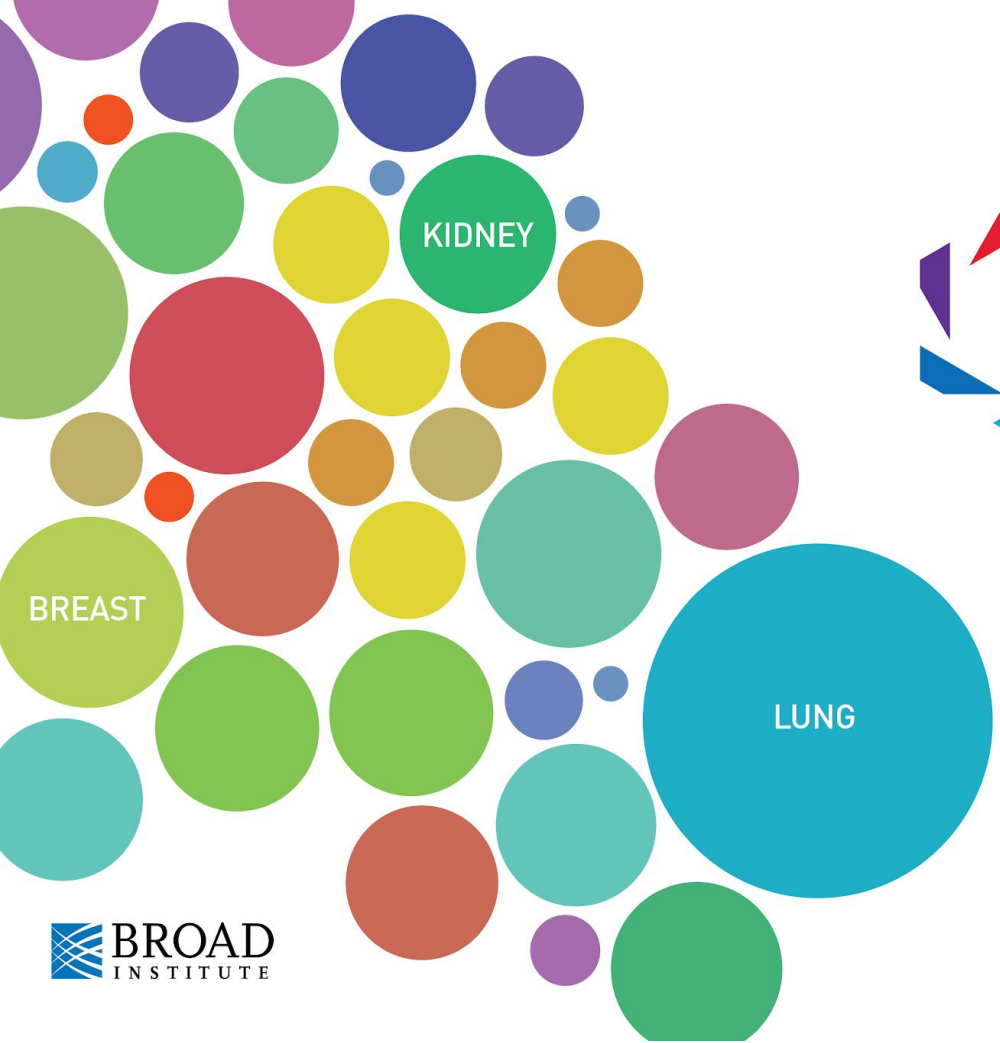


PRISM

MULTIPLEXED CELL LINE PROFILING

BROAD INSTITUTE

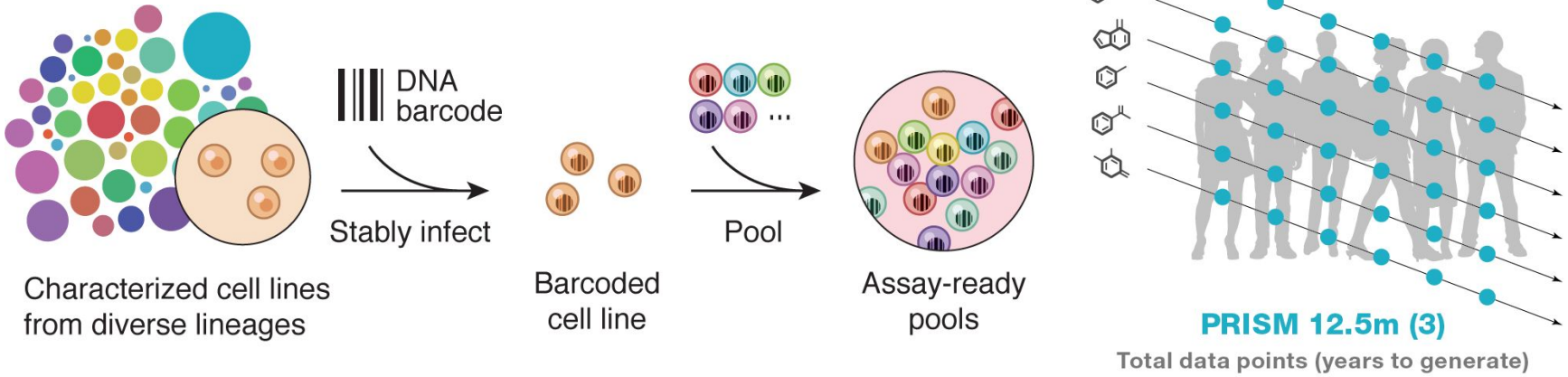
Data processing in MTS screens

PRISM overview

- A viability screen of 489 solid (or 489+278 solid, mixed, and suspension) cell lines in pools
- A primary screen to estimate potency and selectivity
- Does not replace individual cell line dose-response characterization
- Constantly improving (assay, pools, cell lines, etc.)
- Many options for downstream analysis
- Rich genomic and functional characterization of cell lines enables powerful biomarker analyses

