Statistics and machine learning on pharmacogenomics data

Zhaleh Safikhani

zhaleh.safikhani@cyclicarx.com

@zhaleh.julie

Machine Learning Scientist Team Lead at Cyclica
PostGraduate Fellow at Vector institute

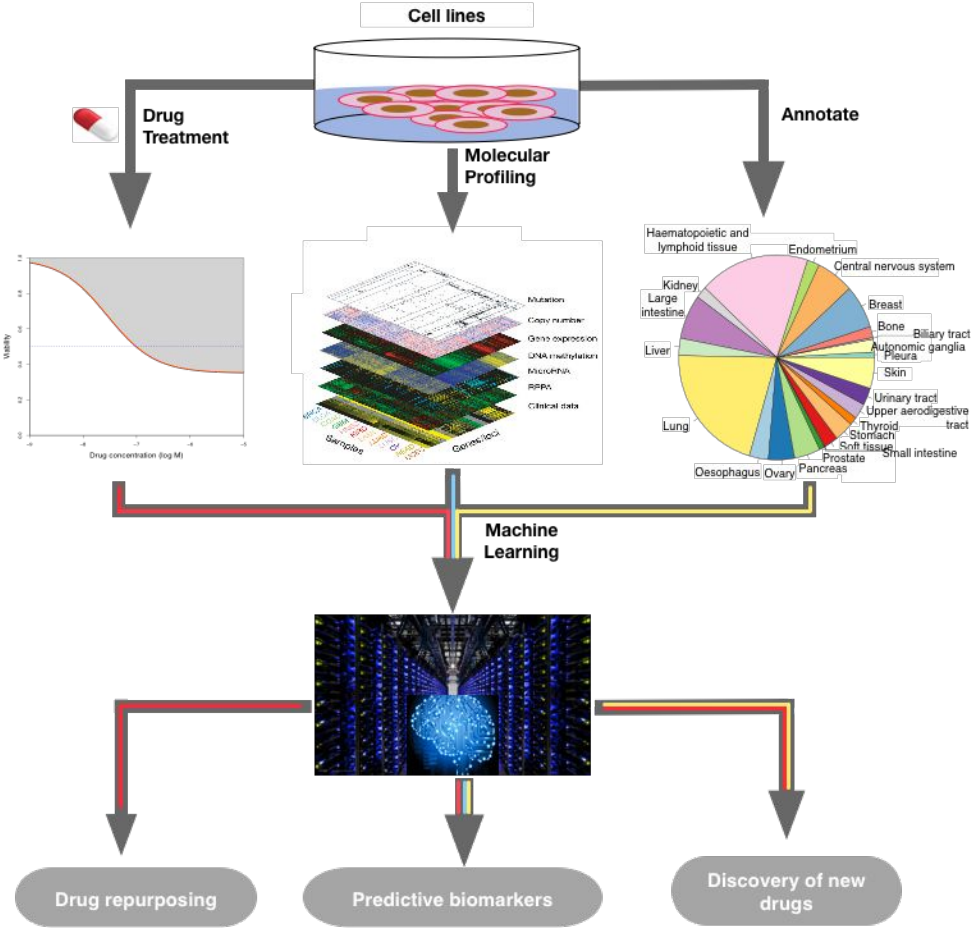
July 21, 2019

ISMB/ECCB 2019

Outline

1. Evaluating reproducibility and handling noise in pharmacogenomics data
2. Meta-analysis across studies
3. Applications of machine learning for drug ranking and predictive modeling

Modeling drug-response data



The PharmacoGx toolbox

CCLE	CTRPv2	gCSI	GDSC1000	FIMM	GRAY	UHNBreast
Mar 2012	Sep 2015	May 2016	June 2016	June 2016	Oct 2017	Ongoing
1061 cell lines 24 drugs	860 cell lines 481 drugs	59 cell lines 16 drugs	1124 cell lines 256 drugs	50 cell lines 52 drugs	71 cell lines 104 drugs	84 cell lines 21+ drugs

Cell Lines	Tissues	Compounds	Dose Response Experiments	Gene-Drug Associations
1,691	**41**	**759**	**650,894**	**200 Million+**

Software package

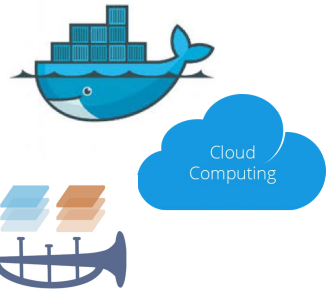
PharmacoGx: an R package for analysis of large pharmacogenomic datasets

Petr Smirnov, Zhaleh Safikhani, Nehme El-Hachem, Dong Wang, Adrian She, Catharina Olsen, Mark Freeman, Heather Selby, Deena M.A. Gendoo, Patrick Grossmann ... Show more

Author Notes

Bioinformatics, Volume 32, Issue 8, 15 April 2016, Pages 1244–1246, https://doi.org/10.1093/bioinformatics/btv723

Software environment



ORCESTRA
(*Fall 2019*)

