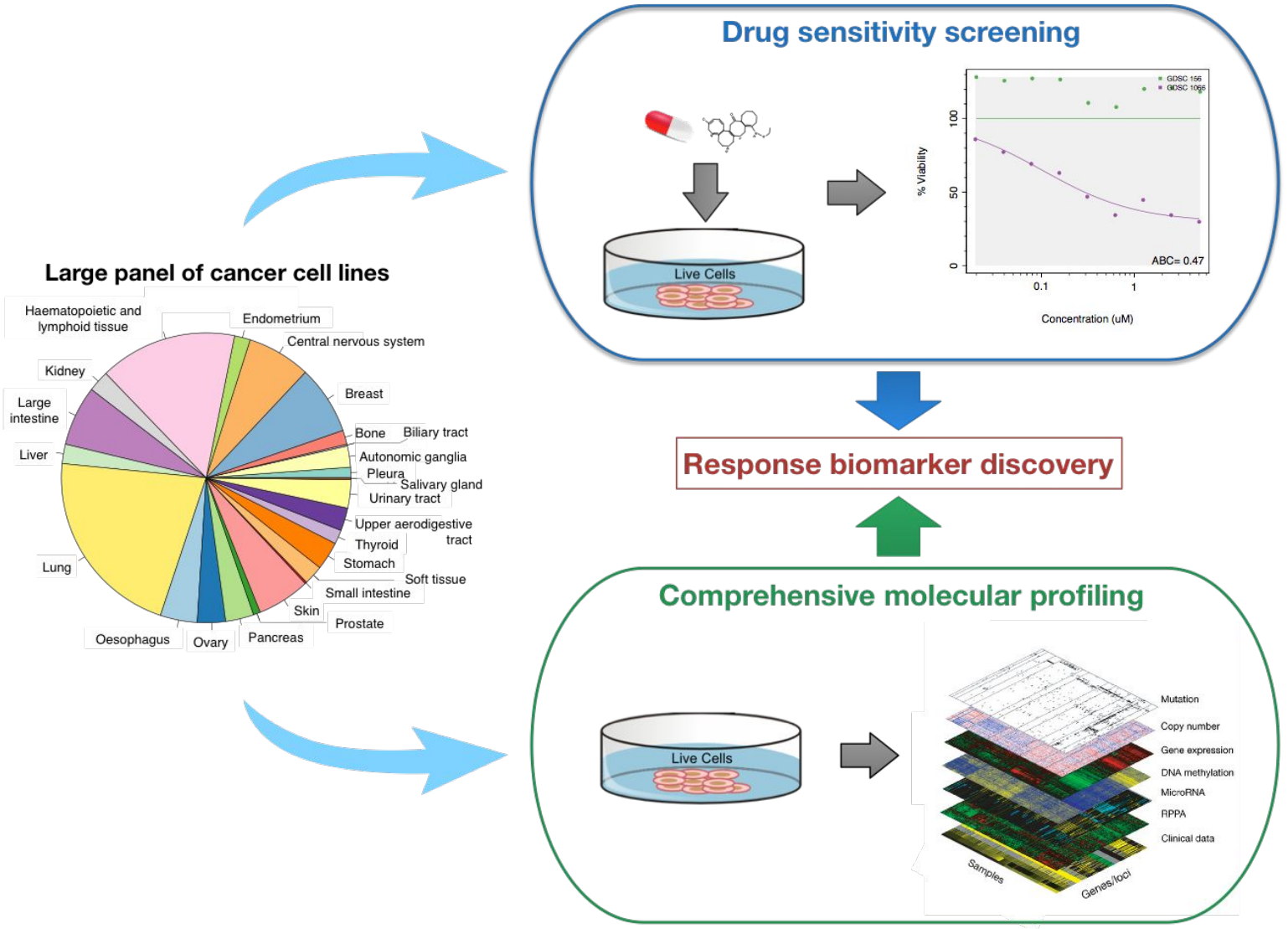
- Data are precious due to limited
  - Amount of samples
  - Resources
  - Budget

"*Anyone who believes in indefinite growth in anything physical, on a physically finite planet, is either mad or an economist.*"