PRISM pooled viability screening

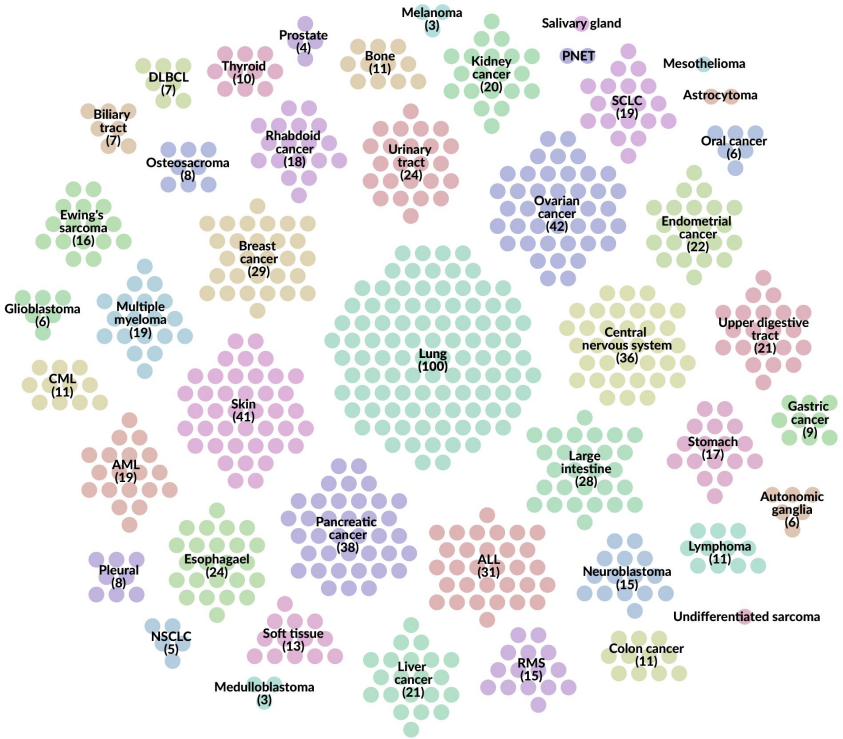
Requires fewer resources and allows for increase in scale



- Cell lines are adapted to RPMI-1640, infected and selected with blasticidin
- Stocks of individual cell lines are made and QC'ed
- Cell lines are pooled based on estimated doubling time in pools of 25 cell lines

PRISM utilizes ~750 cancer cell lines

Solid and Suspension/Hematopoietic cell lines represent the diversity of CCLE

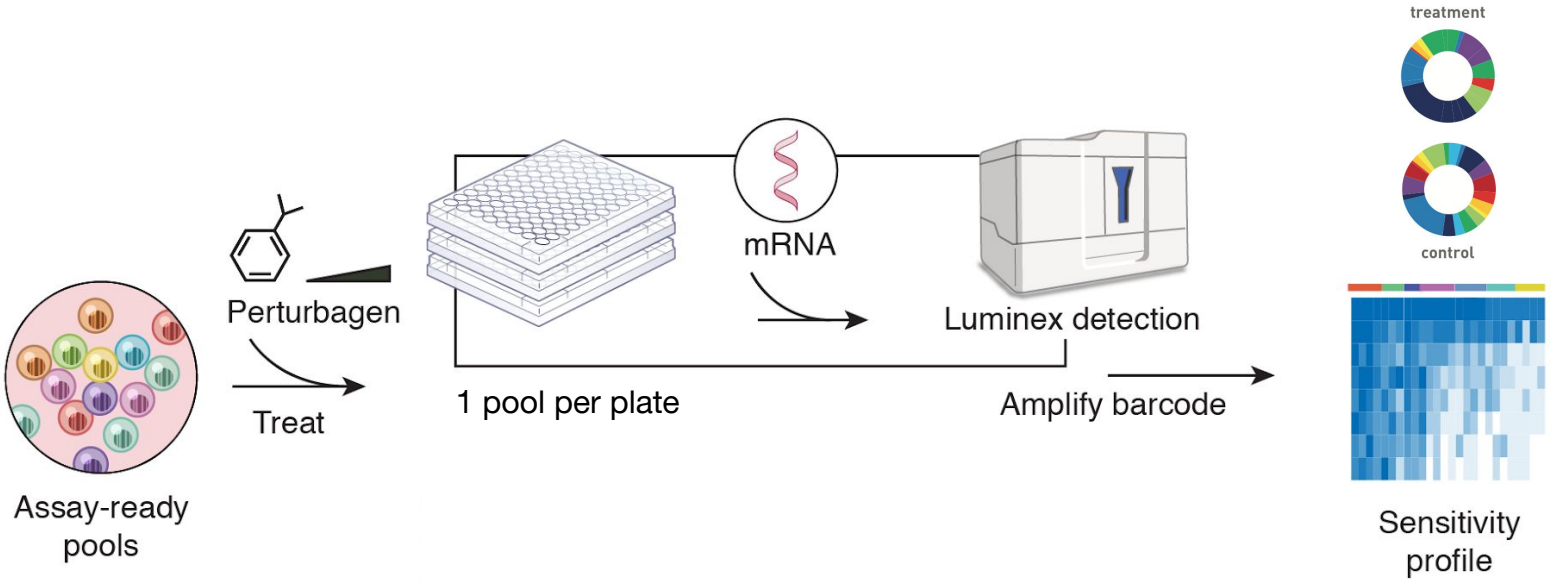


Differences between two cell sets

	PR500	PR300
Kinds of cell lines	Solid/ Adherent	115 hematopoietic, 135 pediatric and 50 solid/ loosely adherent
# of cell lines in cell set	489	278
# of cell lines per well	22-25	20-25
# of cells per cell line	~50	~80
# of cells per well	1250	2000
FBS	10%	20%
Cell thaw procedure	Thawed and treated on same day	Thawed and recovered, then plated 3 days after thaw

