Pan-cancer predictions of MEK inhibitor sensitivity are partially driven by between-tissue differences

John P. Lloyd[1,2,3], Matthew Soellner[2,3*], Sofia D. Merajver[2,3*+], Jun Z. Li[1,3*+]

[1] Department of Human Genetics, University of Michigan, Ann Arbor, MI, 48109

[2] Department of Internal Medicine, University of Michigan, Ann Arbor, MI, 48109

[3] Rogel Cancer Center, University of Michigan

* Co-senior author

+ Co-corresponding author

Correspondence to:

Jun Z. Li

University of Michigan

E-mail: junzli@med.umich.edu

Telephone number: +1-734-615-5754

Sofia D. Merajver

University of Michigan

E-mail: smerajve@med.umich.edu

Telephone number: +1-734- 936-6884

Running title: Tissue effects in pan-cancer drug response

Keywords: Cancer biology, MEK inhibitors, drug response predictions, machine learning

ABSTRACT

Pharmacogenomic databases of drug response and genomics data in tumor cell lines allow the development of pan-cancer (i.e. tissue-agnostic) predictions of drug response. However, it is unclear whether between-tissue differences in both drug response and molecular characteristics bias pan-cancer drug response predictions. Using two datasets containing 346 and 504 cell lines with MEK inhibitor (MEKi) response data and RNA, SNP, and CNV data, we show between-tissue differences produced confounding effects that increase apparent performance in pan-cancer MEKi response predictions. We estimate that 45-58% of the variance in pan-cancer prediction scores was driven by accurate between-tissue predictions, rather than inter-individual predictions within a tissue. Nevertheless, MEKi response is predicted as well or better by pan-cancer approaches compared to tissue-specific approaches. Our results demonstrate strong between-tissue effects in tissue-agnostic drug response prediction models, but also highlight how drug response may invoke shared regulatory mechanisms although different cancer types arise in tissue-specific routes.

INTRODUCTION

Tailoring cancer treatment to the molecular vulnerabilities of individual tumors represents a central goal of precision oncology (1–3). Public repositories of drug response and functional genomics data have facilitated the identification of DNA mutations in well-characterized cancer genes that underlie variable drug response (3, 4), and have expanded to unbiased searches for such biomarkers across the human genome (4, 5). More recent drug response prediction analyses have further incorporated data modalities beyond DNA variants, e.g. gene expression, metabolomics (4, 6, 7). Drug response predictions based on molecular characteristics of primary tumors and auxiliary models are frequently performed within a single cancer type or subtype, as tumors from the same primary tissue are generally expected to harbor shared molecular vulnerabilities to anti-cancer agents – presumably as a result of tissue-specific oncogenic processes. Alternatively, drug response predictions can be developed by considering multiple cancer types jointly, i.e., through pan-cancer analyses. The success of pan-cancer prediction models would suggest diverse cancer types share molecular vulnerabilities despite tissue-specific mechanisms of cancer initiation, progression, or drug-response.

Pharmacogenomic databases of patient-derived cancer cell lines covering a broad range of cancer types represent reusable pre-clinical models for deep interrogation of the cell lines' innate characteristics that account for their drug response profiles (8–12). Using tumor cell line resources, DNA variants and gene expression levels have been used to develop pan-cancer drug-response prediction models via a wide variety of computational methods, including regularized regression (8, 13), random forests (14, 15), neural networks (16), network modeling (17, 18), quantitative structure-activity relationship analysis (19), and deep learning (20). These analyses have also offered many additional insights, including the importance of RNA expression for pan-cancer predictions (14), higher accuracy of multi-gene classifiers (i.e. gene panels) compared to single-gene classifiers (15), and the suitability of cell lines as *in vitro* mimics of primary tumors (13, 14).

Despite the reported progress of pan-cancer drug response predictions, it remains unclear whether between-tissue differences could have contributed to the apparent prediction performance. This concern

stems from the observation that some cancer types may be more sensitive to a certain drug than others, and their tissue-specific molecular properties may drive performance of a pan-cancer panel without accounting for inter-tumor differences within a cancer type. While cancer type-based therapeutic decisions are useful in clinical practice, there is often the additional need to predict variable response among tumors within a cancer type. We set out to examine the relative importance of between- and within-cancer type signals in pan-cancer drug response prediction models. Further, a systematic investigation into whether pan-cancer approaches are more accurate than drug response predictions tailored to specific cancer types or subtypes is currently unavailable. Here, we analyzed data from ten cancer types that were well-represented in two public cell line datasets (11, 12) and examined the between-tissue and within-tissue performances of pan-cancer drug response predictions for MEK inhibition. In addition to investigating tissue-level effects, we evaluated cross-MEK inhibitor prediction models and provided *in silico* replication by applying prediction models across datasets. Based on our results, we highlight key considerations for deploying pan-cancer drug response prediction frameworks.

RESULTS

MEK inhibitor sensitivity across cancer types in two public cell line datasets

To evaluate pan-cancer drug response predictions in pre-clinical tumor models, we utilized publicly available datasets of tumor cell lines from a prior publication (referred to as Klijn 2015) (11) and the Cancer Cell Line Encyclopedia database (CCLE) (12). Klijn 2015 and CCLE include 349 and 503 tumor cell lines from >10 cancer types (**Figure 1A**), respectively, that have drug response data and RNA and DNA characterization, with 154 cell lines in common between the two datasets (**Figure 1B**). Among the 5 and 24 drugs screened in Klijn 2015 and CCLE datasets, respectively, MEK inhibition was the sole mechanism of action in common, with one MEK inhibitor (MEKi) present in both datasets (PD-0325901; referred to as PD-901) and an additional MEKi unique to each dataset (Klijn 2015: GDC-0973; CCLE: Selumetinib). We therefore prioritized predicting MEKi response, as prediction models could be

