Poster #44

Building patient-specific predictors of drug responses from cell line genomics

Christopher W. Szeto Ph.D., Stephen C. Benz Ph.D., Charles J. Vaske Ph.D., Shahrooz Rabizadeh Ph.D., Patrick Soon-Shiong M.D.

NANTOMICS

Abstract

Here we demonstrate a method for using cell line genomics and drug-response data to build robust therapy outcome predictors. We present here a case study of using this system to predict Dasatinib response in glioblastoma multiforme (GBM) patients.

First, we infer pathway-level knowledge of cancer cell lines by integrating multiple genome wide assays into curated pathways (PARADIGM[1]). Next we use a high-throughput machine-learning library (topmodel) to build, analyze, and rank, thousands of candidate predictive models of drug response. Then, we ensure our most accurate predictive models can be extrapolated to patient samples by using statistical methods to bound our predictions with confidence intervals.

Using these methods we identify a predictive task, Dasatinib sensitivity prediction, that is especially predictable using genome-wide assays (77% accuracy in cross-validation). Among the datasets used to predict Dasatinib sensitivity, PARADIGM inferred pathway activities are more predictive than other data types. A statistically significant proportion of the cell lines that scored most highly sensitive were neural cell lines, suggesting some subset of gliomas may be uniquely responsive to Dasatinib.

We show that there is a proportional number of GBM patients that do conform to the Dasatinib-sensitive profile derived in cell lines.

Datasets & Tools

	Types	Number
Genomic datasets	CCLE expression CCLE copynumber CCLE expression paradigm CCLE copynumber paradigm CCLE expression & copynumber paradigm sanger expression sanger copynumber sanger expression paradigm sanger copynumber paradigm sanger_expression & copynumber paradigm	10 (8320 samples)
Drugs	17-AAG 681640 A-443654 A-770041 … WZ-1-84 XMD8-85 Z-LLNle-CHO ZM-447439	139
Classifiers	Linear kernel SVM First order polynomial kernel SVM Second order polynomial kernel SVM Ridge regression Lasso Elastic net Sequential minimal optimization Random forest J48 trees Naive bayes JRip rules HyperPipes NMFpredictor	13
Feature selections	Four levels of variance filters	4

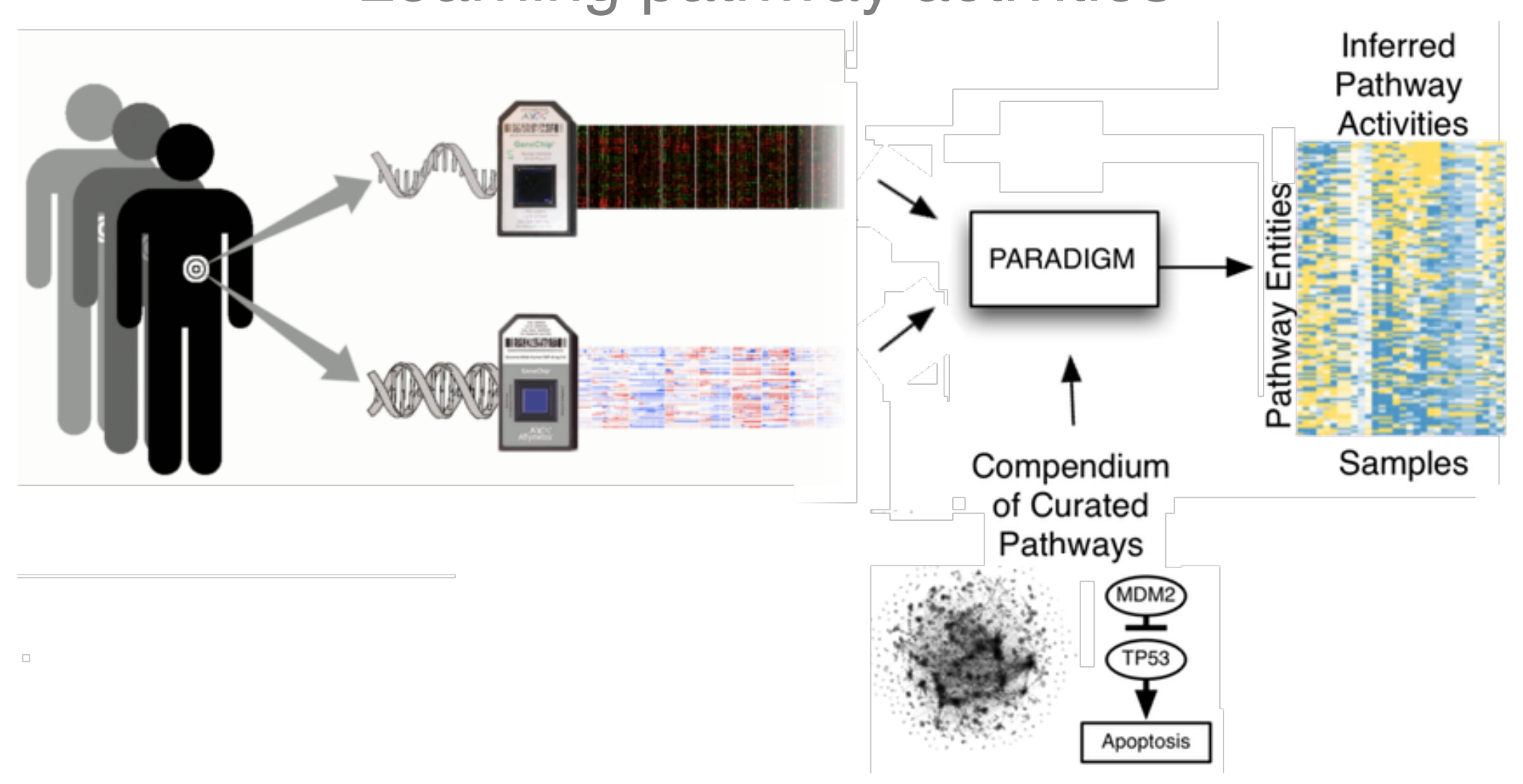
29,352 fully trained drug response models built