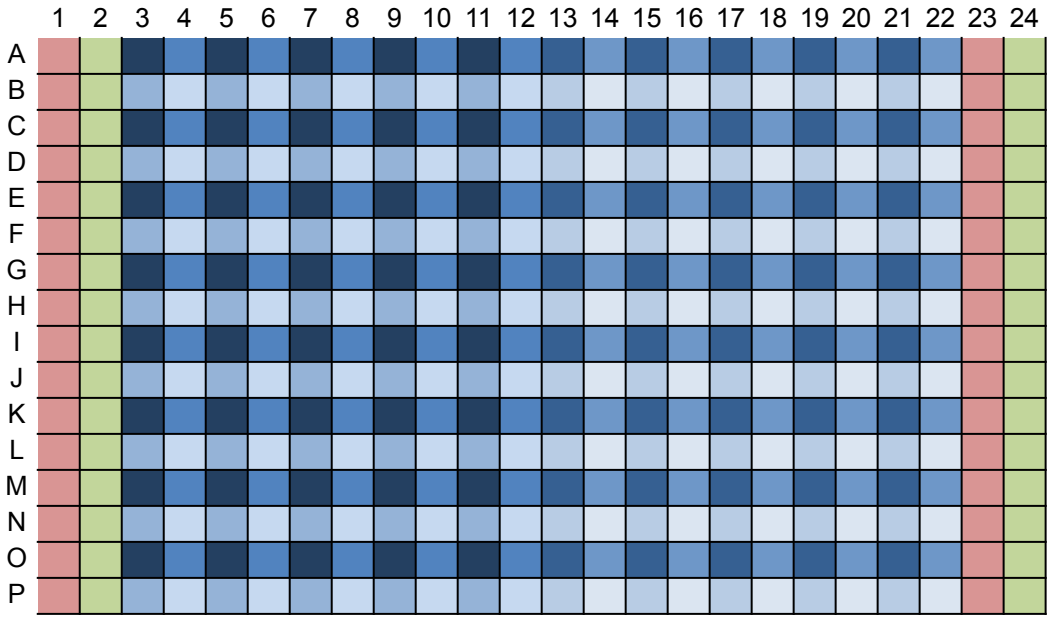
PRISM viability 5-day assay

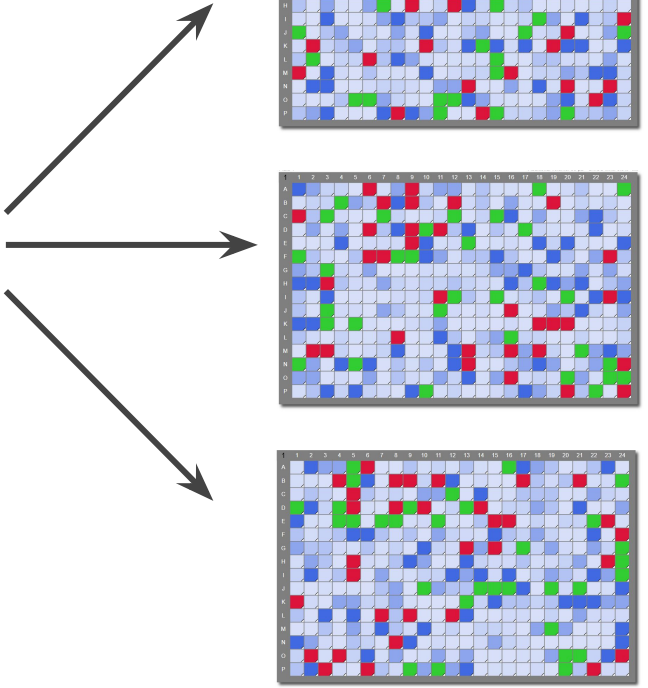
Pooled cytotoxicity assay on up to ~750 barcoded cell lines



Cells are plated into assay-ready plates

Cells are seeded onto plates containing (randomized) compounds

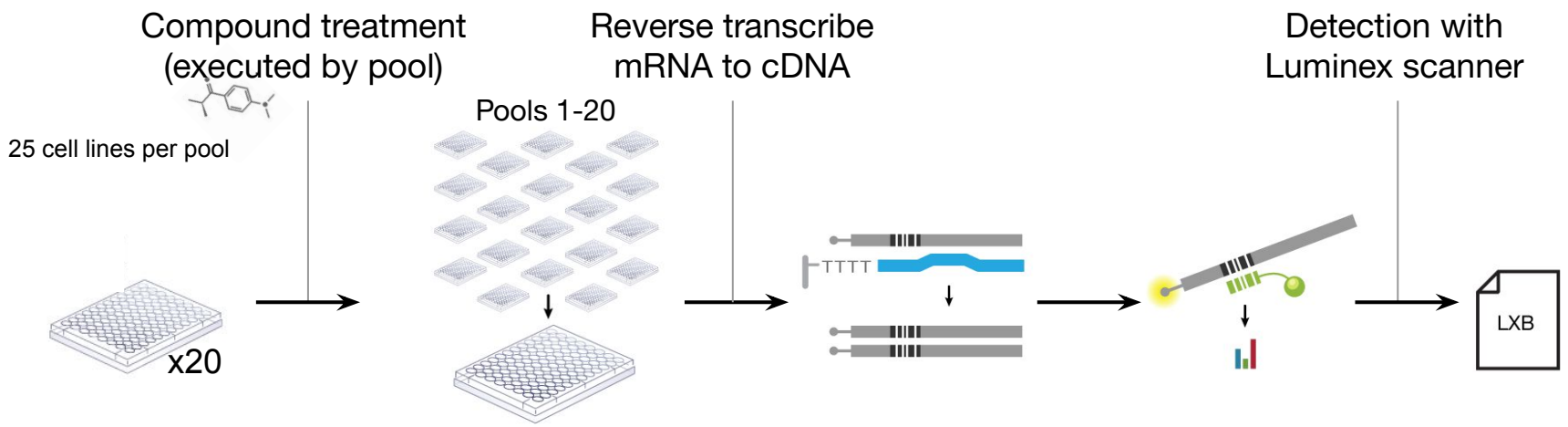




Empty/Positive Control
DMSO
Compound

mRNA detection: Luminex assay

Lysate generation, amplification and detection



Data pre-processing

- Raw files from the luminex scanner: CSV and one LXB per well.
  - Each LXB contains fluorescence values (FI) for each detected analyte. FI values indicate how much amplicon was bound to a given analyte (bead), which in turn corresponds to the unique 24nt barcode that the cell line was labelled with.
  - The CSV files contain summary statistics such as the bead counts and Median Fluorescence Intensity (MFI) for each analyte in each well.

