Crawford et al. Genome Biology (2022) 23:137
https://doi.org/10.1186/s13059-022-02705-y

Genome Biology

RESEARCH Open Access

Widespread redundancy in -omics profiles of cancer mutation states

Jake Crawford[1], Brock C. Christensen[2,3], Maria Chikina[4] and Casey S. Greene[5,6*]

* Correspondence: casey.s.greene@cuanschutz.edu
[6]Center for Health AI, University of Colorado School of Medicine, Aurora, CO, USA
Full list of author information is available at the end of the article

Abstract

Background: In studies of cellular function in cancer, researchers are increasingly able to choose from many -omics assays as functional readouts. Choosing the correct readout for a given study can be difficult, and which layer of cellular function is most suitable to capture the relevant signal remains unclear.

Results: We consider prediction of cancer mutation status (presence or absence) from functional -omics data as a representative problem that presents an opportunity to quantify and compare the ability of different -omics readouts to capture signals of dysregulation in cancer. From the TCGA Pan-Cancer Atlas that contains genetic alteration data, we focus on RNA sequencing, DNA methylation arrays, reverse phase protein arrays (RPPA), microRNA, and somatic mutational signatures as -omics readouts. Across a collection of genes recurrently mutated in cancer, RNA sequencing tends to be the most effective predictor of mutation state. We find that one or more other data types for many of the genes are approximately equally effective predictors. Performance is more variable between mutations than that between data types for the same mutation, and there is little difference between the top data types. We also find that combining data types into a single multi-omics model provides little or no improvement in predictive ability over the best individual data type.

Conclusions: Based on our results, for the design of studies focused on the functional outcomes of cancer mutations, there are often multiple -omics types that can serve as effective readouts, although gene expression seems to be a reasonable default option.

Background

Although cancer can be initiated and driven by many different genetic alterations, these tend to converge on a limited number of pathways or signaling processes [1]. As driver mutation status alone confers limited prognostic information, a comprehensive understanding of how diverse genetic alterations perturb central pathways is vital to precision medicine and biomarker identification efforts [2, 3]. While many methods exist to distinguish driver mutations from passenger mutations based on genomic sequence characteristics [4–6], until recently, it has been a challenge to connect driver

mutations to downstream changes in gene expression and cellular function within individual tumor samples.

The Cancer Genome Atlas (TCGA) Pan-Cancer Atlas provides uniformly processed, multi-platform -omics measurements across tens of thousands of samples from 33 cancer types [7]. Enabled by this publicly available data, a growing body of work on linking the presence of driving genetic alterations in cancer to downstream gene expression changes has emerged. Recent studies have considered Ras pathway alteration status in colorectal cancer [8]; alteration status across many cancer types in Ras genes [9, 10], *TP53* [11], and *PIK3CA* [12]; and alteration status across cancer types in frequently mutated genes [13]. More broadly, other groups have drawn on similar ideas to distinguish between the functional effects of different alterations in the same driver gene [14], to link alterations with similar gene expression signatures within cancer types [15], and to identify trans-acting expression quantitative trait loci (trans-eQTLs) in germline genetic studies [16].

These studies share a common thread: they each combine genomic (point mutation and copy number variation) data with transcriptomic (RNA sequencing) data within samples to interrogate the functional effects of genetic variation. RNA sequencing is ubiquitous and cheap, and its experimental and computational methods are relatively mature, making it a vital tool for generating insight into cancer pathology [17]. Some driver mutations, however, are known to act indirectly on gene expression through varying mechanisms. For example, oncogenic *IDH1* and *IDH2* mutations in glioma have been shown to interfere with histone demethylation, which results in increased DNA methylation and blocked cell differentiation [18–21]. Other genes implicated in aberrant DNA methylation in cancer include the TET family of genes [22] and *SETD2* [23]. Certain driver mutations, such as those in DNA damage repair genes, may lead to detectable patterns of somatic mutation [24]. Additionally, correlation between gene expression and protein abundance in cancer cell lines is limited, and proteomics data could correspond more directly to certain cancer phenotypes and pathway perturbations [25]. In these contexts and others, integrating different data modalities or combining multiple data modalities could be more effective than relying solely on gene expression as a functional signature.

Here, we compare -omics data types profiled in the TCGA Pan-Cancer Atlas to evaluate use as a multivariate functional readout of genetic alterations in cancer. We focus on gene expression (RNA sequencing data), DNA methylation (27K and 450K probe chips), reverse phase protein array (RPPA), microRNA expression, and mutational signatures data [26] as possible readouts. Prior studies have identified univariate correlations of CpG site methylation [27, 28] and correlations of RPPA protein profiles [29] with the presence or absence of certain driver mutations. Other relevant past work includes linking point mutations and copy number variants (CNVs) with changes in methylation and expression at individual genes [30, 31] and identifying functional modules that are perturbed by somatic mutations [32, 33]. However, direct comparison among different data types for this application is lacking, particularly in the multivariate case where we consider changes to -omics-derived gene signatures rather than individual genes in isolation.