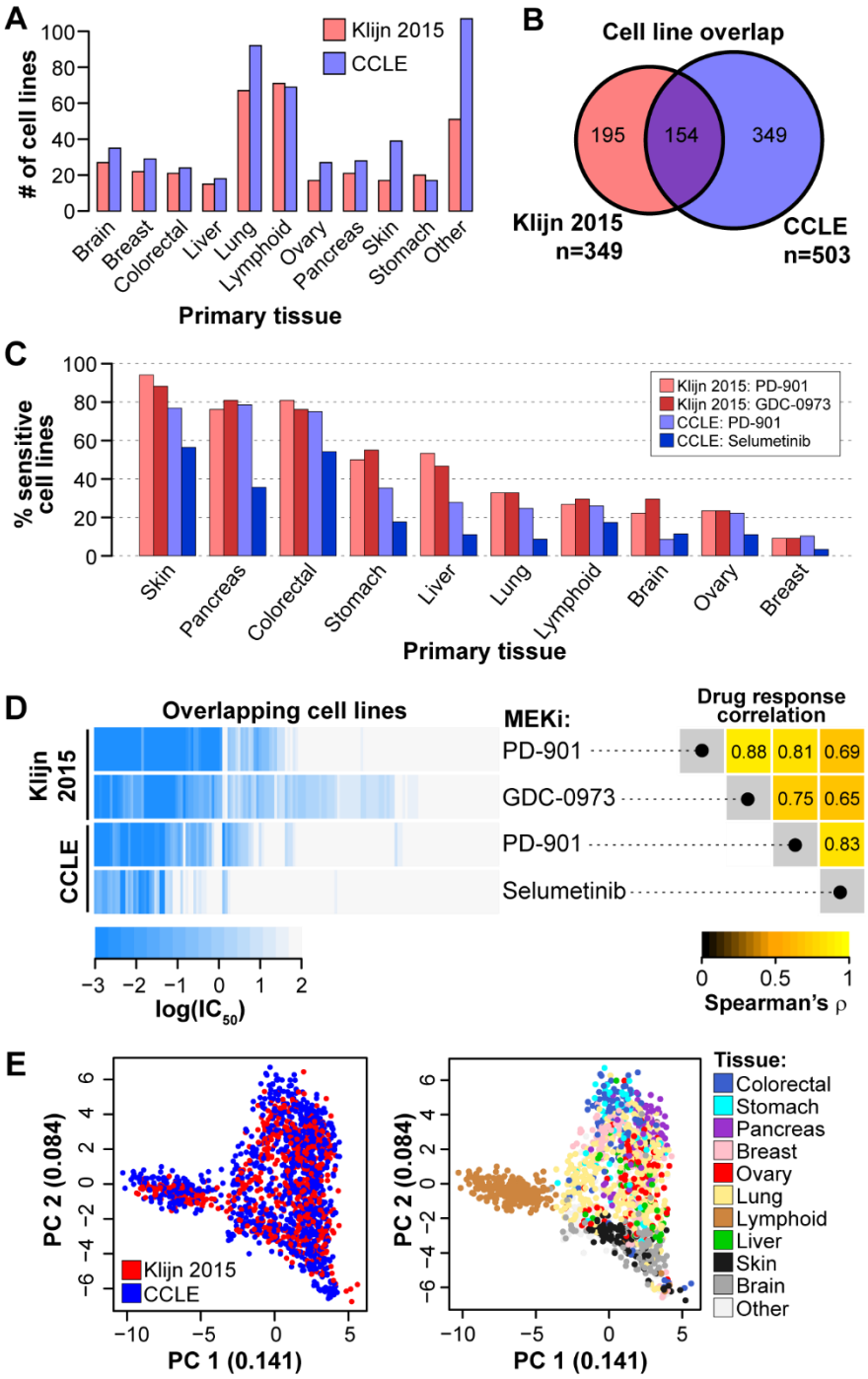

Figure 1 . Overview of pan-cancer cell lines and MEK inhibitor response in the Klijn 2015 and Cancer Cell Line Encyclopedia datasets. **(A)** Counts of tumor cell lines with MEKi response data from 10 prevalent cancer types (n ≥15 in both datasets). **(B)** Venn diagram of the overlap in cell lines with drug response and RNA expression data. The 154 cell lines common to both datasets were excluded from feature selection and model building steps, unless otherwise noted. **(C)** Proportion of MEKi-sensitive cell lines stratified by tissue. Cell lines were considered sensitive based on a threshold of $IC_{50} \leq 1$ nM. **(D)** Similarity in response to distinct MEK small molecule inhibitors across datasets. Left panel: Heatmap of MEKi response in $\log(IC_{50})$. Each column corresponds to a single cell line with drug response data for both MEKi in both datasets and columns were hierarchically clustered (dendrogram not shown). Right panel: Rank correlation across MEK inhibitors and datasets for the data in the left panel. **(E)** Scatterplots of the first two dimensions from principal component (PC) analysis on transcriptome data. Each point represents a patient-derived tumor cell line and is shaded based on dataset (left panel) or tissue-of-origin of the primary tumor (right panel).

evaluated for consistency across both Klijn 2015 and CCLE datasets. Moreover, MEK inhibition has shown promise for pan-cancer drug response predictions (14).

The molecular features of tumor cell lines in the Klijn 2015 and CCLE analyzed here included treatment naïve RNA expression levels (n=29,279 genes; data from RNA-seq), single nucleotide polymorphisms (SNPs; n=12,399 mutated genes; Klijn 2015: SNP microarray; CCLE: exome-seq), and

copy number variations (CNVs; n=4,578 mutated genes; SNP microarray). We mapped SNP and CNV features to the gene level, where they indicated whether an annotated gene contained a coding SNP (binary; gene mutated vs. gene not mutated) or was duplicated/deleted (gene amplified vs. gene deleted vs. no CNV). SNPs included only missense and nonsense mutations and SNPs and CNVs were filtered for known germline variants, but likely contained a mixture of germline and somatic mutations.

First, we compared MEKi responses of individual cell lines across cancer types and found clear differences. Skin, pancreatic, and colorectal cell lines were generally sensitive to MEK inhibition, while lymphoid, brain, ovary, and breast cell lines were generally resistant (**Figure 1C**). We also observed cancer types with more mixed responses to MEKi, including lung, stomach, and liver (**Figure 1C**). For cancer types with mixed response, i.e. wide range of drug responses observed within a cancer type, it is challenging to make treatment decisions based on cancer type alone and there is enormous value and opportunity to develop molecular predictors for individualized prediction. We further compared how similar the 154 cell lines in common between Klijn 2015 and CCLE responded to different MEK inhibitors. MEKi response in overlapping cell lines was highly correlated across distinct MEKi and between datasets (Spearman's ρ=0.65 to 0.88; **Figure 1D**). The observation that cell lines respond similarly to different MEK inhibitors indicates cross-MEKi predictions are feasible, provided the two compounds are chemically similar.

We also compared gene expression profiles between cancer types in each of the two datasets with principal component analysis (PCA) on standardized RNA expression levels (see Methods; **Figure 1E**). Cell lines from Klijn 2015 and CCLE were present in overlapping distributions in PC1/PC2 space (**Figure 1E**), a characteristic that we took advantage of to apply prediction models across datasets (see below). Cell lines from the same primary tissue tend to be present in similar regions of the PC1/PC2 space (**Figure 1E**) and have correlated RNA expression levels (**Figure S1**), highlighting that cell lines derived from the same primary tissue have similar transcriptomic features. Given the effects of primary tissue on both response to MEK inhibition (**Figure 1C**) and transcriptomic profiles (**Figure 1E; Figure S1**), it is plausible that between-tissue differences will be modeled in pan-cancer drug response

predictions that consider RNA data, even in cases where tissue labels are not included in a prediction model. The Klijn 2015 and CCLE datasets provide the needed resources to establish pan-cancer MEKi drug response predictions and investigate possible tissue-level confounding factors.

Pan-cancer machine learning predictions of MEK inhibitor sensitivity

With the data described in the previous section, one can establish the molecular features most strongly correlated with drug response, and further, develop predictors of drug response consisting of multiple molecular features. For the latter, we adopted two ways of using the drug response variables by either taking log(IC_{50}) as a continuous variable or dichotomizing drug response to a binary variable. For the former (continuous variables), we developed prediction models using regularized regression and random forest regression algorithms; and for the later (dichotomized variables), we adopted logistic regression and binary random forest algorithms. In total, four prediction algorithms were used. For each algorithm, we developed prediction models by considering the drug response data for one MEKi along with associated molecular features – e.g. response to PD-901 as assessed by Klijn 2015 compared with the Klijn 2015 molecular features (**Figure 2A**). One prediction model was built for each MEKi in both datasets, resulting in four prediction models per algorithm, and models were applied both within and between the Klijn 2015 and CCLE datasets. To assess performance, we established two validation sets: 1) for within-dataset predictions, model performance was reported for the 154 cell lines in common between the two datasets, which were excluded during all aspects of model generation and 2) for between-dataset predictions, performance was reported across all cell lines as no cell lines in the dataset the model was applied to were used during model development. An example schematic of model training while considering PD-901 in the Klijn 2015 dataset and model application both within and across datasets is shown in **Figure 2A**.

The resulting MEKi response prediction scores from the regularized and logistic regression algorithms for all combinations of training and testing data along with observed log(IC_{50}) values are shown in **Figure 2B**. Predicted MEKi response scores for both regression- and classification-based predictions are highly correlated (**Figure 2B**). Model performance was evaluated with rank correlation

(Spearman's ρ) between observed and predicted log(IC_{50}) values and by calculating area under the receiver operating characteristic curve (auROC) on the distributions of prediction scores between cell lines categorized as sensitive ($IC_{50} \leq 1\mu M$) and resistant ($IC_{50} > 1\ \mu M$). The resulting Spearman's ρ ranged from 0.43 to 0.64 and auROC ranged from 0.74 to 0.86 for regularized and logistic regression

