CRCM
Centre de Recherche
en Cancérologie de Marseille

Precision and recall oncology: combining multiple gene mutations for improved identification of drug-sensitive tumours

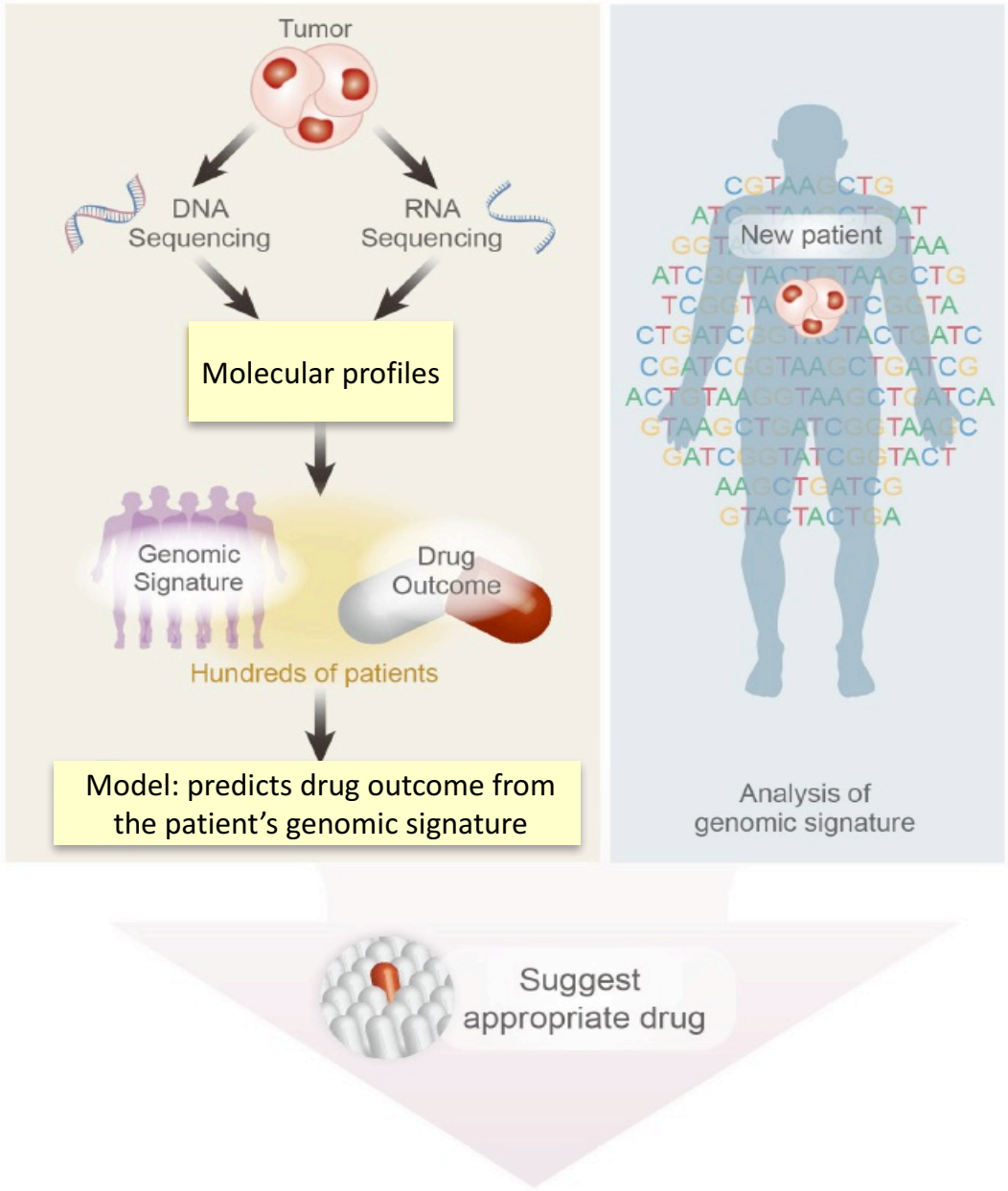
Dr Pedro Ballester

INSERM Group Leader
Cancer Research Centre of Marseille, France

pedro.ballester@inserm.fr
twitter.com/pjballester
linkedin.com/in/pedroballester/

ITMO Cancer session – CIRM conference (Marseille, France) – July 2018

Precision oncology: marker discovery

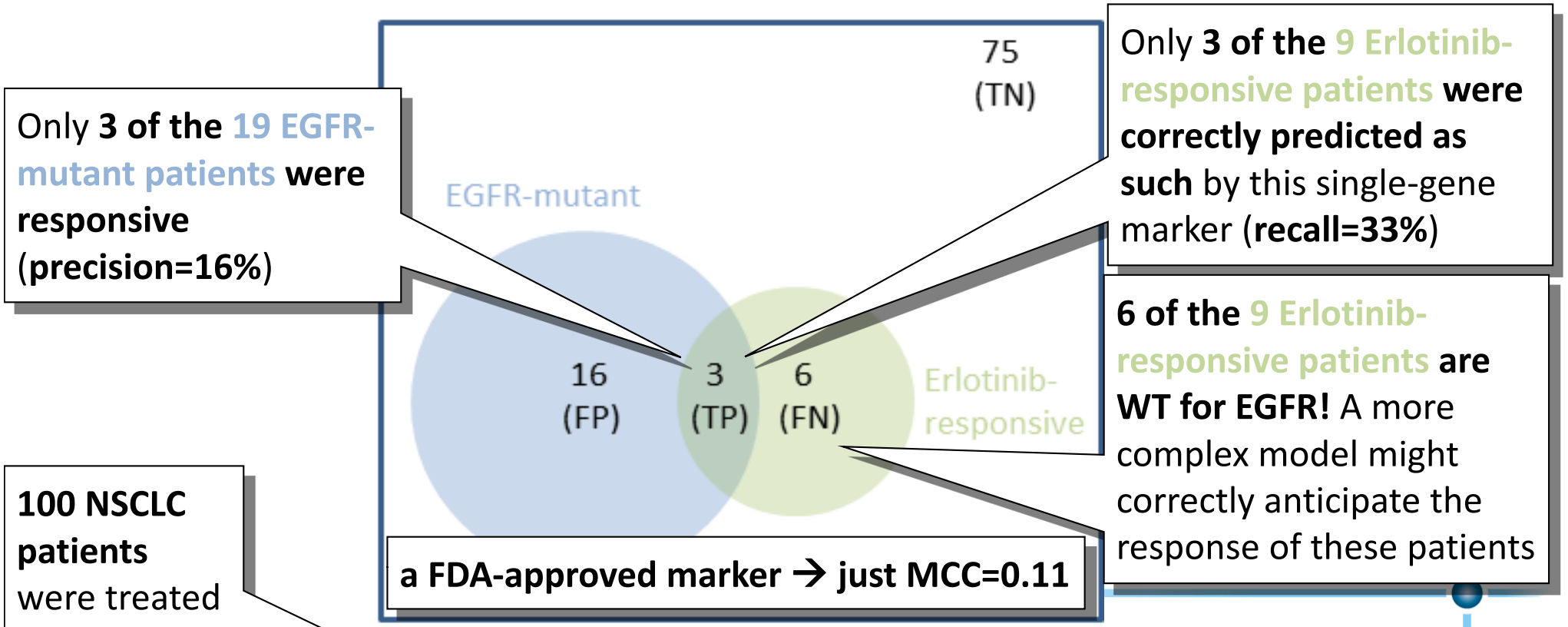


- The right treatment for each patient? → predict which patients respond to a drug
- Model: predict drug response from a molecular profile of the patient's tumour (e.g. a single-nucleotide variant or SNV)
- Most studies, employ 1D model: a given tumour SNV as a single-gene marker (the actionable mutation)

