CSC 2431
Machine Learning in Computational Biology

Pharmacogenomics

Benjamin Haibe-Kains

Principal Investigator, Bioinformatics and Computational Genomics Laboratory

The Princess Margaret Cancer Centre
University Health Network

Assistant Professor, Medical Biophysics, University of Toronto

UNIVERSITY OF TORONTO

Research interests

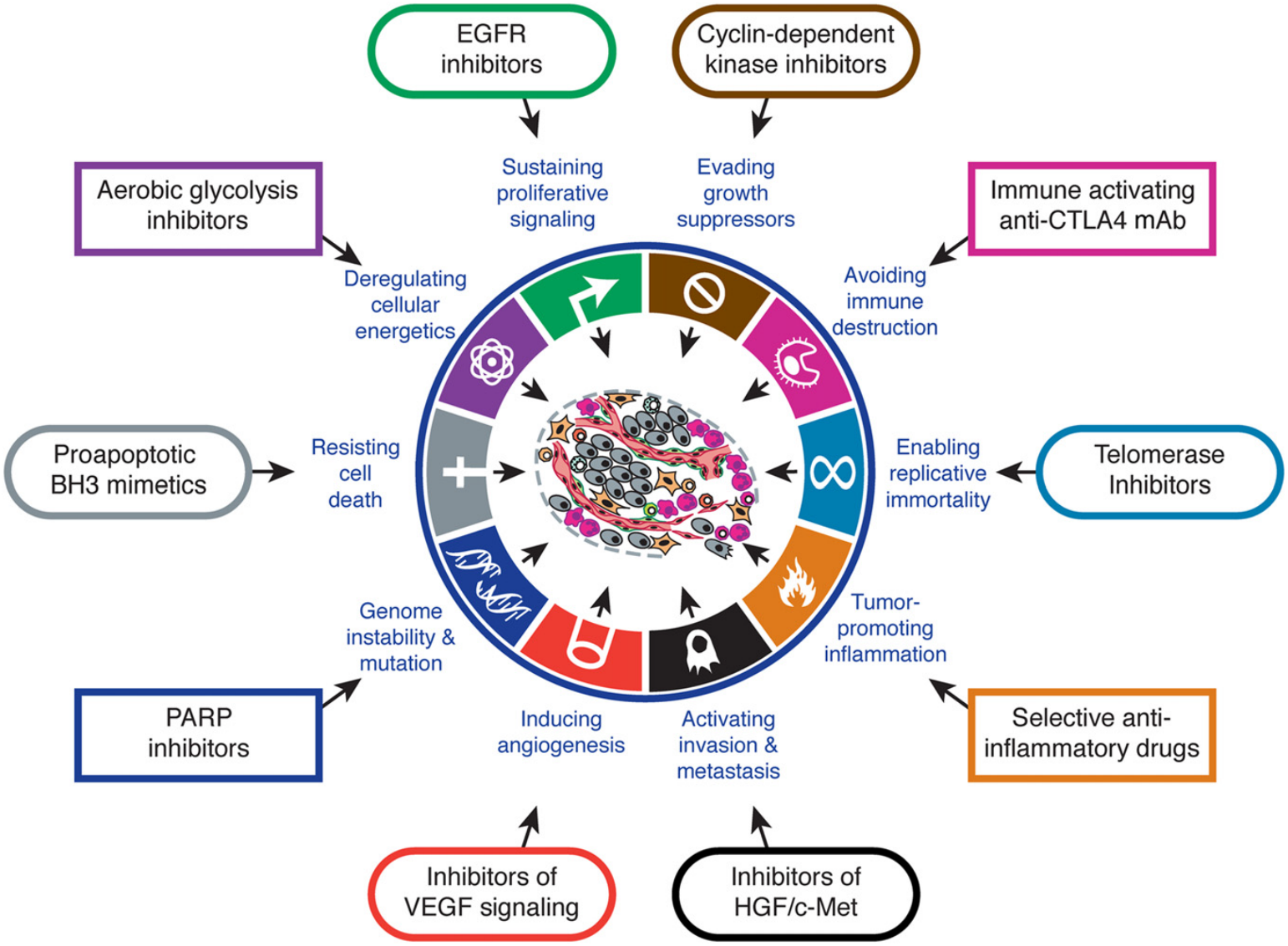
- Website: **http://www.pmgenomics.ca/bhklab/**

- Molecular subtyping
- Prognostic and predictive biomarkers (“signatures”)
- Pharmacogenomics:
  - Drug response prediction
  - Drug repurposing
  - Drug combination/synergy
- Tumor epi-stroma crosstalk

Contents

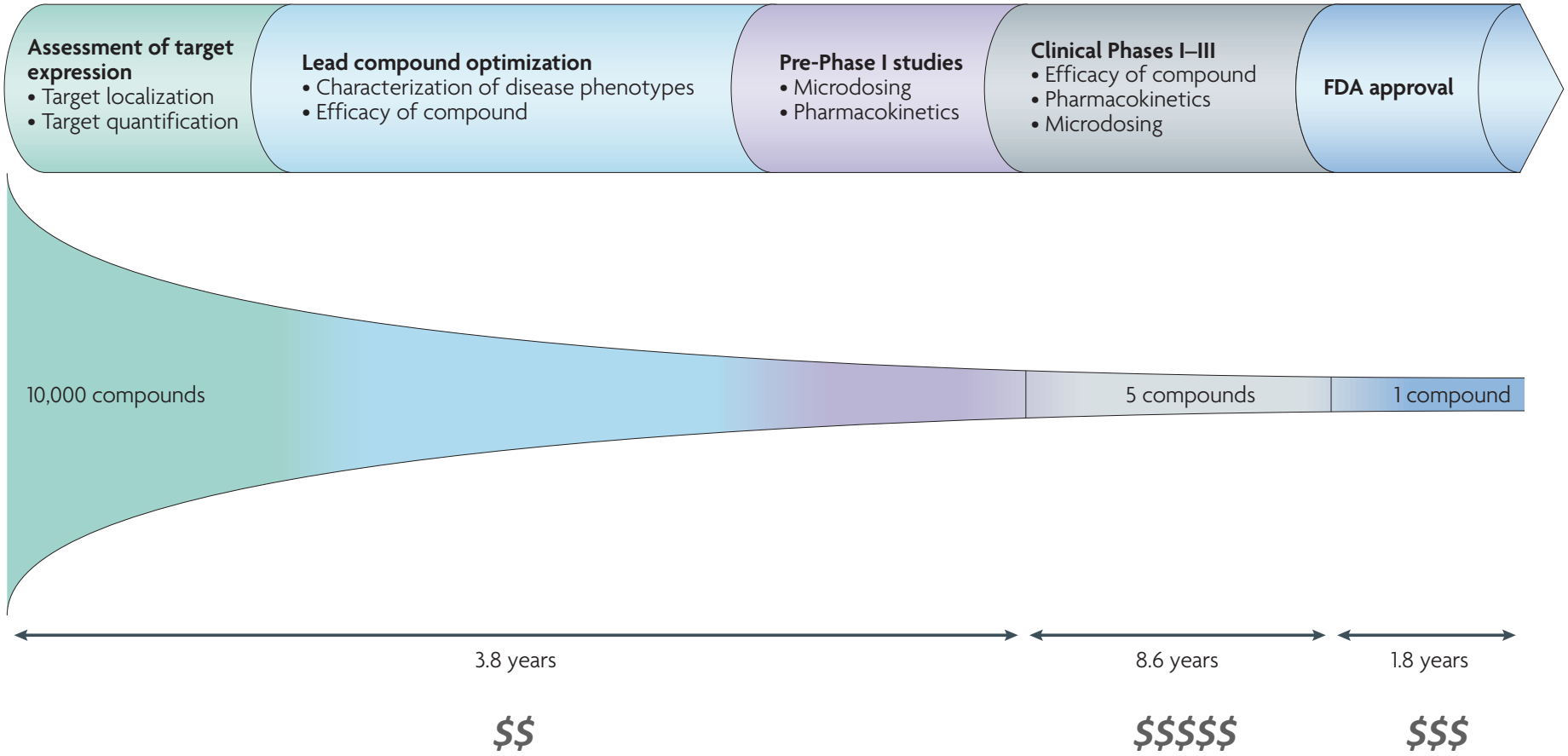
- Therapeutic strategies in cancer
- Biomarkers
- Basics of pharmacogenomic studies
- Machine learning in biomarker discovery
  - Previous efforts
  - Challenges
- Conclusion

Therapeutic strategies in cancer



Hanahan and Weinberg, Cell, 2011

Drug development

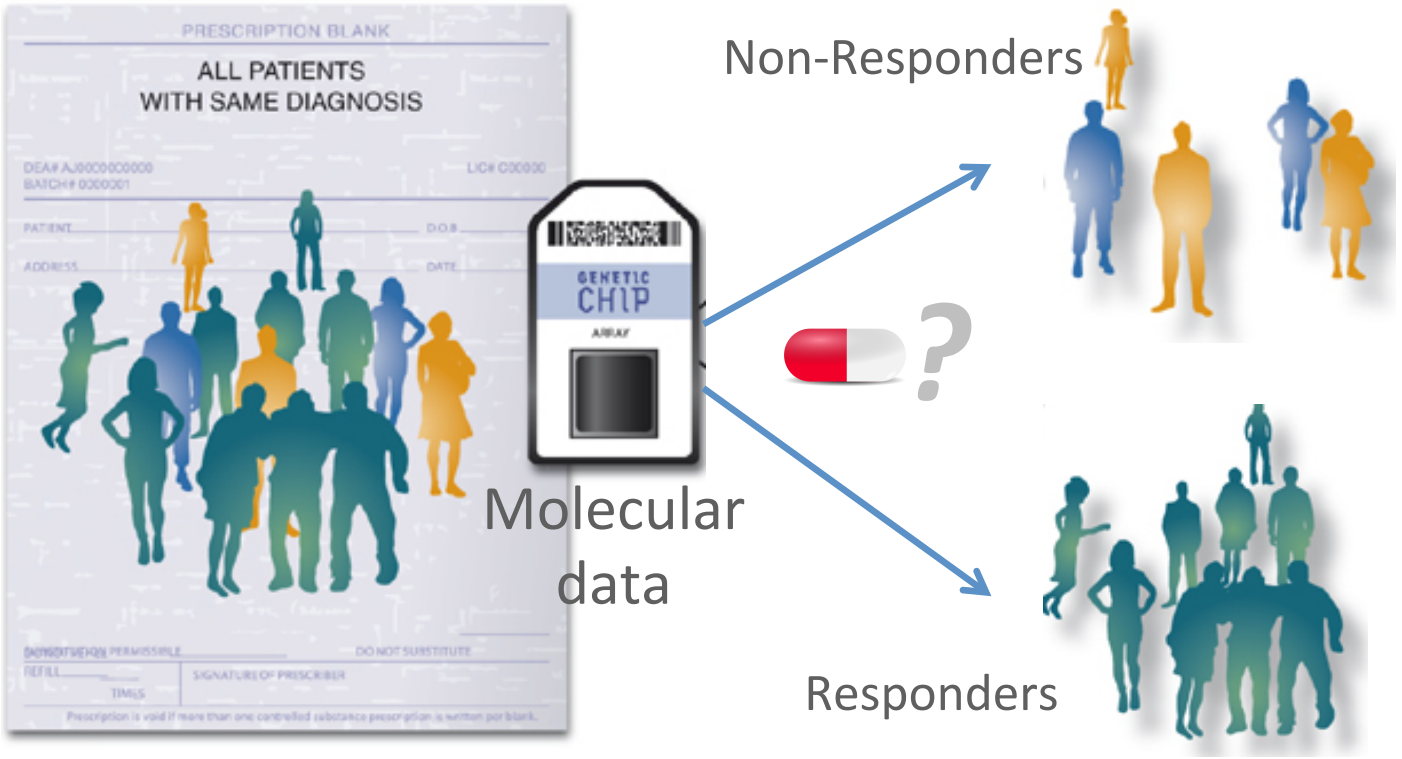


Willman et al., Nature Reviews Drug Discovery 2008

Model systems used for biomarker discovery

- Cost for developing new (targeted) drugs is high and number of drugs being approved is dramatically slowing down

  → ***Need for companion tests to identify patients who are likely to respond to targeted therapies***



Biomarkers

- A **biomarker** is anything that can be used as a measurable indicator of
  - a particular disease state → diagnostic or prognostic biomarker
  - a particular response to a drug → predictive biomarker

- Biomarkers can be measured using different technologies
  - DNA sequencing, gene/protein expression, DNA methylation, etc.
  - Low-throughput: usually univariate biomarkers
  - High-throughput: multivariate biomarkers, often referred to as *signatures* (10-100 features)

Examples of univariate biomarkers

- Genetic aberrations:
  - ✓ BRAF V600E mutation predicts response to vemurafenib and dabrafenib in melanoma
  - ✓ BCR-ABL1 gene fusion predicts response to nilotinib in chronic myelogenous leukemia
- Protein expressions:
  - ✓ ER expression predicts response to tamoxifen in breast cancer
- Gene/transcript expressions:
  - ✓ NQO1 expression predicts response to tanespimycin in multiple cancer types
  - ✓ ERBB2 expression predicts response to lapatinib in breast cancer