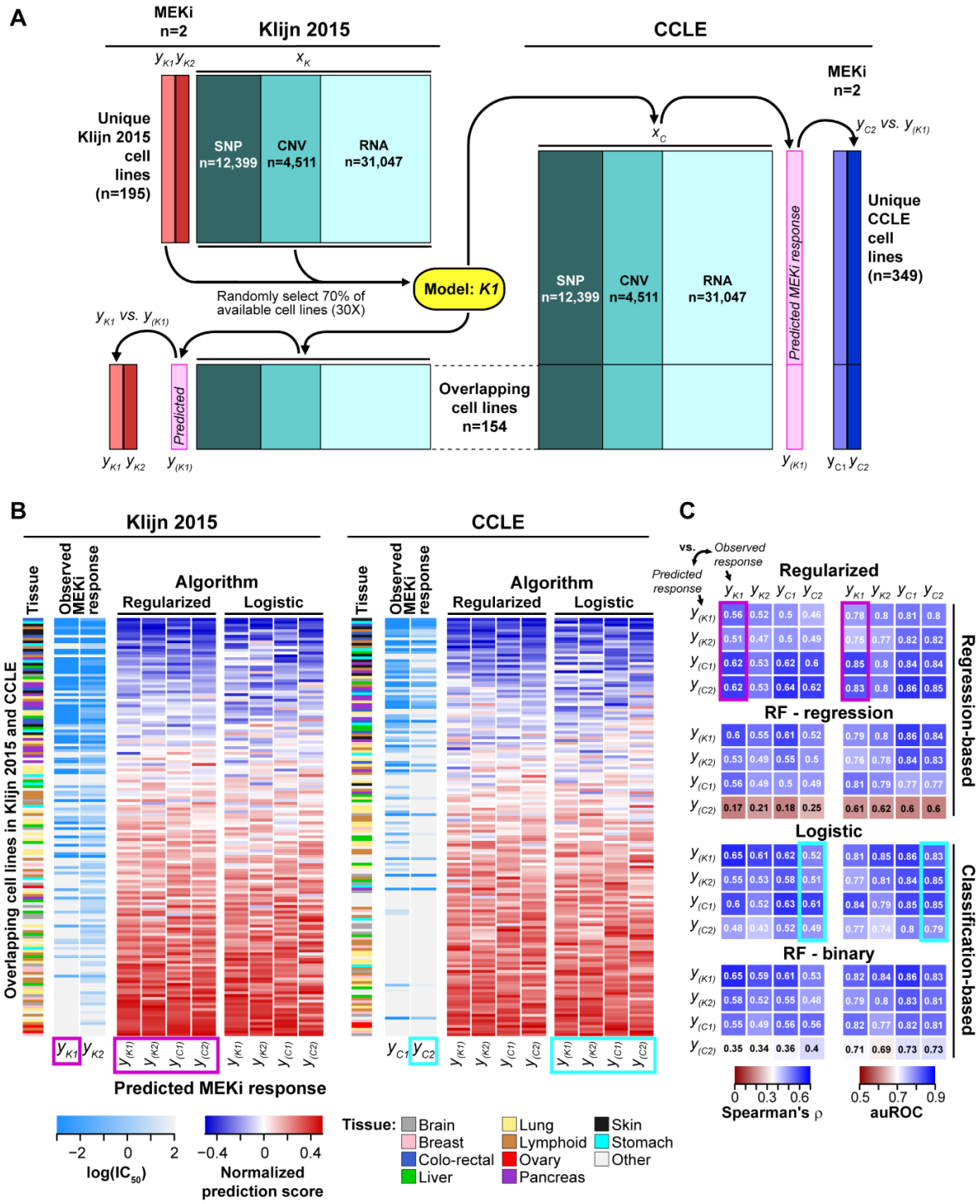

Figure 2. Pan-cancer machine learning predictions of MEKi response. **(A)** Schematic of machine learning experiments depicting the generation of a prediction model (yellow, *K1*) that considers data from the Klijn 2015 dataset: response to MEKi PD-901 (y_{K1}, light red) and DNA and RNA features (x_K). Cell lines in common between the two datasets were excluded from model building. Prediction models were built on 70% of available cell lines (selected randomly), repeated 30 times, and the final predicted drug response of a given cell line was calculated as the average of the 30 models. Note that the remaining 30% of cell lines that were not selected for model training were not used to assess model performance. The resulting prediction models are applied to within-dataset and across-dataset RNA and DNA data (x_K and x_C) to generate predicted drug response scores ($y_{(K1)}$). Predicted drug response values were then compared with observed drug response to evaluate model performance (within-dataset: y_{K1} vs. $y_{(K1)}$; cross-dataset: y_{C2} vs. $y_{(K1)}$). **(B)** Tissue-of-origin, observed MEKi response, and normalized MEKi predicted response from regularized regression and logistic regression algorithms for the 154 cell lines in common between Klijn 2015 (left panel) and CCLE (right panel). Two magenta boxes highlight scores from four regularized regression models applied to the Klijn 2015 data ($y_{(K1)}$ to $y_{(C2)}$) and compared with observed response to PD-901 in Klijn 2015 (y_{K1}). Two cyan boxes highlight the scores from four logistic regression models applied to the CCLE data ($y_{(K1)}$ to $y_{(C2)}$) and compared with observed response to Selumetinib in CCLE (y_{C2}). **(C)** Performance by Spearman's ρ (left panel) and area under the receiver operating characteristic curve (auROC; right panel) for all combinations of observed and predicted MEKi response and molecular features from both datasets, for two regression-based algorithms (regularized and random forest) and two binary classification algorithms (logistic and random forest). auROC was calculated by classifying cell lines with $IC_{50} \leq 1$ μM and $IC_{50} > 1$ μM as separate classes (sensitive and resistant, respectively). Magenta and cyan boxes indicate the performance metrics resulting from the color-matched comparisons in **(B). RF:** random forest.

algorithms, depending on training and testing data (**Figure 2C**). Our results are therefore consistent with those previous that show response to anti-cancer drugs can be successfully predicted based on pan-cancer datasets (8, 14, 15, 20, 21), and our framework allowed us to further assess cross-MEKi and cross-dataset predictions.

With a set of pan-cancer MEKi response predictions, we assessed the performance of individual machine learning algorithms from the perspectives of cross-MEKi and cross-dataset predictions. Random forest models (both regression and binary) that were trained while considering Selumetinib and CCLE RNA and DNA features exhibited lower performance relative to other algorithms and training sets ($y_{(C2)}$ in **Figure 2C**). Otherwise, regularized regression and logistic regression algorithms performed similarly (Mann Whitney U test, all $p \geq 0.86$; **Figure 2C**). We also found that models built with one MEKi predicted response to multiple MEK inhibitors similarly (U test, all $p \geq 0.11$; cross-MEKi examples: y_{K1} *vs.* $y_{(K2)}$ or y_{C2} *vs.* $y_{(C1)}$ in **Figure 2C**), and that prediction models applied across datasets performed similarly to those applied within datasets (U test, all $p \geq 0.06$; cross-dataset examples: y_{C1} *vs.* $y_{(K1)}$ or y_{C2} *vs.* $y_{(K2)}$ in **Figure 2C**). Accurate cross-dataset predictions provide strong *in silico* replication, as MEKi response and functional genomics data (RNA expression and DNA variants) in the Klijn 2015 and CCLE datasets were collected independently. Given the similarity in overall performance, we focus solely on cross-dataset predictions from the simpler regularized regression and logistic regression algorithms for analyses in the following sections.

Between- and within-tissue performance of pan-cancer MEKi response predictions

We evaluated how predicted MEKi responses differ across cancer types by averaging predicted values of cell lines and assessing performance within each of 10 cancer types. We found that mean observed and mean predicted log(IC_{50}) values for cell lines derived from the same primary tissue were highly correlated across tissues (ρ range: 0.88–0.98; **Figure 3A**; **Figure S2A**). Pan-cancer prediction models are therefore accurately capturing average drug responses of tissues, as tissues that tend to be sensitive or resistant to MEK inhibition exhibited generally low or high prediction scores, respectively. We next calculated within-tissue performance metrics from pan-cancer prediction models by considering only cell lines from a single cancer type (**Figure 3A**; **Figure S2B**). Based on hierarchical clustering of the resulting within-tissue prediction performances, we identify five tissues whose response to MEK inhibition was accurately predicted by pan-cancer prediction models: liver, ovary, lymphoid, stomach, and skin (mean $\rho = 0.59$; ρ range: 0.51-0.64; **Figure S2B**). The remaining five tissues (pancreas, brain, colorectal, breast, and lung) exhibited lower prediction performance (mean $\rho = 0.28$; ρ range: 0.11-0.40; **Figure S2B**). We conclude that whether a pan-cancer MEKi response prediction approach can be considered viable will vary based on which tissue is considered, with half of the tissues evaluated here achieving successful inter-individual predictions ($\rho \geq 0.51$).