Cho et al. *Mol. Cells* 2016; 39(2): 77-86

Actionable mutations: important limitations

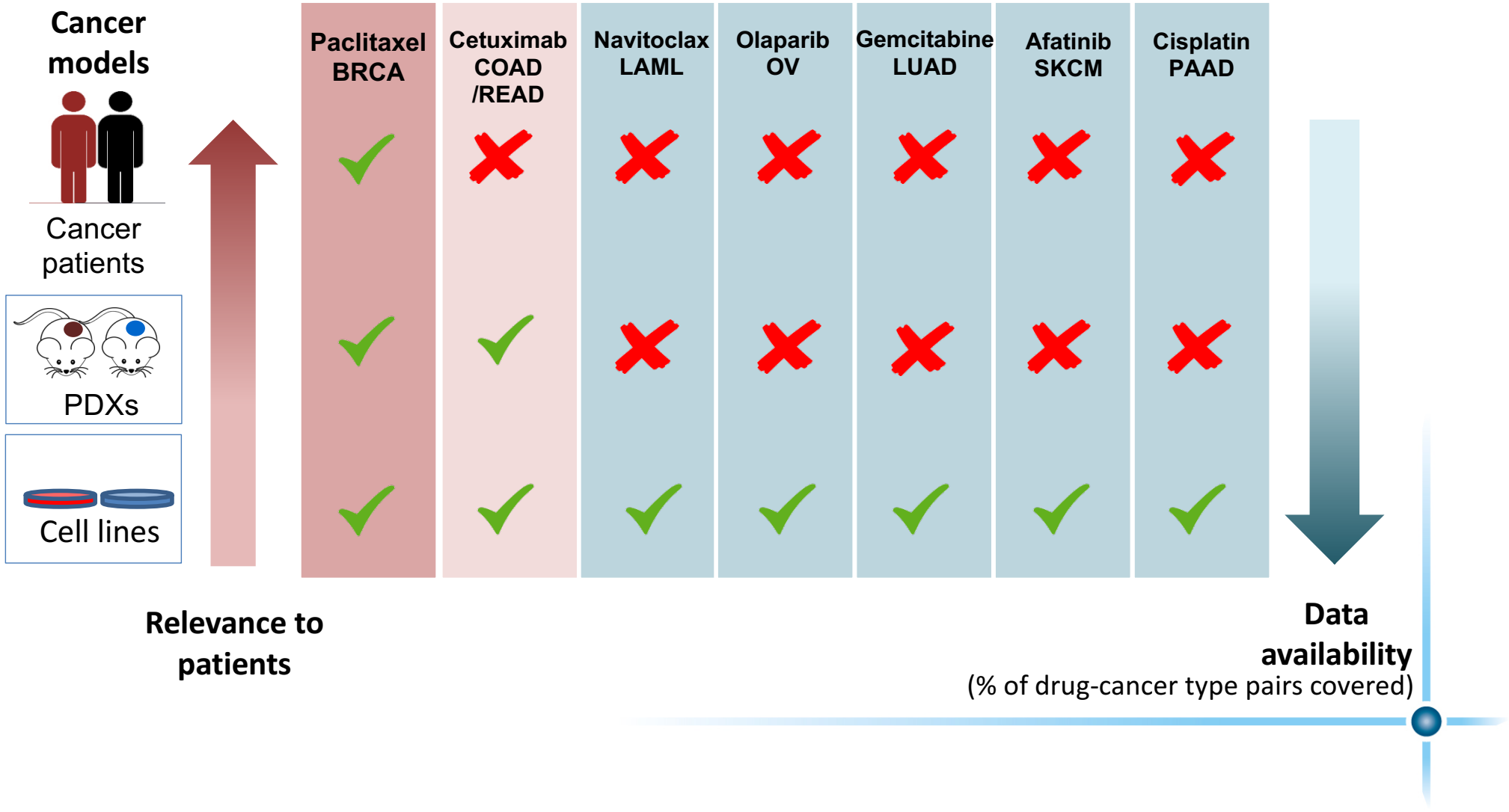
1. Single-gene markers have only been found for a few binomials drug-cancer type (e.g. Erlotinib-NSCLC) → **Few patients benefit**
2. Even when found (e.g. the EGFR^{L858R} SNV for Erlotinib-NSCLC) → simple 1D model usually **modest prediction** of drug response



Data from: Tsao, M.-S. et al. *N. Engl. J. Med.* 353, 133–144 (2005).

Complete and curated clinical data is scarce

Many drug-cancer type pairs lack either the responses of the patients to the drug or the genomic profiles of their tumours → **need preclinical data**



Single- & multi-gene predictors on cell lines

- GDSC: searching for new **single-gene markers**
- Generating new data sets and their systematic analysis:
  - drugs are screened on a large-panel of cancer cell lines
  - a phenotypic readout is made to assess the intrinsic cell sensitivity or resistance to the tested drug
  - a molecular profile of the untreated cell line is determined (e.g. a set of mutations for selected genes)
  - Parametric statistical test to identify significant associations
- **Multi-gene predictors**, as this using multi-task learning:

OPEN ACCESS Freely available online

PLOS ONE

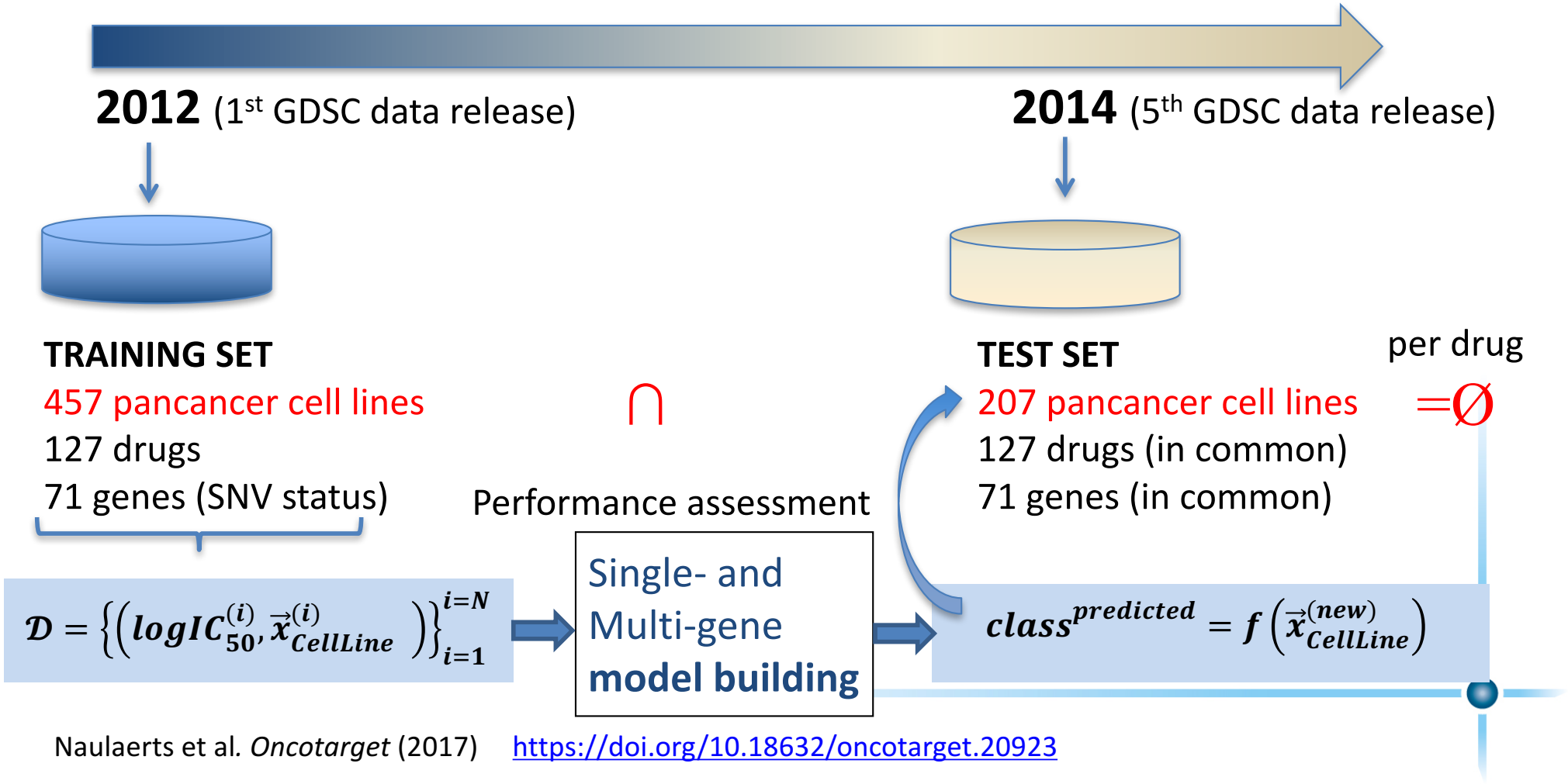
Machine Learning Prediction of Cancer Cell Sensitivity to Drugs Based on Genomic and Chemical Properties

Michael P. Menden[1], Francesco Iorio[1,2], Mathew Garnett[2], Ultan McDermott[2], Cyril H. Benes[3], Pedro J. Ballester[1]*, Julio Saez-Rodriguez[1]*

https://doi.org/10.1371/journal.pone.0061318

Benchmark single- vs multi-gene predictors

Q: Will **combining multiple somatic mutations** result in better **prediction of which cancer cell lines are sensitive to a given drug?**