146,760 additional evaluation models built (5-fold CV)

176,112 total models analyzed

Machine Learning Pipeline

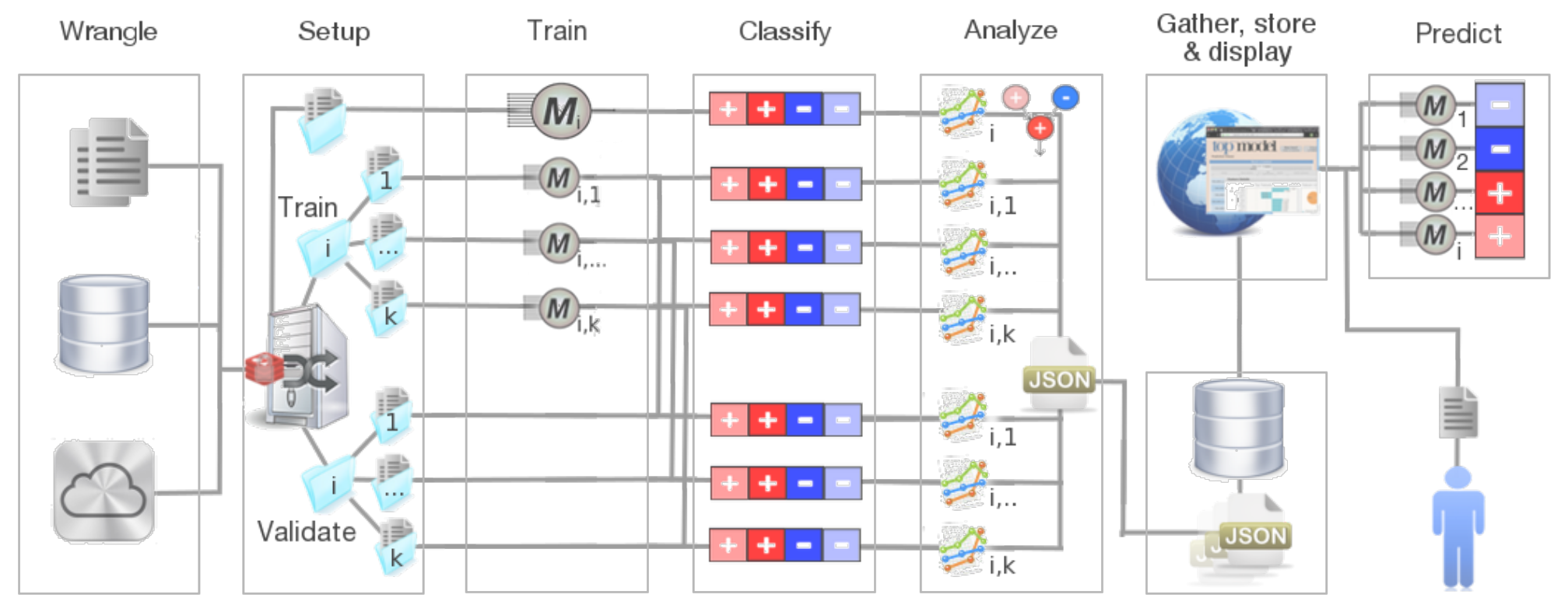
Learning pathway activities



Genomic-scale data are collected from individual cancer samples via microarray or sequencing technology. Several independent assays may be performed on the same samples; for example, both expression profiling and copy-number estimation.