Data pre-processing

- Initial quality control filters:
  - *Bead count*. The number of beads detected is used to filter out poor-signal wells. Wells with a bead count median < 20 are removed.
  - *Control signal intensity*. Spiked-in pooled lysate controls are used to assess technical quality of the assay process on a per-well basis from lysate pooling through detection. Wells with median signal < 600 units are removed.
- If the total number of wells that fail exceeds 10% (i.e >39 wells) the entire plate is re-processed

Data processing

- Our immediate goal is to make all the data processing and downstream analysis reproducible and transparent
  - https://github.com/broadinstitute/prism_data_processing
- *Caveat*: the pipeline is evolving each run (e.g., based on feedback we receive from collaborators)
  - Any form of feedback is very welcome!

Data Processing - Overview

Luminex machine reports **median fluorescence intensity** (MFI) for each **analyte** (barcoded **bead**)

log2

Compound x dose x replicate

Cell lines / analytes / beads

log2(MFI)

Normalize with respect to inert barcodes in same well on the same plate

Compound x dose x replicate

Cell lines / analytes / beads

Barcode normalized log2(MFI)

Filter cell lines that fail QC

Compound x dose x replicate

Cell lines / analytes / beads

Barcode-normalized log2(MFI)

Normalize with respect to median of DMSO wells on the same detection plate

Compound x dose x replicate

Cell lines / analytes / beads

log2(viability)

Correct pool-based batch effects. ComBat is applied to each compound-dose combination independently

Compound x dose x replicate

Cell lines / analytes / beads

ComBat-corrected log2(viability)

Take median of replicates

Cell lines / analytes / beads

Compound x dose

Median log2(viability)

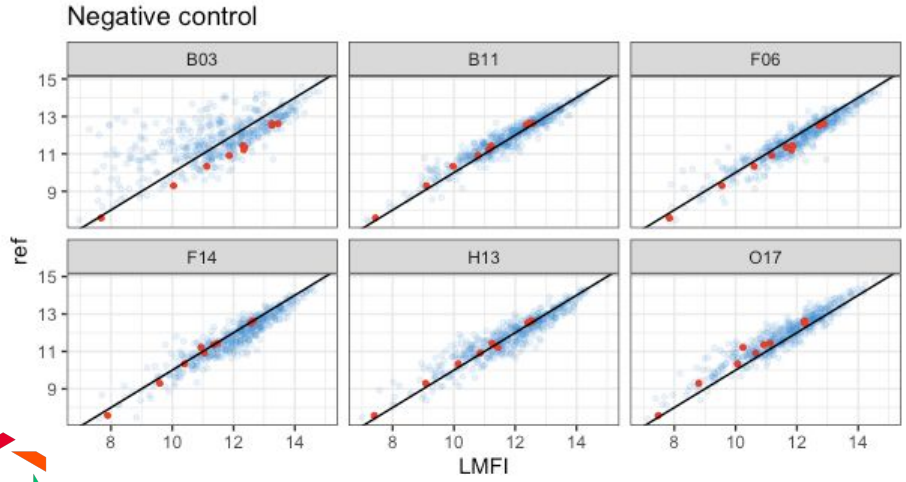
Fit dose response curves

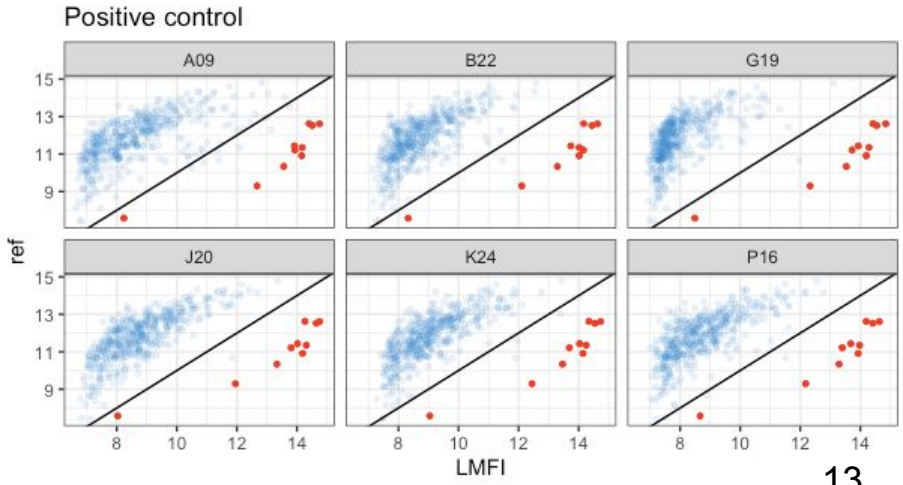
Cell lines x Compound

DRC parameters (AUC, IC50)

Invariant normalization

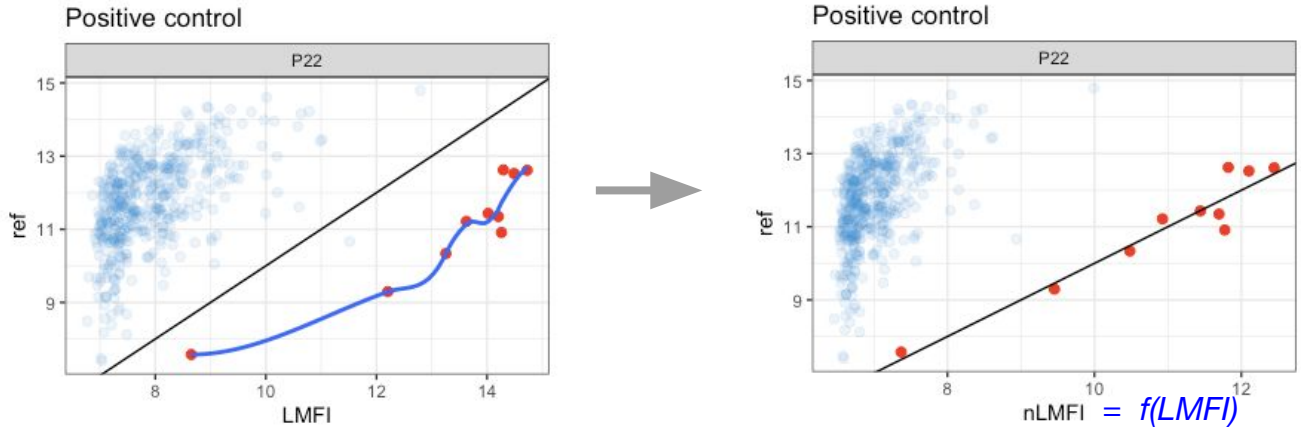
- 10 control barcodes are spiked in after lysate generation.
- Goal: correcting for amplification and detection artifacts.
- In the raw data files, control barcodes are denoted in the pool_id column (= “CTLBC”).