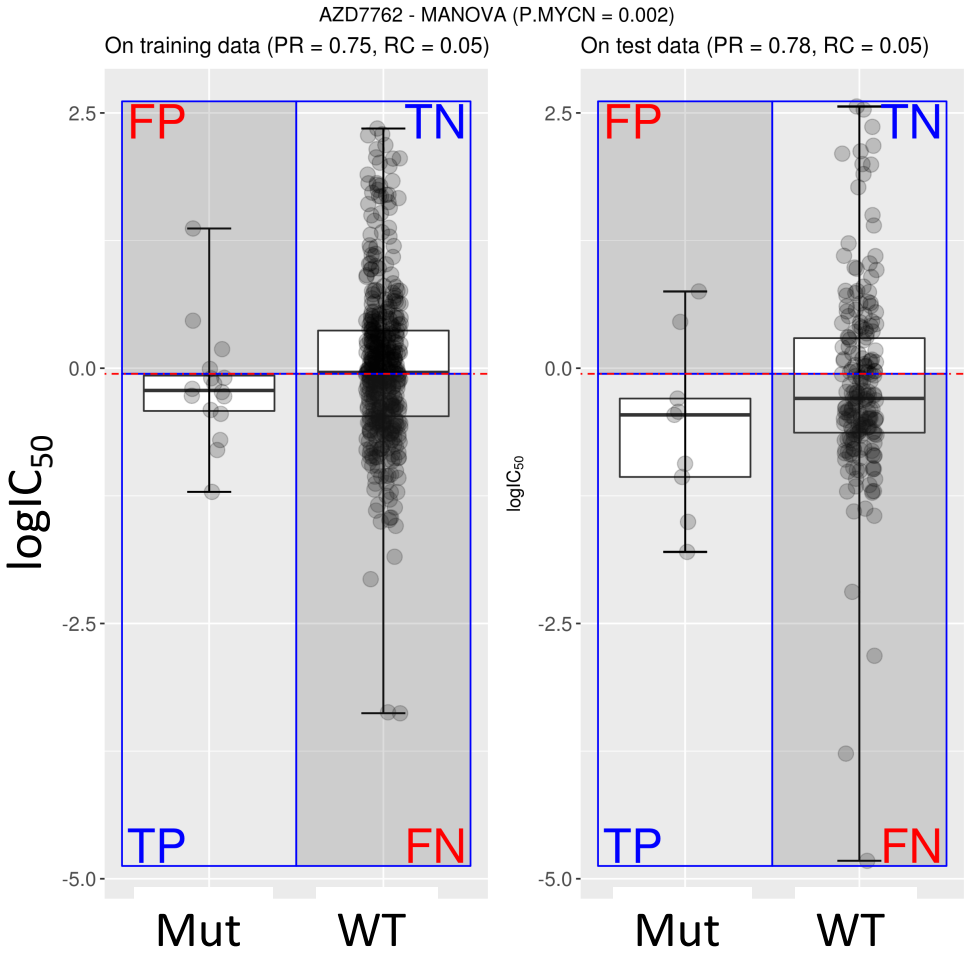


Naulaerts et al. *Oncotarget* (2017) https://doi.org/10.18632/oncotarget.20923

Predictive performance: drug AZD7762

Single-gene
(MANOVA, genomic)

on **training set** | on **test set**



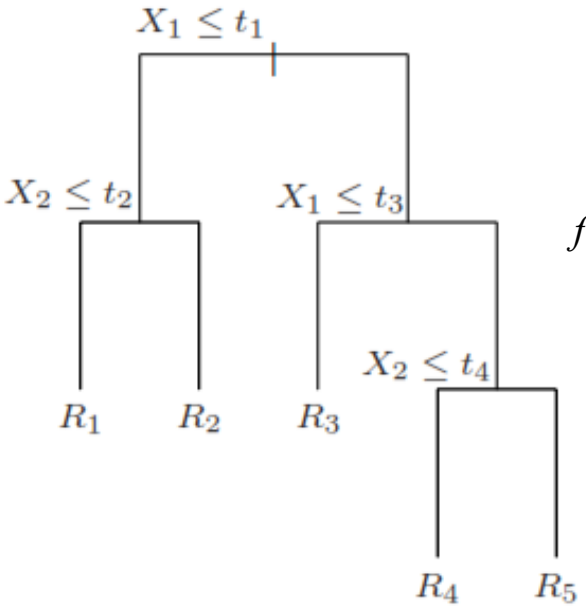
Best single-gene marker from Garnett et al. 2012 Nature:

- MYCN-mutant cell lines are predicted to be sensitive to this drug (P=0.002).
- Sensitivity threshold in red (median IC50 on training set)
- MCC = f(FP, FN, TP, TN)
- Training set: MCC = 0.10
- Test set: MCC = 0.07

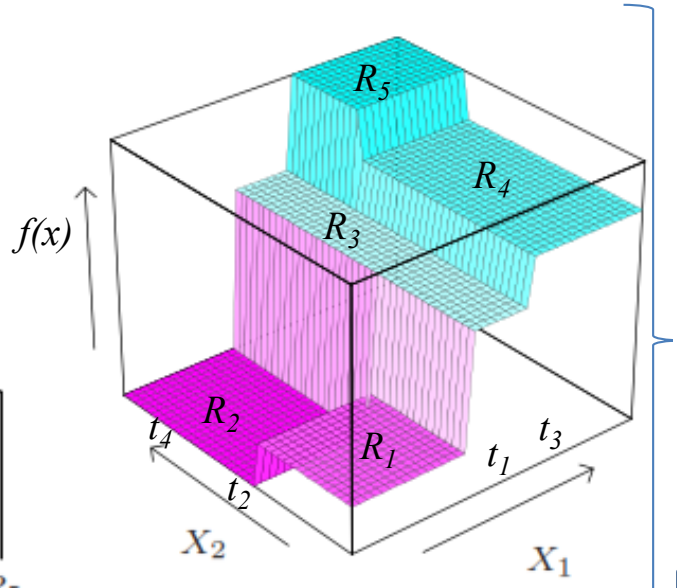
Machine Learning w/ built-in feature selection

Some ML algorithms (e.g. regularisation or tree-based) **discard irrelevant features as a byproduct** ← high-dimensionality

Random Forest (RF) without tuning
(B = 1000 trees, m by 10CV on training set, classification)



$$f(x) = \sum_{m=1} c_m I(x \in R_m)$$



$$\hat{c}_m = \text{ave}(y_i | x_i \in R_m)$$

588 15. Random Forests

Algorithm 15.1 *Random Forest for Regression or Classification.*

1. For $b = 1$ to B:
   (a) Draw a bootstrap sample $\mathbf{Z}^*$ of size N from the training data.
   (b) Grow a random-forest tree T_b to the bootstrapped data, by recursively repeating the following steps for each terminal node of the tree, until the minimum node size n_{min} is reached.
      i. Select m variables at random from the p variables.
      ii. Pick the best variable/split-point among the m.
      iii. Split the node into two daughter nodes.
2. Output the ensemble of trees $\{T_b\}_1^B$.

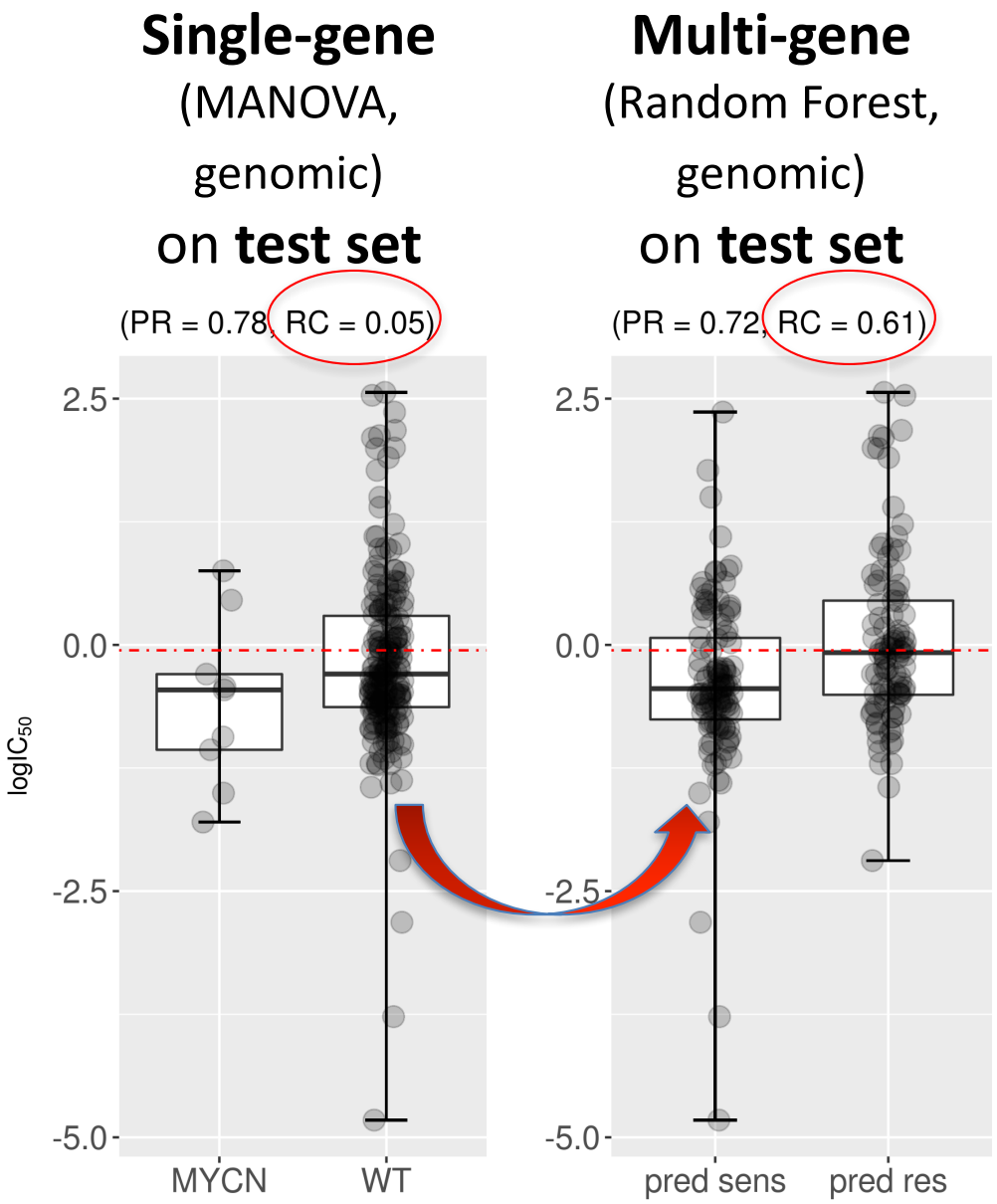
To make a prediction at a new point x:

Regression: $\hat{f}_{\text{rf}}^B(x) = \frac{1}{B}\sum_{b=1}^B T_b(x)$.