These data are integrated in a factor-graph-based model of the central dogma, and linked together into networks of known pathways (PARADIGM). The most likely state for these networks given the -omics data evidence is estimated, and reported as inferred pathway activities.

Building response predictors



topmodel code can access data stored in various formats including flatfiles, databases, or from the cloud.

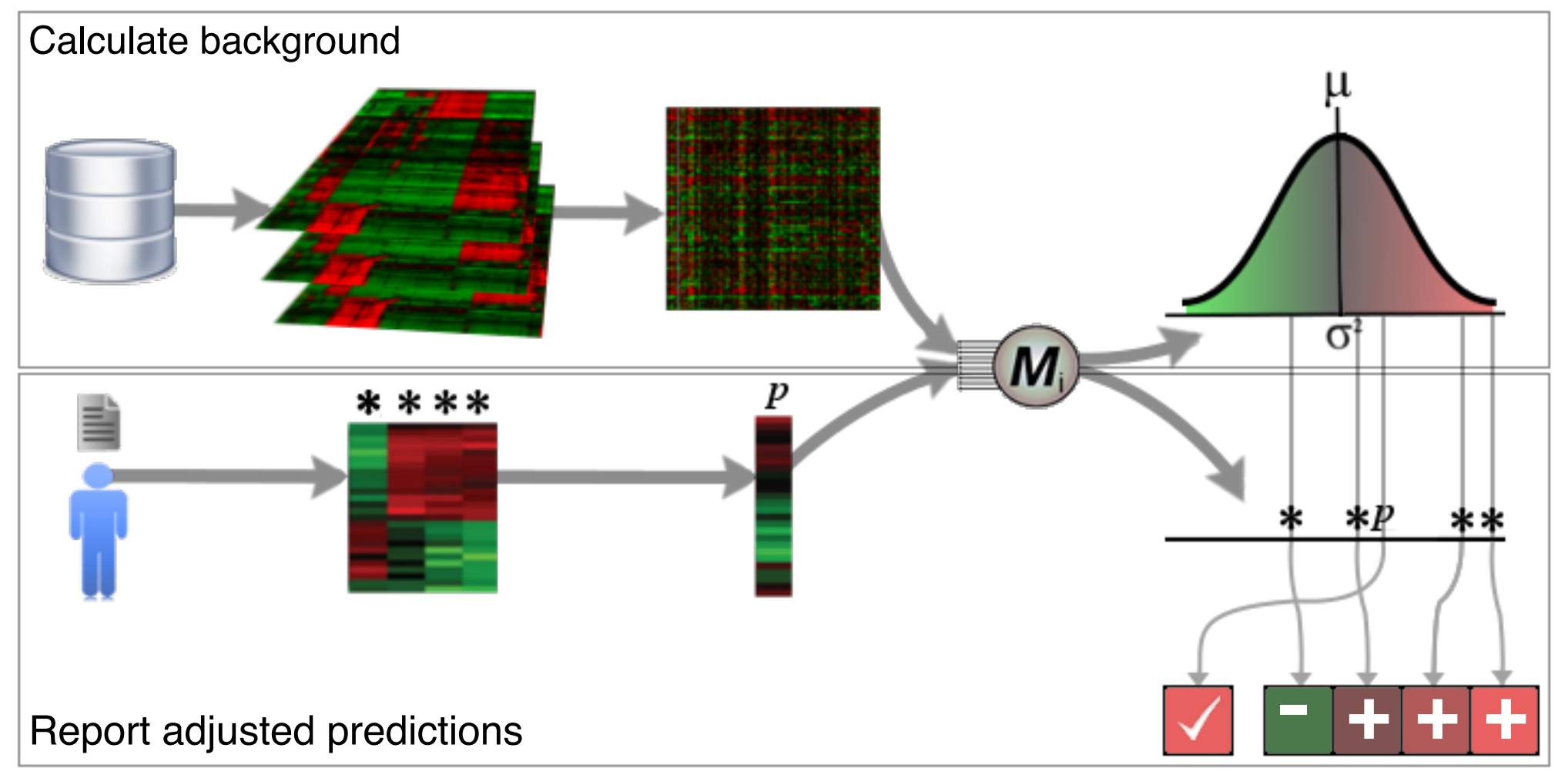
Data and metadata undergo multithreaded preprocessing, then split into training and validation folds. The data is then written to the file formats required by individual machine-learning packages.