Model systems used for biomarker discovery

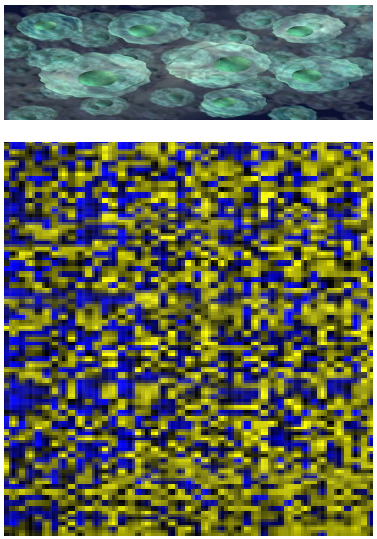
- *In situ*: Patient tumors

- *In vivo*:
  - Patient-derived xenografts (PDX)
  - Genetically-engineered mouse models (GEMM)

- ***In vitro*: Cancer cell lines**
  - **Cheap and high-throughput**
  - **Sensitivity data available for many approved and experimental drugs**
  - **But no cell lines are like real tumors**

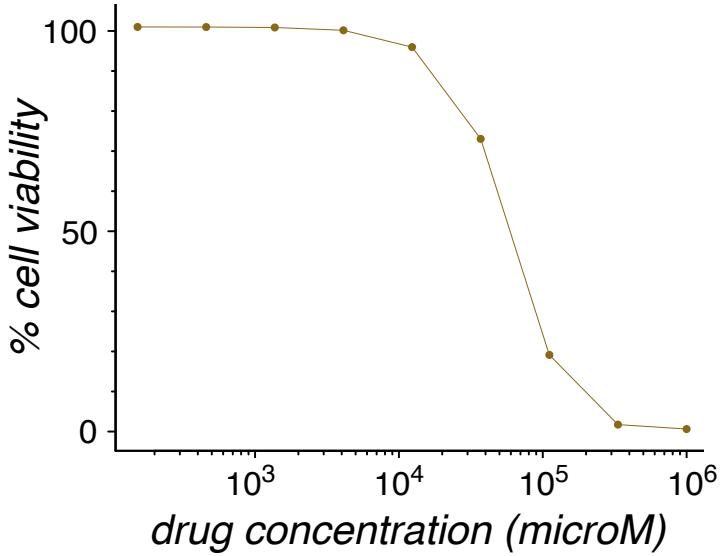
Cell-based drug screening studies

cell lines



molecular data

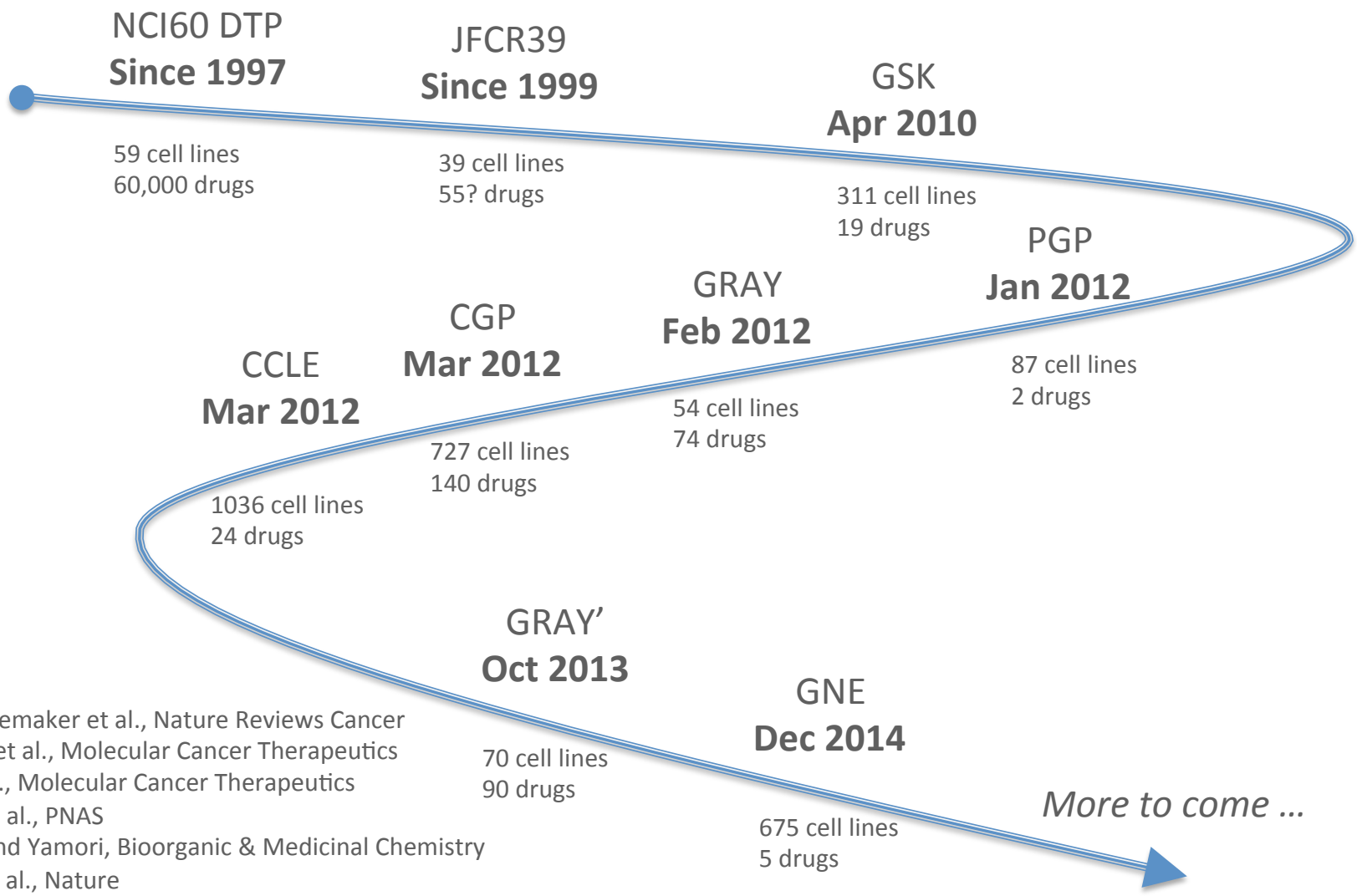
drugs



sensitivity (growth inhibition)

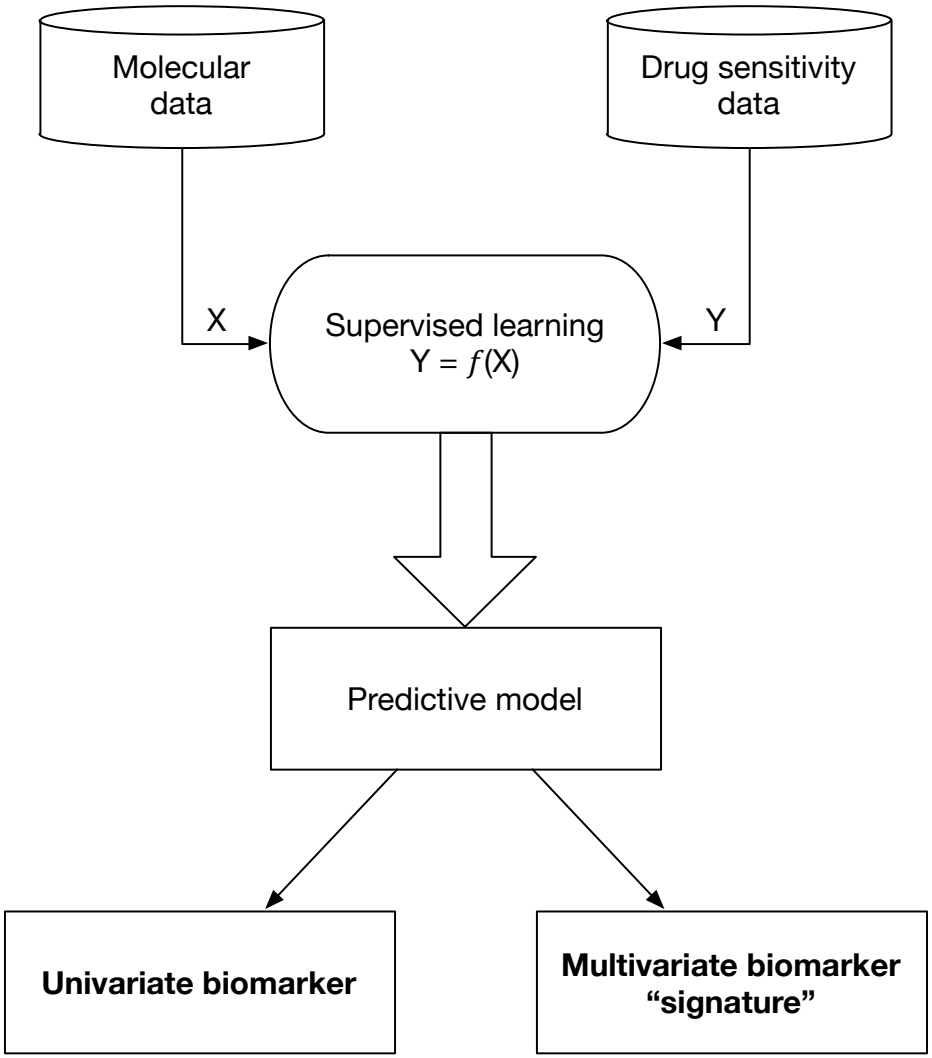
→ resistant vs. sensitive cell lines

Pharmacogenomic datasets



NCI60 DTP: Shoemaker et al., Nature Reviews Cancer
GSK: Greshock et al., Molecular Cancer Therapeutics
PGP: Hook et al., Molecular Cancer Therapeutics
GRAY: Heiser et al., PNAS
JFCR39: Kong and Yamori, Bioorganic & Medicinal Chemistry
CGP: Garnett et al., Nature
CCLE: Barretina et al., Nature
GRAY updated: Daemen et al., Genome Biology
GNE: Klijn et al., Nature Biotechnology

Machine learning in biomarker discovery



First data mining approaches on NCI60

- **COMPARE algorithm** (Paull et al, *JNCI* 1998): unsupervised learning method to cluster drugs with similar IC_{50} measures in the NCI60 panel, implemented in **CellMiner** (Shankavaram et a, *BMC Genomics* 2009)

- **Simple supervised learning** (Staunton, *PNAS* 2001):
  1. Define drug phenotypes (resistant vs. sensitive)
  2. Select most relevant features (high correlation between gene expressions and drug phenotypes)
  3. Build ensemble of univariate models (weighted voting classification)
  4. Assess accuracy in cross-validation

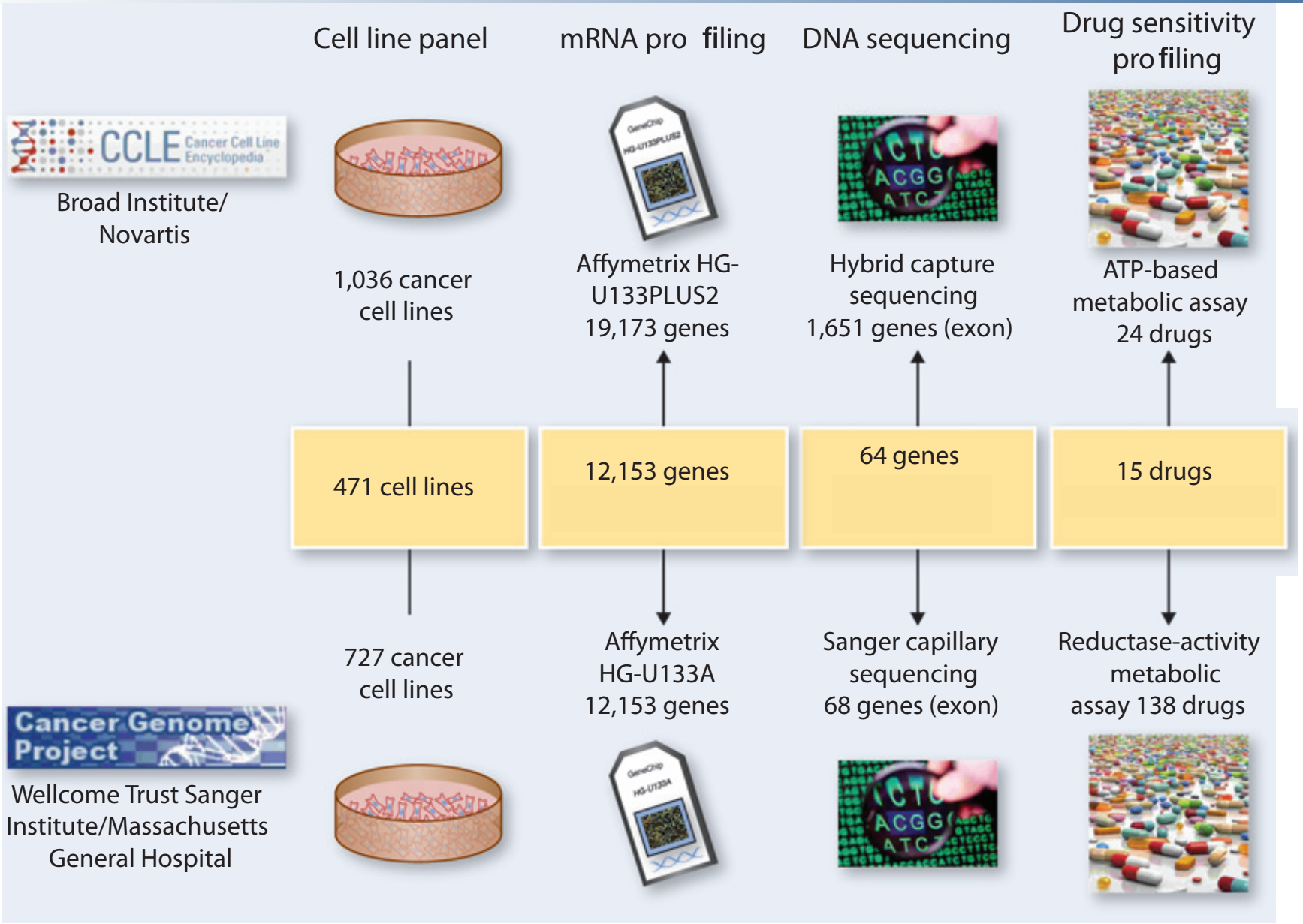
Building and testing predictors with CGP/CCLE