Classification: Let $\hat{C}_b(x)$ be the class prediction of the bth random-forest tree. Then $\hat{C}_{\text{rf}}^B(x) = \textit{majority vote}\ \{\hat{C}_b(x)\}_1^B$.

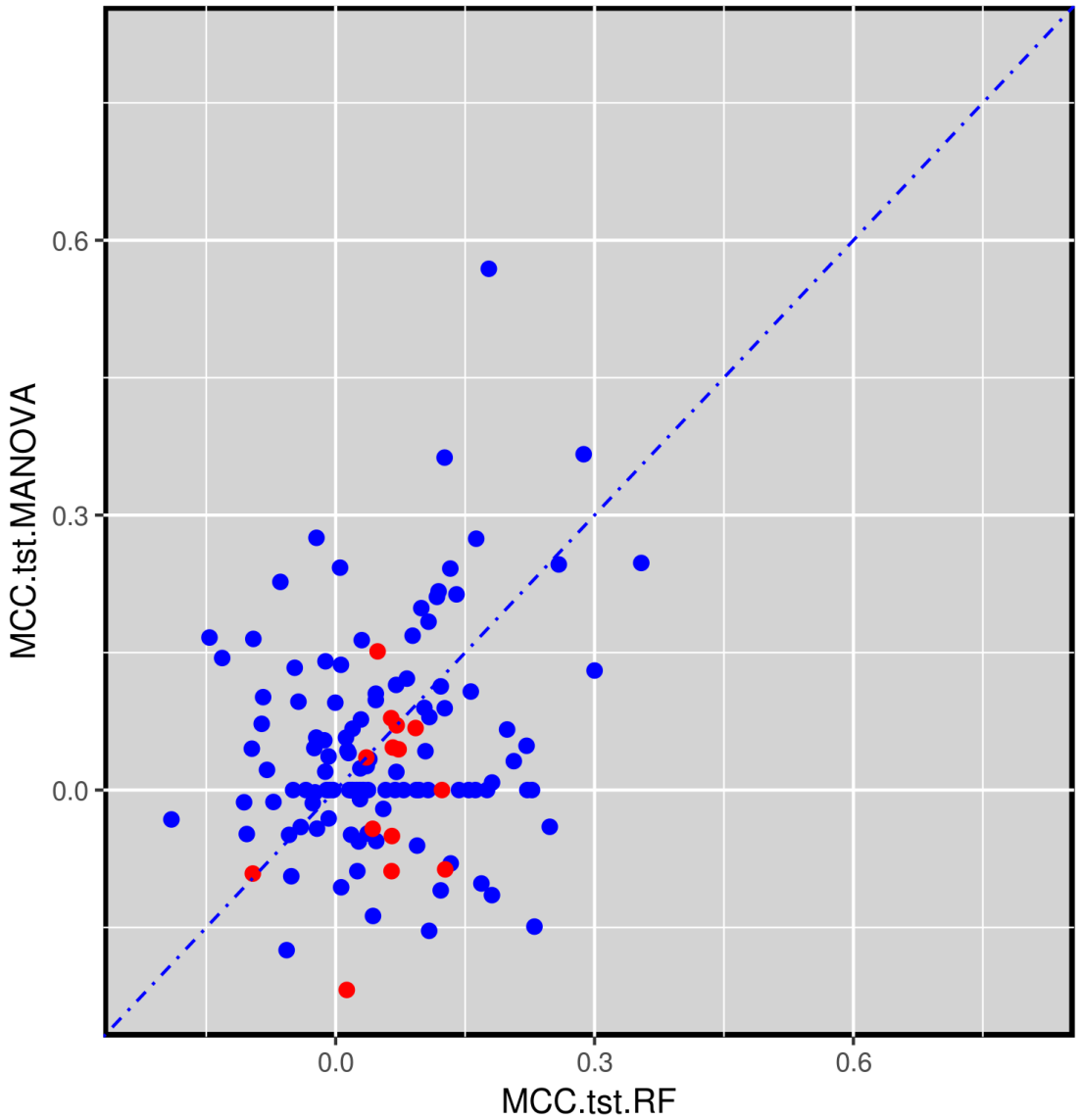
Hastie et al. (2009) "The Elements of Statistical Learning: Data Mining, Inference, and Prediction."

Predictive performance: drug AZD7762



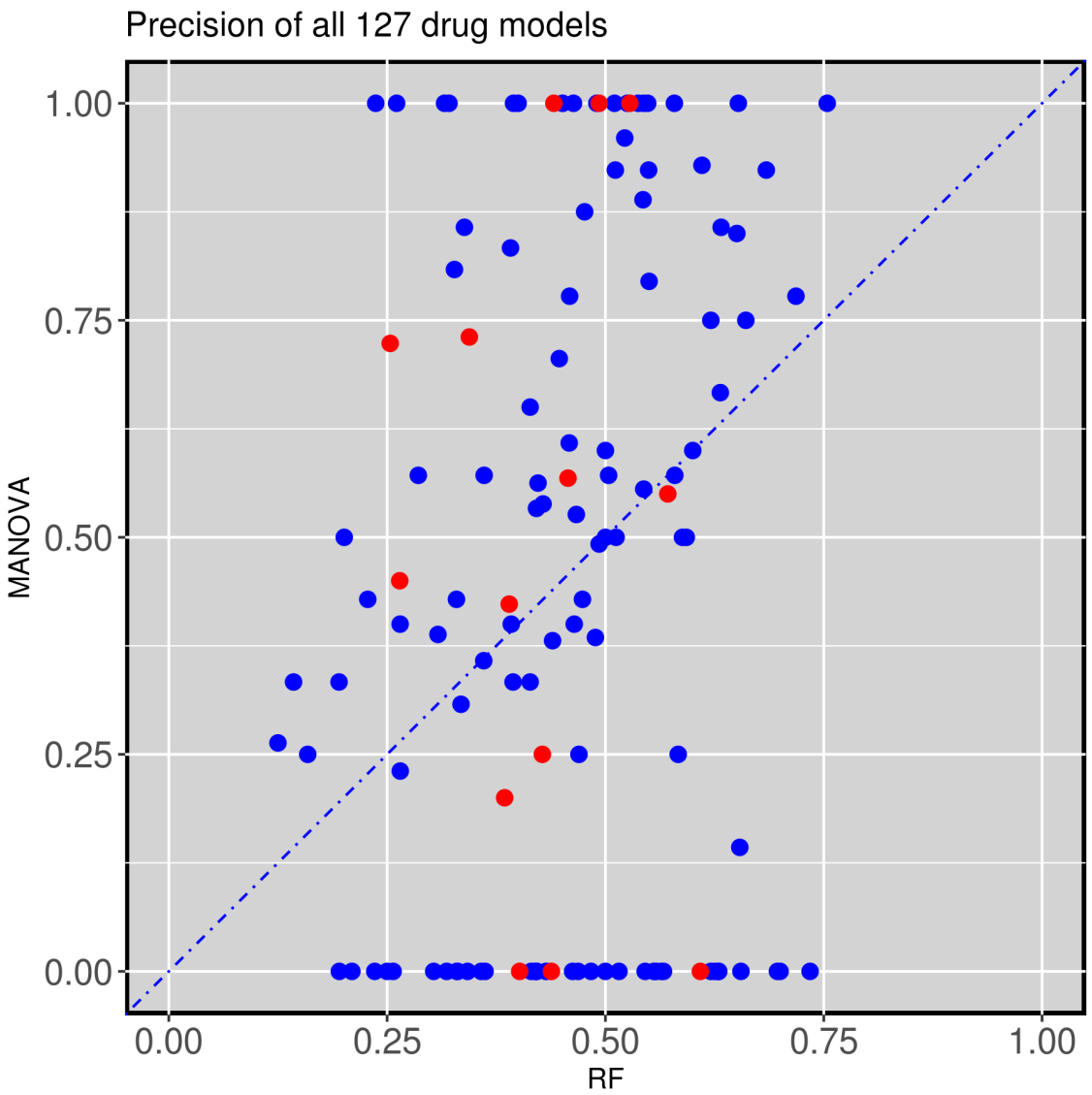
- **=test set (=training set** too)
- For this drug, **multi-gene RF performs 3X better than best single-gene marker** (0.20 vs 0.07 MCC)
- Best marker for this drug (P=0.002) only 0.07 MCC: It is common, **hard problem!**
- Considered **v. good** (PR=0.78)
- Multi-gene: **RC=0.05 → 0.61**