We select a collection of potential cancer drivers with varying functions and roles in cancer development. We use mutation status in these genes as labels to train classifiers,

using each of the data types listed as training data, in a pan-cancer setting; we follow similar methods to the elastic net logistic regression approach described in Way et al. 2018 [9] and Way et al. 2020 [13]. We show that there is considerable predictive signal for many genes relative to a cancer-type corrected baseline and that gene expression tends to provide good predictions of mutation state across most genes. Surprisingly, we find that for a variety of genes, multiple data types are approximately equally effective predictors. We observe similar results for pan-cancer survival prediction across the same data types with little separation between the top-performing data types. In addition, we observe that combining data types into a single multi-omics model for mutation prediction provides little, if any, performance benefit over the most performant model using a single data type. Our results will help to inform the design of future functional genomics studies in cancer, suggesting that for many strong drivers with clear functional signatures, different -omics measurements can provide similar information content.

Results

Using diverse data modalities to predict cancer alterations

We collected five different data modalities from cancer samples in the TCGA Pan-Cancer Atlas, capturing five steps of cellular function that are perturbed by genetic alterations in cancer (Fig. 1A). These included gene expression (RNA-seq data), DNA methylation (27K and 450K Illumina BeadChip arrays), protein abundance (RPPA data),

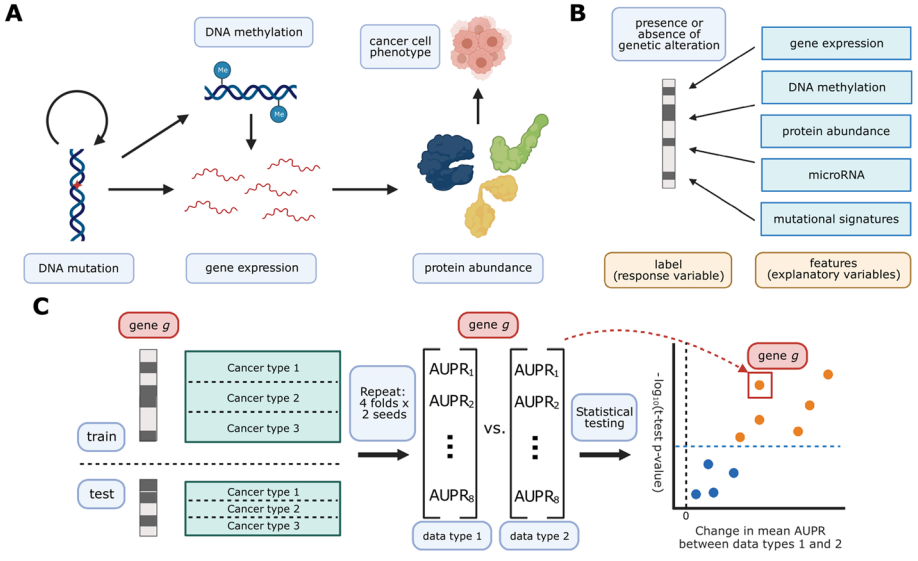

Fig. 1 **A** Cancer mutations can perturb cellular function via a variety of cellular processes. Arrows represent major potential paths of information flow from a somatic mutation in DNA to its resulting cell phenotype; circular arrow represents the ability of certain mutations (e.g., in DNA damage repair genes) to alter somatic mutation patterns. Note that this does not reflect all possible relationships between cellular processes: for instance, changes in gene expression can lead to changes in somatic mutation rates. **B** Predicting presence/absence of somatic alterations in cancer from diverse data modalities. In this study, we use functional readouts from TCGA as predictive features and the presence or absence of mutation in a given gene as labels. This reverses the primary direction of information flow shown in **A**. **C** Schematic of evaluation pipeline

microRNA expression data, and patterns of somatic mutation (mutational signatures). To link these diverse data modalities to changes in mutation status, we used elastic net logistic regression to predict the presence or absence of mutations in cancer genes, using these readouts as predictive features (Fig. 1B). We evaluated the resulting mutation status classifiers in a pan-cancer setting, preserving the proportions of each of the 33 cancer types in TCGA for eight train/test splits (4 folds × 2 replicates) in each of approximately 250 cancer genes (Fig. 1C).

We sought to compare classifiers against a baseline where mutation labels are permuted (to identify genes whose mutation status correlates strongly with a functional signature in a given data type) and also to compare classifiers trained on true labels across different data types (to identify data types that are more or less predictive of mutations in a given gene). To account for variation between dataset splits in making these comparisons, we treat classification metrics from the eight train/test splits as performance distributions, which we compare using *t*-tests. We summarize performance across all genes in our cancer gene set using a similar approach to a volcano plot, in which each point is a gene. In our summary plots, the *x*-axis shows the magnitude of the change in the classification metric between conditions, and the *y*-axis shows the *p*-value for the associated *t*-test (Fig. 1C).