- The Cancer Genome Project (**CGP**) initiated by the Sanger Institute
  - **138** drugs
  - **727** cancer cell lines

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

- Both used regularized regression (*elasticnet*) with mutations, CNVs and gene expressions used as input
- Several summary measures of the dose-response curves were used as output

CCLE ∩ CGP


Cell line panel
mRNA pro filing
DNA sequencing
Drug sensitivity pro filing
CCLE Cancer Cell Line Encyclopedia
Broad Institute/ Novartis
1,036 cancer cell lines
Affymetrix HG-U133PLUS2 19,173 genes
Hybrid capture sequencing 1,651 genes (exon)
ATP-based metabolic assay 24 drugs
471 cell lines
12,153 genes
64 genes
15 drugs
727 cancer cell lines
Affymetrix HG-U133A 12,153 genes
Sanger capillary sequencing 68 genes (exon)
Reductase-activity metabolic assay 138 drugs
Cancer Genome Project
Wellcome Trust Sanger Institute/Massachusetts General Hospital

CCLE ∩ CGP: Shared drugs

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

Studies leveraging CGP/CCLE

JAMIA

Comparison and validation of genomic predictors for anticancer drug sensitivity

Simon Papillon-Cavanagh,[1] Nicolas De Jay,[1] Nehme Hachem,[1] Catharina Olsen,[2] Gianluca Bontempi,[2] Hugo J W L Aerts,[3,4] John Quackenbush,[4] Benjamin Haibe-Kains[1]

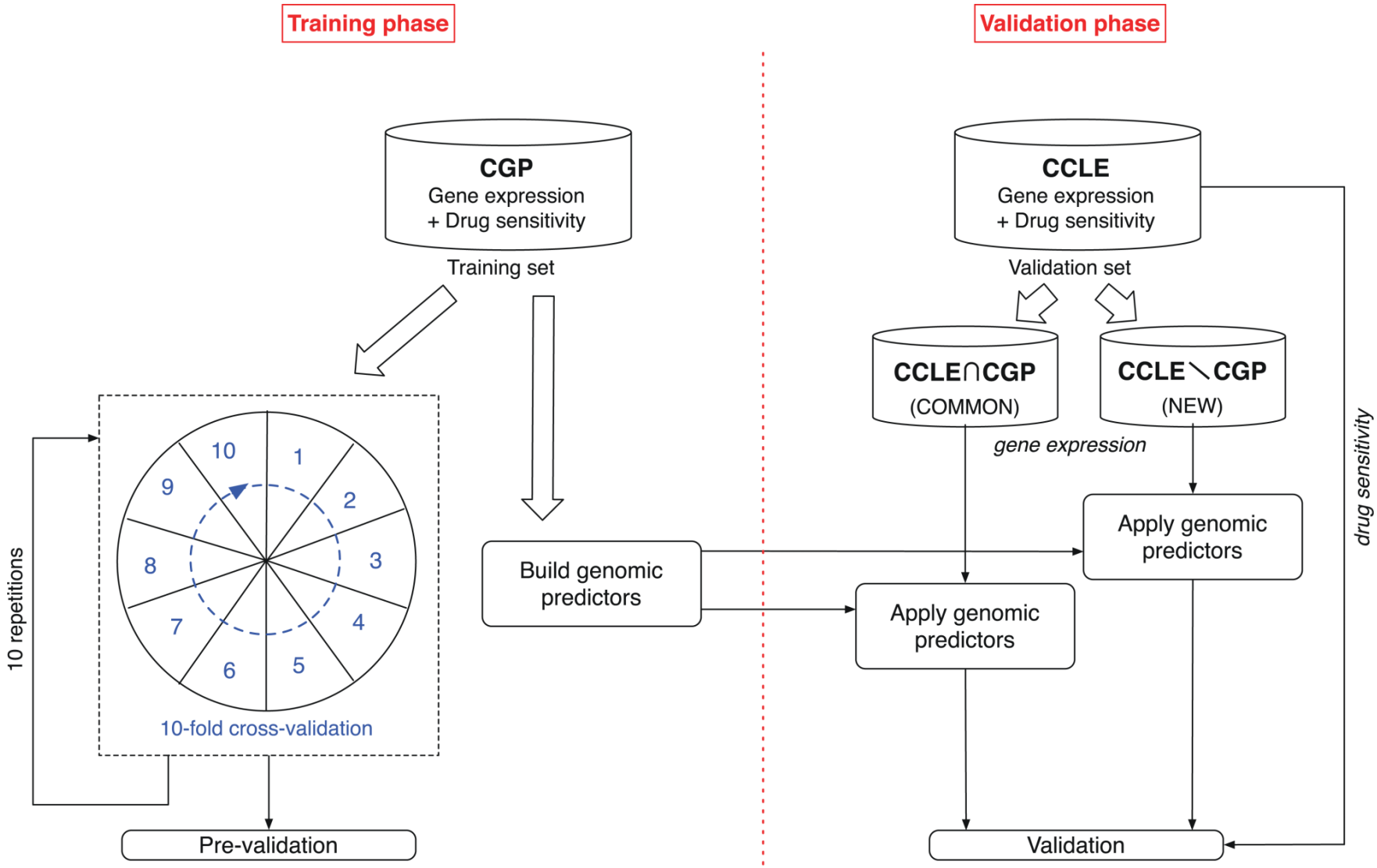
Papillon-Cavanagh S, *et al*. *J Am Med Inform Assoc* 2013;**0**:1–6. doi:10.1136/amiajnl-2012-001442

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]

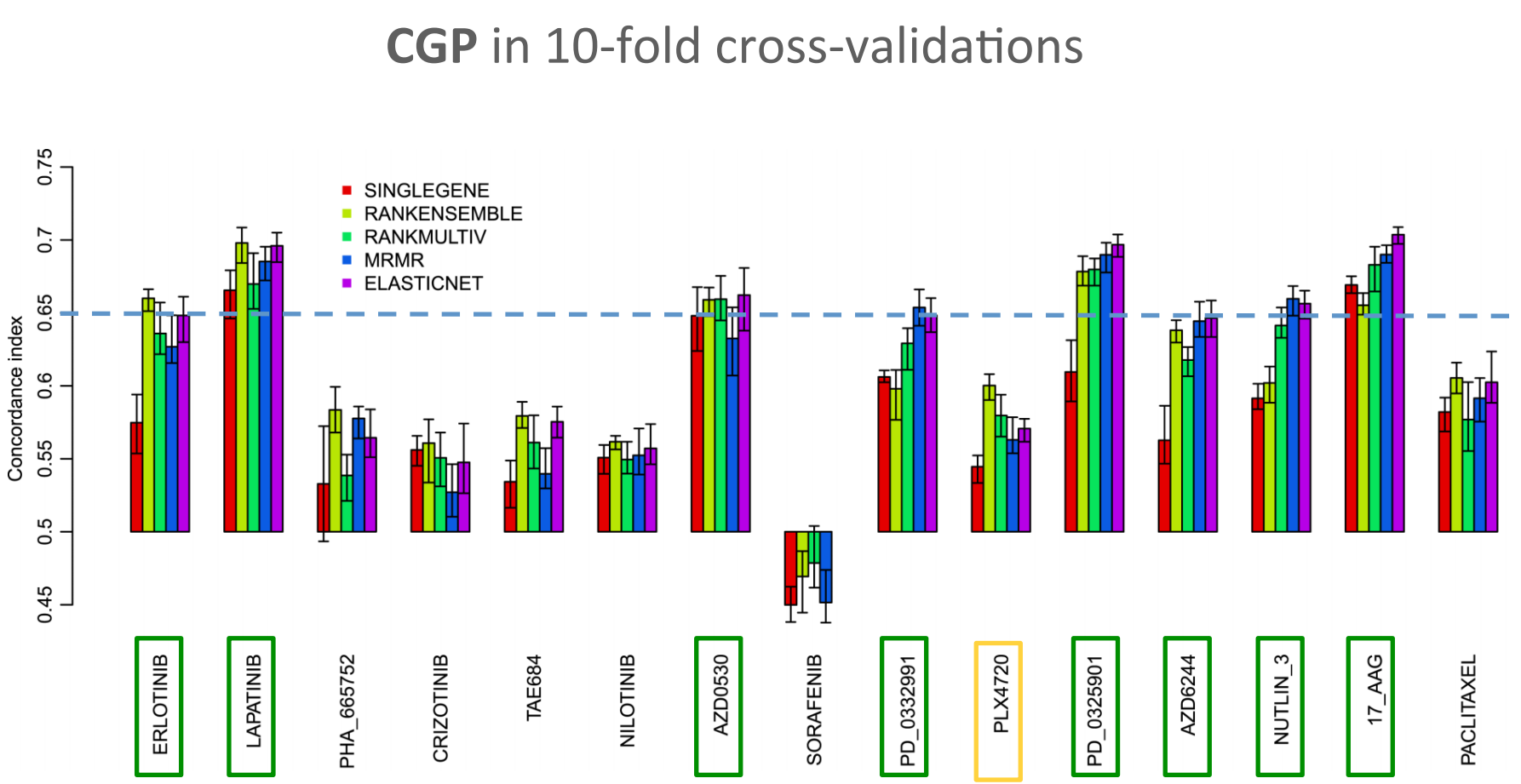
Papillon-Cavanagh et al, experimental design



Papillon-Cavanagh et al, modeling

- **SINGLEGENE**: Univariate linear regression model with the gene the most correlated to drug sensitivity [$-\log_{10}(IC_{50})$]
- **RANKENSEMBLE**: Average of the predictions of the top 30 models
- **RANKMULTIV**: Multivariate model with the top 30 genes
- **MRMR**: Multivariate model with the 30 most correlated and less redundant genes
- **ELASTICNET**: Regularized multivariate model (L1/L2 penalization)

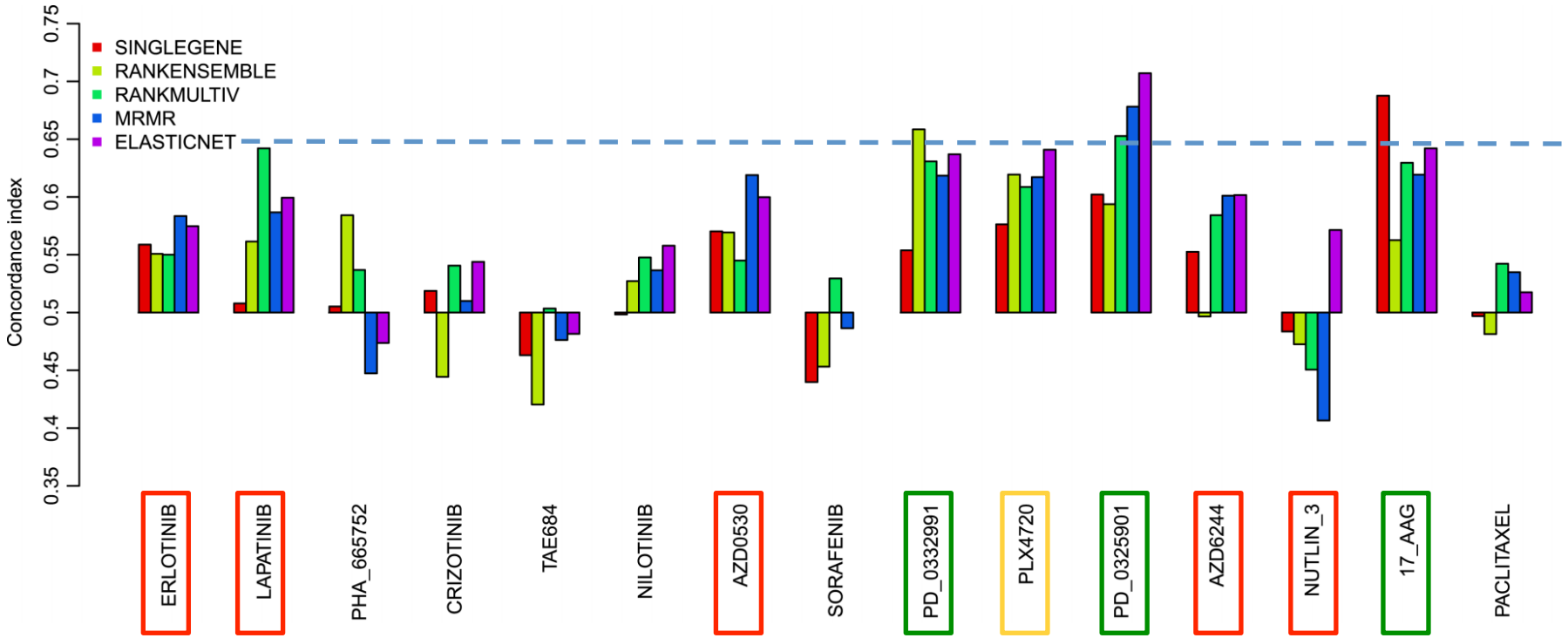
Papillon-Cavanagh et al, results in cross-val