Accurate between-tissue drug response predictions can improve performance metrics when cell lines from different tissues are considered together, even in cases where within-tissue predictions fail. To illustrate this phenomenon, we show the prediction performance for a combination of two tissues: brain and pancreas (**Figure 3B**). Within-tissue performances from pan-cancer predictions for both brain and pancreas cell lines are poor (both $\rho \leq 0.1$; **Figure 3B; Figure S2B**). However, prediction performance increased dramatically when the two tissues are considered together ($\rho = 0.48$; **Figure 3B**), as both observed and predicted log(IC_{50}) for pancreas cell lines are low on average, while both observed and predicted log(IC_{50}) for brain cell lines are high. Thus, accurate between-tissue predictions can erroneously inflate performance metrics when multiple tissues are considered together.

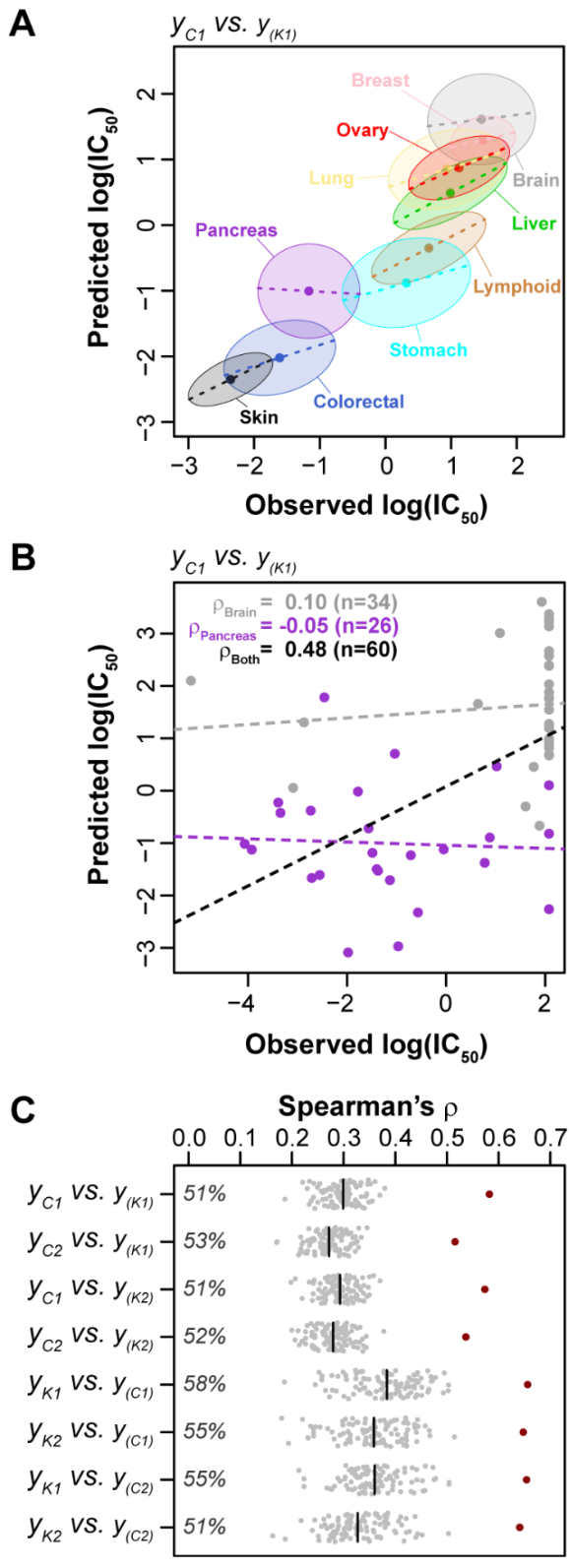

Figure 3. Performance of pan-cancer MEKi response predictions within cancer types. **(A)** Visualization of within-tissue mean observed and mean predicted log(IC_{50}) values from pan-cancer prediction models along with within-tissue pan-cancer prediction performance (**Fig. 2**). Dots at the center of ellipses indicate the mean observed and predicted log(IC_{50}) values for a given tissue. Maximum ellipsis length scales with the range of MEKi responses for a given tissue, where tissues with larger response ranges are associated with larger ellipses (e.g. stomach) while tissues with a smaller response range are associated with smaller ellipses (e.g. breast). The slope of dashed lines, as well as the tilt and width of associated ellipses, correspond to within-tissue prediction performance of the pan-cancer prediction model. Lines at angles closer to 45° and thinner ellipses (more cigar shaped) both indicate better prediction performances. **(B)** Performance of pan-cancer prediction models for a combination of brain and pancreas tissues. Rank correlation (Spearman's ρ) is shown for brain cell lines (gray), pancreas cell lines (purple), and both brain and pancreas combined (black). Dashed lines indicate lines of best fit. The increase in performance for the combined tissues (ρ=0.48) compared to each individual tissue (both ρ<0.10) highlights the confounding effect that accurate between-tissue predictions can exert on overall prediction performance. **(C)** Performance of predictions after shuffling existing pan-cancer prediction scores (**Fig. 2**) within a cancer type 100 times (gray points; e.g. a lung cell line was randomly assigned the prediction score of a different lung cell line). **Vertical black lines:** median performance from 100 iterations of shuffling scores within tissues. **Red points:** performance when scores are not shuffled. The percentages on the left side are calculated by median tissue-shuffled performance / non-shuffled performance × 100. The percentage is interpreted as the proportion of variance in pan-cancer MEKi response scores that can be attributed to accurate between-tissue predictions, rather than inter-individual predictions.

To estimate the effect of accurate between-tissue predictions on the overall performance of the pan-cancer predictions, we developed a set of MEKi response prediction scores based on tissue-of-origin,

rather than individual tumor characteristics. Tissue-based prediction scores were generated by shuffling existing MEKi prediction scores within tissues (i.e. a lung cell line randomly assigned the existing prediction score from another lung cell line). As a result, any tissue-scale effects within the prediction model will be maintained, while the effects from characteristics of individual tumors are necessarily lost. The performance of the resulting tissue-shuffled prediction scores across all 10 tissues were compared to non-shuffled prediction performance (**Figure 3C**). For regularized regression predictions, the mean rank correlation between observed log(IC_{50}) and tissue-shuffled scores is 0.32, while the mean non-shuffled correlation was 0.61. Similar results are also observed for logistic regression predictions (mean tissue-shuffled auROC=0.65; mean non-shuffled auROC=0.83; **Figure S2C**). Overall, depending on algorithm and training/testing set, we found that performance metrics from tissue-shuffled predictions are 42% to 58% lower than those from non-shuffled predictions (**Figure 3C**; **Figure S2C**). We therefore conclude that roughly half of the performance of the pan-cancer MEKi response predictions developed here can be attributed to accurate between-tissue MEKi response predictions, rather than prediction of inter-individual differences within a tissue.

Comparisons of pan-cancer and tissue-specific MEKi response predictions

Consistent with previous studies (8, 14, 15, 20, 21), we demonstrated that pan-cancer predictions of drug sensitivity can be successful (**Figure 2**). However, it is unknown whether pan-cancer prediction models can outperform those generated by considering a single cancer type. To address this question, we carried out a systematic comparison of tissue-specific prediction models (i.e. trained and tested with cell lines from a single cancer type) and pan-cancer prediction results for the same 10 cancer types (**Figure 4**). As models were developed within a single cancer type, fewer cell lines are available compared to pan-cancer models. Due to these sample size constraints, we considered only regularized regression prediction models trained on CCLE data, which includes more cell lines, with models applied cross-dataset to Klijn 2015 data.