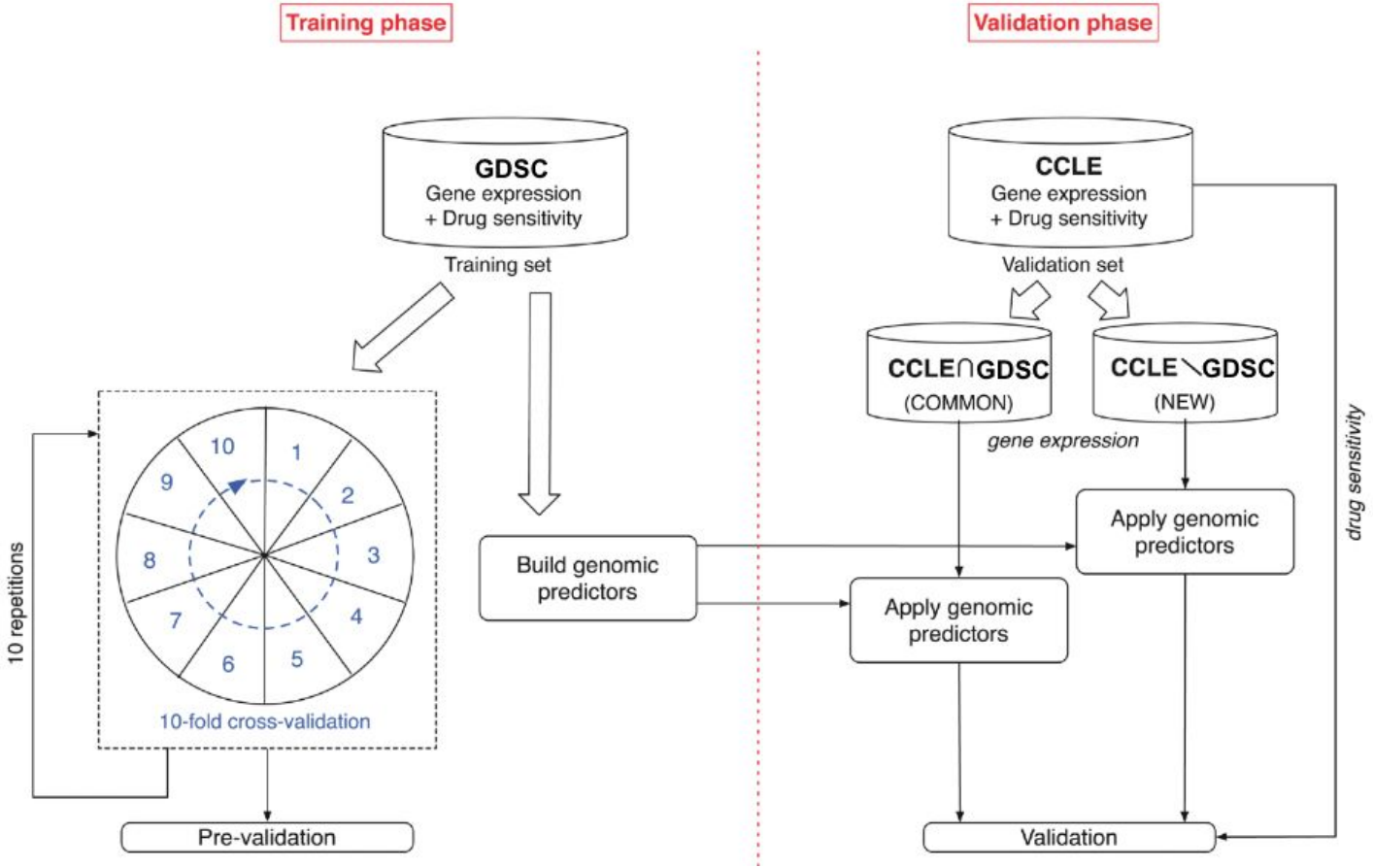
Web-application

PharmacoDB: an integrative database for mining *in vitro* anticancer drug screening studies

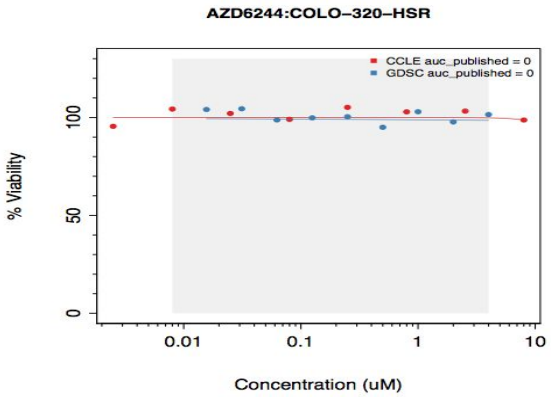
Petr Smirnov, Victor Kofia, Alexander Maru, Mark Freeman, Chantal Ho, Nehme El-Hachem, George-Alexandru Adam, Wail Ba-alawi, Zhaleh Safikhani, Benjamin Haibe-Kains ✉

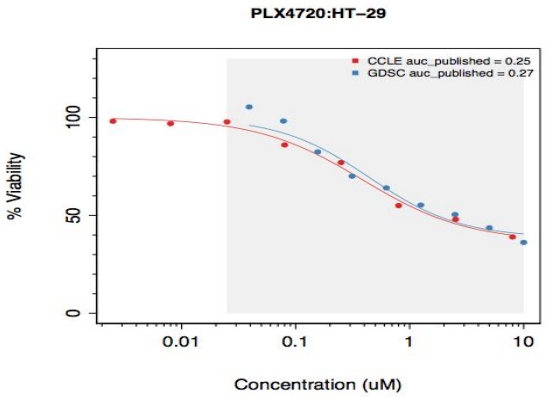
Nucleic Acids Research, Volume 46, Issue D1, 4 January 2018, Pages D994–D1002, https://doi.org/10.1093/nar/gkx911