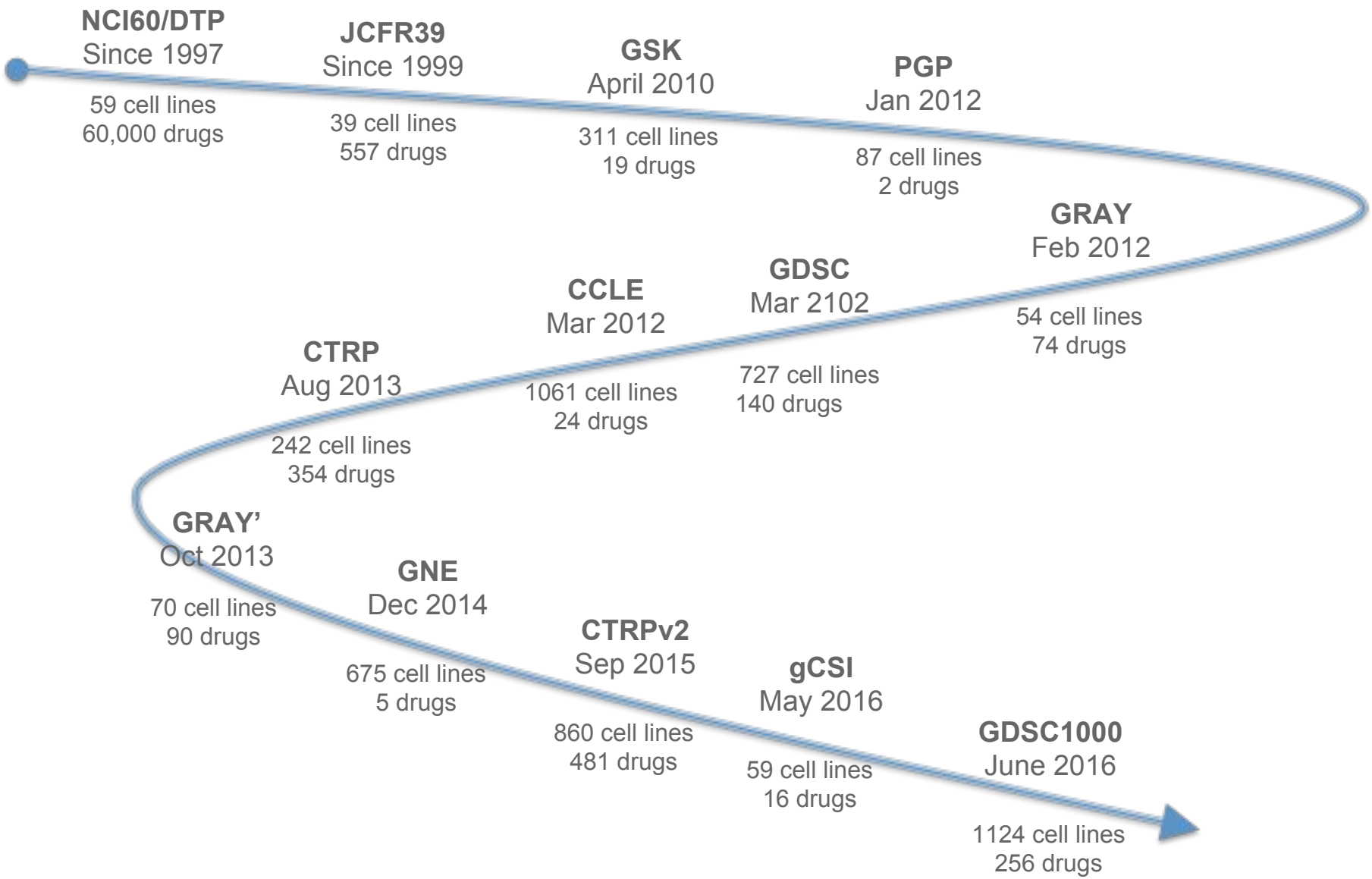
— Kenneth E. Boulding

- Benefits of sharing data and code
  - Replicability
  - Reproducibility
  - Reusability
  - Post-publication peer review

High-throughput *in vitro* drug screening



Long history of data sharing in pharmacogenomics



More to come...

Predictors trained on one dataset hardly validate on an independent set

JAMIA

Comparison and validation of genomic predictors for anticancer drug sensitivity

Simon Papillon-Cavanagh,[1] Nicolas De Jay, Gianluca Bontempi,[2] Hugo J W L Aerts,[3,4] J

Papillon-Cavanagh S, *et al. J A*

Pacific Symposium on Biocomputing 2014

SYSTEMATIC ASSESSMENT OF ANALYTICAL METHODS FOR DRUG SENSITIVITY PREDICTION FROM CANCER CELL LINE DATA*

IN SOCK JANG[1], ELIAS CHAIBUB NETO, JUSTIN GUINNEY, STEPHEN H. FRIEND, ADAM A. MARGOLIN[1]

Dong *et al. BMC Cancer* (2015) 15:489
DOI 10.1186/s12885-015-1492-6

BMC Cancer

RESEARCH ARTICLE **Open Access**

CrossMark

Anticancer drug sensitivity prediction in cell lines from baseline gene expression through recursive feature selection

Zuoli Dong[1†], Naiqian Zhang[1†], Chun Li[2], Haiyun Wang[3], Yun Fang[1], Jun Wang[1*] and Xiaoqi Zheng[1*]

Bioinformatics

OXFORD JOURNALS

Improved large-scale prediction of growth inhibition patterns using the NCI60 cancer cell line panel

Isidro Cortés-Ciriano[1], Gerard J.P. van Westen[2], Guillaume Bouvier[1], Michael Nilges[1], John P. Overington[2], Andreas Bender[3]* and Thérèse E. Malliavin[1]*

Comparative studies

2013

nature

ANALYSIS RESEARCH

Inconsistency in large pharmacogenomic studies

Benjamin Haibe-Kains[1,2], Nehme El-Hachem[1], Nicolai Juul Birkbak[3], Andrew C. Jin[4], Andrew H. Beck[4]*, Hugo J. W. L. Aerts[5,6,7]* & John Quackenbush[5,8]*

2015

nat

Revisiting inconsistency in large pharmacogenomic studies

Pharr

two c

Zhaleh Safikhani, Mark Freeman, Petr Smirnov, Nehme El-Hachem, Adrian She, Rene Quevedo, Anna Goldenberg, Nicolai Juul Birkbak, Christos Hatzis, Leming Shi, Andrew H Beck, Hugo JWL Aerts, John Quackenbush, Benjamin Haibe-Kains

doi: http://dx.doi.org/10.1101/026153

The Cancer Cell L

CSH Cold Spring Harbor Laboratory

bioRχiv beta

THE PREPRINT SERVER FOR BIOLOGY

2016

nature

Reprodu

pro

www.impac

Assessment of pharmacogenomic agreement [version 1; referees: 1 approved]

Zhaleh Safikhani[1,2], Nehme El-Hachem[3], Rene Quevedo[1,2], Petr Smirnov[1], Anna Goldenberg[4,5], Nicolai Juul Birkbak[6], Christopher Mason[7-9], Christos Hatzis[10,11], Leming Shi[12,13], Hugo JWL Aerts[14,15], John Quackenbush[14,16], Benjamin Haibe-Kains[1,2,5]

Peter M.

Jeff Settle

Integrating heterogeneous drug sensitivity data from cancer pharmacogenomic studies

Nikita Pozdeyev[1], Minjae Yoo[1], Ryan Mackie[1], Rebecca E. Schweppe[1], Aik Choon Tan[1,*], Bryan R. Haugen[1,*]