Papillon-Cavanagh et al, test set #1

Trained on CGP, tested on CCLE

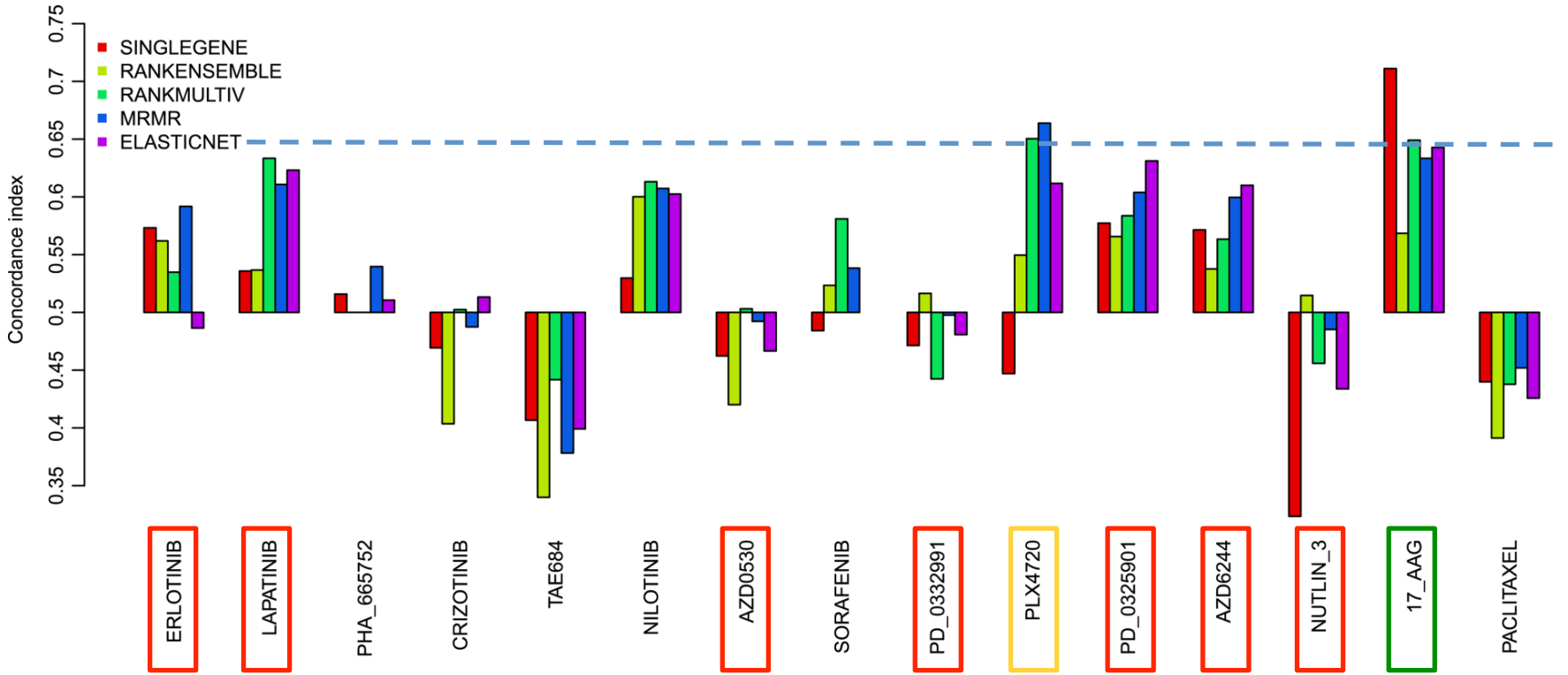
Common cell lines



Papillon-Cavanagh et al, test set #2

Trained on CGP, tested on CCLE

New cell lines



Jang et al, experimental design

- Evaluation of >110,000 predictive modeling strategies to establish best practices in biomarker discovery

- 5 experimental factors:
  - *Molecular features*
  - *Compound*
  - *Response summary*
  - *Continuous vs. categorical models*
  - *Algorithms*

- All combinations are tested in 5-fold cross-validation

Jang et al, results

- Factors explaining most of the variance:
  - the type of *molecular features* used to build the model
  - the *compound* being predicted by the model
  - then *algorithm*
- Strong interaction between compound and response summary factors in CCLE but not in CGP
- Gene expression data are the most predictive
- Elasticnet and ridge regression are the best performers

NCI-DREAM Drug Sensitivity Prediction Challenge

- Training set: multiassay molecular profiling of **35** breast cancer cell lines (mutations, CNV, DNA methylation, gene and protein expressions) treated with **28** drugs
- Test set: [illegible]es
- Go[illegible] the 28
- Pre[illegible]ch compounds
- Performance estimator: weighted average of probabilistic concordance indices across drugs

ANALYSIS

_computational BIOLOGY

A community effort to assess and improve drug sensitivity prediction algorithms

James C Costello[1,2,13,14], Laura M Heiser[3,14], Elisabeth Georgii[4,14], Mehmet Gönen[4], Michael P Menden[5], Nicholas J Wang[3], Mukesh Bansal[6], Muhammad Ammad-ud-din[4], Petteri Hintsanen[7], Suleiman A Khan[4], John-Patrick Mpindi[7], Olli Kallioniemi[7], Antti Honkela[8], Tero Aittokallio[7], Krister Wennerberg[7], NCI DREAM Community[9], James J Collins[1,2,10], Dan Gallahan[11], Dinah Singer[11], Julio Saez-Rodriguez[5], Samuel Kaski[4,8], Joe W Gray[3] & Gustavo Stolovitzky[12]

VOLUME 32 NUMBER 12 DECEMBER 2014 NATURE BIOTECHNOLOGY

NCI-DREAM, the top predictor is quite fancy

- Bayesian multitask multiple kernel learning method that leveraged four machine-learning principles:
  - **kernelized regression** computes outputs from similarities between cell lines
  - **Bayesian inference** to learn drug-specific parameters of the kernelized regression
  - **multiview learning** to combine different "views" of the data (data discretization, pathway-based summarization, data combination, …)
  - **multitask learning** to simultaneously model kernel weights based on drug sensitivities across all the drugs

NCI-DREAM, the second best predictor is less fancy

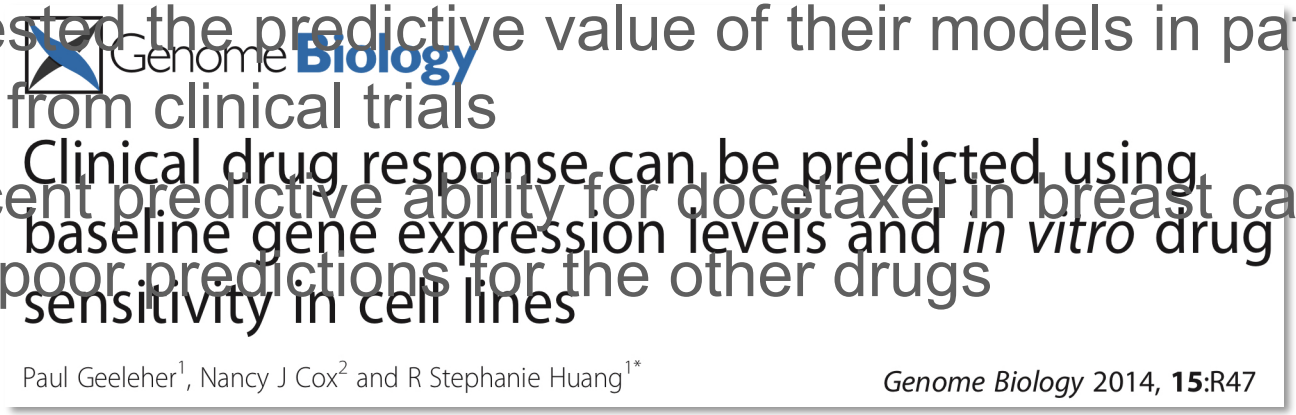
- Ensemble combination of the random forests trained on each data type
- Data type-specific random forest predictions weighted by their R^2

- As usual, ensemble classifiers combining all predictions performed better (wisdom of crowds)
- Gene expression data were the most informative followed by methylation
- Other teams found the proteomic (RPPA) data being very predictive too

Cell lines to patients

- Geeleher et al. used CGP to build predictors of drug response
  - *Ridge* and *elasticnet* regression models
  - *Lasso* logistic regression model using extreme drug phenotypes

- They tested the predictive value of their models in patient tumors from clinical trials
  - Decent predictive ability for docetaxel in breast cancer but poor predictions for the other drugs



Challenges

- Over 1000 PubMed articles just for gene expression-based biomarkers in cancer cell lines
- But few biomarkers translated into clinical settings …
- Challenges:
  - Analysis reproducibility and validation
  - Cell line identity crisis
  - (In)Consistency in drug phenotypes
  - Experimental protocols
  - Summary measures for drug phenotypes

Analysis reproducibility and validation

- Irreproducible research, see Anil Potti's scandal at Duke University

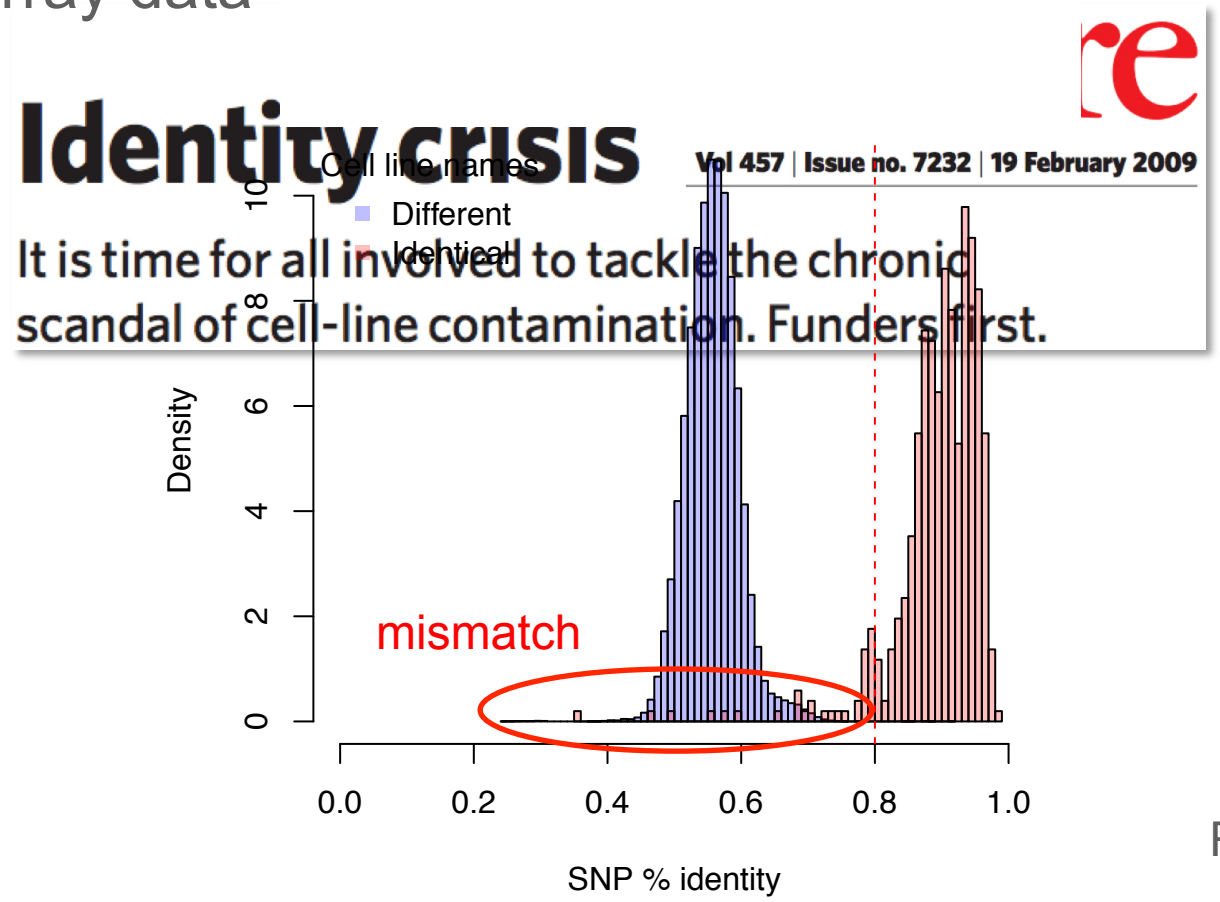
  (Potti et al, *Nat Med* 2011) ***retracted**

- Unsuccessful validation of MPI predictors built on the NCI60 cell line panel to predict therapy response

  (Wang *et al.*, *JNCI* 2013)

Cell line identity crisis

- Mislabeling or contamination of the cancer cell lines
- One can check the identity of the cell lines by leveraging the SNP array data



Rene Quevedo

(In)Consistency of pharmacogenomic data

- **Input data** → Molecular profiles
  - For each [cell line, gene], are *molecular measurements* consistent across studies?