Training and testing of biomarkers predictors



Challenges of Assessment

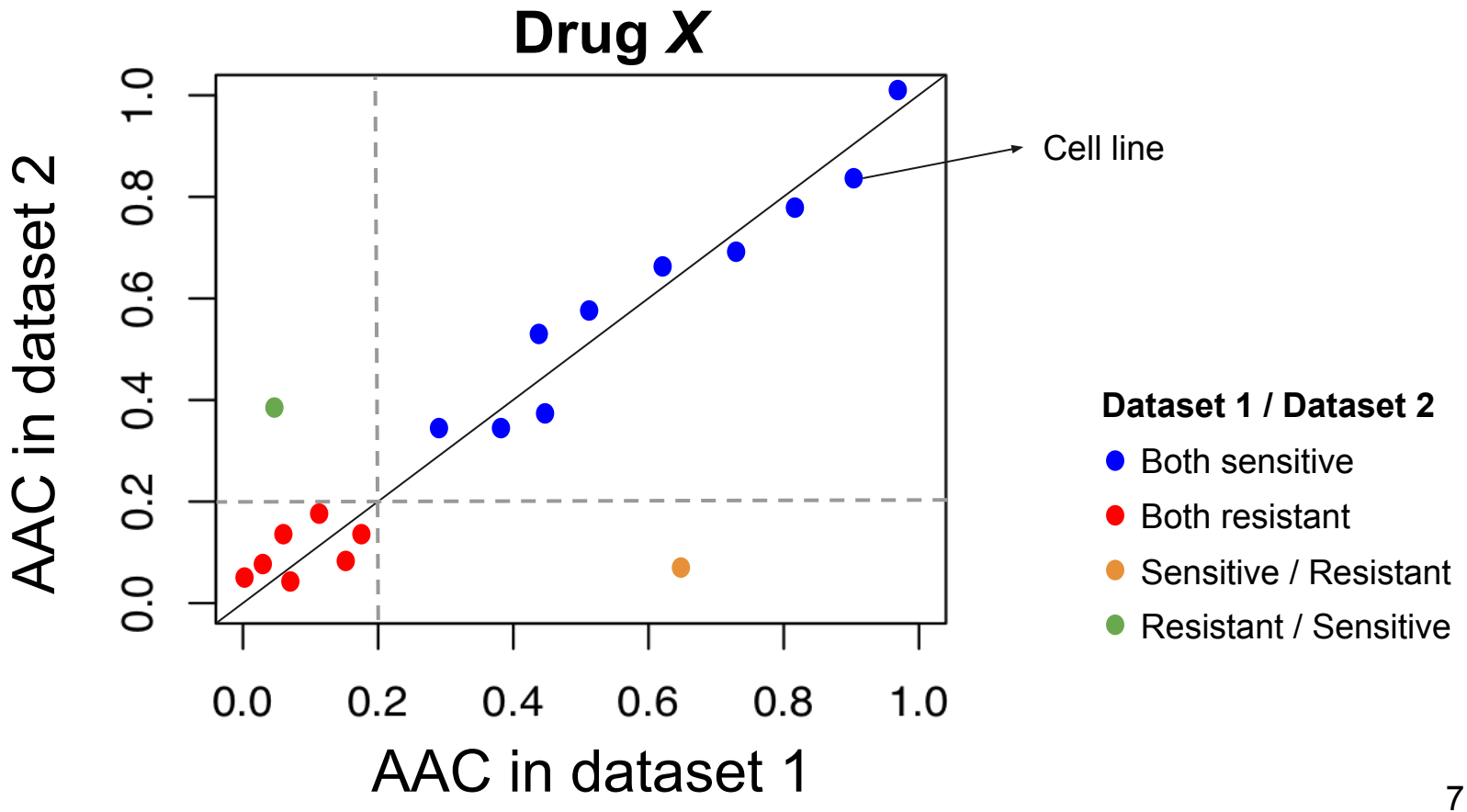




Highly consistent

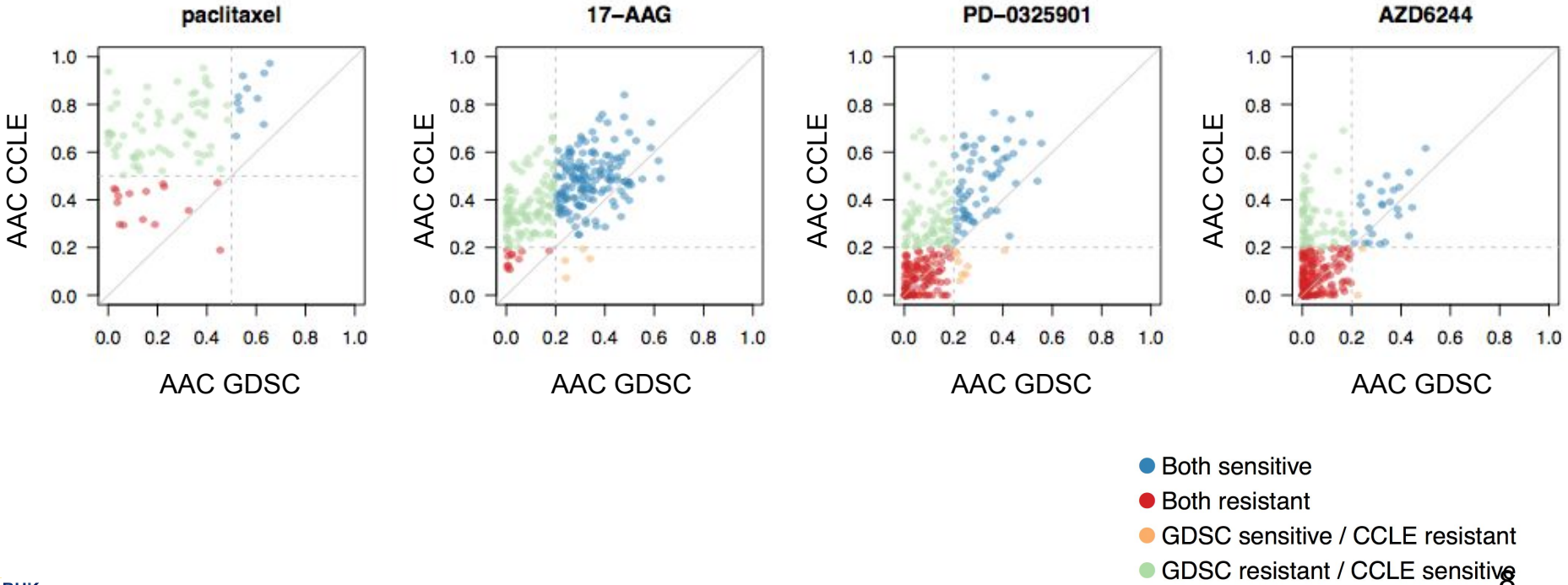
Consistency of drug response across datasets

Ideal case...



Drugs with “broad” effects

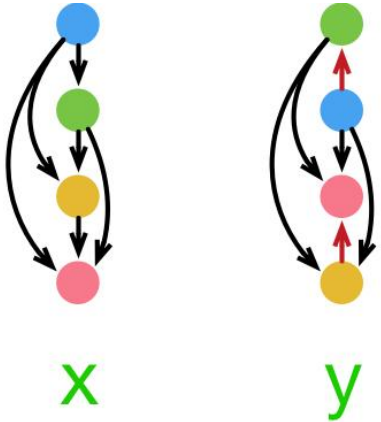
Broad effects = high variance in drug response