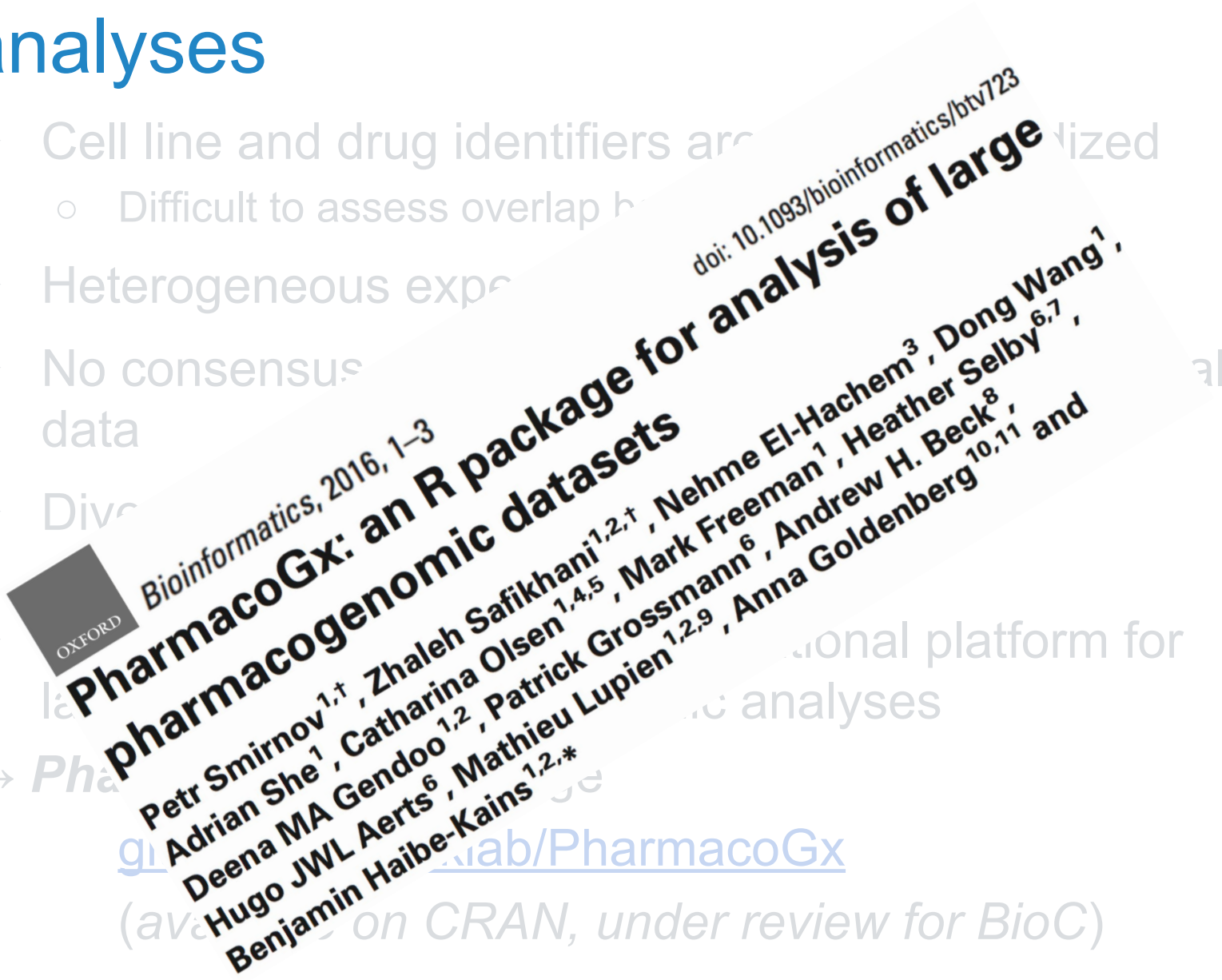
Challenges in pharmacogenomic analyses

- Cell line and drug identifiers ar… …lized
  - Difficult to assess overlap b…
- Heterogeneous expo…
- No consensus … …al data
- Div…
- … …ional platform for la… …c analyses

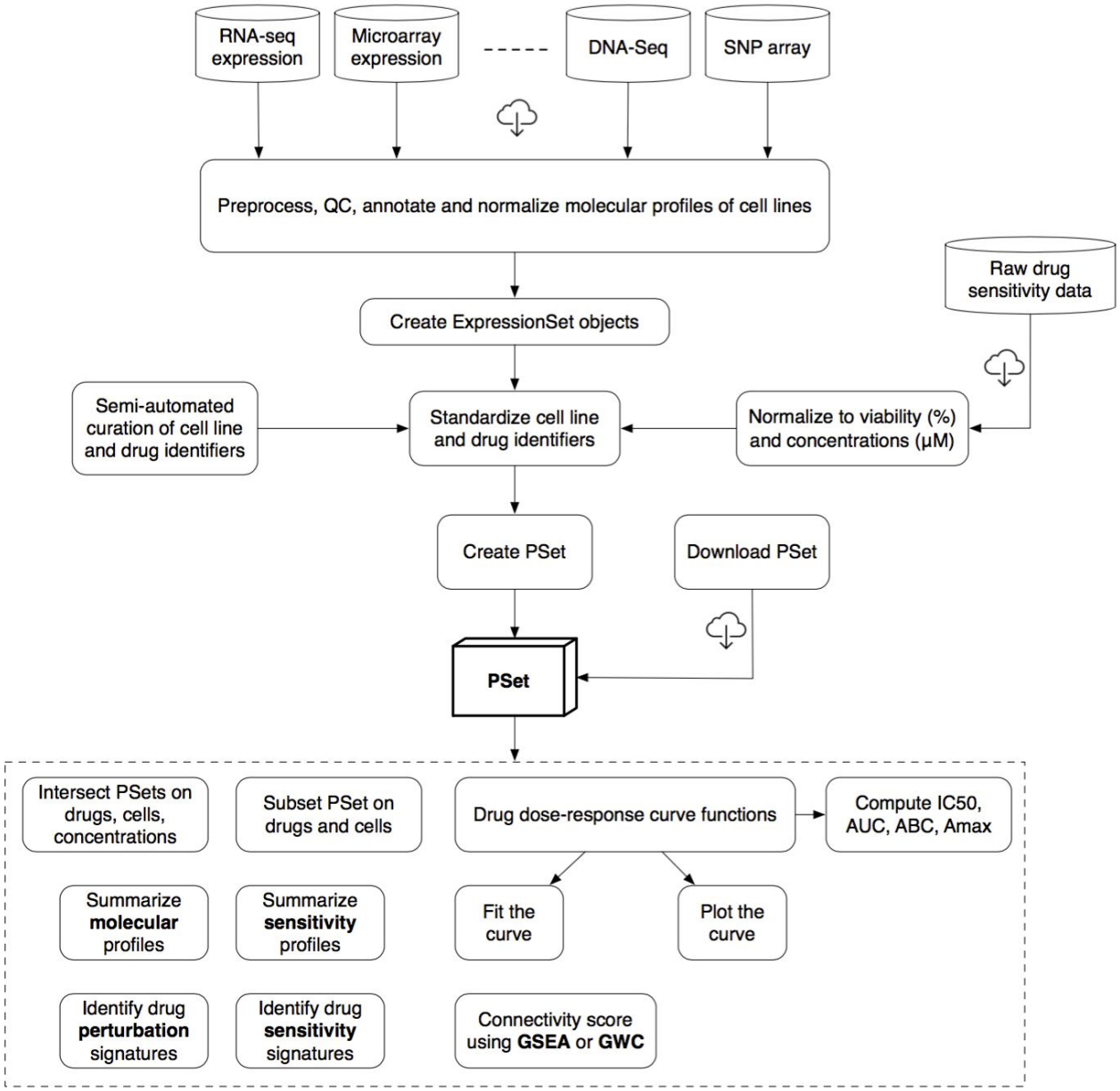
→ ***Pha…***

gi… …ab/PharmacoGx

(*ava… on CRAN, under review for BioC*)



PharmacoGx in a nutshell



PharmacoSet S4 class

@ `annotation`:

- $ `name`: Acronym of the pharmacogenomic dataset.
- $ `dateCreated`: When the object was created.
- $ `sessionInfo`: Software environment used to create the object.
- $ `call`: Set of parameters used to create the object.