- **Output data** → Drug phenotypes
  - For each [cell line, drug], are *sensitivity measurements* consistent?

nature

RESEARCH ANALYSIS

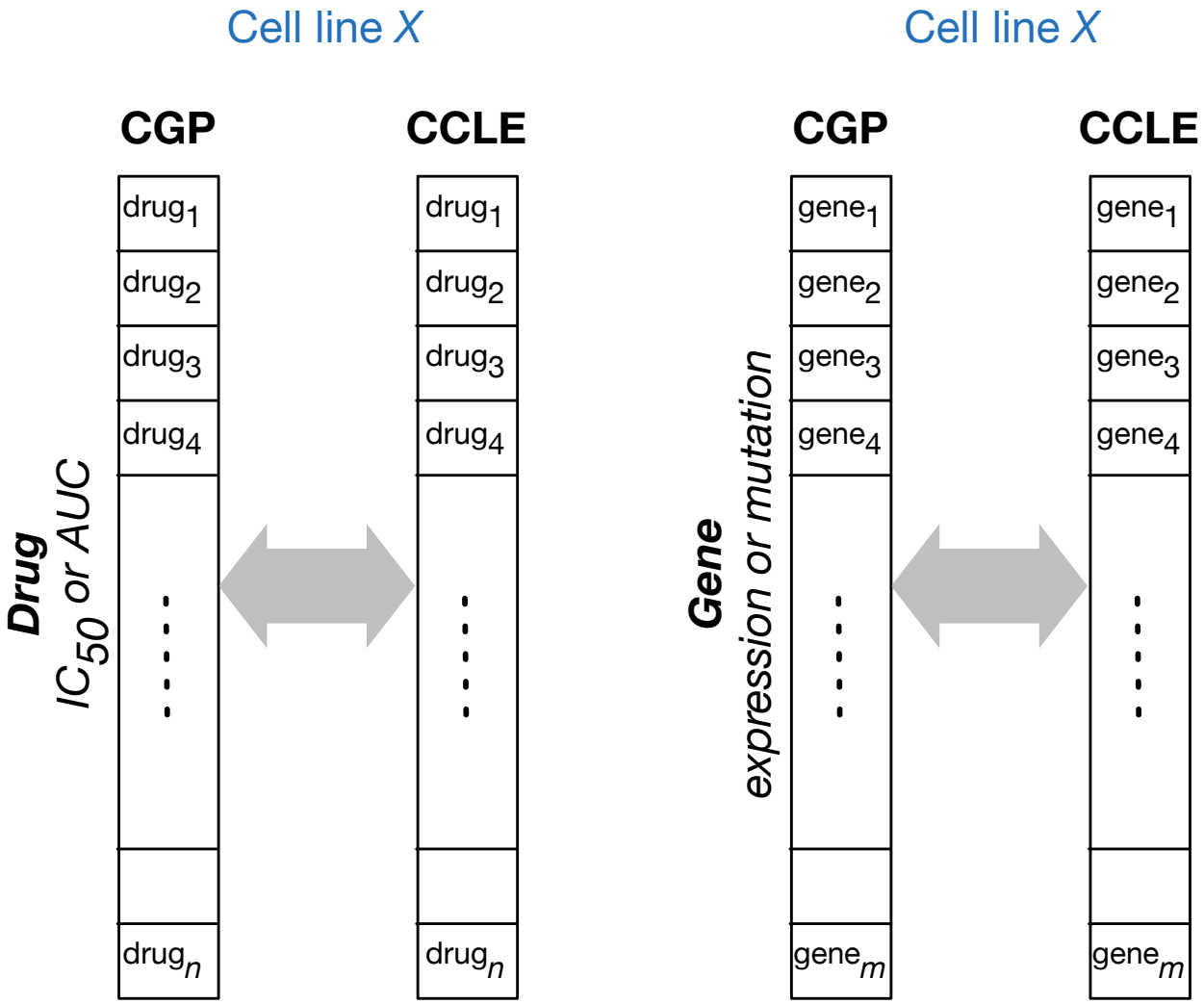
Inconsistency in large pharmacogenomic studies

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]*

Nehme El-Hachem

Concordance “between cell lines”

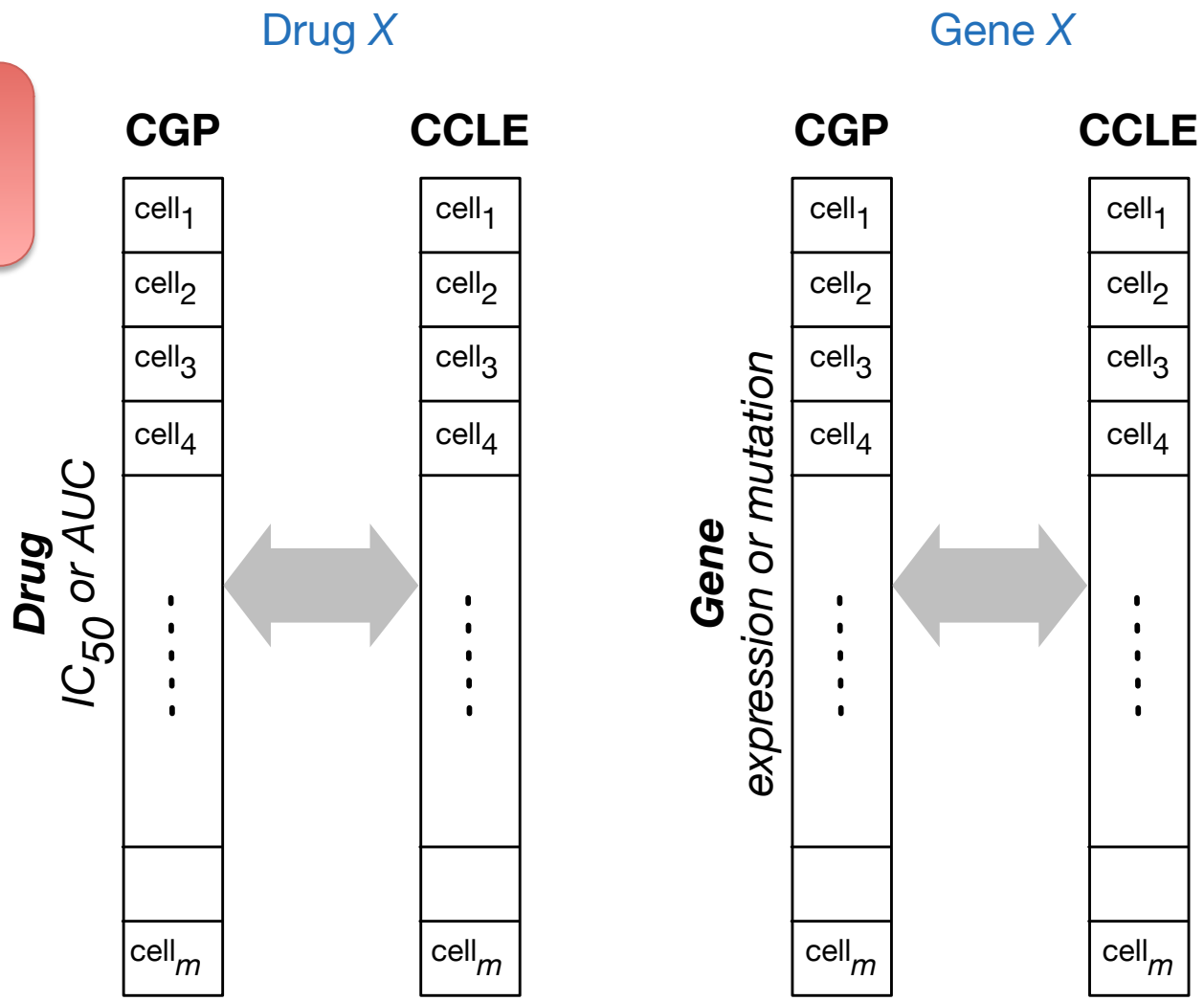
- Identical cell lines in CGP and CCLE should exhibit similar profiles



Concordance “across cell lines”

- Identical drugs or genes in CGP and CCLE should exhibit similar measurements in cell lines

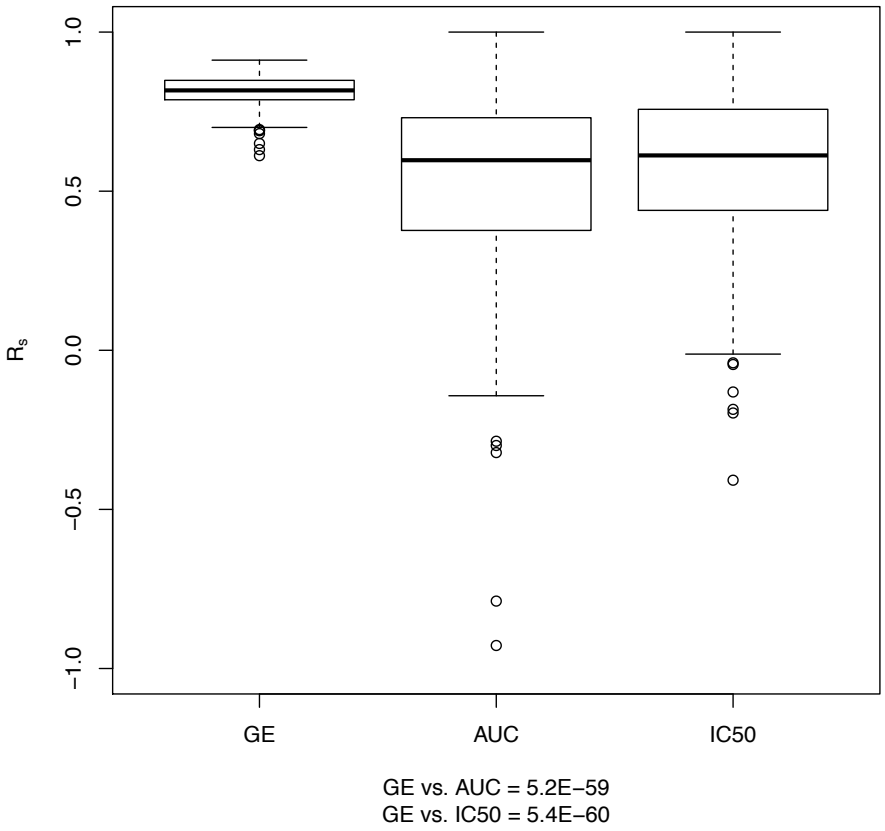
Most relevant for biomarker discovery



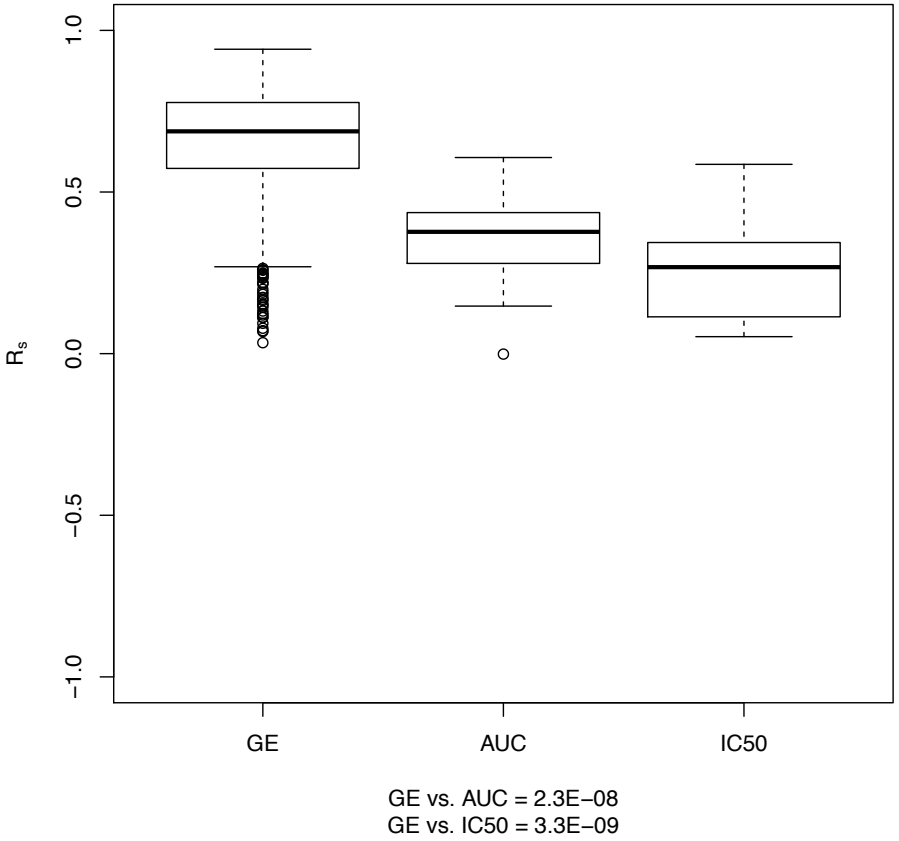
Concordance of transcriptomic and pharmacologic data