Concordance Index

Concordance index (CI) is a generalization of the AUROC

- Comparison of all pairs of cell lines



Concordant pairs = 4

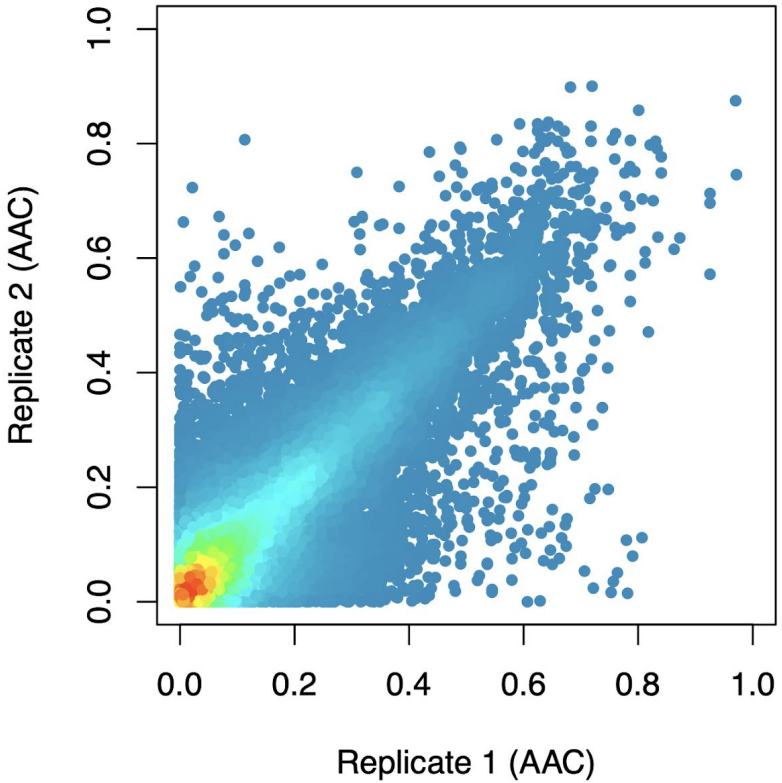
Discordant pairs=2

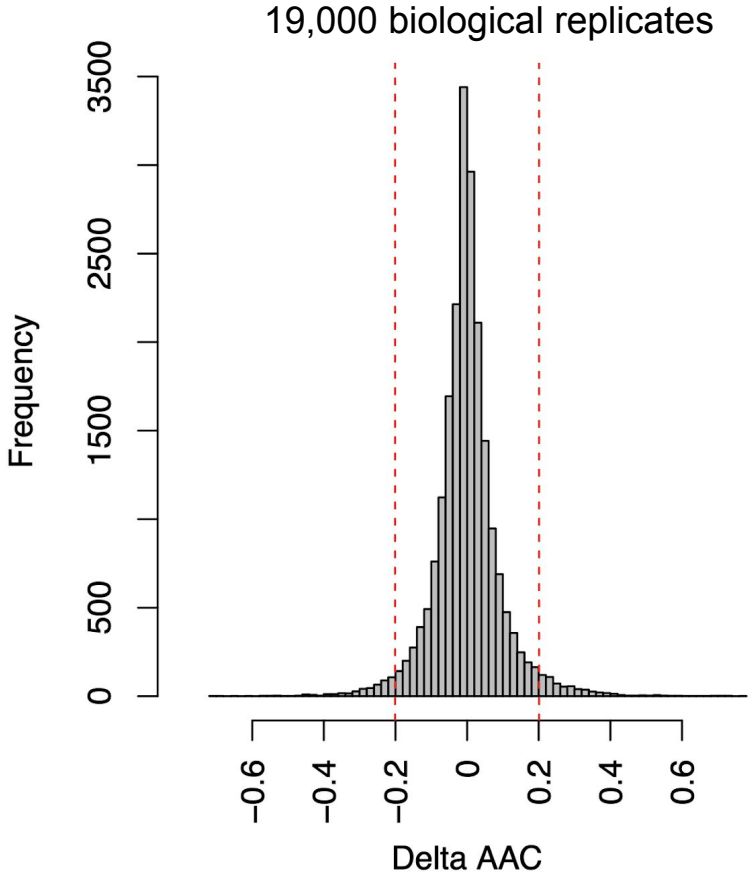
Concordance index= 4/(4+2)=0.67

- The **concordance index (CI)** is the probability that two randomly-chosen cell lines are ***ranked*** identically by two assays
- Interpretable scale: [0, 1]
  - CI = 0 denotes a perfect *inverse* consistency
  - CI = 0.5 denotes a random association (no consistency)
  - CI = 1 denotes a perfect consistency

Drug response is noisy

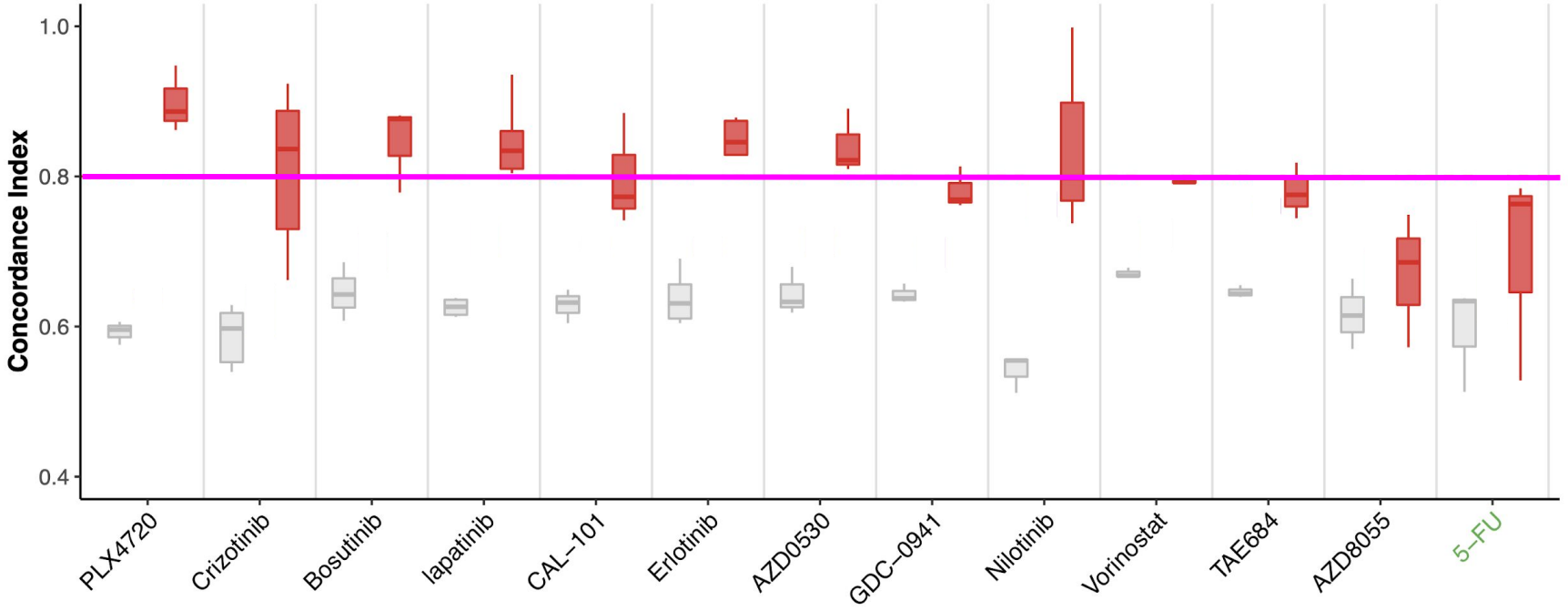
robust Concordance index = 0.92





Concordance across multiple datasets

[CCLE, CTRPv2, gCSI, GDSC1000]