@ `datasetType`: Either 'sensitivity', 'perturbation', or 'both'

@ `cell`: data frame annotating all cell lines investigated in the study.

@ `drug`: data frame annotating all the drugs investigated in the study.

@ `sensitivity`:

- $ `n`: Number of experiments for each cell line treated with a given drug
- $ `info`: Metadata for each pharmacological experiment.
- $ `raw`: All cell viability measurements at each drug concentration from the drug dose-response curves.
- $ `phenotype`: Drug sensitivity values summarizing each dose-response curve (IC_{50}, AUC, etc.)

@ `perturbation`:

- $ `n`: Number of experiments for each cell line perturbed by a given drug, for each molecular data type
- $ `info`: 'The metadata for the perturbation experiments is available for each molecular type by calling the appropriate info function'

@ `molecularProfiles`: List of `ExpressionSet` objects containing the molecular profiles of the cell lines, such as mutations, gene expressions, or copy number variations.

→ MultiAssayExperiment

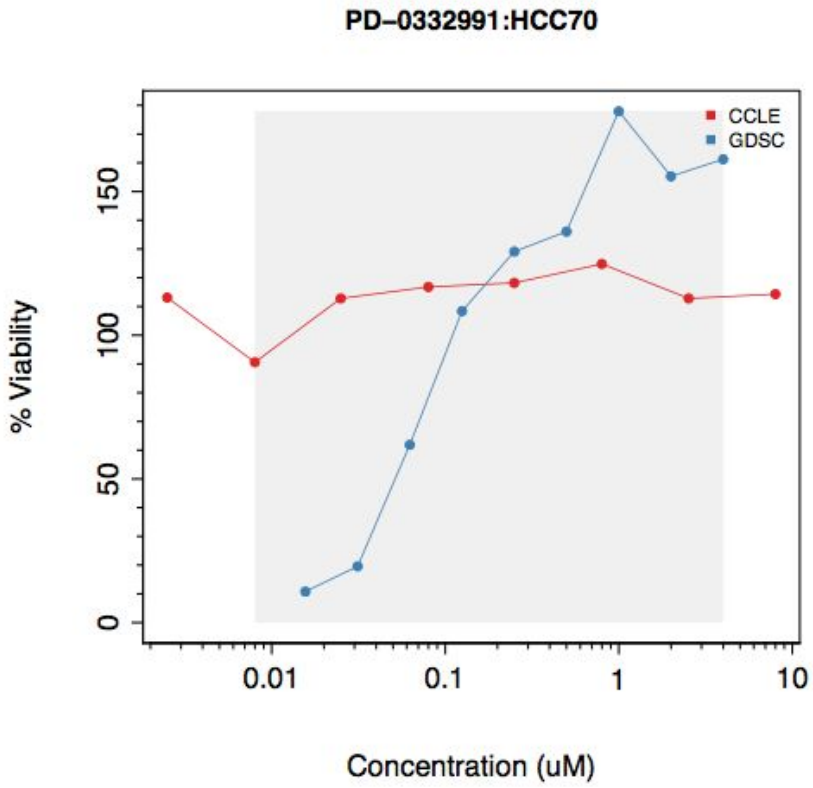
PharmacoGx enables meta-analysis

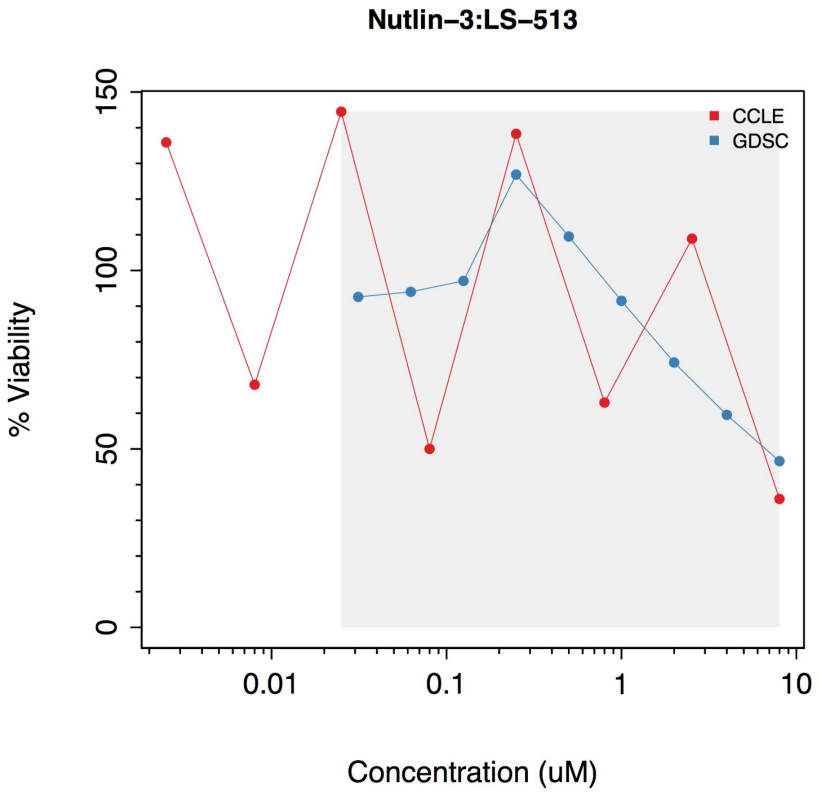
- Cellosaurus to uniquely identify and annotate cell lines and tissues
  - web.expasy.org/cellosaurus/
- Drugs annotated with PubChem ID, InChiKey and SMILES
  - Exact and fuzzy matching based on structure similarity
- Ensembl annotations for omics profiles
- Functions to download, curate, subset, and summarize pharmacogenomic studies
  - *DownloadPSet()*
  - *IntersectPSets()*
  - *SubsetTo()*
  - *summarize*()*