Spearman correlation

Between cell lines

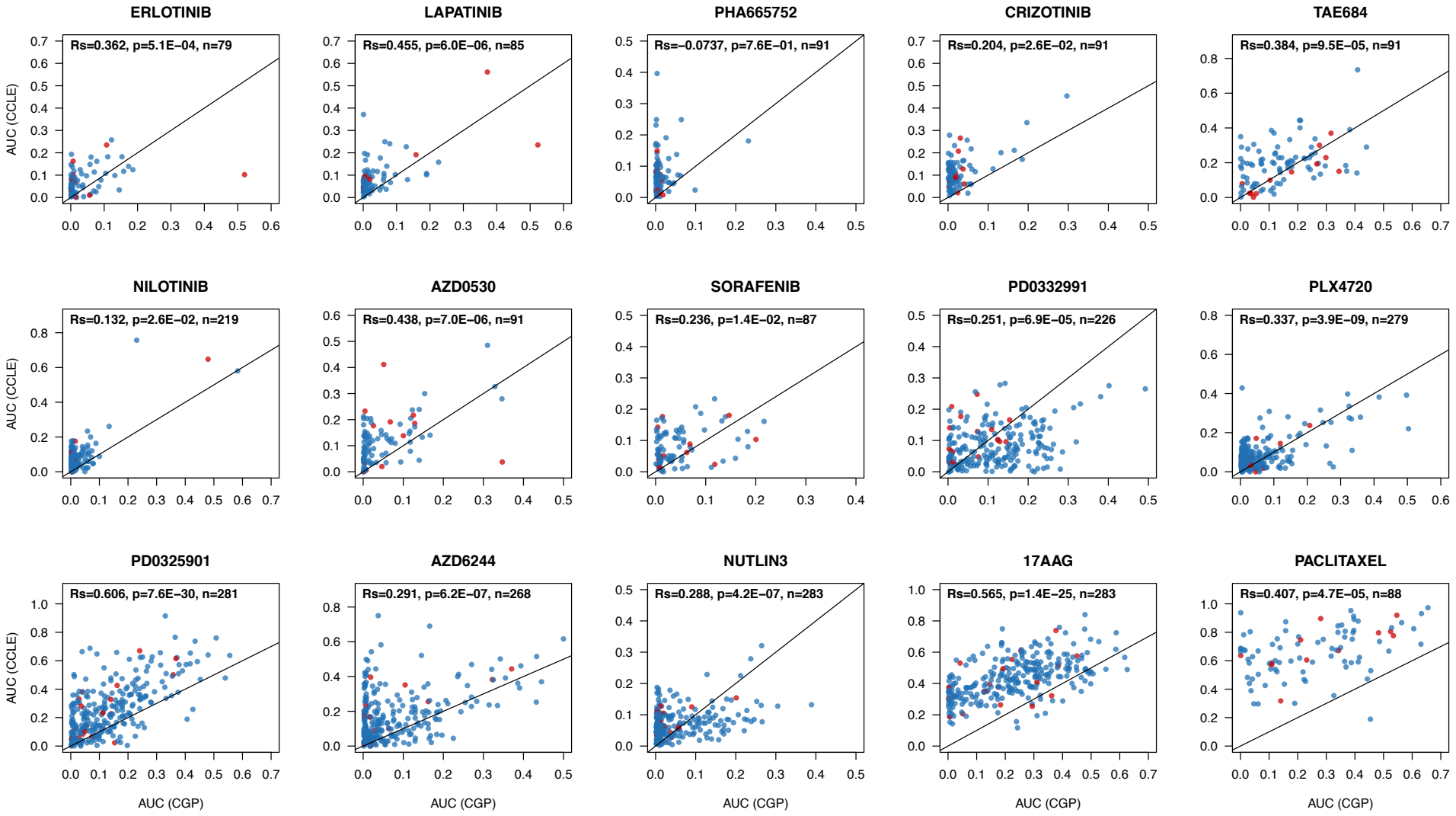


Across cell lines



➔ Gene expression profiles are significantly more correlated than drug sensitivity measures

Across cell lines → published AUC



Consistency of gene-drug associations

- Univariate biomarkers

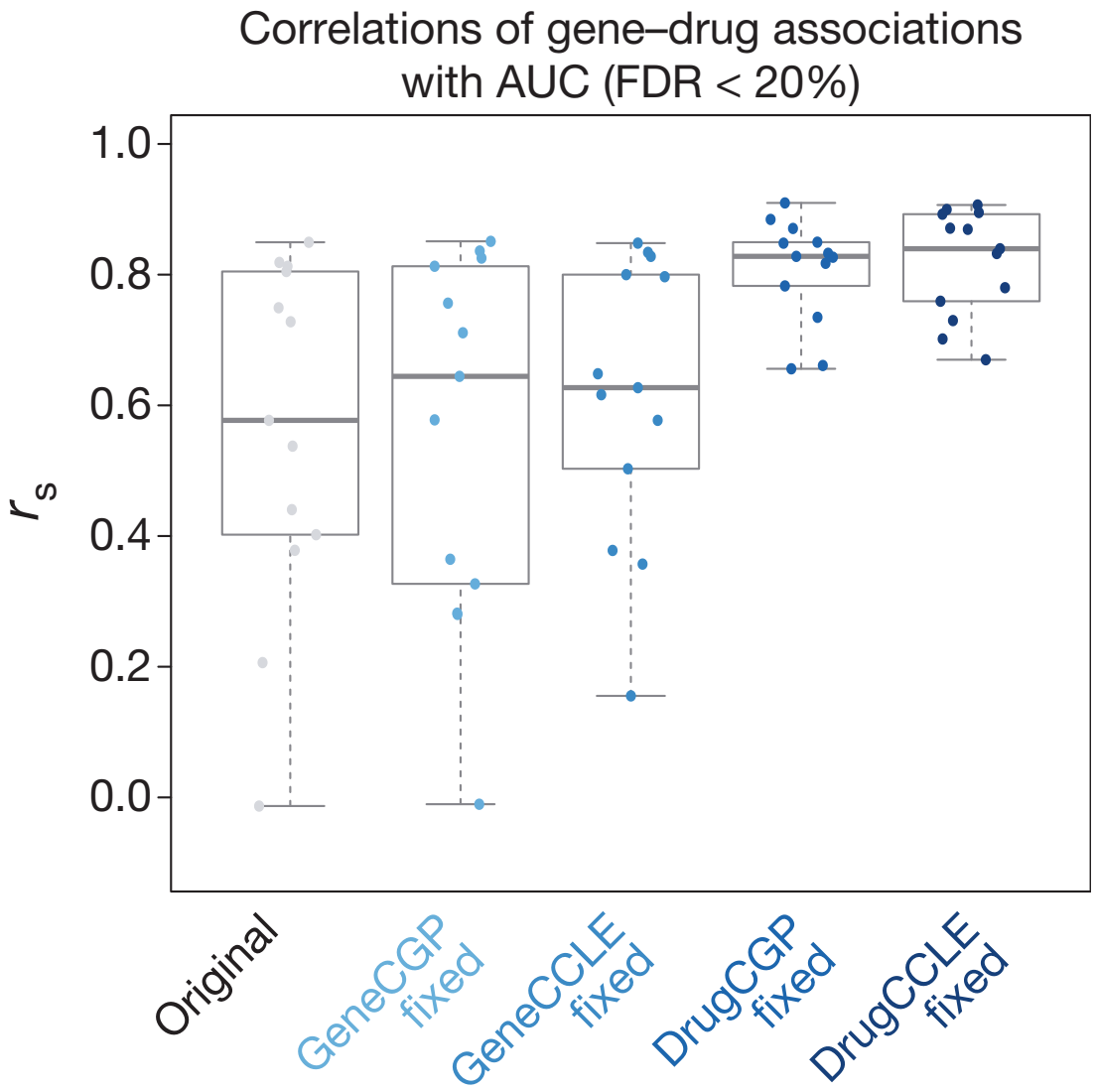
- Model for gene-drug association:

$$Y = \beta_0 + \beta_i G_i + \beta_t T$$

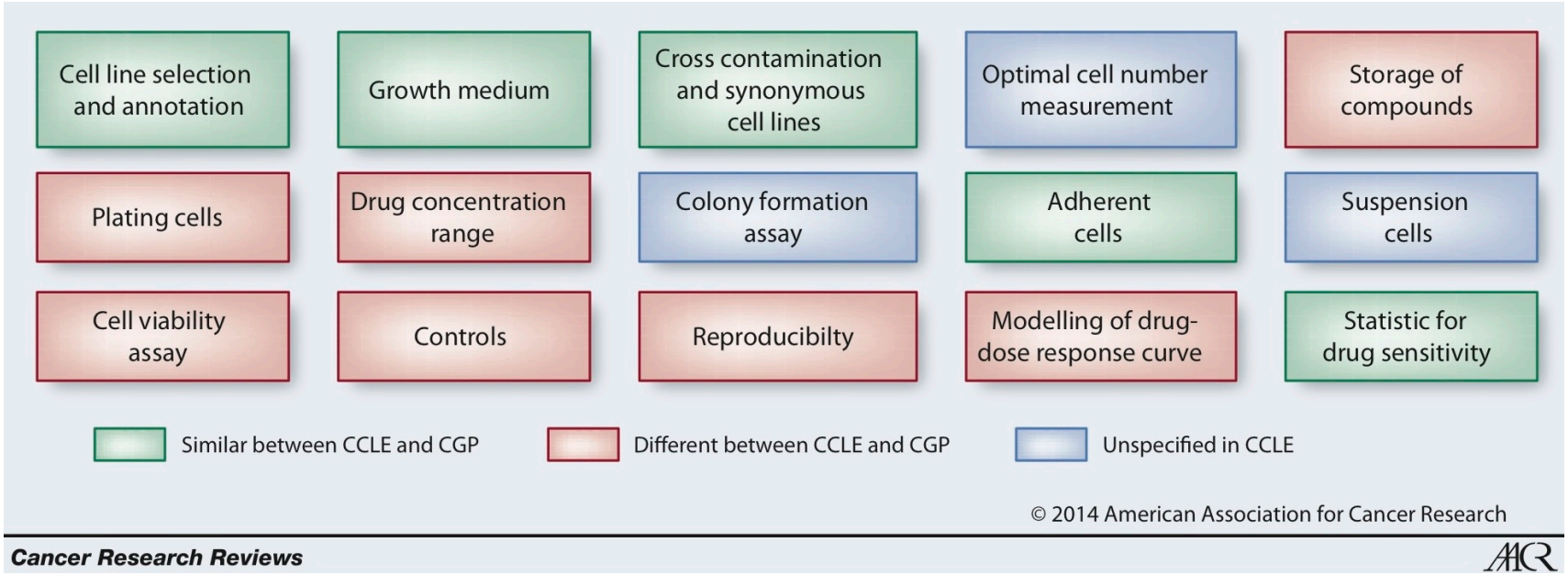
where Y = drug sensitivity (AUC)
G_i = gene expression of gene i
T = tissue type

- Strength and significance of association are given by β_i and its corresponding p-value

Gene expression or drug sensitivity at fault?



Heterogeneous experimental protocols



Review

Cancer Research

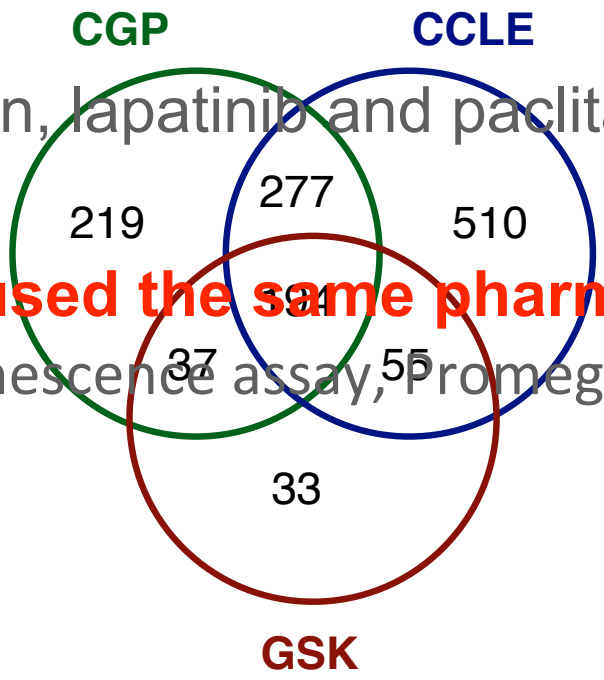
Enhancing Reproducibility in Cancer Drug Screening: How Do We Move Forward?