Based on these comparisons, we found that pan-cancer MEKi response predictions performed similarly or better than tissue-specific predictions for 9 out of 10 tissues (**Figure 4A**–**D**). Pancreas was the only tissue for which tissue-specific prediction models outperformed pan-cancer models and the

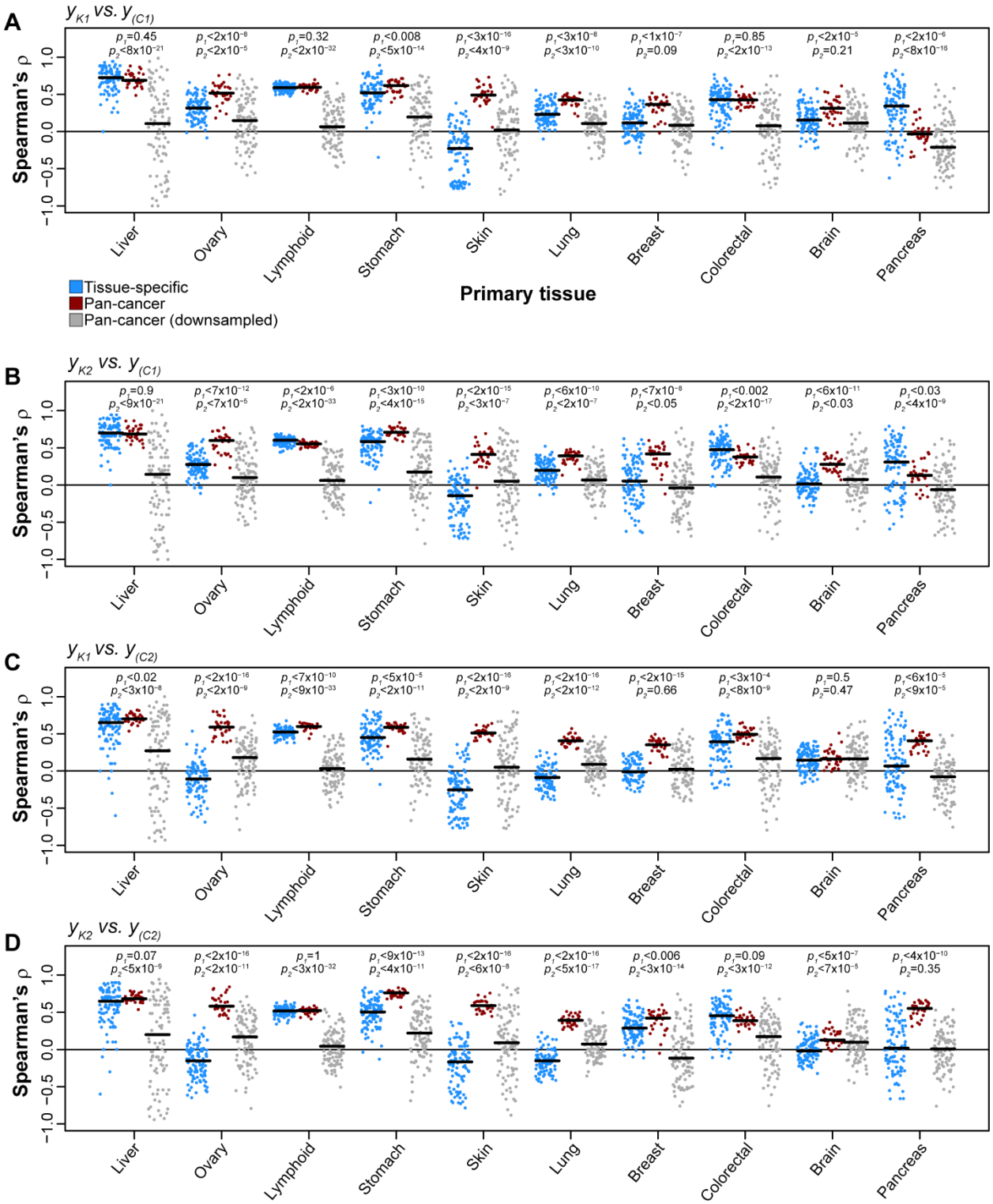

Figure 4. Comparisons of pan-cancer and tissue-specific MEKi response prediction models. For each tissue with ≥15 cell lines in both Klijn 2015 and CCLE datasets, a tissue-specific prediction model was trained and tested by considering only cell lines from a given tissue. Regularized regression prediction models were trained on a random selection of 75% of cell lines in CCLE data of a given tissue type (repeated 100 times) and applied to Klijn 2015 cell lines of the same tissue type. Performance was reported using rank correlation between observed and predicted MEKi drug response for each iteration (blue points). Rank correlation for the 30 iterations of MEKi response predictions based on pan-cancer prediction models (**Fig. 2**) for a given tissue are indicated with red points. Pan-cancer prediction models were trained and tested using many more cell lines than tissue-specific prediction models. A set of pan-cancer prediction models were generated that downsampled the available pan-cancer cell line sets and tissue-specific training sets to equal the sample sizes used for building and evaluating tissue-specific prediction models (gray points). **(A-D)** Results are shown for all combinations of models trained using CCLE data and applied to Klijn 2015 data. **Black horizontal lines**: median performance for a given distribution. $\mathbf{p_1}$: P-value from a Mann-Whitney U test of the difference between tissue-specific and pan-cancer prediction performances for a given tissue. $\mathbf{p_2}$: P-value from a U test of the difference between tissue-specific and downsampled pan-cancer prediction performances.

performance of pancreas predictions was generally low, and varied based on training data (**Figure 4A**–**D**). An important caveat is that pan-cancer prediction models were trained while considering 4- to 28-fold more cell lines than tissue-specific prediction models (e.g. training set size of 68 lung and 12 liver cell lines for tissue-specific models). To account for differences in sample sizes between pan-cancer and tissue-specific predictions, we trained an additional set of pan-cancer prediction models that considered downsampled sets of cell lines equal in size to the training set of the tissue-specific models ("Pan-cancer (downsampled)" in **Figure 4**). These downsampled pan-cancer prediction models rarely outperformed tissue-specific predictions and then, only in cases where tissue-specific models perform particularly poorly (e.g. skin; **Figure 4**). Unsurprisingly, downsampled pan-cancer prediction models performed worse (95th percentile ρ=0.2) than full pan-cancer prediction models (95th percentile ρ=0.67; **Figure 4**). Overall, we find that the accuracy of pan-cancer prediction models relative to tissue-specific models highlights the potential of pan-cancer approaches as tools for precision oncology. However, we emphasize that when pan-cancer prediction models outperform tissue-specific models, it is due to sample sizes available from large pan-cancer datasets.

Estimating sample sizes required for optimal prediction performance

Lower performance among downsampled pan-cancer prediction models (**Figure 4**) raises the question: how many cell lines are needed for optimal predictive performance? To estimate needed sample sizes, we developed regularized regression prediction models from a series of randomly downsampled sets of cell lines ($n_{\text{cell lines}}$ = 20-300; step size = 10; $n_{\text{iterations/step}}$ = 30). Performance of downsampled models was evaluated across cell lines of all tissues (pan-cancer) and within the five tissues that are well-