MCC: single-gene vs multi-gene



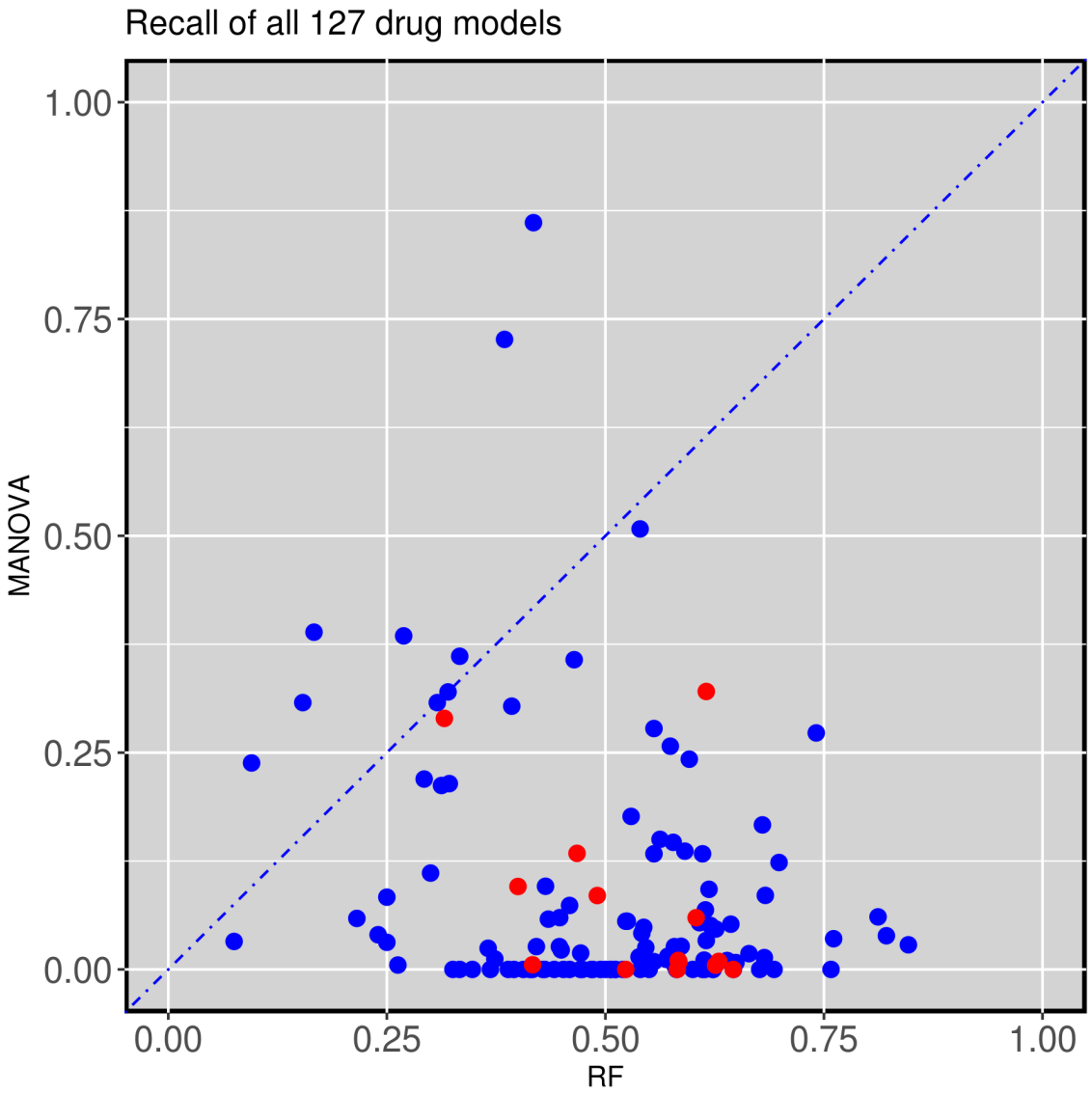
- **MCC**: Matthews Correlation Coefficient
- **Test set MCC** across 127 drugs: **large variability**
- **55% of drugs obtained better MCC** when using multi-gene model
- nine of the 14 cytotoxic drugs (64%) had better MCC by combining multiple genes via RF

PRECISION: single-gene vs multi-gene



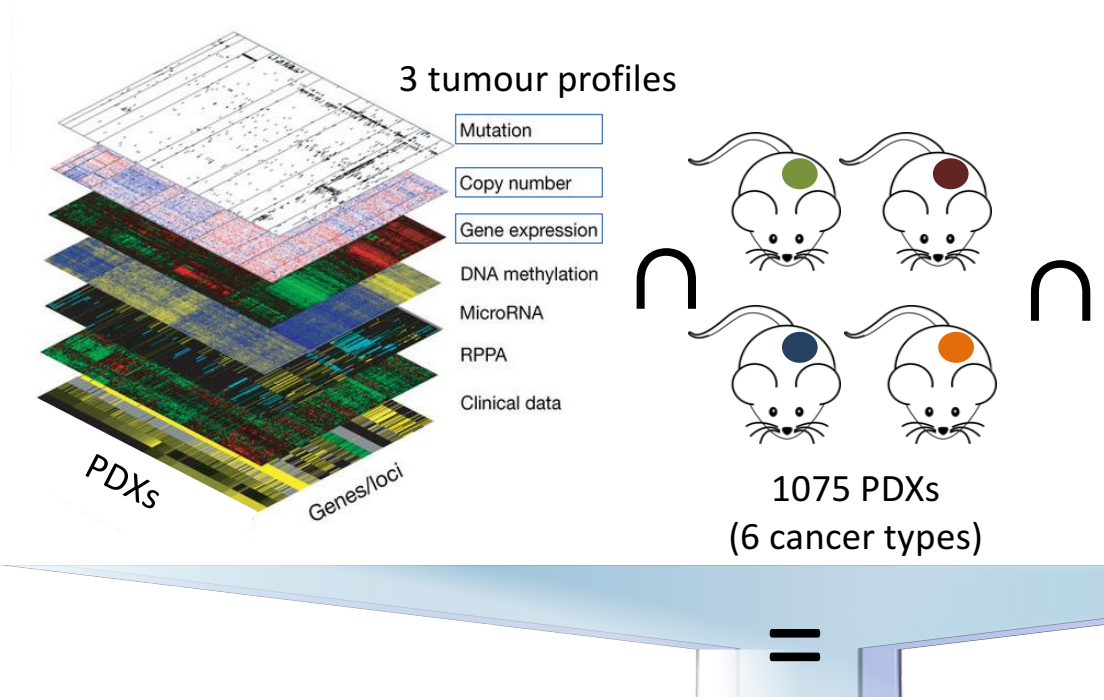
- ***Precision PR**=TP/(TP+**FP**)*
- **PR**: proportion of cell lines predicted sensitive that are actually sensitive
- **Test set PR** across 127 drugs: **large variability**
- **49% of drugs obtained better PR** with multi-gene models

RECALL: single-gene vs multi-gene



- ***Recall RC**=TP/(TP+**FN**)*
- **RC**: proportion of correctly predicted sensitive cell lines
- **Test set RC** across 127 drugs: **large variability with multi-gene model**
- **93% of drugs obtained better RC when using multi-gene models**

Single- & multi-gene predictors on PDX data



NIBR-PDXE: Gao et al. "High-throughput screening using patient-derived tumor xenografts to predict clinical trial drug response" (2015) Nat Med 21: 1318-25

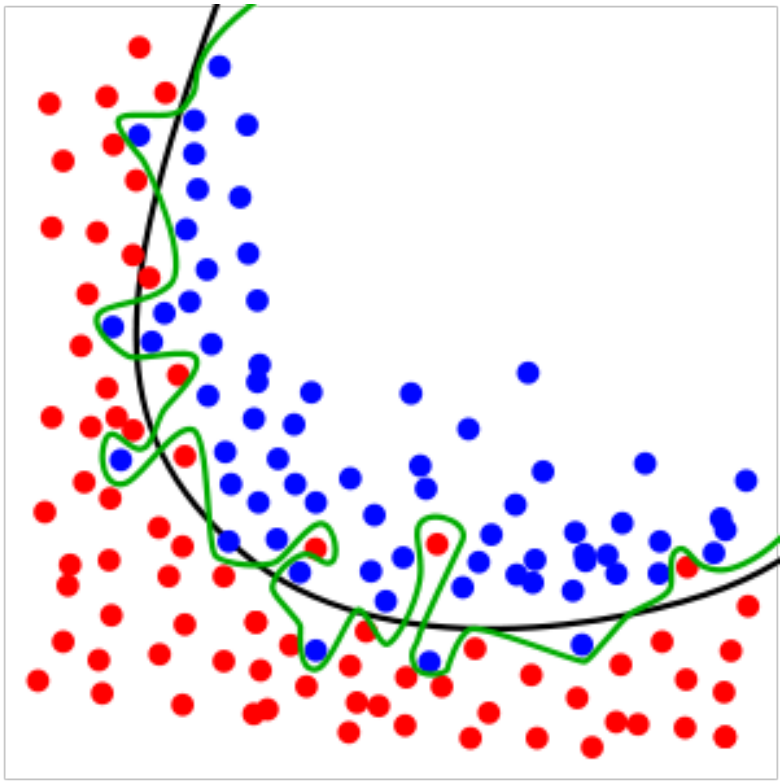
- **Two types** with the highest #s of **treated and profiled PDXs**:
  - breast cancer or BRCA (42 PDXs)
  - colorectal cancer or CRC (50 PDXs)
- **Each type** treated with **13 drug therapies** (mono- or combo)
- **RF-OMC** (Optimal Model Complexity): most predictive features only