Datasets available today:
CMAP, GDSC, CCLE and gCSI

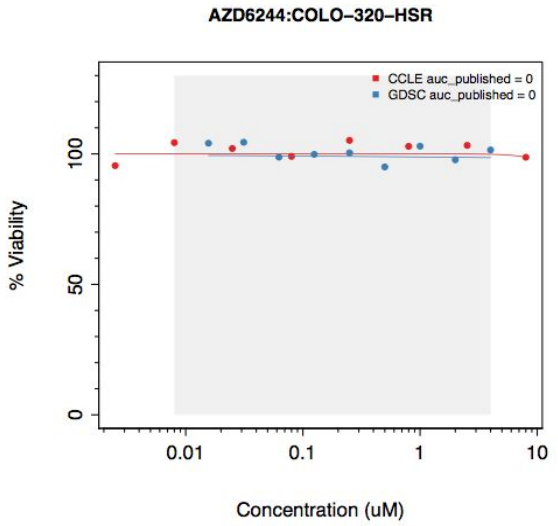
In the oven:
L1000, NCI60, GSK, GNE, CTRPv2, GRAY

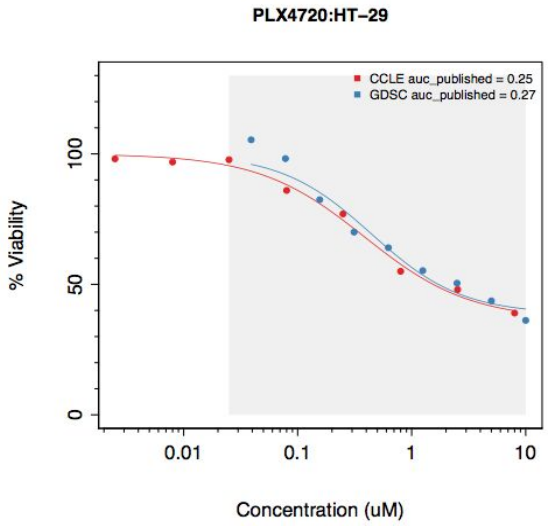
Filtering of noisy dose-response curves





Fitting of drug dose-response curves





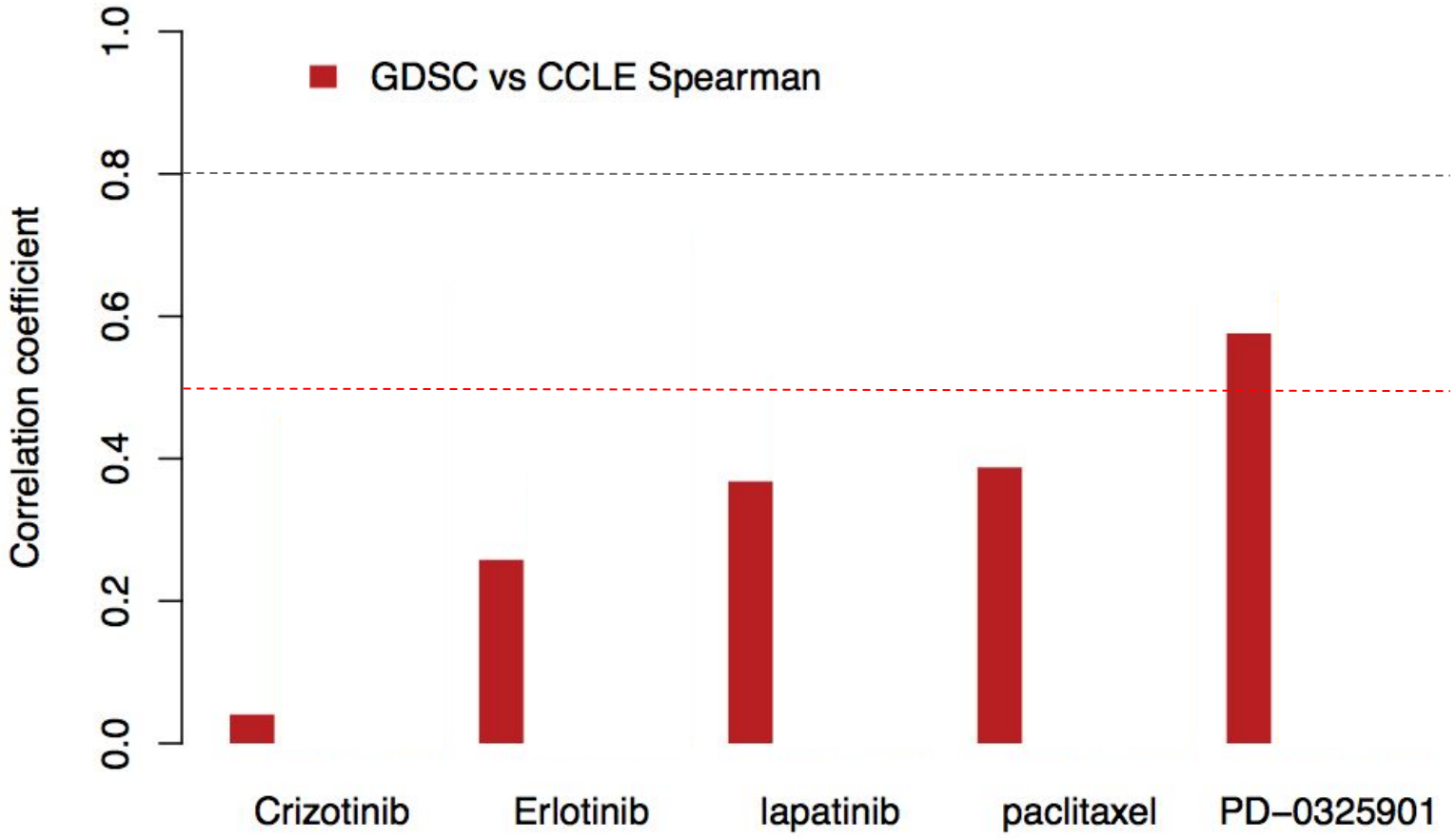
Highly consistent

Correlations of drug sensitivity data

2013 **Inconsistency in large pharmacogenomics studies**

2015 Revisiting inconsistency in large pharmacogenomic studies
Pharmacogenomic agreement between two cancer cell line data sets