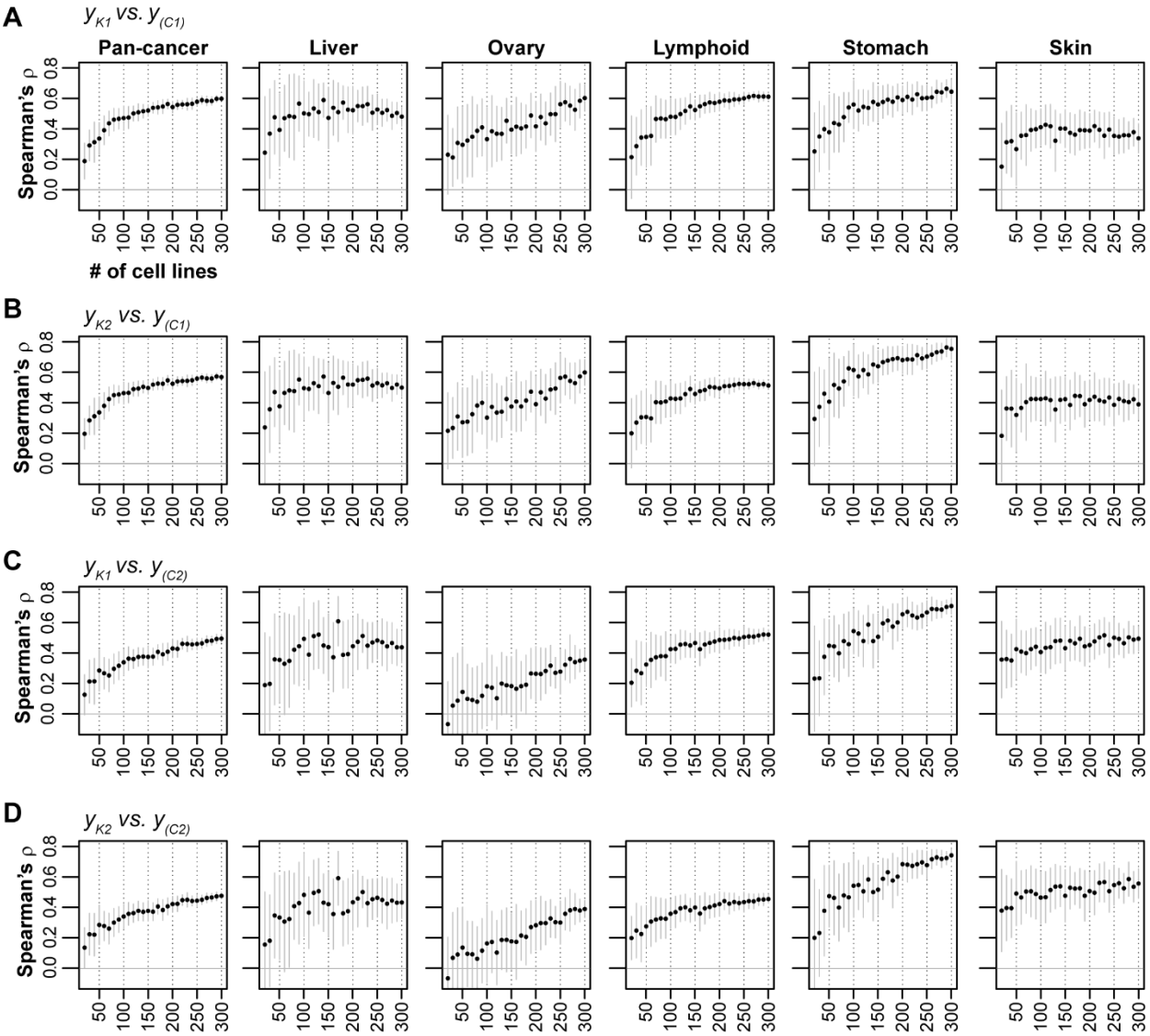

Figure 5. Performance of pan-cancer MEKi response predictions based on downsampled sets of cell lines. Cell lines in the CCLE dataset were randomly downsampled to sample sizes ranging from 20 to 300 in 10 cell line increments (x-axis; $n_{iterations}$=30). Regularized regression prediction models were trained on downsampled cell lines sets and applied to cell lines in the Klijn 2015 dataset. Mean ranked correlation (Spearman's ρ) between observed and predicted logged IC_{50} values across the 30 iterations for each sample size are plotted for performance in predicting Klijn 2015 MEKi drug response. **(A-D)** Results are shown for all combinations of models trained using CCLE data and applied to Klijn 2015 data. **Solid gray vertical lines**: ±standard deviation.

predicted by pan-cancer prediction models (liver, ovary, lymphoid, stomach, and skin; **Supplemental Figure 2B**). For overall pan-cancer prediction performance, we observed an inflection point of largely diminishing returns between 70 and 100 cell lines, depending on the training data (70 cell lines: **Figure 5A,B**; 100 cell lines: **Figure 5C,D**). However, including additional cell lines further increased performance, albeit with marginal increases beyond n=70. For within-tissue predictions, inclusion of more than 100 cell lines led to diminishing performance returns for liver, lymphoid, stomach, and skin

cell lines. Including additional cell lines led to marginally increased performance for lymphoid and stomach cell lines, while predictions for cell lines derived from liver and skin cell lines saturate at n=100 (**Figure 5**). Notably, performance for ovary cell lines showed no inflection point of diminishing returns (**Figure 5**), suggesting ovarian cancers may be well-positioned to benefit from drug response predictions generated from large, pan-cancer datasets. Overall, we find that ≥70-100 samples are needed to provide robust pan-cancer prediction performance in most cases.

DISCUSSION

Pan-cancer drug response predictions that identify shared molecular vulnerabilities in heterogeneous cancer types have the potential to contribute to the advancement of precision oncology approaches. In this study, we highlight tissue-level effects that can bias the interpretation of performance for pan-cancer drug response predictions, show that pan-cancer prediction models provide additional information beyond these tissue-level effects, and demonstrate that pan-cancer prediction models typically outperform tissue-specific prediction models, although the performance advantage usually disappears at comparable sample sizes. We highlight that when pan-cancer approaches to drug response prediction are carefully implemented and interpreted, they are a powerful tool for selecting effective anti-cancer therapies, not only based on the cancer type but also for individual variability of drug response within a cancer type. This capability is especially valuable for the cancer types that manifest a wide range of responses to MEK inhibition (e.g. lung, liver, and stomach).

Using public pharmacogenomic datasets, we established robust sets of pan-cancer drug response predictions focused on MEK inhibition. We highlight the accuracy of cross-MEKi predictions, which demonstrates MEKi response prediction models are largely agnostic to which MEK inhibitor is being considered, provided they are chemically similar. Ultimately, MEKi prediction models or biomarker panels could reasonably be applied to a preferred MEKi. We also note that the datasets considered here included lymphoid cell lines, a non-epithelial tissue type that is biologically and clinically distinct from the majority epithelial-origin cell lines that dominate both datasets. Importantly, pan-cancer prediction

models can accurately predict this highly distinct tissue type. This suggests lymphoid cell lines may respond to MEK inhibition by invoking biological pathways similar to those seen in epithelial cells and therefore genetic or epigenetic variability in other cancer types can account for response in lymphoid cancers.

Although the pan-cancer MEKi response predictions were established in a tissue-agnostic manner (i.e. tissue labels are not provided during the generation of prediction models), we show that between-tissue effects are present in the resulting prediction models. First, prediction performance is variable between tissues, and as a result, a pan-cancer prediction score should not be used to inform treatment for an under-performing tissue. For example, the pan-cancer MEKi response predictions for pancreas, brain, or colorectal tissues provide little more information than knowing the general drug sensitivity of the cancer type. Critically, accurate between-tissue drug response predictions are contributing to the overall success of pan-cancer prediction, even in cases where within-tissue predictions fail (**Figure 3B,C**; **Figure S2C**). As a result, reports of overall prediction performance across all cancer types in combination will include both between-tissue and within-tissue effects. To account for the between-tissue effects, additional reporting of prediction performance within cancer/tissue types is recommended.

The utility of pan-cancer drug response prediction models will depend on whether they can outperform cancer type-specific prediction models. Based on a systematic evaluation, we found that pan-cancer prediction models of MEKi response perform similarly or better than tissue-specific prediction models when the former are allowed to use the large sample sizes available from public datasets, indicating that pan-cancer predictions could be useful for precision oncology approaches. However, we further show that pan-cancer prediction models trained with sample sizes comparable to cancer-specific models offer almost no advantage. We emphasize that large public data repositories could offer useful drug response predictions even in cases where a specific cancer type is underrepresented or absent. When assessing needed sample sizes, we show that inclusion of more than 100 cell lines frequently led to diminishing performance returns. This knowledge is useful when designing future experiments and deciding how to invest resources to investigate many other cancer models and drug families.

The results described here are based on an important drug class (MEK inhibitors) and the reproducibility for tissue-scale effects should be tested for additional drug classes. However, we posit that the confounding effects arise from tissue type influencing both drug response and molecular characteristics (e.g. RNA levels and DNA variants). As a result, we emphasize dual considerations for pan-cancer drug response prediction approaches. First, pan-cancer approaches are useful for uncovering common biological mechanisms shared by many cell types when challenged with a drug. Second, tissue-level effects will likely exert confounding effects and potentially inflate naïvely reported performance. To maximally benefit, pan-cancer approaches must be carefully designed and evaluated to account for tissue-level effects, and existing pan-cancer drug response predictions reported across a pan-cancer dataset should therefore be interpreted with caution.

With the ever-increasing availability of large pharmacogenomic datasets and introduction of basket clinical trials that select patients based on tumor features rather than cancer type, big data approaches that seek pan-cancer prediction models and biomarker sets will be positioned to match patients with effective treatments. Conceptualizing between-tissue and within-tissue differences of the statistical signal from drug response predictions is essential for contextualizing "precision" in precision oncology, allowing the joint pursuit of both group-based (i.e. tissue class) and individual-based treatment decisions. Thus, accounting for tissue-level effects when designing pan-cancer drug response prediction models or biomarker identification experiments will be required to develop truly personalized treatment based on the characteristics of individual tumors.