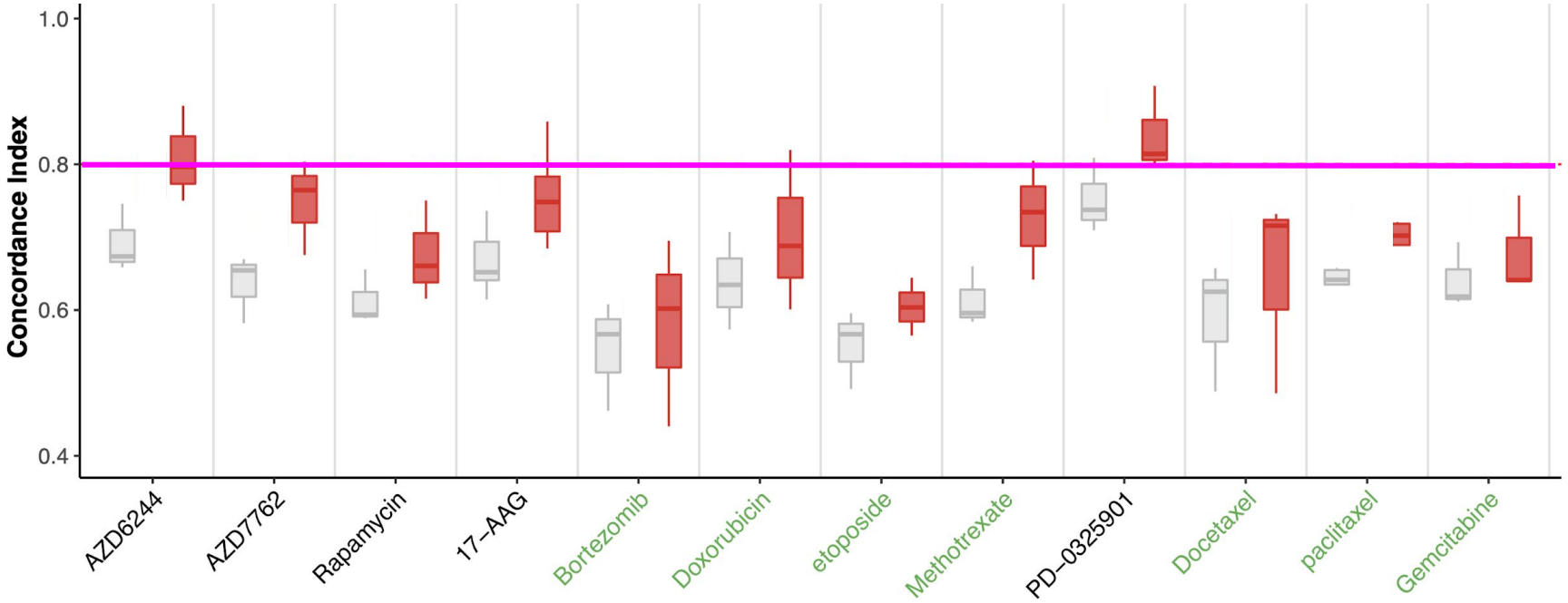


Targeted therapies
Chemotherapies

CI rCI

Concordance across multiple datasets

[CCLE, CTRPv2, gCSI, GDSC1000]



Targeted therapies
Chemotherapies

CI rCI

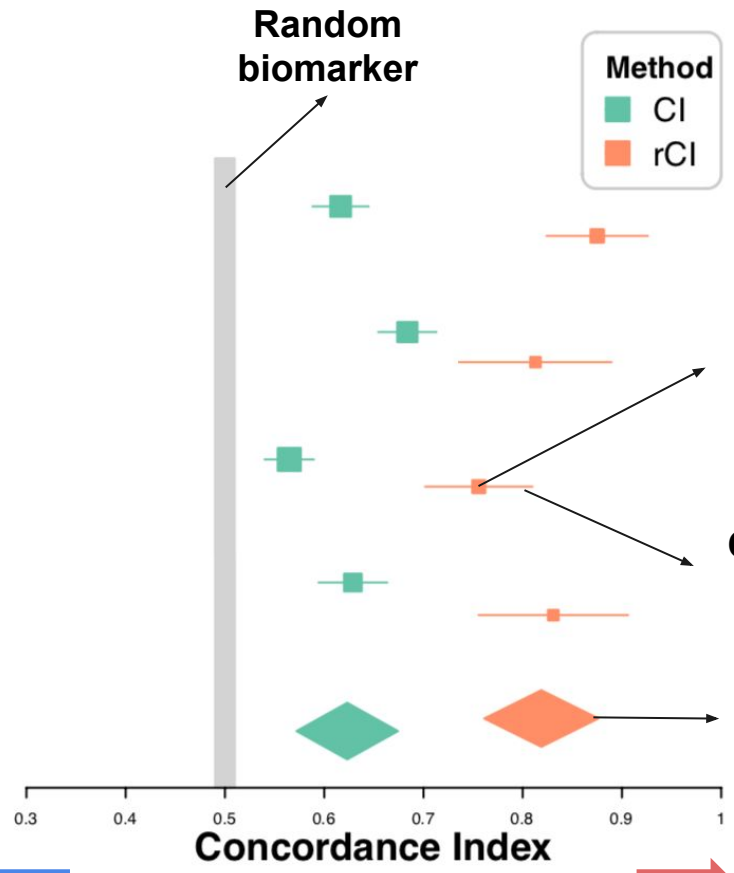
BHK LAB

Meta-analysis + forestplot

Dataset	CI (p)	rCI (p)
CCLE	4.4e−16	<1e−16
gCSI	<1e−16	1.1e−15
CTRPv2	2.3e−07	<1e−16
GDSC1000	3.2e−13	<1e−16
Meta analysis	2.4e−06	<1e−16

Dataset

P-value (significance)



Estimate (effect size)

Confidence interval

Meta-estimate (consensus)