2016 Reproducible pharmacogenomic profiling of cancer cell line panels

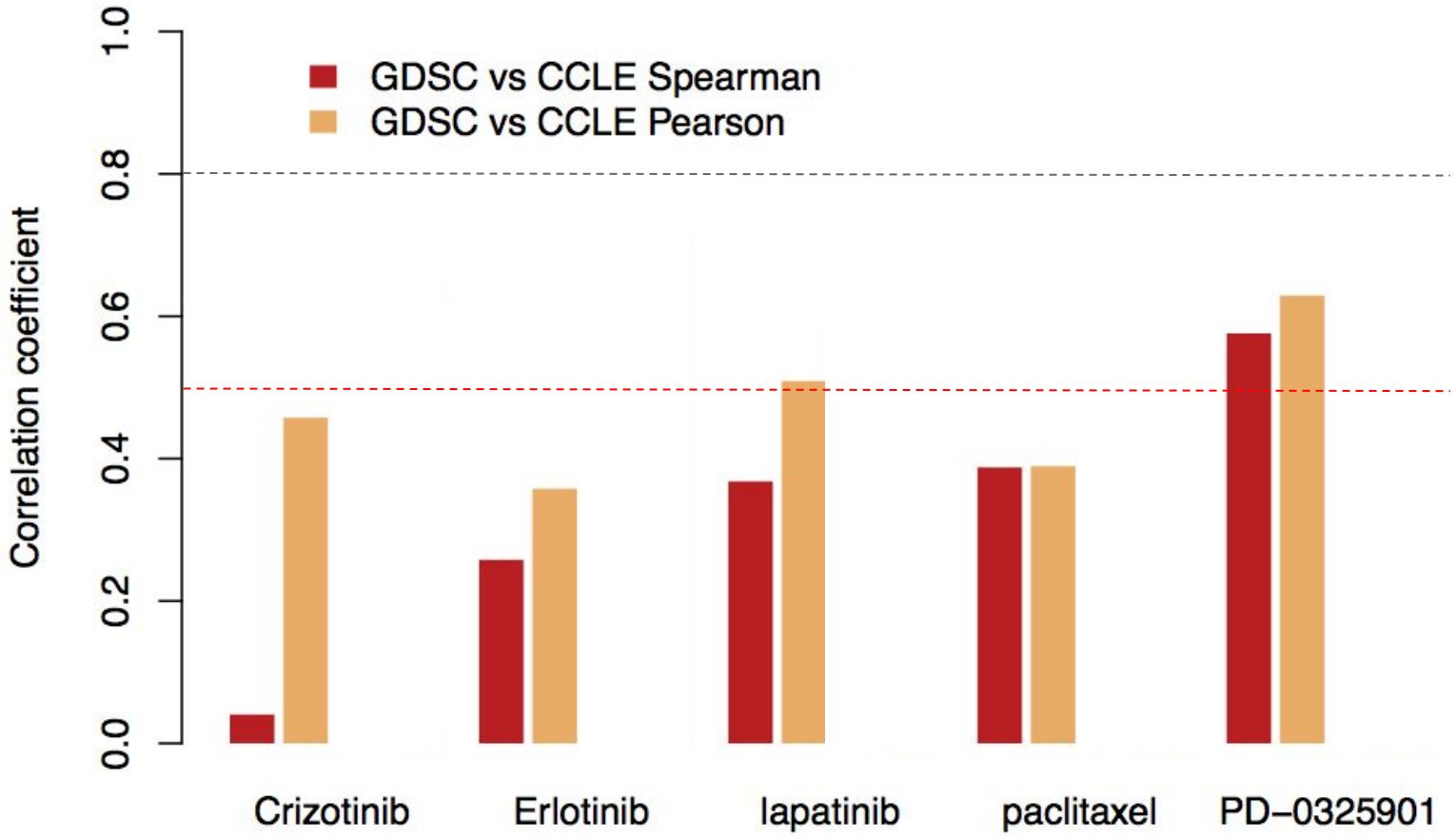


Correlations of drug sensitivity data

2013 Inconsistency in large pharmacogenomics studies

2015 Revisiting inconsistency in large pharmacogenomic studies
Pharmacogenomic agreement between two cancer cell line data sets