Each classifier is trained on training data, and evaluated on validation data. This is performed on a cluster, increasing throughput enormously. In addition to the evaluation models, a fully-trained model is built upon the whole input dataset.

Each algorithm and its parameters are evaluated for accuracy. Those evaluations are collected into a unified file format and stored in a database.

The database of predictors can be queried, and the models used to make predictions on novel datasets.

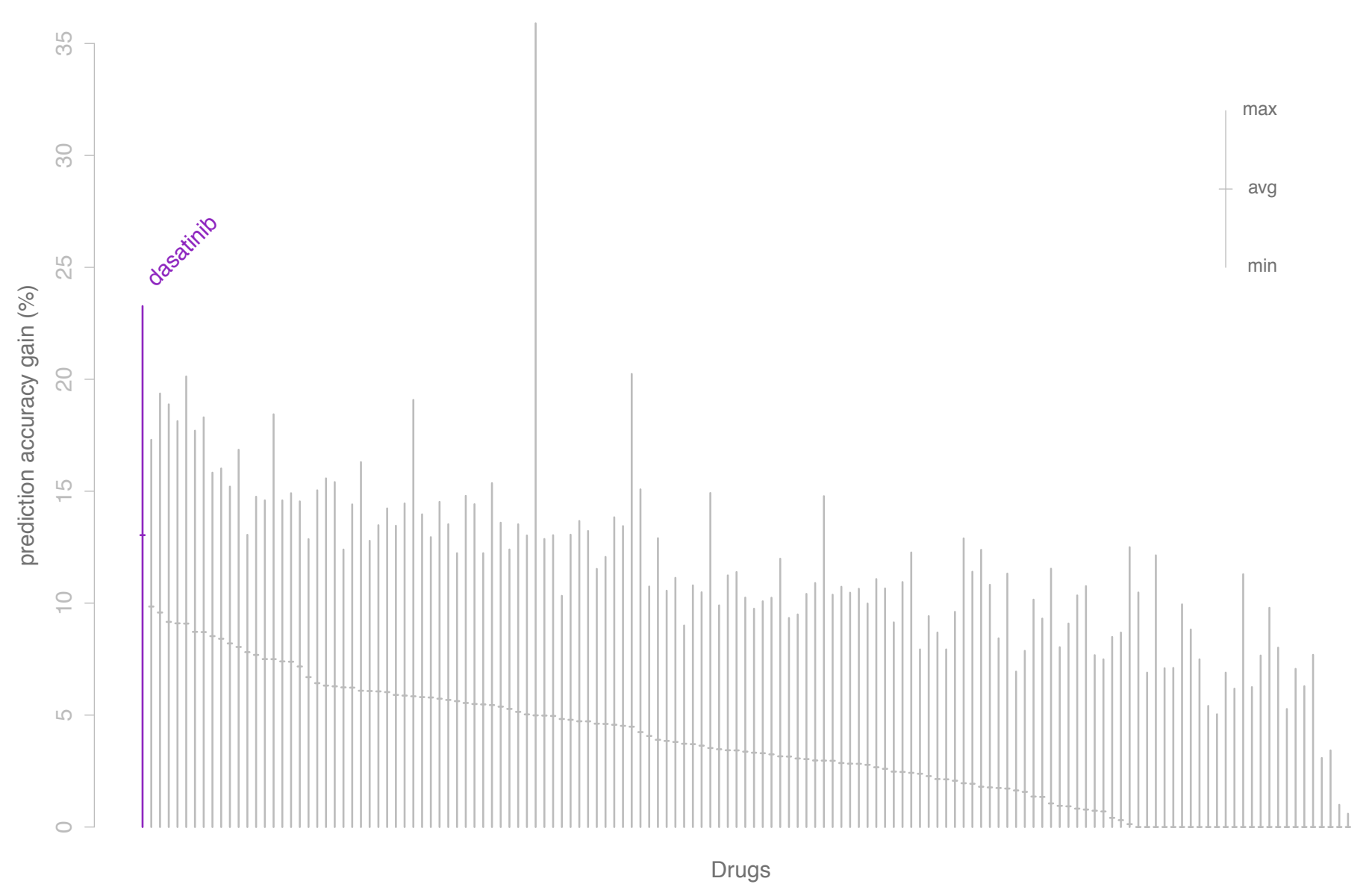
Making robust predictions



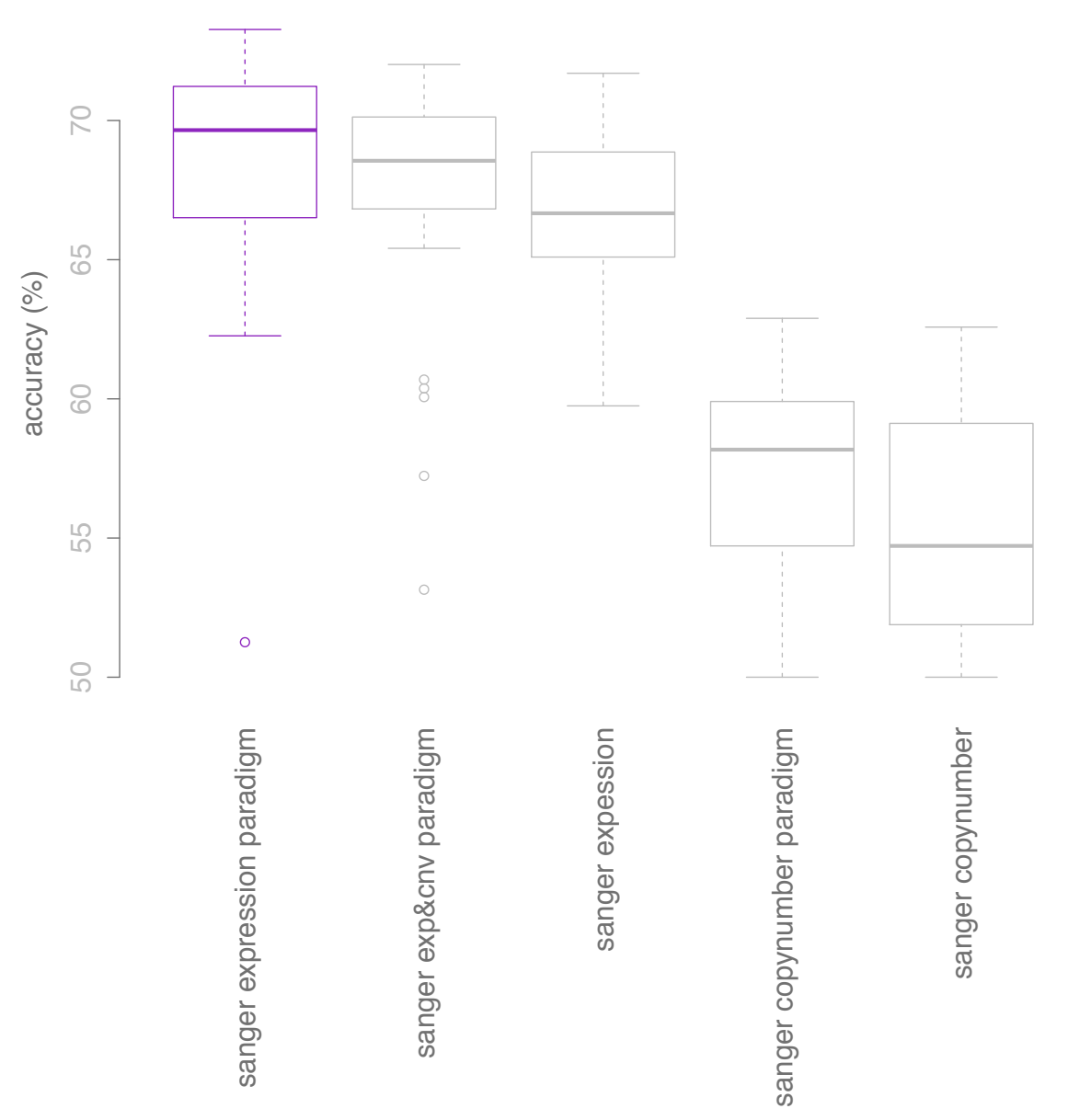
All datasets from one type (e.g. RNAseq expression) are queried from the topmodel database. One thousand randomly selected samples from across these datasets are independently permuted to create a randomized background dataset. The fully-trained top model is used to classify the randomized dataset. The mean and standard deviation for the randomized dataset is recorded for use as a null model.

On novel data, raw prediction scores are adjusted to standard-scores by comparing to the cached mean and standard deviation from the null model. Additionally, permutations (p) are generated from the novel data and classified to ensure the novel data and the background distribution are distributed similarly.