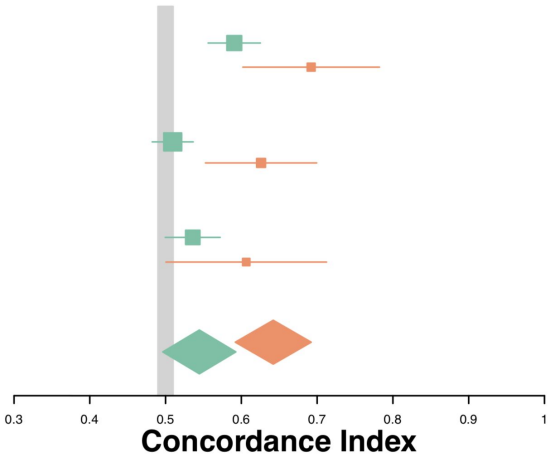
Predictive of resistance

Predictive of sensitivity

EGFR vs Erlotinib

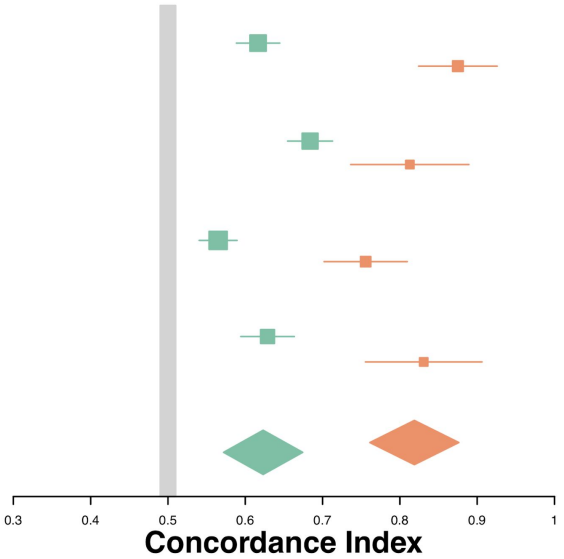
Copy Number Variation

Dataset	CI (p)	rCI (p)
CCLE	_	_
gCSI	2.5e−07	2.9e−05
CTRPv2	4.9e−01	7.6e−04
GDSC1000	5.2e−02	4.9e−02
Meta analysis	7.0e−02	2.7e−08



mRNA Expression

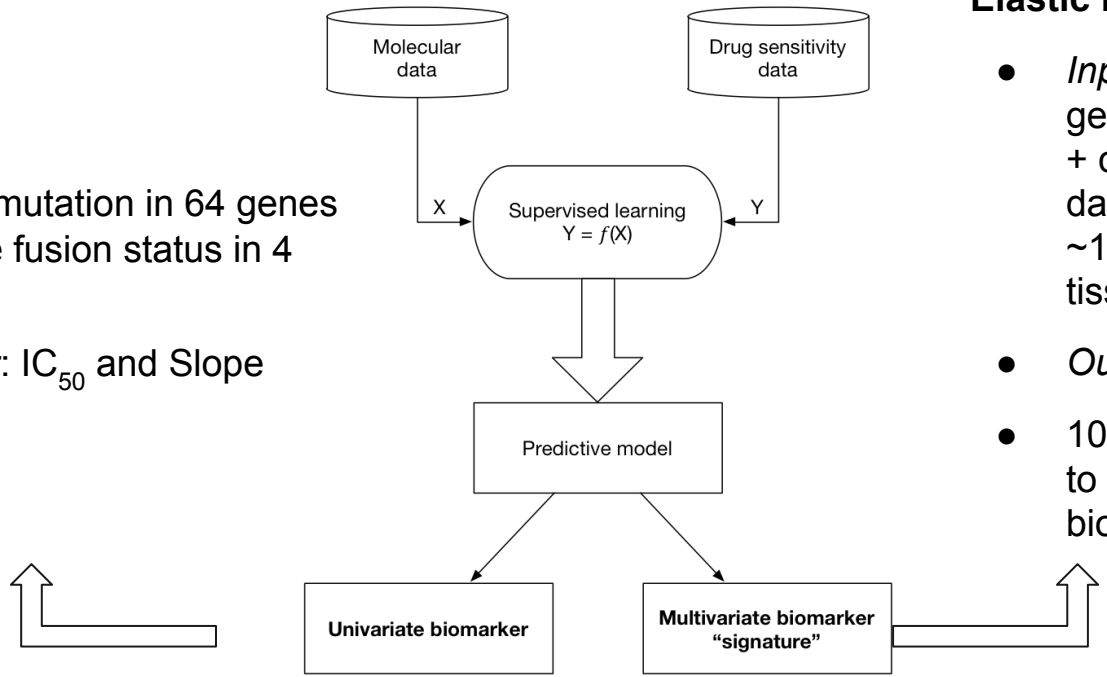
CI (p)	rCI (p)
4.4e−16	<1e−16
<1e−16	1.1e−15
2.3e−07	<1e−16
3.2e−13	<1e−16
2.4e−06	<1e−16



Multivariate models (GDSC)

MANOVA

- *Input*: mutation in 64 genes + gene fusion status in 4 genes
- *Output*: IC_{50} and Slope



Elastic Net

- *Input*: mutation in 64 genes + gene fusion status in 4 genes + continuous copy number data from 426 genes + ~10,000 gene expressions + tissue type
- *Output*: IC_{50}
- 100 x 10-fold cross-validation to assess stability of biomarkers