2016 Reproducible pharmacogenomic profiling of cancer cell line panels

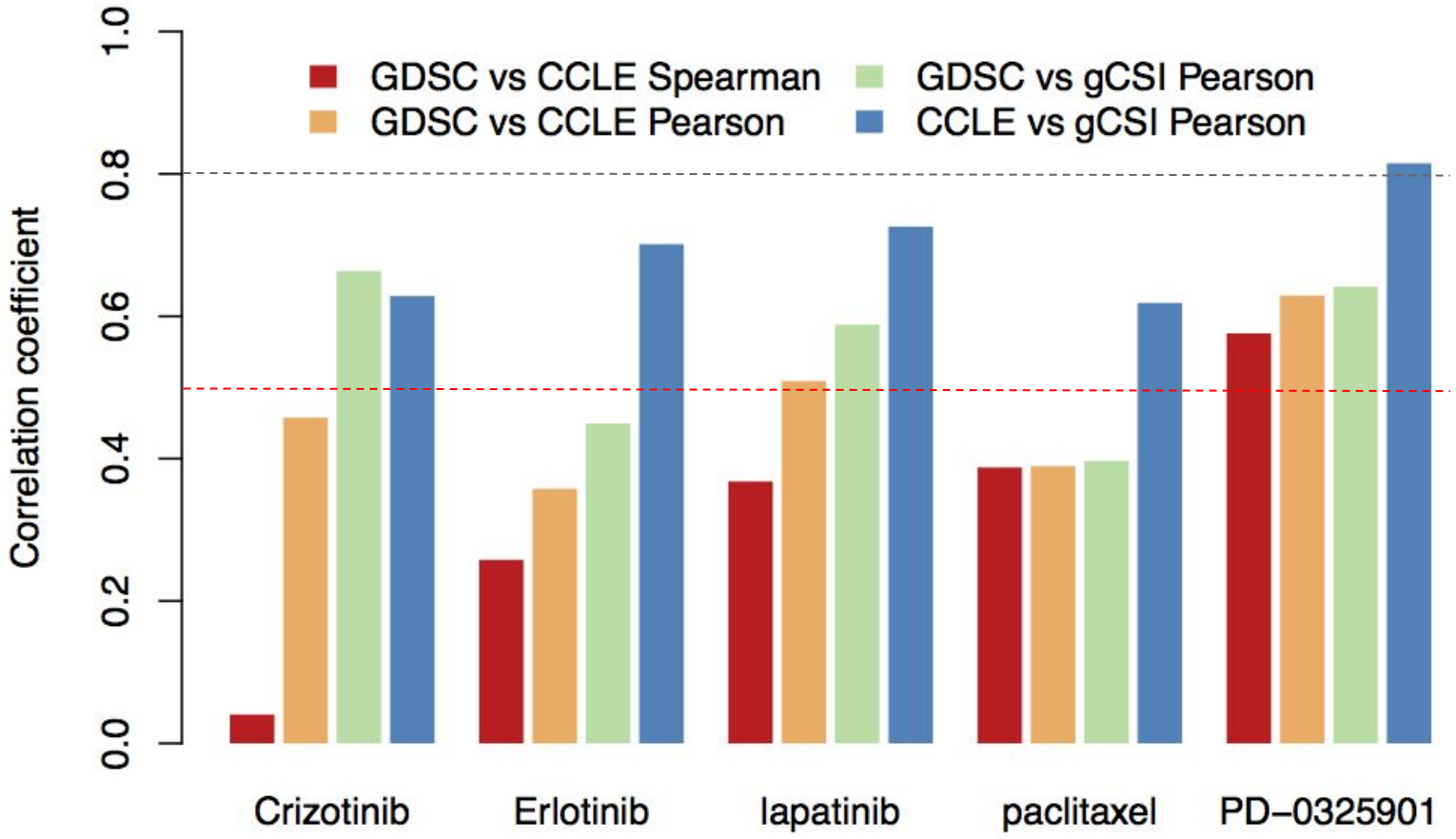


Correlations of drug sensitivity data

2013 Inconsistency in large pharmacogenomics studies

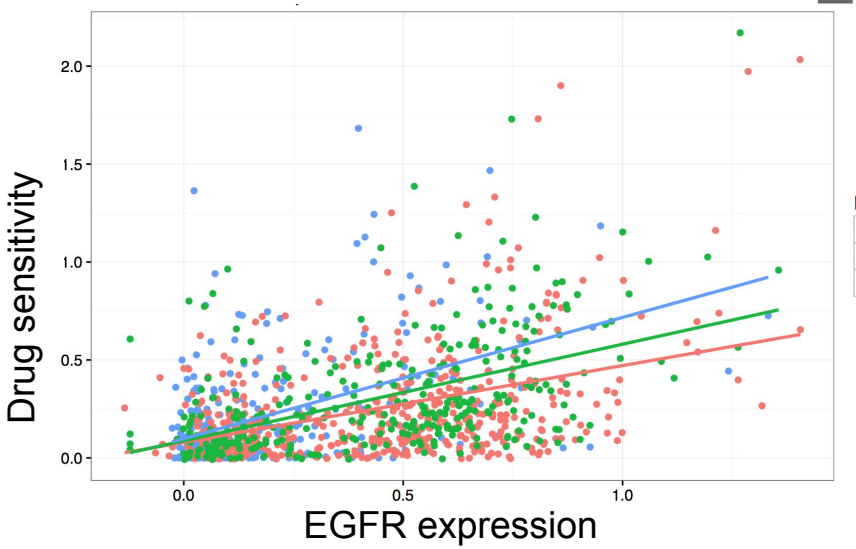
2015 Revisiting inconsistency in large pharmacogenomic studies
Pharmacogenomic agreement between two cancer cell line data sets

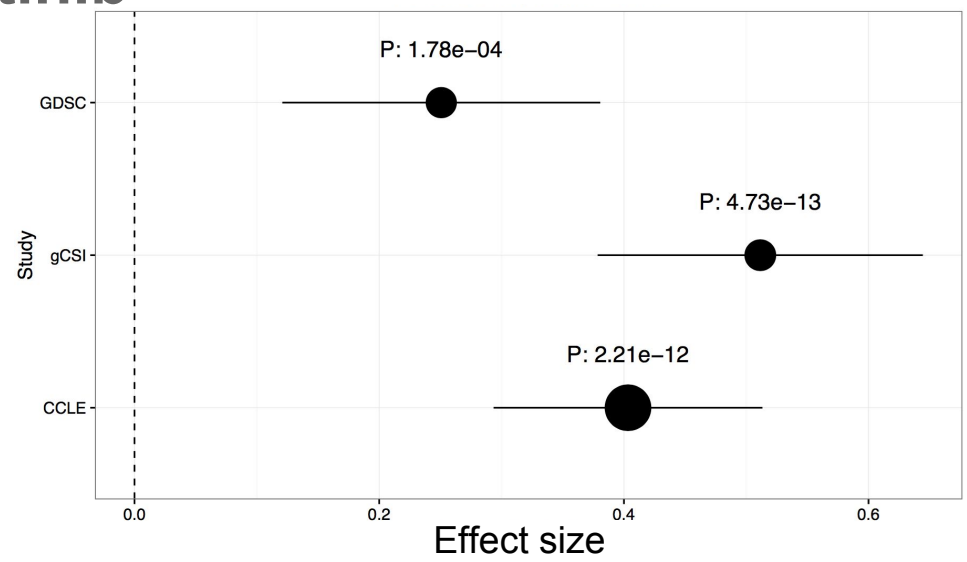
2016 Reproducible pharmacogenomic profiling of cancer cell line panels