METHODS

Cell line annotations and drug response data

Drug response and molecular characterization data were retrieved from two sources: 1) the supplemental material of a prior publication (Klijn 2015) (11) and 2) a tumor cell line database (CCLE) (12). For Klijn 2015, cell line annotations, including identifiers and primary tumor types, were retrieved from

"Supplementary Table 1" in (11) and annotations for CCLE were retrieved from "Supplementary Table 1" in (12). Cell line identifiers were connected across datasets and data modalities (i.e. drug response, RNA expression, and DNA variants) by accounting for differential use of upper/lower case characters and/or non-alphanumeric characters; alphabet characters in cell line identifiers were converted to upper case and identifiers were stripped of non-alphanumeric characters. Klijn et al (11) excluded 65 cell lines from analysis due to SNP similarity to other cell lines or uncertainty in cell line origin; we also excluded these cell lines from our analyses ("In Final Analysis Set" column = FALSE; "Supplementary Table 1" in (11)).

For Klijn 2015, we retrieved drug response data from "Supplementary Table 13" in (11). Drug response was provided for 5 total drugs as IC_{50}, the dose at which 50% of cells are non-viable. Lower IC_{50} values indicate greater sensitivity to a drug than higher values. Among the 5 drugs, we focused on two MEK inhibitors: PD-0325901 (referred to as PD-901) and GDC-0973. For CCLE, drug response data were downloaded in the file "CCLE_NP24.2009_Drug_data_2015.02.24.csv" (database file date: 24 February 2015), which included IC_{50} data for 24 anti-cancer drugs. We used drug response data for two MEK inhibitors in CCLE: PD-901 and AZD-6244 (trade name: Selumetinib). The natural log of IC_{50} values were used for feature selection, parameter sweeps, and model generation and evaluation.

RNA expression data

For Klijn 2015, we analyzed expression data from RNA-seq available from two supplemental datasets, designated as protein-coding ("Supplementary Data 1" in (11)) and non-coding ("Supplementary Data 2" in (11)). In all, there were 25,984 coding genes and 21,398 non-coding genes after excluding genes that were invariable in expression across all 675 cell lines with RNA data. Expression levels were provided as variance stabilization-transformed read counts from the DESeq package (22). We further standardized RNA expression data for each gene by linearly scaling values across cell lines to a range between 0 and 1 and shifting the scaled values by subtracting the scaled mean. Gene identifiers in the protein-coding table were converted from Entrez Gene ID to ENSEMBL format using the org.Hs.eg.db package (23) in R, while gene identifiers in the non-coding table were already provided in ENSEMBL

format. Once gene identifiers for the Klijn 2015 protein-coding and non-coding datasets were both in ENSEMBL format, we removed genes from the non-coding dataset that were found in the coding dataset. Following processing described above, the coding and non-coding expression data were merged and treated as a single dataset for downstream analysis.

For CCLE, we retrieved expression data from RNA-seq analysis as a single table from the CCLE database (file name: "CCLE_RNAseq_genes_rpkm_20180929.txt"; database file date: 29 September 2018). Expression levels for 55,414 genes (after excluding genes that were invariable in expression across all 1,019 CCLE cell lines) were provided as reads per kilobase of transcript per million mapped reads (RPKM). The natural log of CCLE RNA values was taken by log(RPKM+0.001), as the data exhibited a long right tail with a small number of extremely high expression values. CCLE gene identifiers for RNA expression data were provided in ENSEMBL format. CCLE RNA expression data was standardized as described above for the Klijn 2015 dataset (i.e. scaling and shifting). Standardization was performed independently for both Klijn 2015 and CCLE datasets. For downstream analysis (e.g. feature selection, model building), we included only the 31,047 genes with RNA expression data present in both Klijn 2015 and CCLE RNA-seq datasets, while genes present in only one dataset or the other were excluded.

DNA variant data

For Klijn 2015, we retrieved single nucleotide polymorphism (SNP) and copy number variation (CNV) mutation data from the associated supplementary material (SNPs: "Supplementary Data 3" in (11); CNVs: "Supplementary Data 4" in (11)). Klijn 2015 SNP and CNV data were identified from SNP microarray analysis (Illumina HumanOmni2.5_4v1). The provided SNPs had been previously filtered to include only missense and nonsense mutations. In all, there were 127,579 unique SNPs across 14,375 genes for all 675 cell lines with SNP calls. Because any given SNP was rare in the Klijn 2015 dataset, SNPs were mapped up to the gene level; genes then served as binary features (i.e. gene mutated / gene not mutated) without regard to which SNP a gene contained. For Klijn 2015, CNV data were provided based on a gene-by-gene basis as ploidy-adjusted copy number calls from the SNP array data using the PICNIC (24) and GISTIC (25) algorithms. We then classified the provided ploidy-adjusted copy number calls as

amplifications (coded as 1) or deletions (coded as -1) based on thresholds described by Klijn et al (11) (amplifications: ≥1; deletions: ≤-0.75). In total, 18,396 gene amplifications and 42,371 gene deletions for 11,864 unique genes were present across 668 cell lines with CNV calls. Gene identifiers for both the SNP and CNV data in Klijn 2015 were converted from Entrez Gene ID to ENSEMBL format using the org.Hs.eg.db package (23).

For CCLE, we retrieved whole exome sequencing SNP data from the CCLE database (file name: "CCLE_mutation_DepMap_18q3_maf_20180718.txt"; database file date: 18 July 2018). Of the 1,203,976 total SNP calls present, only the 705,056 SNPs annotated as either missense or nonsense mutations were included in further analysis – consistent with the SNPs available in the Klijn 2015 dataset. As described above for the Klijn 2015 dataset, SNPs were mapped up to the gene level where the mutation status of a gene served as a binary feature. Gene identifiers were converted from Entrez Gene ID to ENSEMBL format using the org.Hs.eg.db package (23). CCLE CNV data were downloaded from the CCLE database (file name: "CCLE_MUT_CNA_AMP_DEL_binary_Revealer.gct"; database file date: 29 February 2016). CNVs were provided based on calls from SNP microarray data (Affymetrix Genome-Wide Human SNP Array 6.0) and the GISTIC algorithm (25). The downloaded CNV file provided gene-based amplification and deletion calls, which were assembled into a gene-by-cell line matrix with amplifications coded as 1, deletions coded as -1, and no copy number variation coded as 0. In total, 691,866 gene amplifications and 948,159 gene deletions were called across 1,030 cell lines. Gene symbols in the CCLE CNV data were converted to ENSEMBL IDs using the org.Hs.eg.db package (23). In total, 12,399 genes mutated by SNP and 4,511 genes with CNVs were in common between the two datasets and included for downstream analysis.

Principle component analysis

We performed dimensionality reduction by principle component analysis (PCA) on a subset of genes with highly variable expression patterns. For each gene, the variance and mean were calculated for distributions of the anti-logged of expression values (pre-standardization) minus one. Log variance-to-mean ratio was then calculated as the log of the ratio of variance over mean (values calculated in the

previous step). We ranked expression variability of each gene based on log variance-to-mean ratio and the included the union of the 3,000 most variable genes from the Klijn 2015 and CCLE datasets in dimensionality reduction analysis shown in Figure 1E.

Model building approach

For each algorithm, we developed prediction models by considering the drug response data for one MEKi along with associated molecular features – e.g. response to PD-901 as assessed by Klijn 2015 compared with the Klijn 2015 molecular features (**Figure 2A**). One prediction model was built for each MEKi in both datasets, resulting in four prediction models per algorithm, and models were applied both within and between the Klijn 2015 and CCLE datasets. To assess performance, we established two validation sets: 1) for within-dataset predictions, model performance was reported for the 154 cell lines in common between the two datasets, which were excluded during all aspects of model generation and 2) for between-dataset predictions, performance was reported across all cell lines as no cell lines in the dataset the model was applied to were used during model building. Prediction models were built on 70% of available cell lines (selected randomly), repeated 30 times, and the final predicted drug response of a given cell line was calculated as the average of the 30 models. Note that the remaining 30% of cell lines that were not selected for model training were not used to assess model performance. An example schematic of model training while considering PD-901 in the Klijn 2015 dataset and model application both within and across datasets is shown in **Figure 2A**.