Christos Hatzis[1,3], Philippe L. Bedard[10,11], Nicolai Juul Birkbak[13], Andrew H. Beck[4], Hugo J.W.L. Aerts[5,7], David F. Stern[2,3], Leming Shi[8,14,15], Robert Clarke[9], John Quackenbush[5,6], and Benjamin Haibe-Kains[10,12]

Pharmacological assay

- In 2010, GlaxoSmithKline tested
  - **19** compounds
  - on **311** cancer cell lines

- 194 cell lines in common with CGP and CCLE

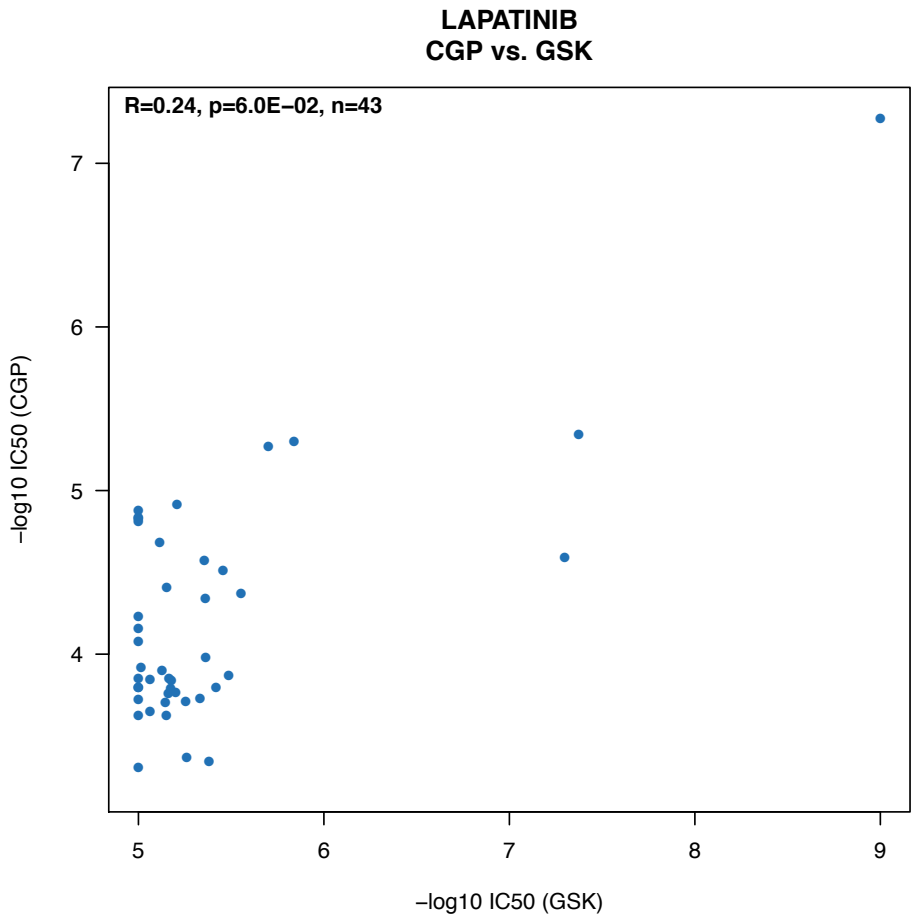


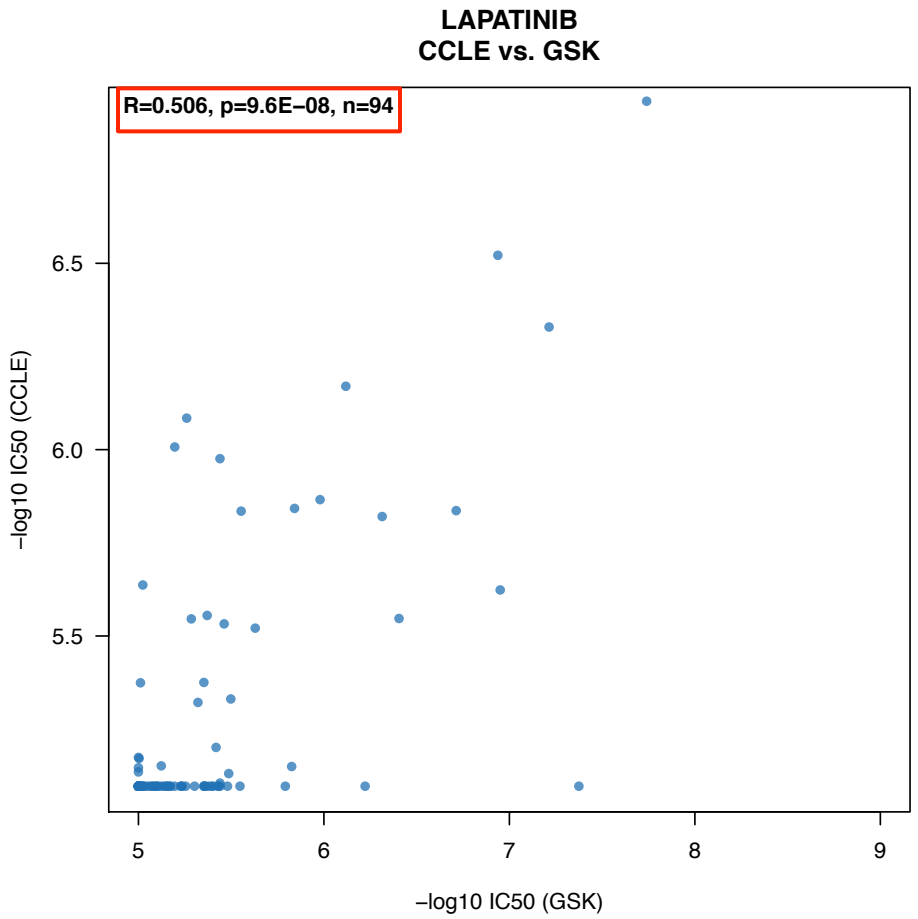
- 2 drugs in common, lapatinib and paclitaxel

- **CCLE and GSK used the same pharmacological assay** (Cell Titer Glo luminescence assay, Promega)