ARTICLE

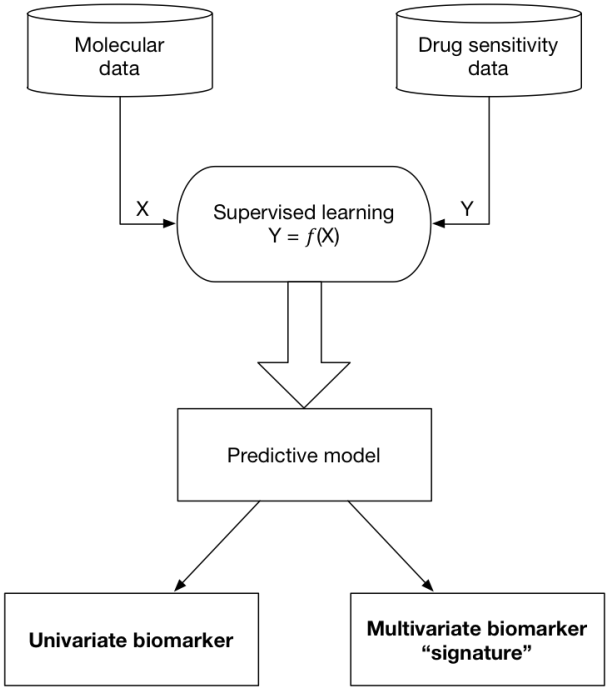
doi:10.1038/nature11005

Systematic identification of genomic markers of drug sensitivity in cancer cells

Mathew J. Garnett[1]*, Elena J. Edelman[2]*, Sonja J. Heidorn[1]*, Chris D. Greenman[1]†, Anahita Dastur[2], King Wai Lau[1], Patricia Greninger[2], I. Richard Thompson[1], Xi Luo[2], Jorge Soares[1], Qingsong Liu[1,3], Francesco Iorio[1,3], Didier Surdez[4], Li Chen[2], Randy J. Milano[2], Graham R. Bignell[1], Ah T. Tam[2], Helen Davies[1], Jesse A. Stevenson[2], Syd Barthorpe[1], Stephen R. Lutz[2],

BHK LAB

Multivariate models (CCLE)



Elastic Net

- *Input*: 50,000 features (mutations + CNV + expressions) across and within tissue types
- *Output*: log IC_{50}, A_{max}, AUC
- 10 x 10-fold cross-validation + bootstrap to assess stability of biomarkers

Naive Bayes classifier with discretized drug sensitivities ...

LETTER

doi:10.1038/nature11003

The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity

Jordi Barretina[1,2,3]†*, Giordano Caponigro[1]*, Nicolas Stransky[1]*, Kavitha Venkatesan[1]*, Adam A. Margolin[1]†*, Sungjoon Kim[2], Christopher J. Wilson[4], Joseph Lehár[4], Gregory V. Kryukov[1], Dmitriy Sonkin[4], Anupama Reddy[4], Manway Liu[4], Lauren Murray[1], Michael F. Berger[1]†, John E. Monahan[1], Paula Morais[1], Jodi Meltzer[1], Adam Korejwa[1], Judit Jané-Valbuena[1,2], Felipa A. Mapa[4], Joseph Thibault[2], Eva Bric-Furlong[4], Pichai Raman[4], Aaron Shipway[2], Ingo H. Engels[2], Jill Cheng[2], Guoying K. Yu[2], Jianjun Yu[2], Peter Aspesi Jr[2], Melanie de Silva[2], Kalpana Jagtap[2], Michael D. Jones[2], Li Wang[2], Charles Hatton[2], Emanuele Palescandolo[2], Supriya Gupta[1], Scott Mahan[1], Carrie Sougnez[1], Robert C. Onofrio[1], Ted Liefeld[1], Laura MacConaill[3], Wendy Winckler[1], Michael Reich[1], Nanxin Li[3], Jill P. Mesirov[1], Stacey B. Gabriel[1], Gad Getz[1], Kristin Ardlie[1], Vivien Chan[4], Vic E. Myer[4], Barbara L. Weber[4], Jeff Porter[4], Markus Warmuth[5], Peter Finan[4], Jennifer L. Harris[2], Matthew Meyerson[1,2,3], Todd R. Golub[1,3,7,8], Michael P. Morrissey[4]*, William R. Sellers[4]* & Levi A. Garraway[1,2,3]*

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**: molecular profiles of 18 breast cancer cell lines
- **Gold standard**: drug sensitivity of these 18 cell lines to the 28 drugs

NCI-DREAM: Top predictor

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

Applications

Exploring new classes of biomarkers

- Changes in alternative splicing of mRNA associated with cancer hallmarks
- RNA-seq enables quantification of isoform-specific expression
  - Recent release of RNA seq profiles for >1000 cell lines
- Opportunity to investigate the associations between isoform expression and drug sensitivity *in vitro*