We established sets of regularized regression, random forest (regression), logistic regression, and random forest (classification) prediction models. Regularized regression and random forest (for both regression and classification) models were trained using the glmnet (26) and randomForest (27) packages in R, respectively. Logistic regression models were trained using the lm() function (family = binomial) in the base installation of R. Regression prediction models (regularized and random forest) were trained on log(IC_{50}) of MEKi response. Classification prediction models (logistic and random forest) were trained on binarized IC_{50} MEKi response, with $IC_{50} \leq 1$ defining sensitive cell lines and $IC_{50} > 1$ defining resistant. For within-dataset predictions, the 154 cell lines in common between the Klijn 2015 and CCLE datasets

were used as a validation set and excluded from all aspects of model building (i.e. including feature selection, parameter tuning, and model training). For cross-dataset predictions, all cell lines in the dataset not used to train the model were used as a validation set, as no overlapping cell lines were used to develop prediction models. Regression prediction models were trained on 70% of available training cell lines (i.e. cell lines unique to the Klijn 2015 or CCLE datasets). Classification prediction models were trained using a balanced set of sensitive and resistant cell lines: 70% of cell lines of the least populated class (typically sensitive) in the training set were randomly selected and matched with an equal number of randomly-selected cell lines of the more populated class. Model generation with randomly-selected training cell lines was repeated 30 times for both regression and classification predictions. Prediction models were applied to within-dataset and cross-dataset validation sets and the final prediction score was calculated as the average of 30 prediction models.

For regularized regression predictions, we tuned the α and λ parameters through a parameter sweep. The α parameter controls whether a ridge regression ($\alpha = 0$), elastic net ($0 < \alpha < 1$), or least absolute shrinkage and selection operator (LASSO; $\alpha = 1$) model is generated. We tested α values from 0 to 1 in 0.1 step increments. The λ parameter controls the strength of the penalty on model β values, and λ values tested were 0.001, 0.01, 0.1, 1, and 10. For logistic regression and random forest (regression and classification) models, we performed feature selection by LASSO prior to model generation. We performed a parameter sweep to test multiple λ values for the LASSO feature selection. λ values tested for LASSO feature selection were $1x10^{-5}$, $5x10^{-5}$, $1x10^{-4}$, $5x10^{-4}$, 0.001, 0.005, 0.01, 0.05, and 0.1. Tuning of α and λ parameters was performed by randomly selecting 70% of training cell lines and applying models trained with different parameter sets to the remaining 30% (testing set), repeated for 100 iterations. Note that the validation cell lines were not used for parameter tuning. Optimal parameter sets were selected based on maximum mean performance in the testing set across the 100 iterations. Performance was measured by Spearman's ρ for regression algorithms and auROC for classification algorithms. **Table S1** shows the optimal parameters selected.

We trained tissue-specific prediction models with CCLE data, randomly selecting 75% of cell lines of a given cancer type, and applied the models to Klijn 2015 data for testing. If a cell line overlapping between Klijn 2015 and CCLE was used to build a tissue-specific model, it was excluded from model testing. Tissue-specific prediction models were the only instance in which validation set cell lines (i.e. the 154 cell lines in common between Klijn 2015 and CCLE) were used to build prediction models, as there were fewer overall cell lines available within cancer types. Parameters selected from pan-cancer prediction models were also used for tissue-specific models.

ACKNOWLEDGEMENTS

Work was supported by funds from the Breast Cancer Research Foundation, Michigan Institute for Data Science, and National Institutes of Health (NIH 1R21CA218498-01 to S.D.M. and NIH R01GM118928-01 to J.Z.L.).

AUTHOR CONTRIBUTIONS

J.P.L., M.S., S.D.M, and J.Z.L designed the research. J.P.L performed the research. J.P.L., M.S., S.D.M, and J.Z.L wrote the manuscript.

COMPETING INTERESTS

The authors declare no competing interests.

REFERENCES

1. L. A. Garraway, J. Verweij, K. V. Ballman, Precision Oncology: An Overview. *Journal of Clinical Oncology* **31**, 1803–1805 (2013).

2. V. Prasad, T. Fojo, M. Brada, Precision oncology: origins, optimism, and potential. *The Lancet Oncology* **17**, e81–e86 (2016).

3. L. Schwartzberg, E. S. Kim, D. Liu, D. Schrag, Precision Oncology: Who, How, What, When, and When Not? *ASCO Educational Book* **37**, 160–169 (2017).

4. C. Kumar-Sinha, A. M. Chinnaiyan, Precision oncology in the age of integrative genomics. *Nature Biotechnology* **36**, 46–60 (2018).

5. N. Goossens, S. Nakagawa, X. Sun, Y. Hoshida, Cancer biomarker discovery and validation. *Transl Cancer Res* **4**, 256–269 (2015).

6. E. G. Armitage, C. Barbas, Metabolomics in cancer biomarker discovery: Current trends and future perspectives. *Journal of Pharmaceutical and Biomedical Analysis* **87**, 1–11 (2014).

7. A. Letai, Functional precision cancer medicine—moving beyond pure genomics. *Nature Medicine* **23**, 1028–1035 (2017).

8. J. Barretina, *et al.*, The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603–607 (2012).

9. M. J. Garnett, *et al.*, Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature* **483**, 570–575 (2012).

10. W. Yang, *et al.*, Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Research* **41**, D955–D961 (2012).

11. C. Klijn, *et al.*, A comprehensive transcriptional portrait of human cancer cell lines. *Nature Biotechnology* **33**, 306–312 (2015).

12. M. Ghandi, *et al.*, Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature* **569**, 503–508 (2019).

13. P. Geeleher, N. J. Cox, R. Huang, Clinical drug response can be predicted using baseline gene expression levels and in vitro drug sensitivity in cell lines. *Genome Biology* **15**, R47 (2014).

14. F. Iorio, *et al.*, A Landscape of Pharmacogenomic Interactions in Cancer. *Cell* **166**, 740–754 (2016).

15. L. Nguyen, C. Dang, P. Ballester, Systematic assessment of multi-gene predictors of pan-cancer cell line sensitivity to drugs exploiting gene expression data. *F1000 Research* **5**, 2927 (2017).

16. M. P. Menden, *et al.*, Machine Learning Prediction of Cancer Cell Sensitivity to Drugs Based on Genomic and Chemical Properties. *PLoS ONE* **8**, e61318 (2013).

17. N. Zhang, *et al.*, Predicting Anticancer Drug Responses Using a Dual-Layer Integrated Cell Line-Drug Network Model. *PLOS Computational Biology* **11**, e1004498 (2015).

18. D. Brubaker, *et al.*, Drug Intervention Response Predictions with PARADIGM (DIRPP) Identifies Drug Resistant Cancer Cell Lines and Pathway Mechanisms of Resistance in *Biocomputing 2014*, (WORLD SCIENTIFIC, 2013), pp. 125–135.

19. M. Ammad-ud-din, *et al.*, Integrative and Personalized QSAR Analysis in Cancer by Kernelized Bayesian Matrix Factorization. *Journal of Chemical Information and Modeling* **54**, 2347–2359 (2014).

20. M. Li, *et al.*, DeepDSC: A Deep Learning Method to Predict Drug Sensitivity of Cancer Cell Lines. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 1–1 (2019).

21. D. G. Covell, Data Mining Approaches for Genomic Biomarker Development: Applications Using Drug Screening Data from the Cancer Genome Project and the Cancer Cell Line Encyclopedia. *PLOS ONE* **10**, e0127433 (2015).

22. S. Anders, W. Huber, Differential expression analysis for sequence count data. 12 (2010).

23. M. Carlson, *org.Hs.eg.db: Genome wide annotation for Human* (2019).

24. C. D. Greenman, *et al.*, PICNIC: an algorithm to predict absolute allelic copy number variation with microarray cancer data. *Biostatistics* **11**, 164–175 (2010).

25. C. H. Mermel, *et al.*, GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biology* **12** (2011).

26. J. Friedman, T. Hastie, R. Tibshirani, Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software* **33** (2010).

27. A. Liaw, M. Wiener, Classification and Regression by randomForest. *R News* **2**, 18–22 (2002).