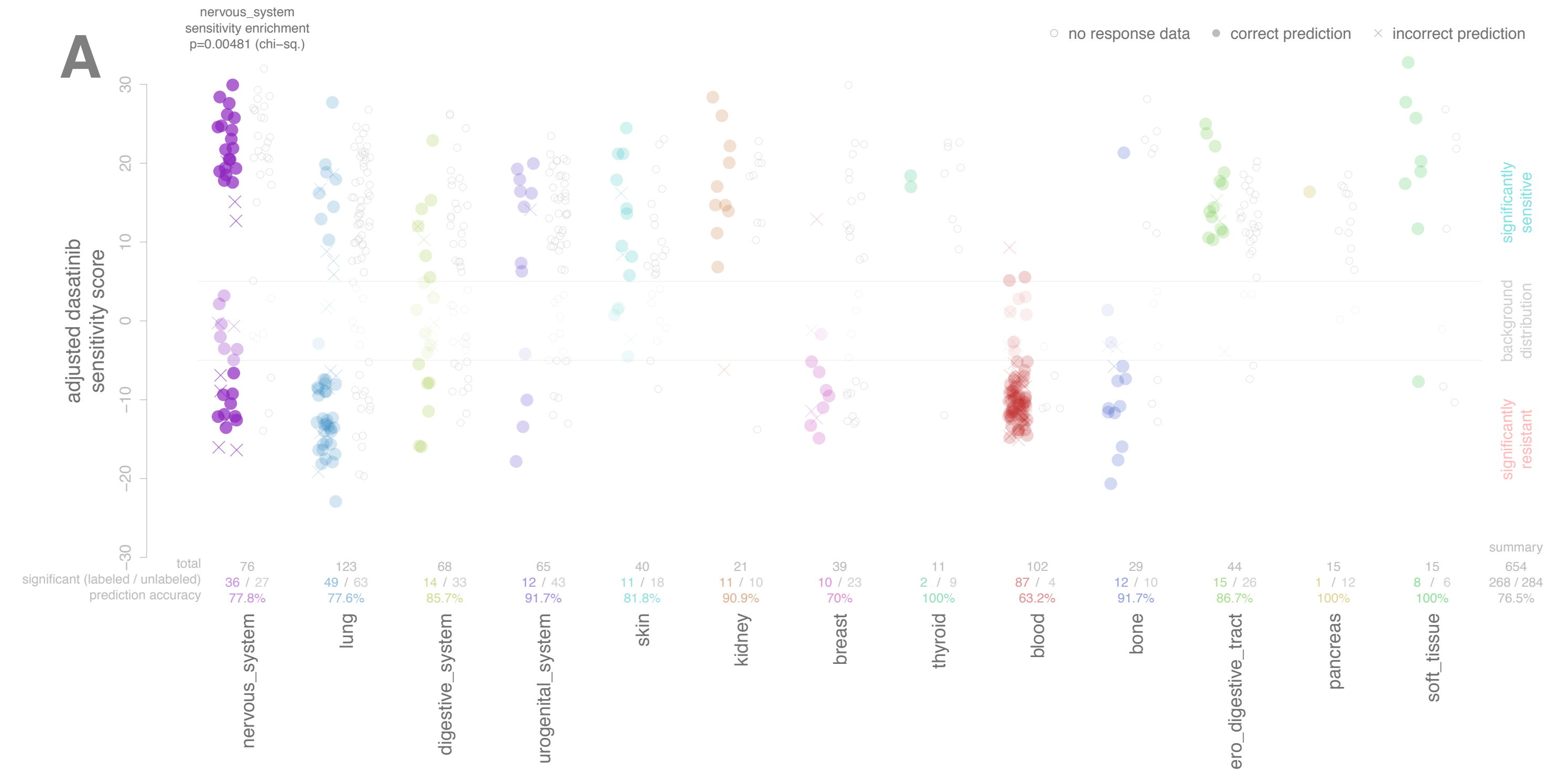
Results

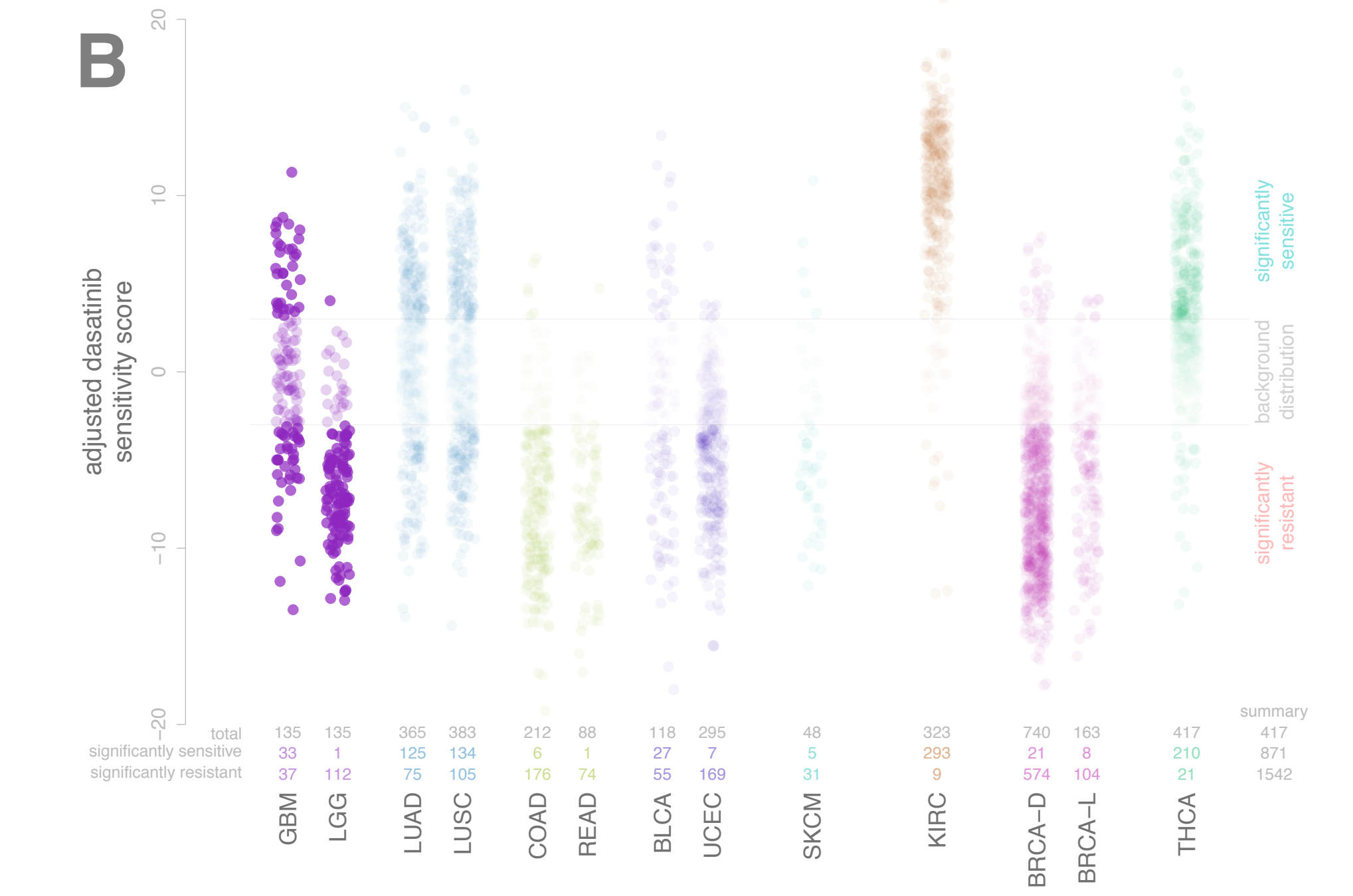


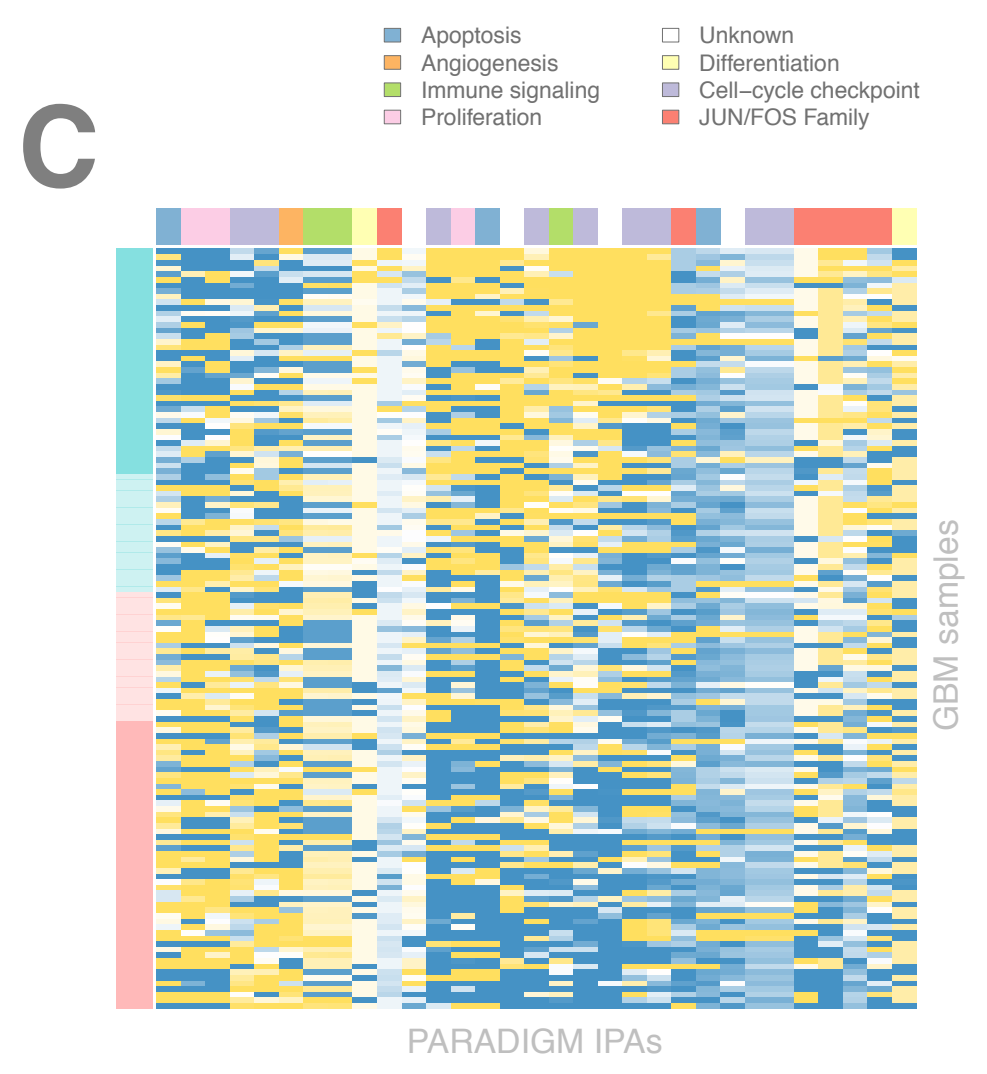
Drug predictability: Max, avg., and min accuracy gain over the majority classifier for each drug, sorted left-to-right by avg. accuracy. Drugs to the left are more consistently accurately predicted. The most consistently correctly predicted drug is Dasatinib.