Invariant normalization

- For each plate, a reference profile is generated (median DMSO log2(MFI)).
- For each profile, we fit a monotonic-increasing smooth transformation (p-spline) using only the spike-ins.
- All cell-line log2(MFI) values pass through this transformation.



Data Processing - Overview

Luminex machine reports **median fluorescence intensity** (MFI) for each **analyte** (barcoded **bead**)

log2

Compound x dose x replicate

Cell lines / analytes / beads

log2(MFI)

Normalize with respect to inert barcodes in same well on the same plate

Compound x dose x replicate

Cell lines / analytes / beads

Barcode normalized log2(MFI)

Filter cell lines that fail QC

Compound x dose x replicate

Cell lines / analytes / beads

Barcode-normalized log2(MFI)

Normalize with respect to median of DMSO wells on the same detection plate

Compound x dose x replicate

Cell lines / analytes / beads

log2(viability)

Correct pool-based batch effects. ComBat is applied to each compound-dose combination independently

Compound x dose x replicate

Cell lines / analytes / beads

ComBat-corrected log2(viability)

Take median of replicates

Compound x dose

Cell lines / analytes / beads

Median log2(viability)

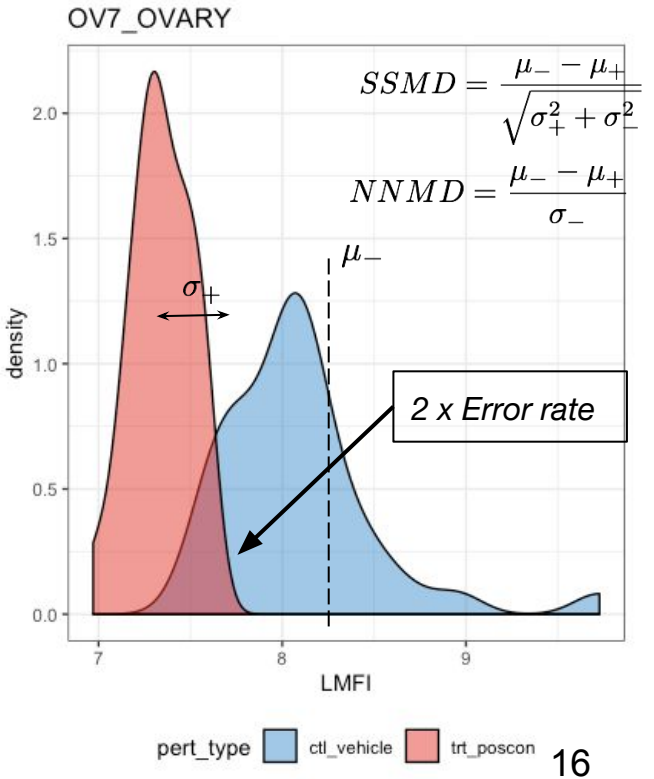
Fit dose response curves

Cell lines x Compound

DRC parameters (AUC, IC50)

QC: Control Separation

- Post-detection, we assess separation between positive and negative controls.
  - Historically, we used SSMD > 2
  - Recent screens have used error rate < 0.05
- Filters applied per detection plate × cell line.
- Compound × cell line plates filtered if not at least 2/3 replicates with good separation.



QC: Control Separation Notes

- NNMD is fairly similar to SSMD except it does not punish high deviation in positive controls.
- Error rate is more interpretable and robust to outliers but oblivious to the dispersion in the negative controls.

Data Processing - Overview

Luminex machine reports **median fluorescence intensity** (MFI) for each **analyte** (barcoded **bead**)

log2

Compound x dose x replicate

Cell lines / analytes / beads

log2(MFI)

Normalize with respect to inert barcodes in same well on the same plate

Compound x dose x replicate

Cell lines / analytes / beads

Barcode normalized log2(MFI)

Filter cell lines that fail QC

Compound x dose x replicate

Cell lines / analytes / beads

Barcode-normalized log2(MFI)

Normalize with respect to median of DMSO wells on the same detection plate

Compound x dose x replicate

Cell lines / analytes / beads

log2(viability)

Correct pool-based batch effects. ComBat is applied to each compound-dose combination independently

Compound x dose x replicate

Cell lines / analytes / beads

ComBat-corrected log2(viability)

Take median of replicates

Cell lines / analytes / beads

Compound x dose

Median log2(viability)

Fit dose response curves

Cell lines x Compound

DRC parameters (AUC, IC50)

Computing Viability: Log Fold-Change

- Log2(fold-change) viabilities computed per plate × cell line combination: $LFC = LMFI - median\,(LMFI_{DMSO})$
  - Does not allow differentiation between cytotoxic and cytostatic responses.
  - High variance (noise) in DMSO complicates interpretation.
  - As a heuristic, we use log2(0.3) (30% viability) as the threshold to consider a cell line as sensitive.