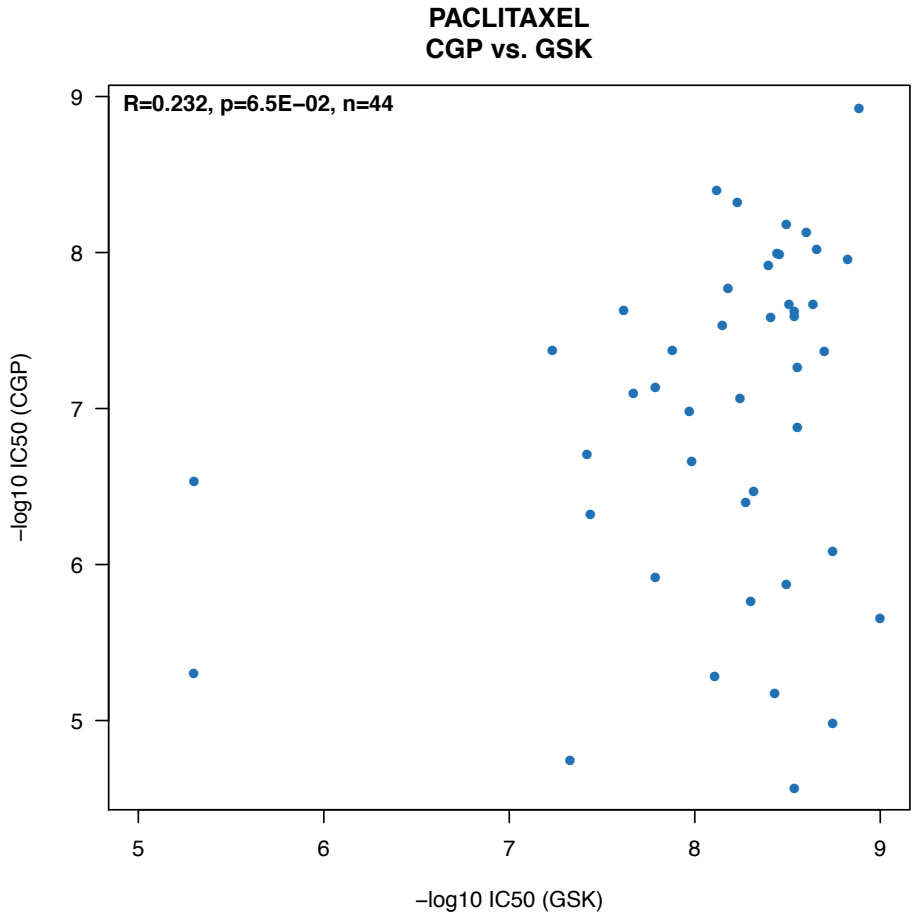
Nehme El-Hachem
Donald Wang

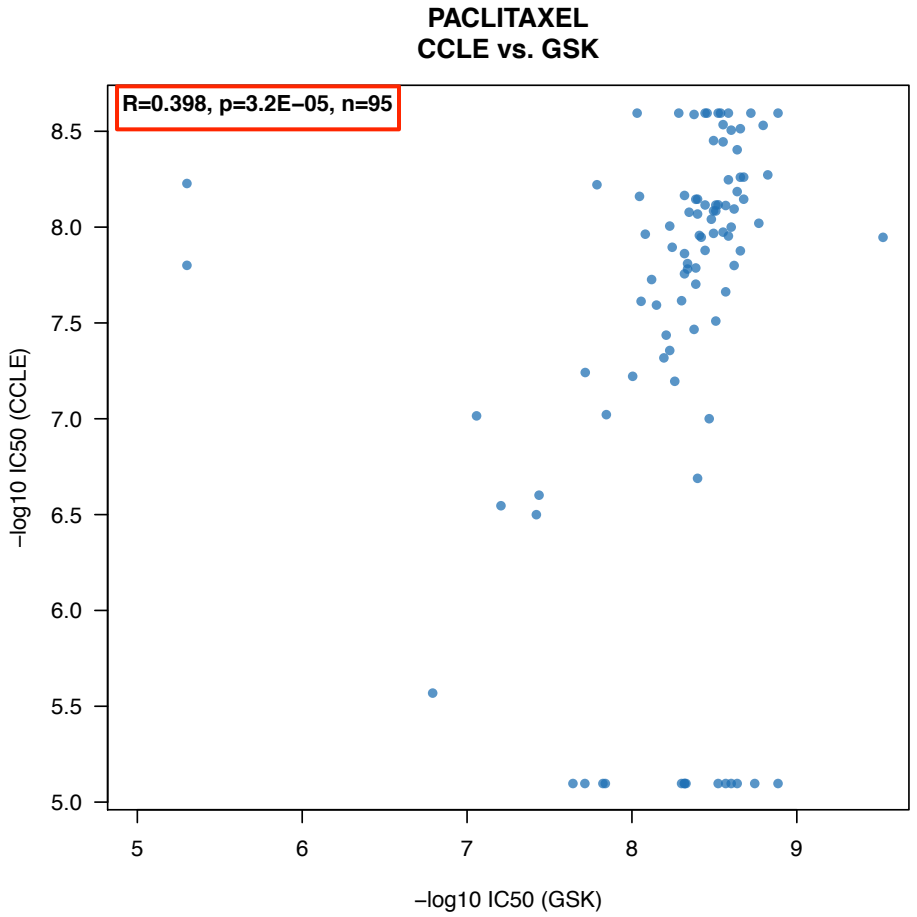
Comparison with GSK for lapatinib



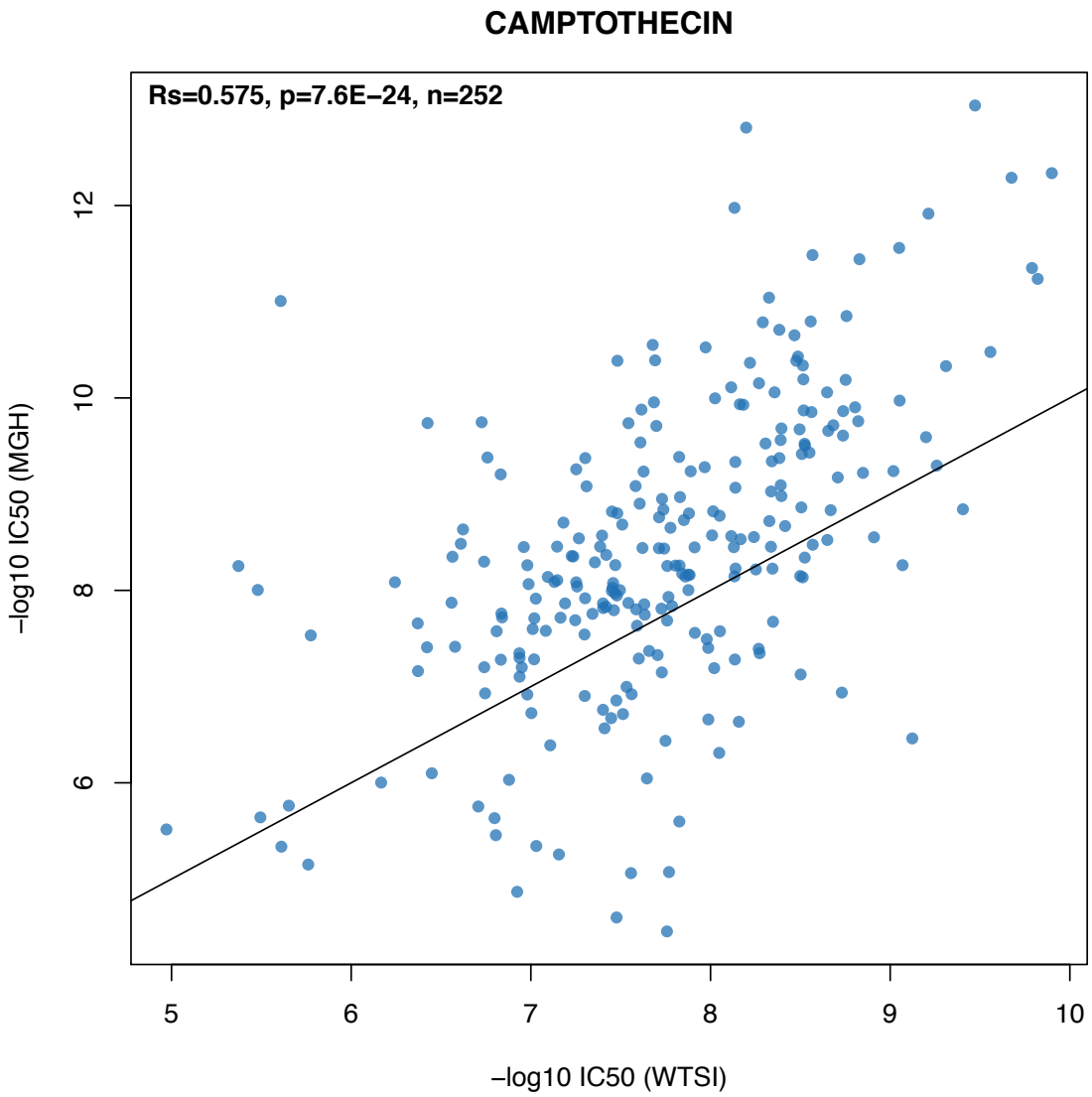


Comparison with GSK for paclitaxel

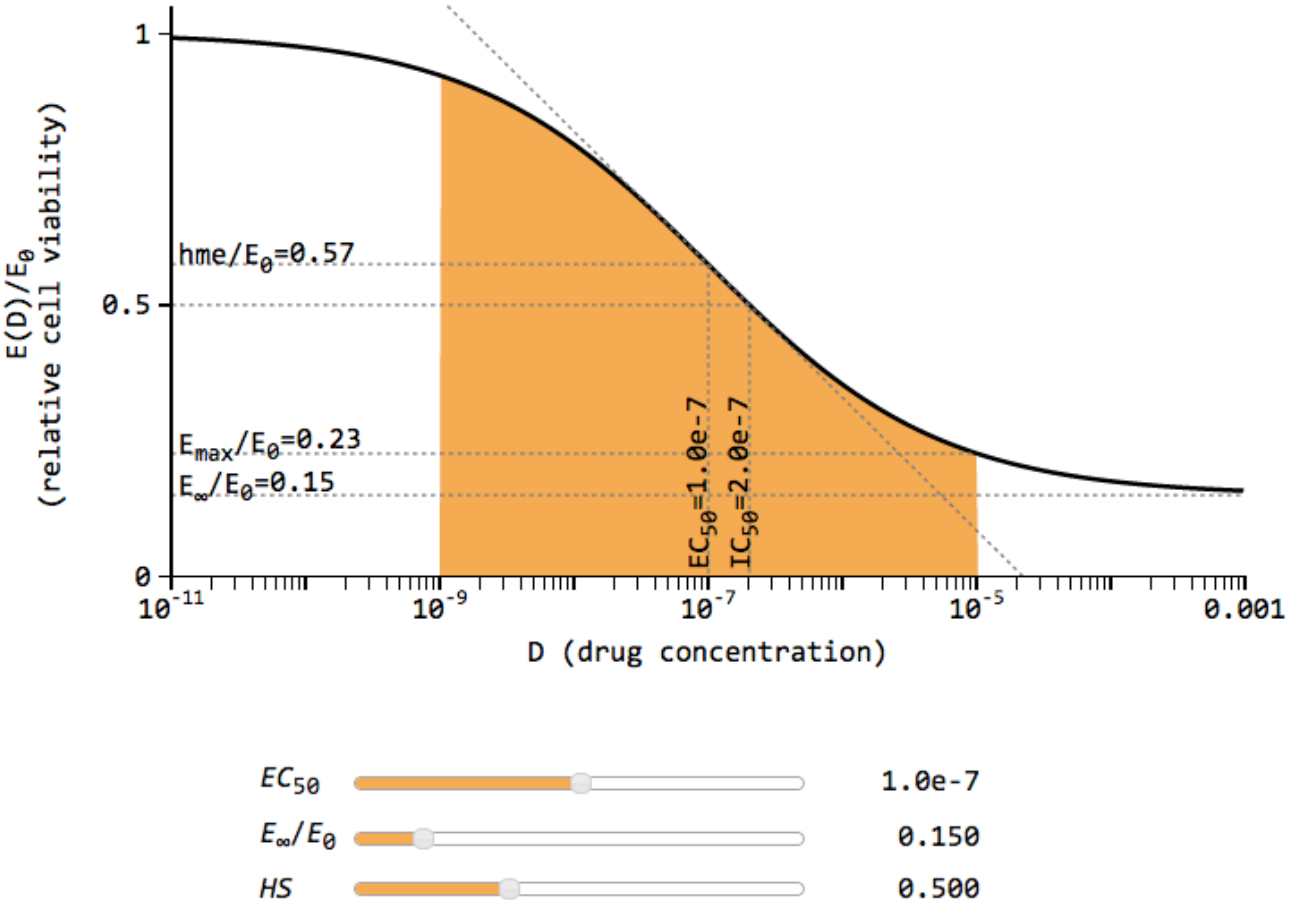




Identical protocols in different laboratories?



Complexity of drug dose-response curves



http://lincs.hms.harvard.edu/explore/10.1038-nchembio.1337/fallahi-sichani-2013/

Complexity of drug dose-response curves

- The choice of drug sensitivity summary measure will strongly affect the biomarker discovery process

nature chemical biology

ARTICLE

PUBLISHED ONLINE: 8 SEPTEMBER 2013 | DOI: 10.1038/NCHEMBIO.1337

Metrics other than potency reveal systematic variation in responses to cancer drugs

Mohammad Fallahi-Sichani[1], Saman Honarnejad[1], Laura M Heiser[2], Joe W Gray[2] & Peter K Sorger[1]

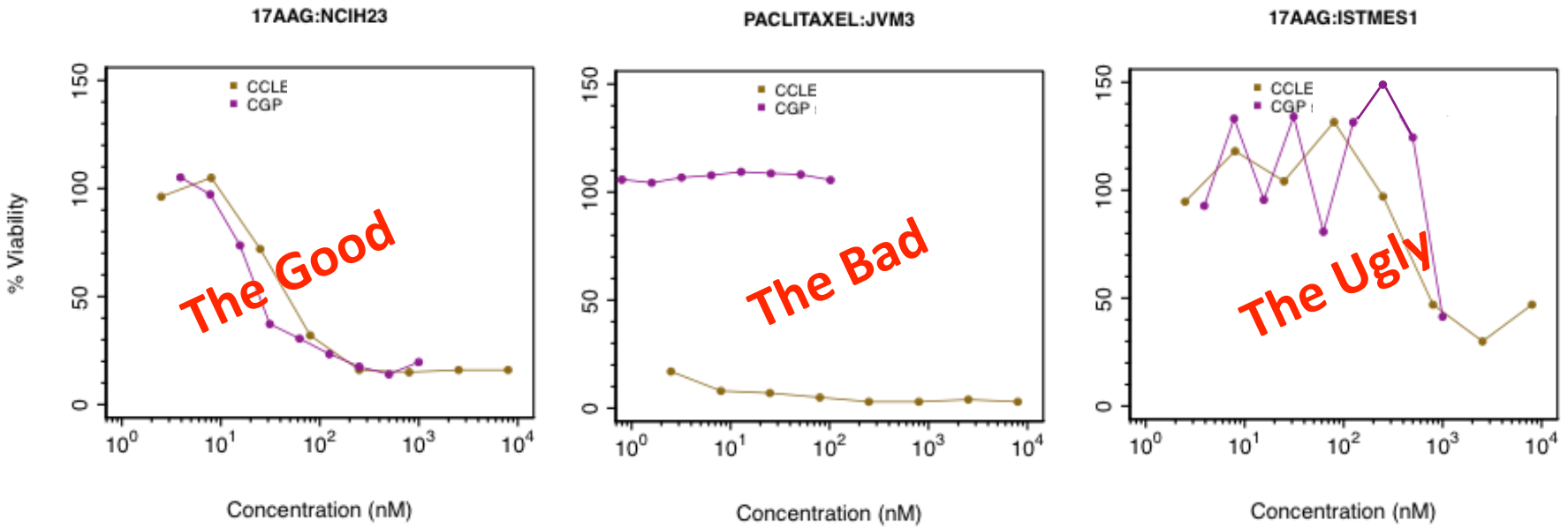
Pacific Symposium on Biocomputing 2014

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]

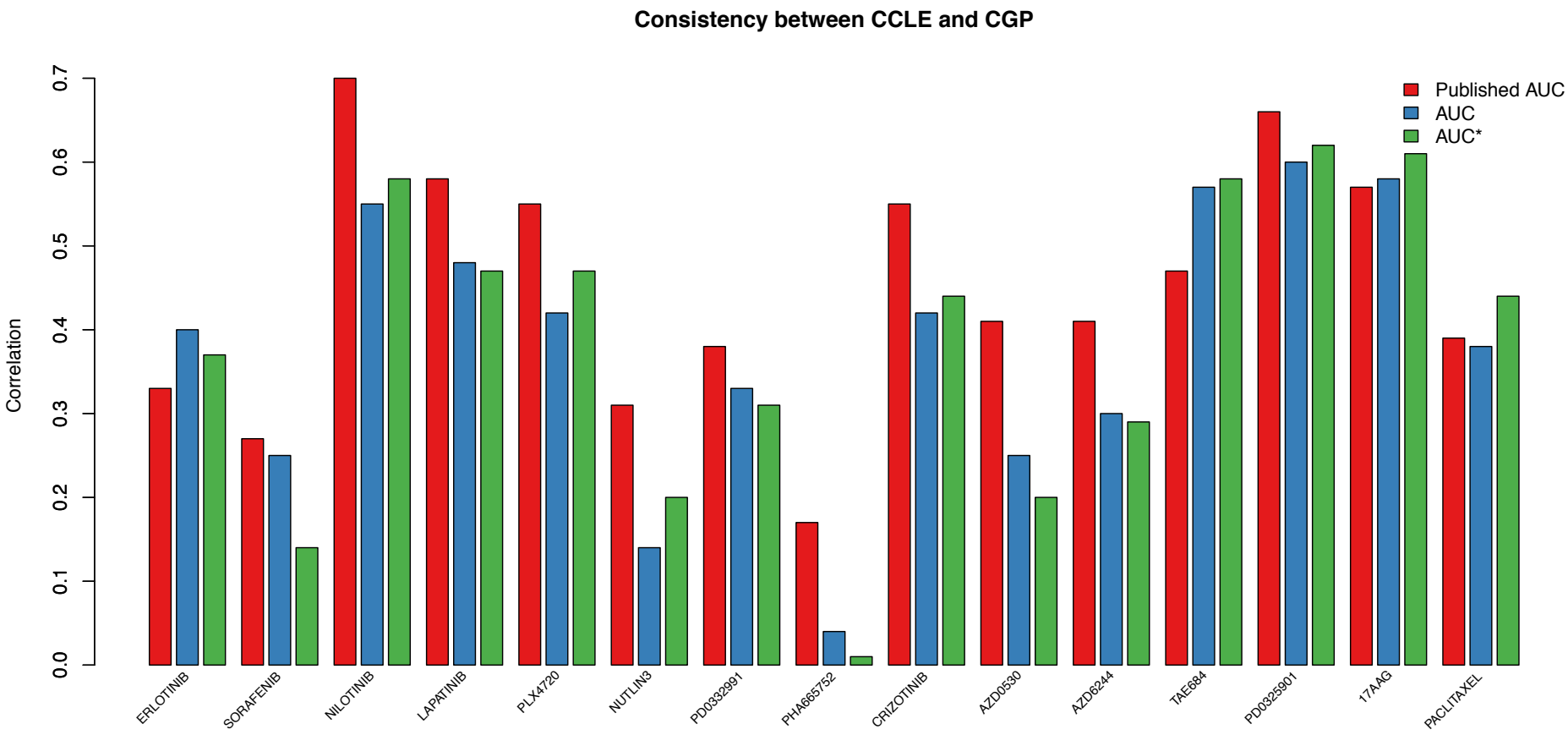
Summarization of drug dose-response curves

- CGP and CCLE used different statistics for drug sensitivity
- We went back to the raw cell viability data to re-estimate AUC using the standard trapezoidal rule



Zhaleh Safikhani

Homogeneous summarization does not help



Zhaleh Safikhani

Conclusions

Pharmacogenomics

- Pharmacogenomics holds great potential for
  - Discovering new targeted therapies
  - Defining drug mechanisms of action
  - Identifying robust biomarkers of drug response

- In this course we focused on “drug sensitivity” datasets as opposed to other pharmacogenomic datasets such as
  - Drug perturbations, aka Connectivity Map (Lamb et al. Science 2006)
  - Pharmacokinetic (PK) data to determine the fate of substances administered externally to a living organism
  - Pharmacodynamic (PD) data to study effects of drugs on the body and the mechanisms of drug action and the relationship between drug concentration and effect.

Take home messages

- Prediction of drug response is challenging
- Not much success to date
  - Even in simple models such as cancer cell lines
  - It will only be more difficult *in vivo*!
- Potential solutions:
  - Improve reproducibility of cell-based HTS studies
  - Perform meta-analysis to extract the consistent signal from the data (*PharmacoGx* coming soon; Petr Smirnov)
- This is just the beginning…
  - How to deal with tumor heterogeneity?
  - How to design effective drug combinations?

Mining existing pharmacogenomic data

Cancer Genome Project

Pfizer

Genentech

S2C

gsk

CCLE Cancer Cell Line Encyclopedia

NCI60

Courtesy of Mathieu Lupien

Acknowledgements

From the lab

- Zhaleh Safikhani
- Petr Smirnov
- Rene Quevedo
- Nehme El-Hachem
- Donald Wang
- Simon Papillon-Cavanagh
- Nicolas de Jay
- Adrian She
- Deena Gendoo

IRCM

- Jacques Archambault

DANA-FARBER CANCER INSTITUTE

- Hugo Aerts
- John Quackenbush

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

- Andrew Beck

FUDAN UNIVERSITY 1905

- Leming Shi

Yale University School of Medicine

- Christos Hatzis

CENTERFORBIOLOGICALSEQUENCEANALYSIS CBS

- Nicolai Juul Birkbak

- Carl Virtanen

The Princess Margaret Cancer Foundation UHN

Thank you for your attention!