High-dimensionality of data is challenging

$$Data = \left\{\left(class^{(i)}, \vec{x}^{(i)}_{tumour}\right)\right\}_{i=1}^{i=N}$$

$$\vec{x}^{(i)}_{tumour} \; either \; \vec{x}^{(i)}_{SNV}, \; \vec{x}^{(i)}_{CNA} \; or \; \vec{x}^{(i)}_{GEX}$$

e.g. while cetuximab-SNV-CRC tested on N=40 PDXs, each PDX profiled for M=15232 genes



https://en.wikipedia.org/wiki/File:Overfitting.svg

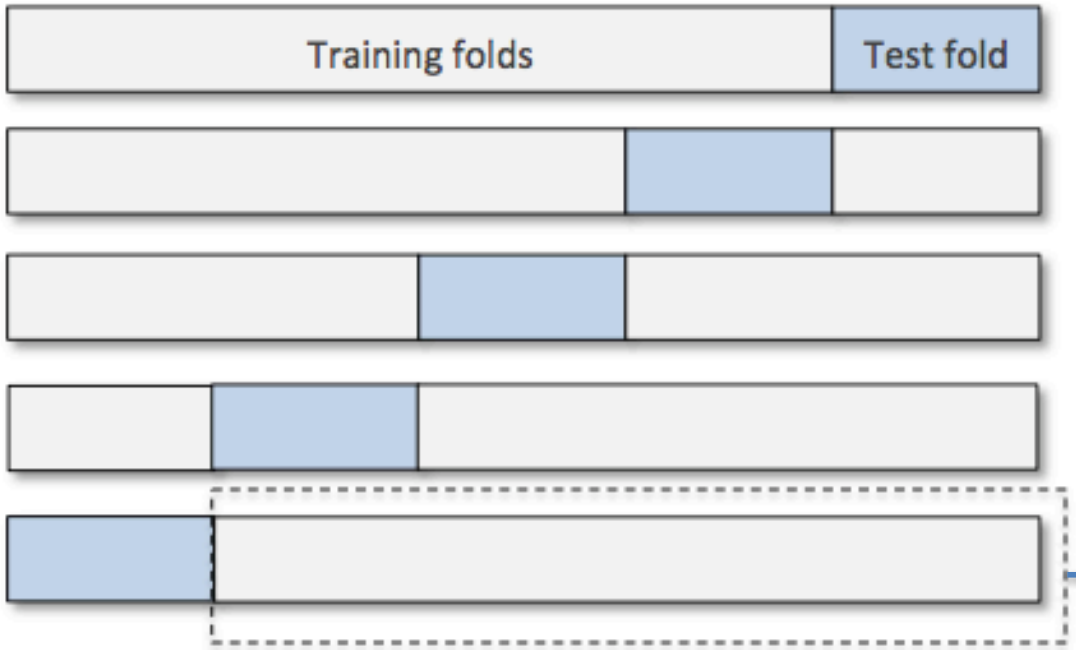
$$\vec{x}^{(i)}_{tumour} = \vec{x}^{(i)}_{SNV} \in \{0,1\}^M \qquad i: 1,..40$$

- **Dimensionality D ~ M / N**
- **Model 1 built on ↑D: too complex for training data → model overfits the data**
- **Model 2: right complexity for training data → more likely to generalise well**
- **Right complexity ~ by ignoring or excluding the many irrelevant features**

Cross-validation (CV) to measure performance

stratified 5-fold CV: every PDX exactly once in a test fold → hence one predicted class per PDX. Also, each PDX has its actual class.

38 BRCA PDX models (with all profiles)



→ 5th fold's training set

		Predicted class: P	Predicted class: N
Actual Class	P	True Positives (TP)	False Negatives (FN)
	N	False Positives (FP)	True Negatives (TN)

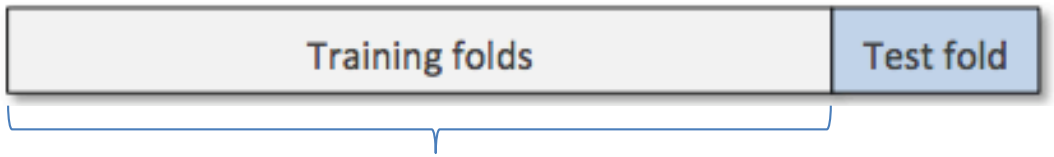
Raschka (2015) "Python machine learning".

Matthews Correlation Coefficient (MCC)

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

Optimal Model Complexity (OMC): motivation

More data or **most informative features** → **OMC**: *a strategy for data-driven identification of the **subset** of most relevant features*



e.g. Binimetinib-GEX-BRCA

For each outer training fold

1. Calculate M p-values between each feature & class across N PDXs
2. Rank all M features by increasing p-value (i.e. decreasing relevance)
3. Consider N/2 nested feature subsets: top 2, top 3, …, top N/2 and M features
4. Among these N/2 models, select that with highest inner CV MCC