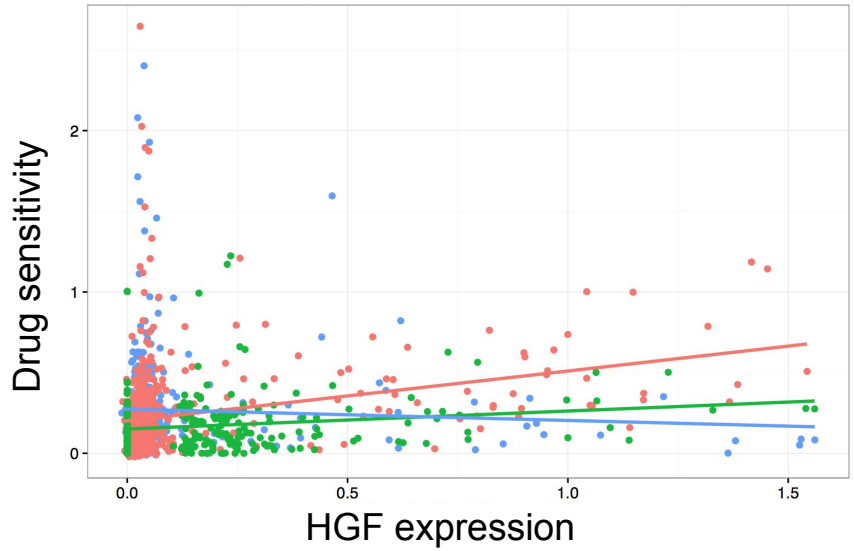
Robust biomarker discovery

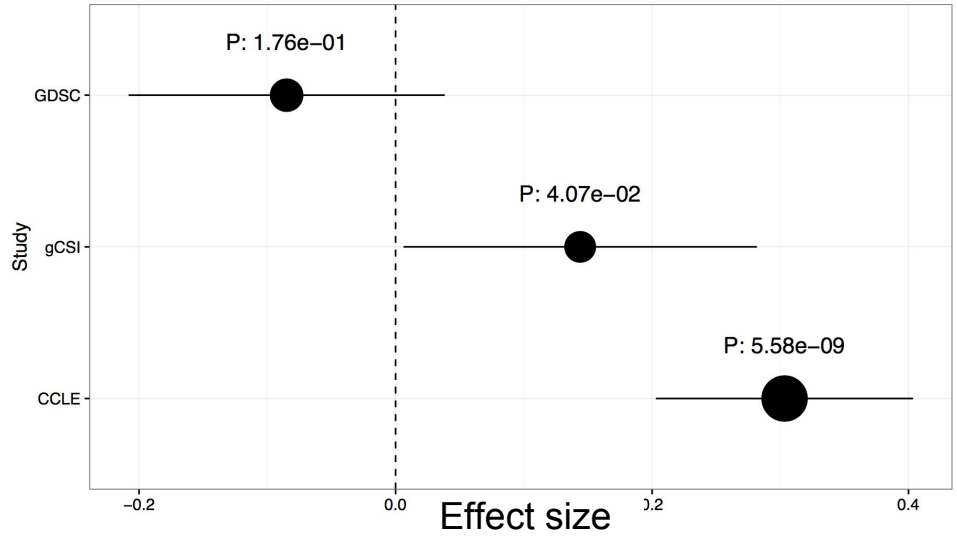
Erlotinib





Crizotinib





Conclusions

- Pharmacogenomics is a hot field, new datasets and new players everyday
  - You can even stay in the game after pissing off the major league :-)
- Great need for standardization
  - Experimental protocols
  - Data processing
  - Annotations
- ***PharmacoGx*** provides a unified platform for meta-analysis of pharmacogenomic studies

Our curation is far from perfect, we need your feedback to make it better!

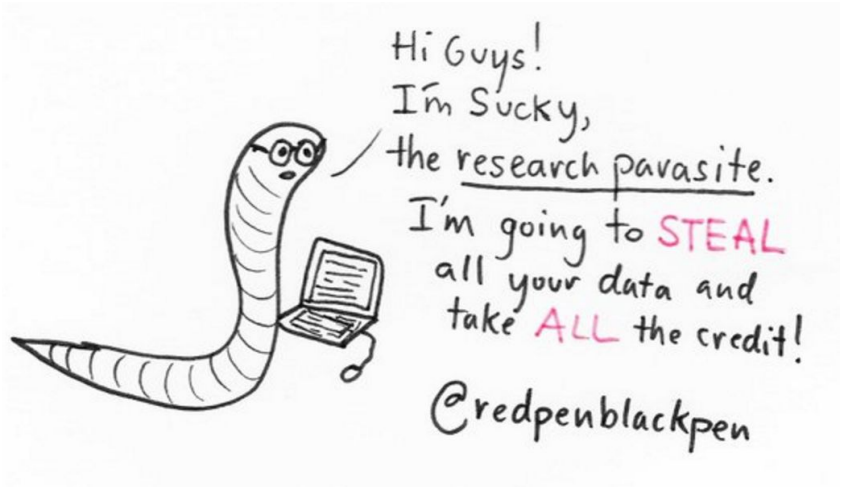
Future directions

- **MultiAssayExperiment (MAE)** to replace the list of ExpressionSet objects and better integrate diverse molecular profiles -- *Workshop session 3*
- ***PharmacoDb***: Companion web-application to faciltate exploration of the large compendium of published pharmacogenomics datasets
- Development of statistical/machine learning methods to jointly analyze heterogeneous pharmacogenomics datasets
- Extension to drug combinations (AstraZeneca-Sanger DREAM Challenge)

PharmacoGx can be safely used by

Data vultures

Data vampires



And research parasites

#IAmAResearchParasite

Research parasites

The NEW ENGLAND JOURNAL of MEDICINE

Data Sharing

Dan L. Longo, M.D., and Jeffrey M. Drazen, M.D.

January 2016

Scientists?

[...] *concern held by some is that a new class of person will emerge — people who had nothing to do with the design and execution of the study but ... other group's data for their own ends, possibly in the research productivity planned by the original investigators, or even use the data to try to disprove what the original investigators had posited. There is concern among some front-line researchers that the system will be taken over by what some researchers have characterized as "research parasites."*

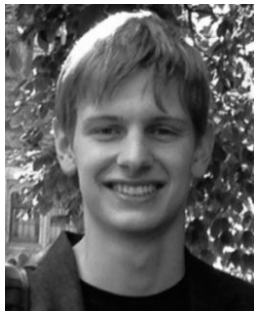
Doing Science?

Acknowledgements

BHK lab

Princess Margaret Cancer Centre

- **Zhaleh Safikhani**
- **Petr Smirnov**
- **Nehme El-Hachem**
- **Mark Freeman**
- **Ali Madani**



Collaborators

- John Quackenbush
- Christos Hatzis
- Christopher Mason
- Leming Shi
- Anna Goldenberg
- Nicolai Juul-Birkbak
- Andrew Beck
- Hugo Aerts

Canadian Cancer Society

OICR

NSERC CRSNG

CIHR IRSC

The Cancer Research Society

The Princess Margaret Cancer Foundation UHN

Thank you
for your attention!

Questions?