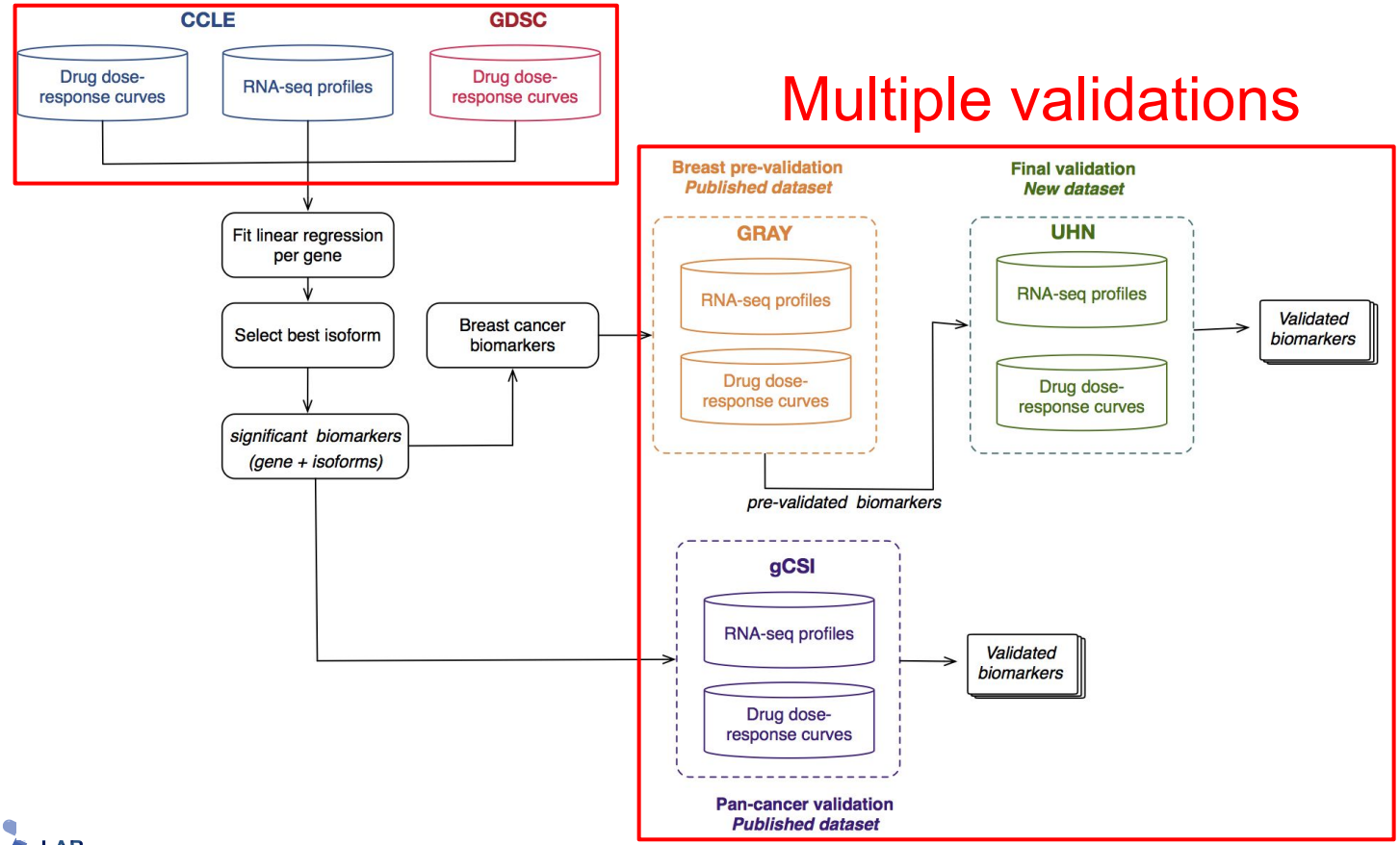
Reproducibility of results and analyses

https://codeocean.com/2018/05/03/gene-isoforms-as-expression-based-biomarkers-predictive-of-drug-response-in-vitro-lsqb-pmid-colon-29066719-rsqb/code

Gene isoforms as expression-based biomarkers predictive of drug response in vitro

Zhaleh Safikhani[1,2], Petr Smirnov[1], Kelsie L. Thu[1,3], Jennifer Silvester[1,3], Nehme El-Hachem[3], Rene Quevedo[1,2], Mathieu Lupien[1,2], Tak W. Mak[1,2,4], David Cescon[1,4,5] & Benjamin Haibe-Kains[1,2,6,7]

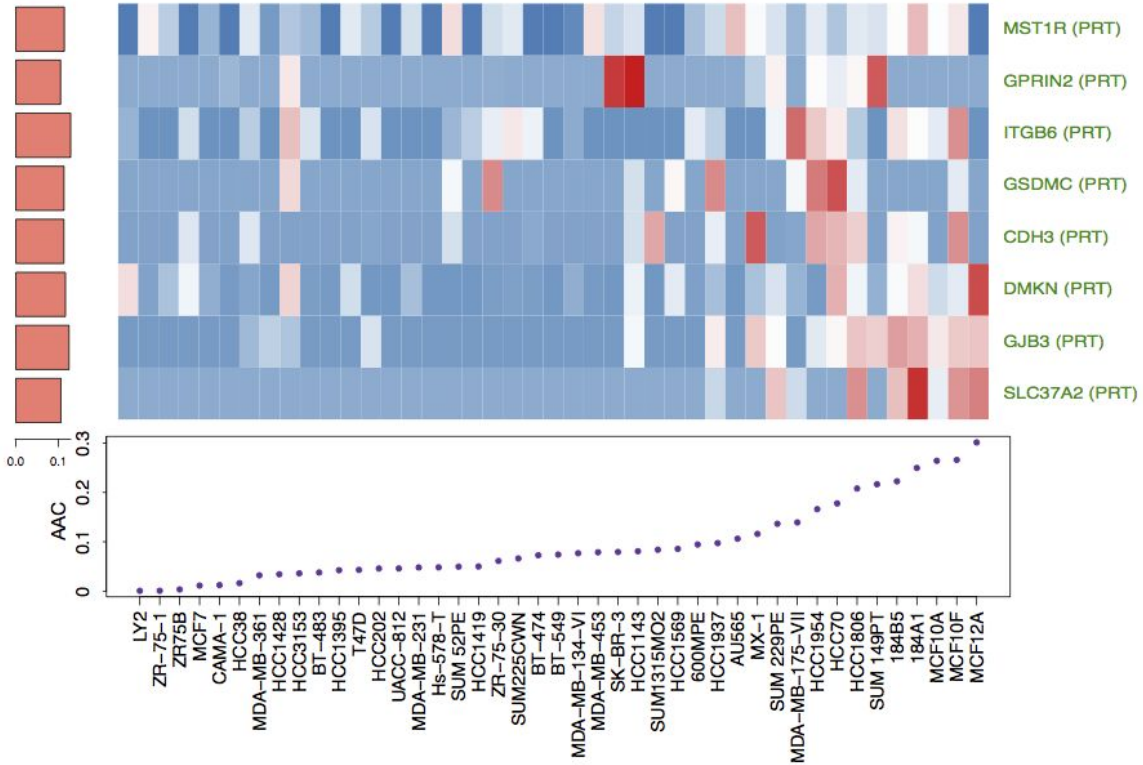
Meta analysis



Validation in breast cancer (GRAY)

Isoform-specific

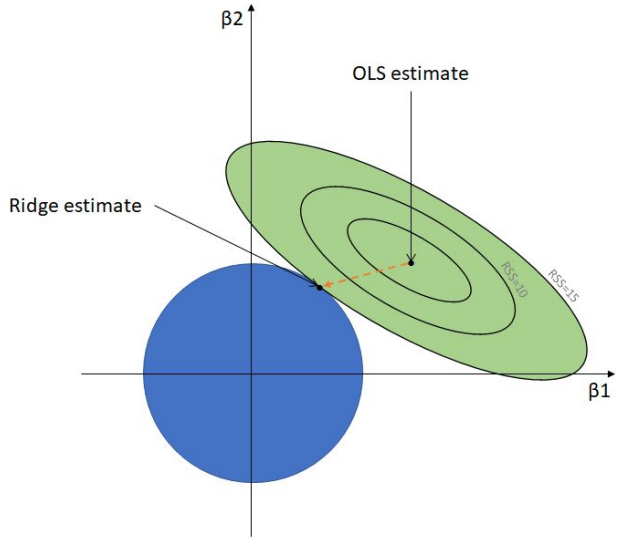
Erlotinib



Multivariate model building

- mRMR feature selection (Minimum Redundancy Maximum Relevance)

- Ridge Regression
- Cross validation and Shuffling



$$\hat{\beta}^{ridge} = \underset{\beta \in \mathbb{R}}{argmin} \|y - XB\|_2^2 + \lambda \|B\|_2^2$$

BHK LAB

Performance assessment in cross validation