Data Processing - Overview

Luminex machine reports **median fluorescence intensity** (MFI) for each **analyte** (barcoded **bead**)

log2

Compound x dose x replicate

Cell lines / analytes / beads

log2(MFI)

Normalize with respect to inert barcodes in same well on the same plate

Compound x dose x replicate

Cell lines / analytes / beads

Barcode normalized log2(MFI)

Filter cell lines that fail QC

Compound x dose x replicate

Cell lines / analytes / beads

Barcode-normalized log2(MFI)

Normalize with respect to median of DMSO wells on the same detection plate

Compound x dose x replicate

Cell lines / analytes / beads

log2(viability)

Correct pool-based batch effects. ComBat is applied to each compound-dose combination independently

Compound x dose x replicate

Cell lines / analytes / beads

ComBat-corrected log2(viability)

Take median of replicates

Compound x dose

Cell lines / analytes / beads

Median log2(viability)

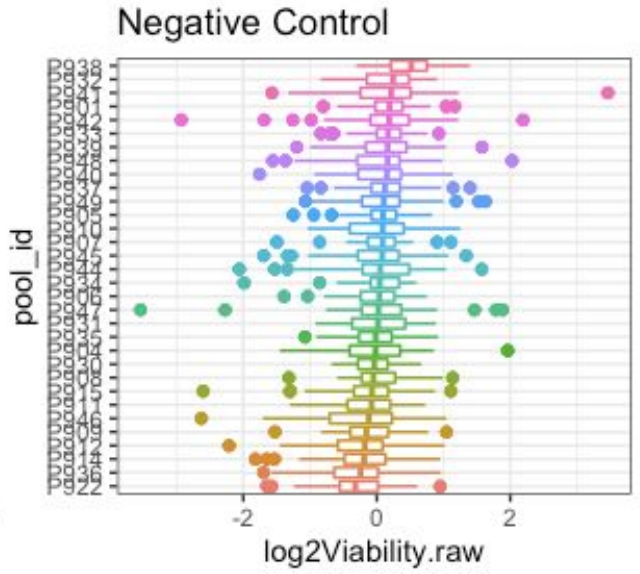
Fit dose response curves

Cell lines x Compound

DRC parameters (AUC, IC50)

Pool Effects: ComBat

- Pool effects in viability profiles are often often as strong as the biological biomarkers.
- Use ComBat[1] to correct for pool effects, so that each pool has similar mean and variance.



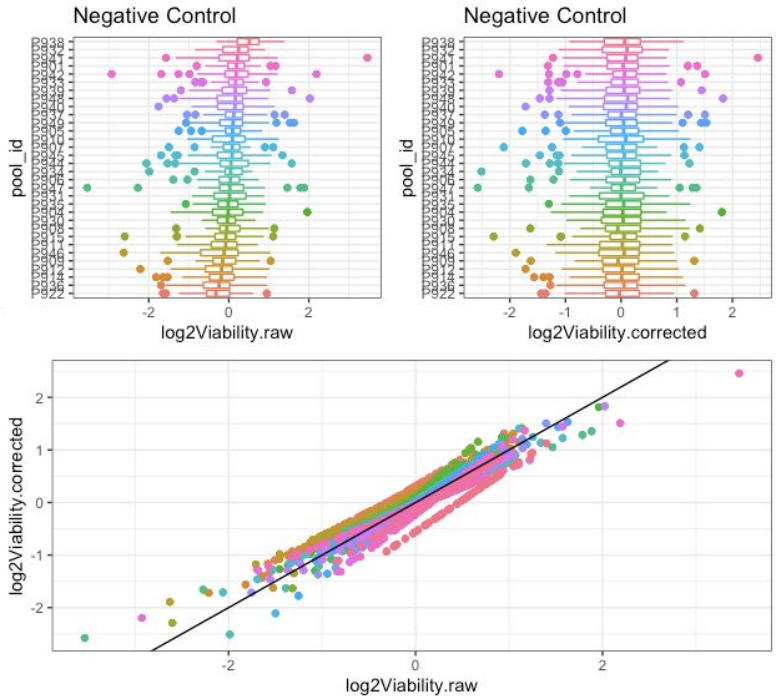
$$Y_{ijg} = \alpha_g + X\beta_g + \gamma_{ig} + \delta_{ig}\varepsilon_{ijg},$$

i : pool id, g : probe (dummy),
j : cell line

[1] Johnson, WE, Rabinovic, A, and Li, C (2007). Adjusting batch effects in microarray expression data using Empirical Bayes methods. *Biostatistics* **8**(1):118-127.

Pool Effects: ComBat

- ComBat is a pretty light-handed correction.
  - Most points are around the main diagonal in the pre- vs post-correction scatter.
- Note: the pool effects are driven by two main artifacts:
  - Assay plates.
  - Growth rate differences.



ComBat: Notes

- The risk of ComBat is if a compound truly selectively kills a particular pool.
  - Deconvolving the artifacts and the true signal is not trivial.
  - When designing pools, we are careful to ensure particular lineages or genomic features are not overrepresented.

- Note: that this is the first part PR300 and PR500 data touches each other.

Data Processing - Overview

Luminex machine reports **median fluorescence intensity** (MFI) for each **analyte** (barcoded **bead**)

log2

Compound x dose x replicate

Cell lines / analytes / beads

log2(MFI)

Normalize with respect to inert barcodes in same well on the same plate

Compound x dose x replicate

Cell lines / analytes / beads

Barcode normalized log2(MFI)

Filter cell lines that fail QC

Compound x dose x replicate

Cell lines / analytes / beads

Barcode-normalized log2(MFI)

Normalize with respect to median of DMSO wells on the same detection plate

Compound x dose x replicate

Cell lines / analytes / beads

log2(viability)

Correct pool-based batch effects. ComBat is applied to each compound-dose combination independently

Compound x dose x replicate

Cell lines / analytes / beads

ComBat-corrected log2(viability)

Take median of replicates

Compound x dose

Cell lines / analytes / beads

Median log2(viability)