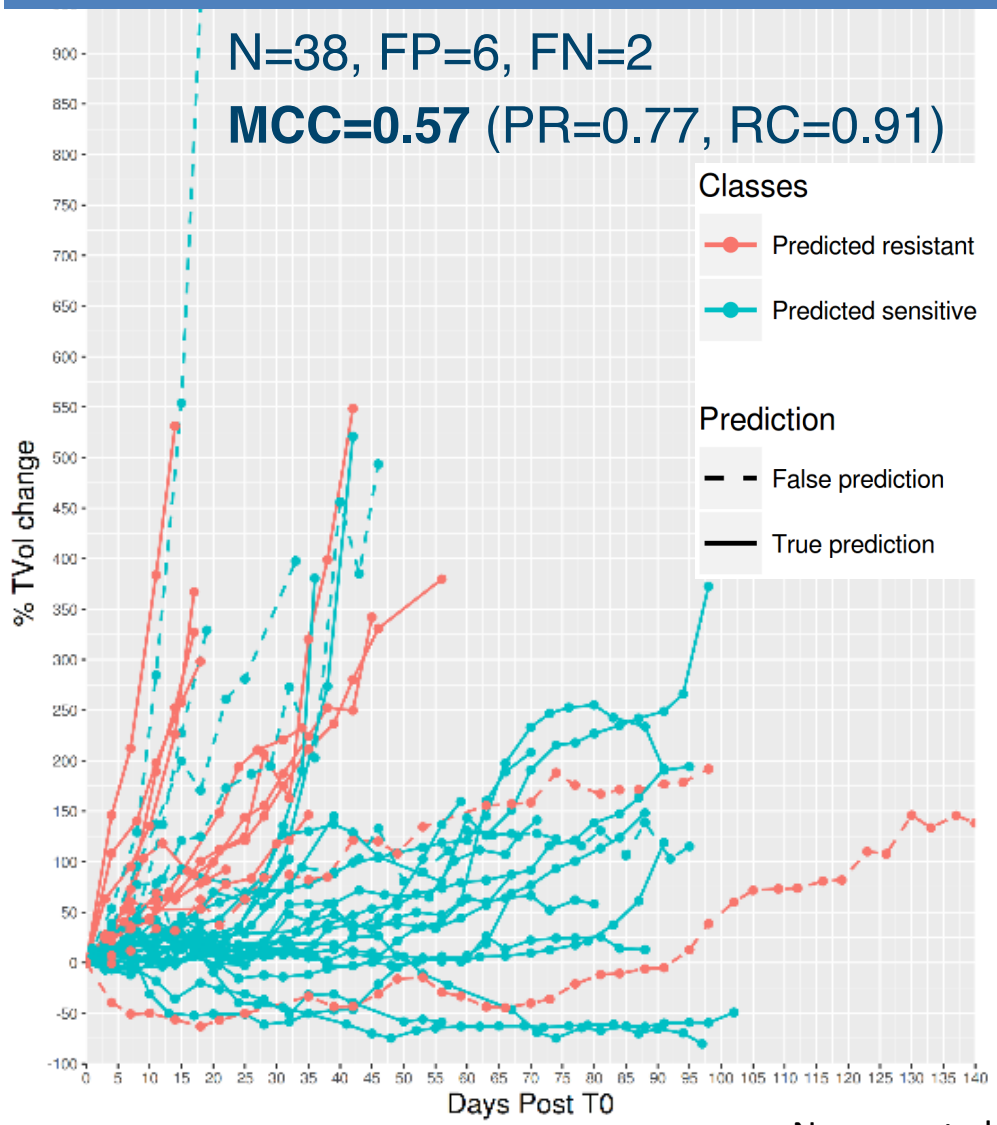
For the corresponding outer test fold

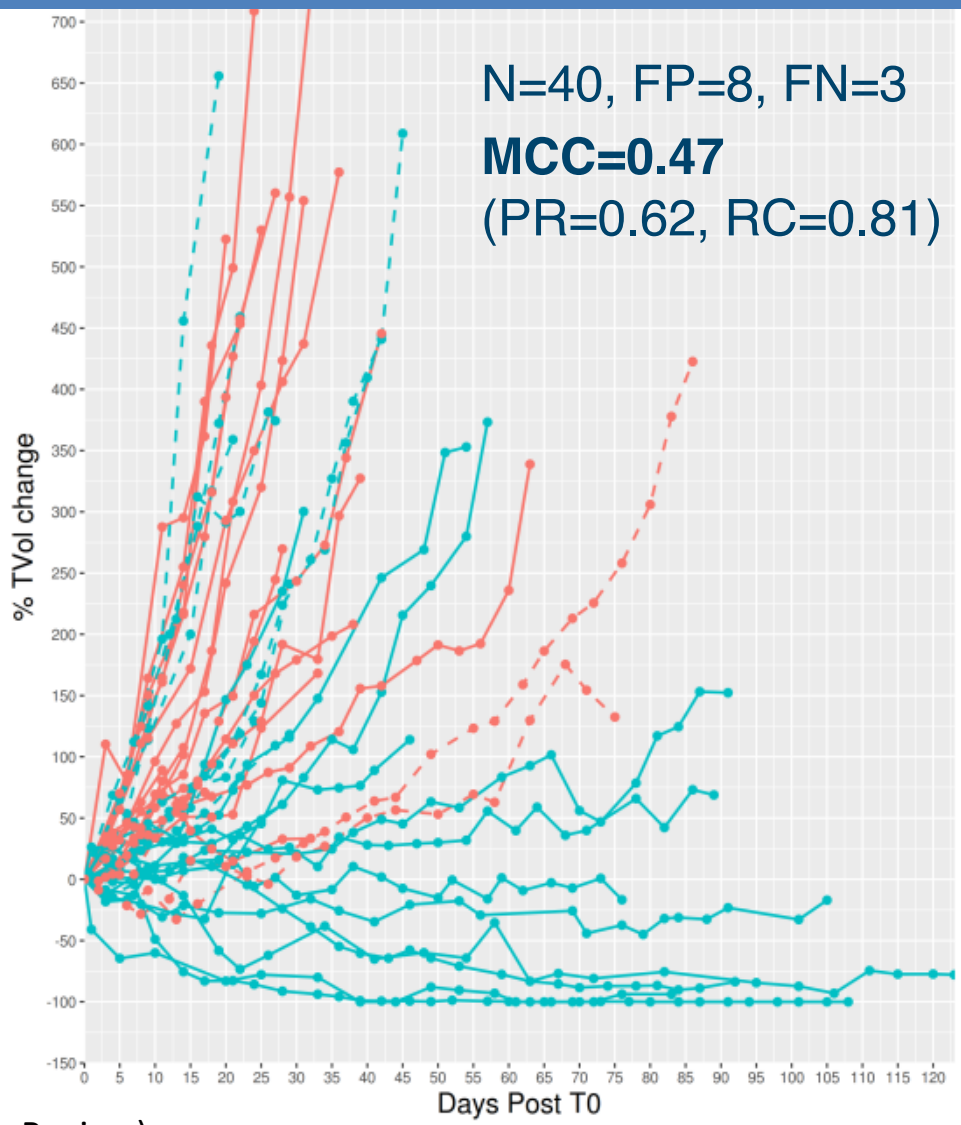
1. Use the selected model (e.g. RF-top7 feats) to predict the class of test PDXs

Note

- **Nested CV ↔ a single CV using a model optimised for each training fold**
- **No information from the test folds is used for model training or selection!**

Visualising nested CV performance (RF-OMC)





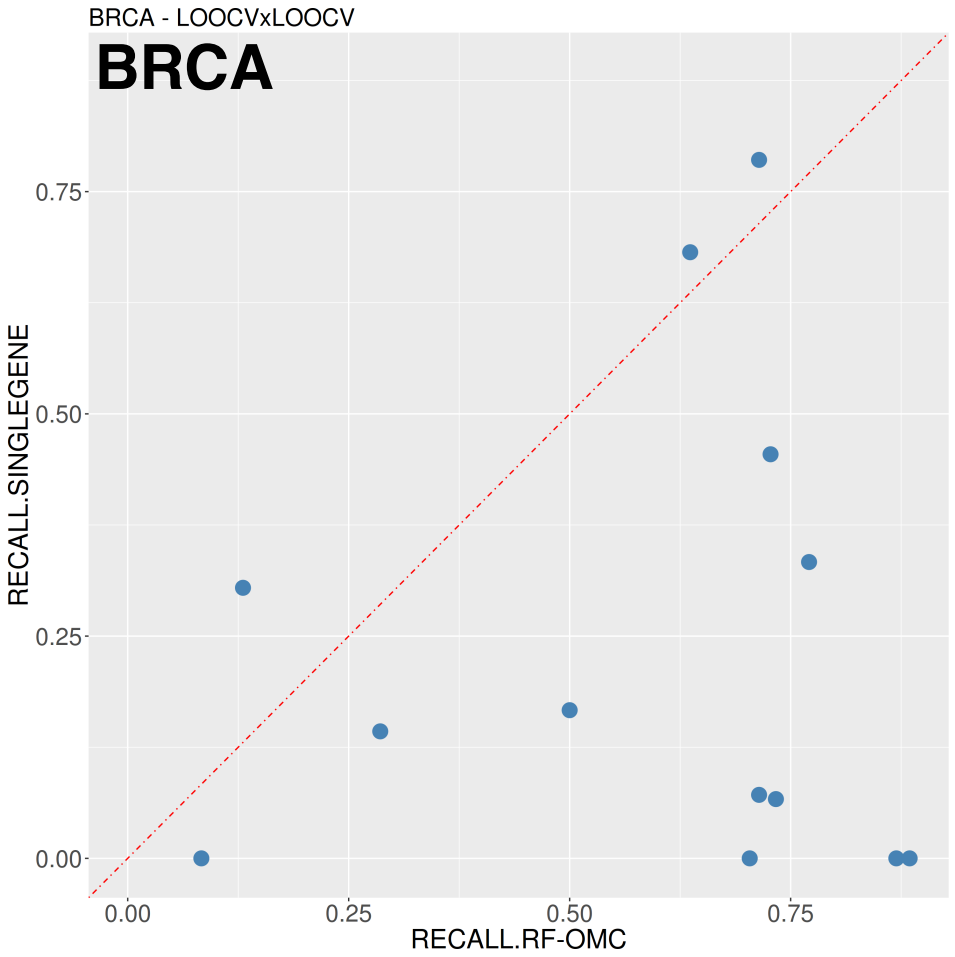
Nguyen et al. (In Review)
https://doi.org/10.1101/277772

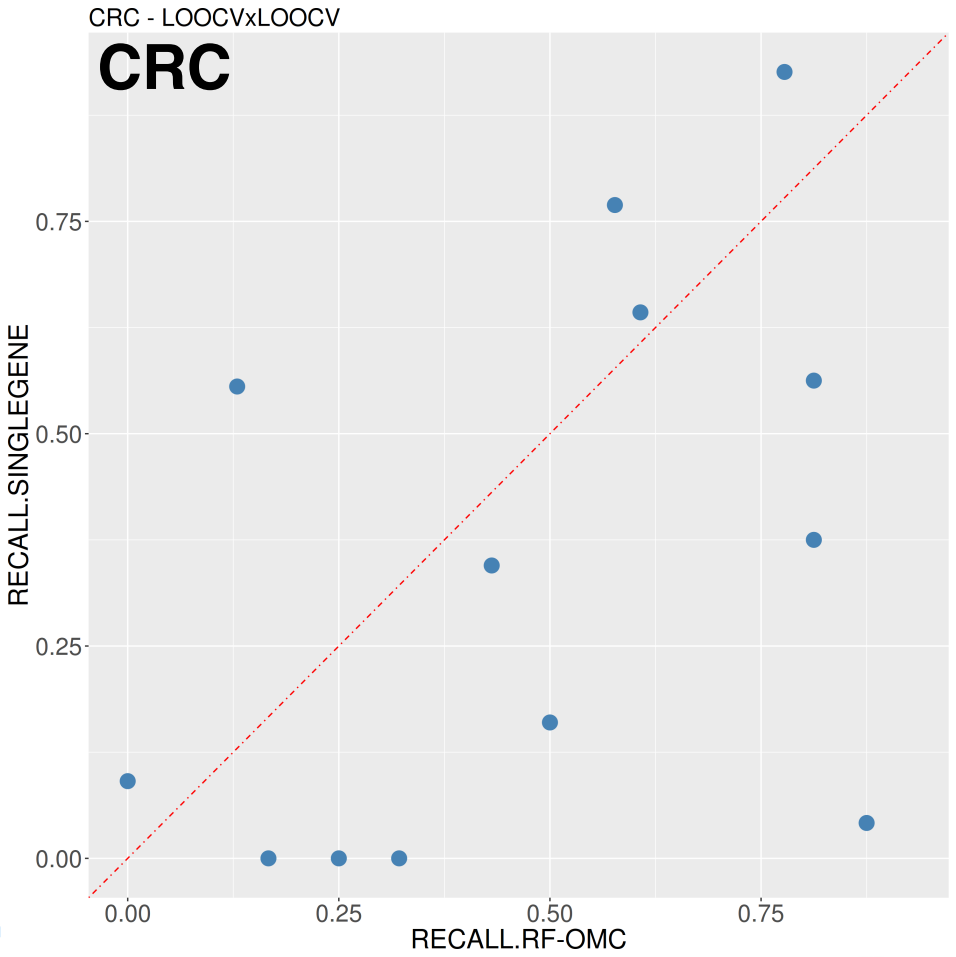
binimetinib-GEX-BRCA,
# of selected features = 14