Predictability of Dasatinib by different data types: PARADIGM inferred pathway activities based on expression was consistently the most predictive data type.







(A) Dasatinib sensitiiy by cell line type: Adjusted sensitivity scores for both labeled and unlabeled cell lines, and their respective accuracy in cross-validation. Note that cell lines correctly predicted as sensitive are enriched for neural system cell lines.

(B) Predicted Dasatinib sensitivity in primary patients: Adjusted sensitivity scores for TCGA samples in tissues that correspond to the training cell line panel. Note that tissue-effects behave similarly between cell line and patient data. Similarly to neural system lines, GBM samples are predicted to contain responder and non-responder subsets.

(C) TCGA GBM Dasatinib sensitivity diagnostic panel: PARADIGM inferred pathway activities projected onto the predictive model eigenvector. Note that a subset of genes associated with cell-cycle checkpoints are upregulated in sensitive samples. Conversely some genes associated with proliferation, angiogenesis and immune signaling are upregulated in resistant patients.

Discussion

Presented here is a rational, data-driven method for stratifying individual patients into responders and non-responders to oncotherapeutics.

This approach:

- uses best-in-class genome-wide assays, pathway analysis, and machine learning techniques in combination
- is not rate-limited by laboriously identifying drug-target interactions biochemically
- is agnostic to tissue of origin; Potential for rational drug reuse in novel contexts
- can recognizing when a prediction challenge is not surmounted by the given data

We present this approach with a potentially clinically-actionable demonstration: Identifying a subset of GBM patients who may respond to Dasatinib.

Dasatinib sensitivity in GBM xenografts (in combination with bevacizumab) has been demonstrated by others[2], leading to clinical trials in GBM patients being conducted.

At the time of writing 11 clinical trials are being conducted testing Dasatinib response in glioma patients. To the author's knowledge none of these trials use any biomarkers to recruit or stratify participants.

One such trial recently showed a very small proportion of GBM patients (3/50) have an increased 6mo PFS[3].

References

[1] C. J. Vaske, S. C. Benz, J. Z. Sanborn, D. Earl, C. Szeto, J. Zhu, D. Haussler, and J. M. Stuart. Inference of patient-specific pathway activities from multi- dimensional cancer genomics data using PARADIGM. Bioinformatics, 26:i237– 245, Jun 2010.

[2] D. Huveldt, L. J. Lewis-Tuffin, B. L. Carlson, M. A. Schroeder, F. Rodriguez, C. Giannini, E. Galanis, J. N. Sarkaria, P. Z. Anastasiadis. Targeting Src Family Kinases Inhibits Bevacizumab-Induced Glioma Cell Invasion. PLOSone, DOI: 10.1371/journal.pone.0056505, Feb 2013

[3] NCI, Dasatinib in Treating Patients With Recurrent Glioblastoma Multiforme or Gliosarcoma. ClinicalTrials.gov Identifier: NCT00423735, Last updated Dec. 2014