Fit dose response curves

Cell lines x Compound

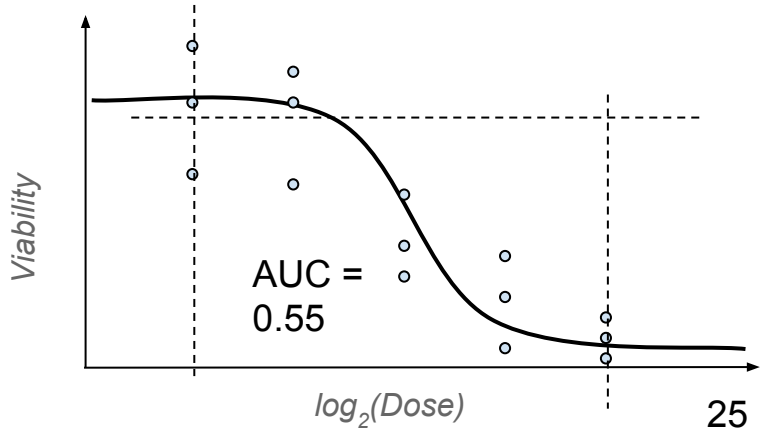
DRC parameters (AUC, IC50)

Curve fitting

- useS the 4-parameter log-logistic model (R packages dr4pl or drc):

$$Viability(Dose) = LowerLimit + \frac{UpperLimit - LowerLimit}{1 + (Dose/EC50)^{-Slope}}$$

- Compute AUC and IC50
  - AUCs are more stable.
  - IC50s are universal: comparable across compounds, not assays.



PRISM sensitivity profiles are correlated with DepMap feature sets to identify biomarkers

depmap

Genomics

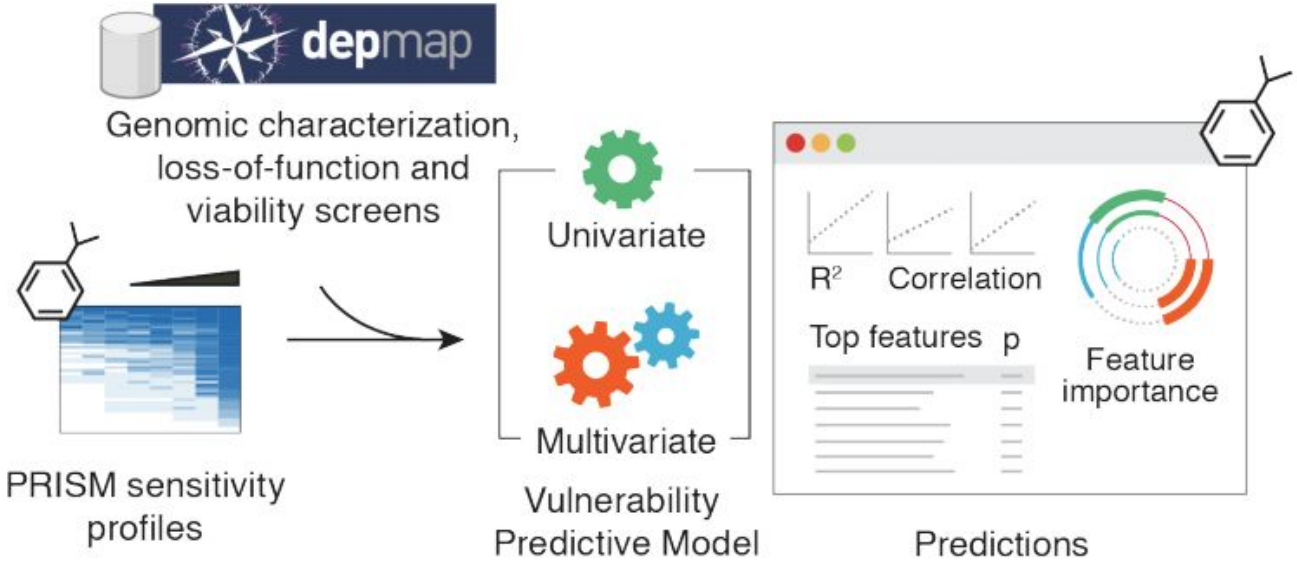
Mutations
Copy number
Methylation

Expression

Single gene-expression
Pathway level-expression
mRNA

Other

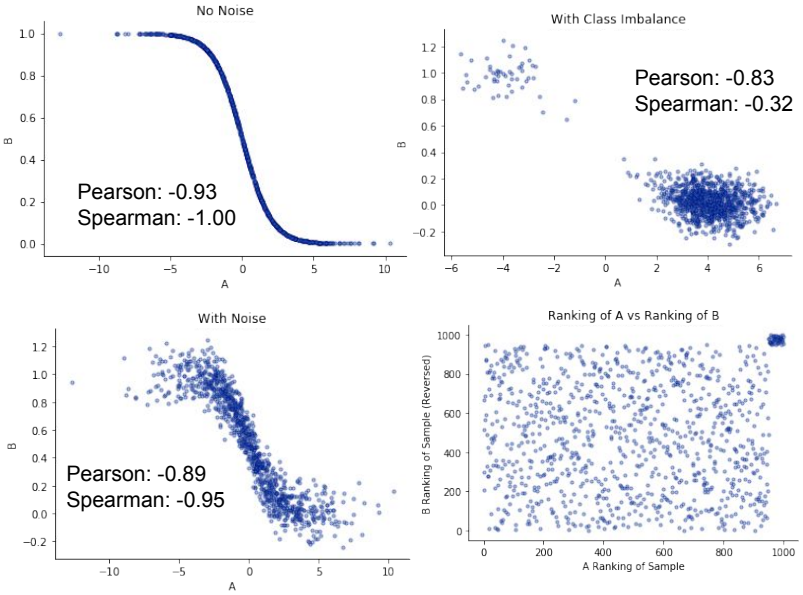
Protein (RPPA)
Metabolomics
Achilles (CRISPR/RNA)
PRISM Repurposing



Univariate Analysis

- For each depmap.org dataset* and PRISM profile**:
  - Compute Pearson correlations between each feature and the profile
- Compute p-values:
  - Fisher's z-transform, limma
  - Bootstrap: not feasible for bulk computations!
- Correct for multiple hypotheses:
  - Benjamini-Hochberg

Josh Dempster



* mutation, gene expression, CRISPR, lineage, metabolomics, miRNA, copy number, protein (RPPA)
** log2(Viability), AUC, log2(IC50)

Multivariate Analysis

- We train random forest and elastic net (α=0.5) models using:
  - CCLE features (copy-number alterations, RNA expression, mutation status, and lineage annotation)
  - CCLE features + RPPA (protein) + CRISPR + metabolomics + miRNA.
- As with univariate analyses, we analyze data for both individual concentration points and overall fit curves

Multivariate Analysis

- For each model, we report the cross-validated R-squared values and Pearson scores (the correlation between the model predictions and PRISM profiles). These describe how accurate the model is.
  - For each feature of each model, the feature importances are computed (after normalizing to the sum of the importances to 1 in each model) and tabulated along with the accuracy measures.