cetuximab-SNV-CRC,
# of selected features = 4

RECALL: single-gene vs multi-gene

The proportion of sensitive PDXs that are correctly predicted as sensitive (**recall** or sensitivity) of the best single-gene marker was generally lower: **same conclusion on these two cancer types as with *in vitro* data**





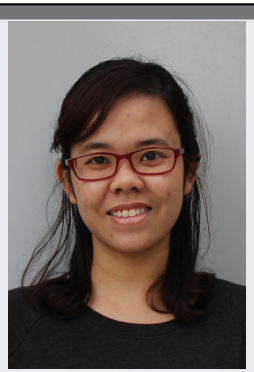
Nguyen et al. (In Review) https://doi.org/10.1101/277772

Summary

- Multi-gene often more predictive than single-gene (shown: *in vitro* pancancer & *in vivo* cancer-specific)
- Also, **multi-gene models generally have higher recall**
- With few exceptions, **single-gene markers have low recall: responsive tumours w/out marker are missed!**
- Consequently, **combining the mutational status of multiple genes via ML** should be always considered.
- RF-OMC → predictors with just 2-20 gene alterations (↓features, beneficial for clinical implementation and interpretability)
- Apply to other tumour profiles (e.g. miRNA, DNA methy)

Acknowledgements

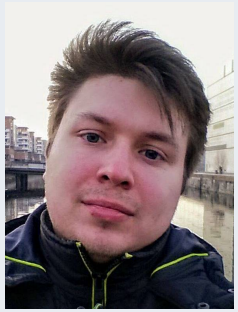
Group



Linh Nguyen
(PhD student)

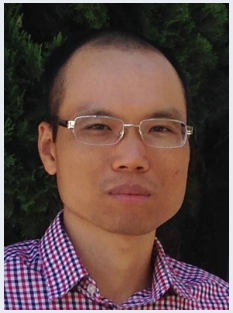
Alexandra Bomane
(PhD student)

Stefan Naulaerts
(Postdoc)



Pavel Sidorov
(Postdoc)

Michal Zulcinski
(MSc student)



Cuong Dang
(former Postdoc)

Funding

Instituts thématiques Inserm
Institut national de la santé et de la recherche médicale

CAMPUS FRANCE
campusfrance.org

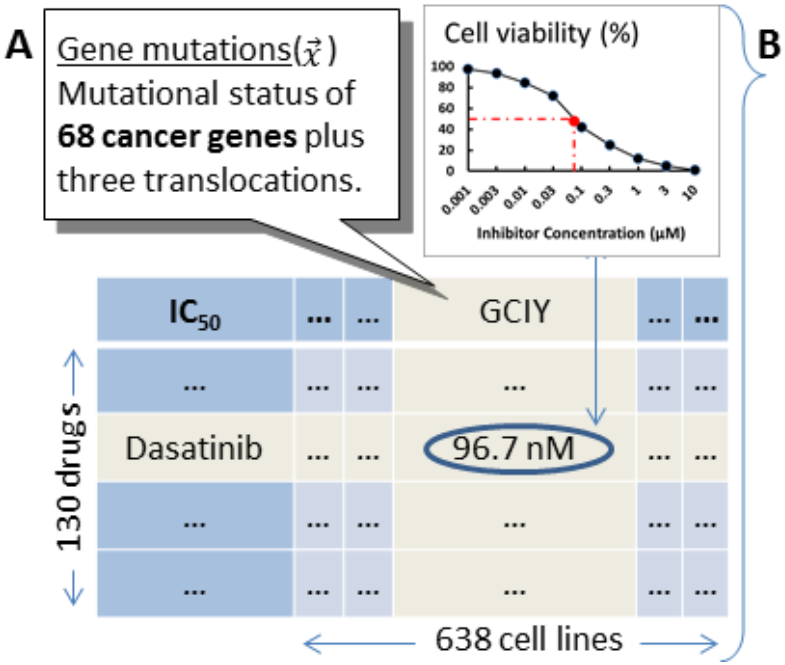
PACA Canceropole
PROVENCE-ALPES CÔTE D'AZUR

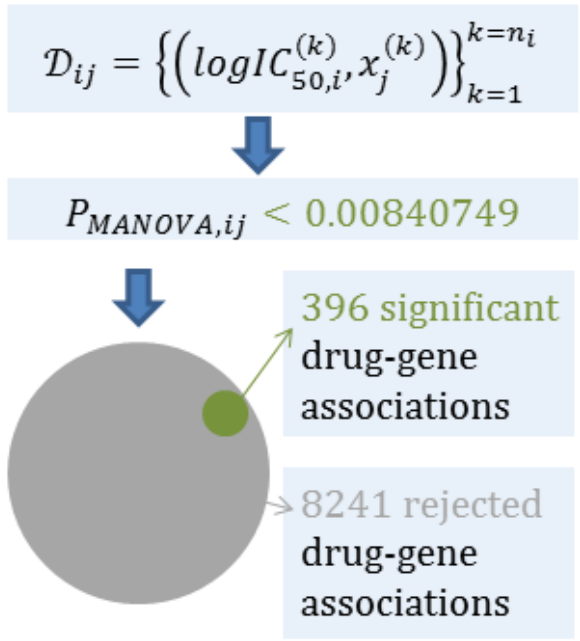
INSTITUT PAOLI-CALMETTES

Ministry of Education and Training
Vietnam International Education Development
VIED

AGENCE NATIONALE DE LA RECHERCHE
ANR

GDSC: single-gene markers of drug response





Garnett, et al. (2012) Nature

Genomics of Drug Sensitivity in Cancer **GDSC data**

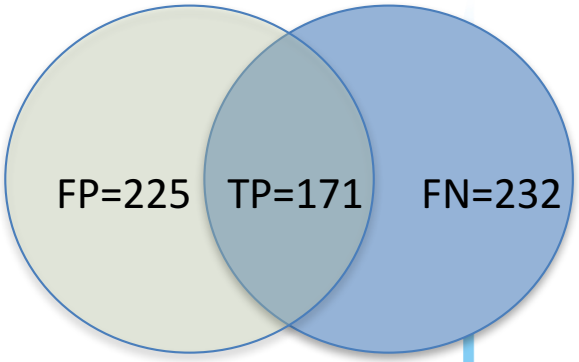
most drugs have either **weakly significant markers of response** (yet potentially useful) **or no found markers at all**

A **parametric test** makes **strong modelling assumptions** (e.g. normality and equal variances of residuals in MANOVA), but drug responses across cell lines are often skewed, contain outliers and/or have different variances → **Impact?**

Compare w/non-parametric test on the same dataset → FPs, FNs

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{2}\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

$$P_{\chi^2} = pdf_{\chi^2}(\chi^2, df = 1)$$



Dang et al. *BMC Medical Genomics* (2018) https://doi.org/10.1186/s12920-018-0336-z

Single-gene vs multi-gene expression (GEX)

Question: Will **combining transcriptomic features** result in better **prediction of which cancer cell lines are sensitive to a given drug?**

2012 (1st GDSC data release)

2014 (5th GDSC data release)

TRAINING SET
366 cell lines
127 drugs
13,321 GEx values

$\cap$

TEST SET
203 cell lines
127 drugs (in common)
13,321 GEx values (in common)

per drug
$= \emptyset$

Performance assessment

$$\mathcal{D} = \left\{ \left(logIC_{50}^{(i)}, \vec{x}_{CellLine}^{(i)} \right) \right\}_{i=1}^{i=N}$$

$$class^{predicted} = f\left(\vec{x}_{CellLine}^{(new)} \right)$$

Nguyen et al. *F1000Research* (2016) https://f1000research.com/articles/5-2927

Baseline: Prior Probability (PP)

Higher MCC = more unlikely due to chance, but quantify

Training set → # of **sensitive (S)** and **resistant (R)** PDXs

Test set → **For each PDX,** generate a random number Z in (0,S+R) and use it to **predict its class.** For example,