Future work

- Pipeline optimization:
  - Better documentation and reproducible code.
  - Collect more data on whether to combine PR500 and PR300 for biomarker analyses or not.
- Dose-response curves:
  - Constraints:
    - Fixed or free asymptotes?
    - Capping AUCs?
    - Corrected or raw viability values?
- Batch/pool correction?

Future work

- Reliable estimation of growth rate inhibition metrics.
  - Traditional viability metrics are susceptible to variation in the doubling time of cell lines.
  - GR metrics are a nice alternative:

$$GR(c) = 2^{\frac{\log_2(x(c)/x_0))}{\log_2(x_{ctrl}/x_0)}} - 1$$

    - Challenges:
      - large variation in day 0 abundances
      - assumes constant growth rate throughout assay duration
      - logistical challenges

Future work

- Better characterization of assay noise.
  - MFI values are overdispersed; Naive Bayes model may capture the noise structure better.
- Better account of uncertainty.
  - Confidence intervals and standard errors via bootstrap.
- Better biomarker models.

Acknowledgements

PRISM Team


Jen Roth


Ginevra Botta


Melissa Ronan

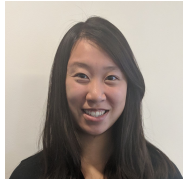

Danny Rosenberg


Maggie LeMaitre


Li Wang


Jordan Rossen


Mustafa Kocak


Vickie Wang


Andrew Boghossian


Hannah Miller


Caryn Liu


Kevin Larpenteur


Luc De Waal


Tenzin Sangpo


Massami Laird


Matt Rees

CMap Team


Aravind Subramanian


Nicholas Lyons


Sarah Johnson


David Peck

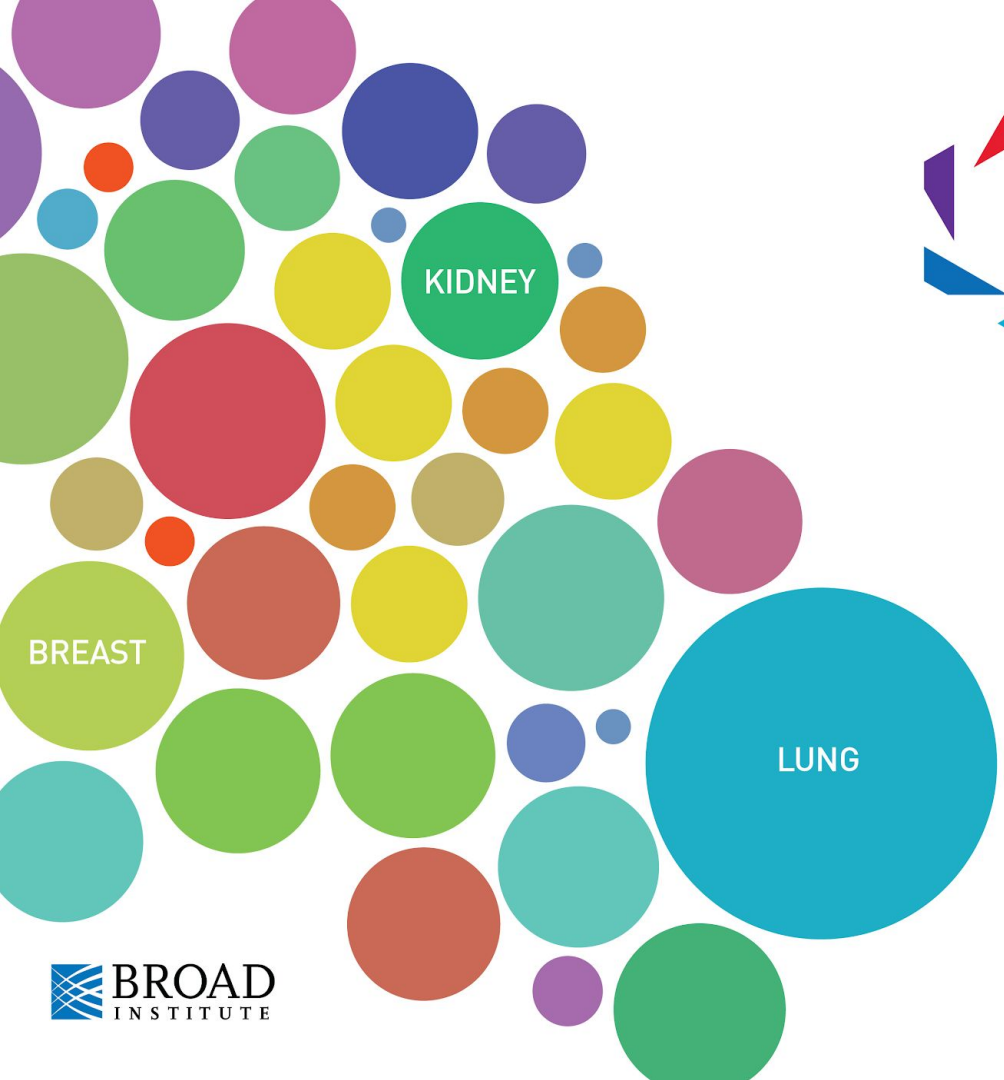

John Davis


Evan Lemire

Compound Management/ Analytical Chemistry


Anita Vrcic

Ryan Babcock
Joshua Sacher
Ryan Philbin



PRISM

MULTIPLEXED CELL LINE PROFILING

prism@broadinstitute.org
broadinstitute.org/prism

BROAD INSTITUTE