Selection of cancer-related genes improves predictive signal

As a baseline, we evaluated prediction of mutation status from gene expression data across several different gene sets. Past work has evaluated mutation prediction for the top 50 most mutated genes in TCGA [13], and we sought to extend this to a broader list of gene sets. To evaluate whether using known cancer-related genes tends to improve prediction, we compiled a set of cancer-related genes (n=268) from Vogelstein et al. [34], Bailey et al. [35], and the COSMIC Cancer Gene Census [36]. We compared performance on this curated gene set with performance on an equal number of genes sampled randomly after applying a mutation frequency threshold (n=268, see the "Methods" section for sampling details) and an equal number of the most mutated genes in TCGA (n=268). For all gene sets, we used only the set of TCGA samples for which both gene expression and somatic mutation data exists, resulting in a total of 9074 samples across all 33 cancer types. This set of samples was further filtered for each target gene to cancer types containing at least 15 mutated samples and at least 5% of samples mutated for that cancer type. As an alternate approach, we tried including/excluding entire genes using similar filters, and the results were consistent across filtering strategies (Additional file : Fig. S4). We then evaluated the performance for each target gene in each of the three gene sets.

Overall, genes from the cancer-related gene set were more predictable than randomly chosen genes or those selected by total mutation count (Fig. 2A). In total, for a significance threshold of $\alpha = 0.001$, 120/268 genes (44.8%) in the cancer-related gene set are significantly predictable from gene expression data, compared to 14/268 genes (5.22%) in the random gene set and 80/268 genes (29.9%) in the most mutated gene set. Of the 14 significantly predictable genes in the random gene set, 13 of them are also in the cancer-related gene set (highlighted with 'X' in Fig. 2B), and of the 80 significantly predictable genes in the most mutated gene set, 26 of them are also in the cancer-related

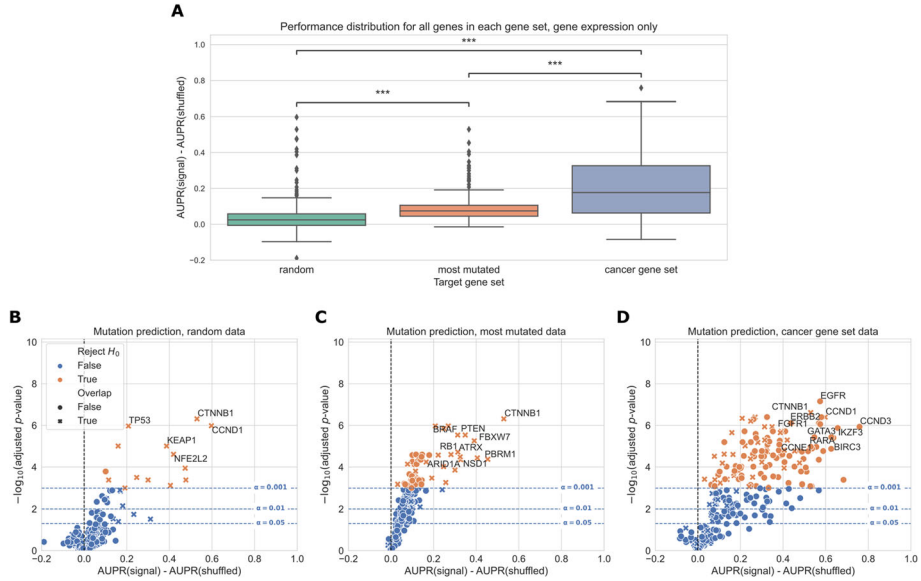

Fig. 2 A Overall distribution of performance across three gene sets, using gene expression (RNA-seq) data to predict mutations. Each data point represents the mean cross-validated AUPR difference, compared with a baseline model trained on permuted mutation presence/absence labels, for one gene in the given gene set; notches show bootstrapped 95% confidence intervals. "random" = 268 random genes, "most mutated" = 268 most mutated genes, and "cancer gene set" = 268 cancer-related genes from curated gene sets. Significance stars indicate results of Bonferroni-corrected pairwise Wilcoxon tests: $^{**}p < 0.01$, $^{***}p < 0.001$, ns: not statistically significant for a cutoff of p = 0.05. **B**–**D** Volcano-like plots showing mutation presence/absence predictive performance for each gene in each of the three gene sets. The *x*-axis shows the difference in mean AUPR compared with a baseline model trained on permuted labels, and the *y*-axis shows *p*-values for a paired t-test comparing cross-validated AUPR values within folds. Points (genes) marked with an "X" are overlapping between the cancer gene set and either the random or most mutated gene set

gene set (highlighted in red in Fig. 2C). These results suggest that selecting target genes for mutation prediction based on prior knowledge of their involvement in cancer pathways and processes, rather than randomly or based on mutation frequency alone, can improve predictive signal and identify more highly predictable mutations from gene expression data.

Gene expression predicts cancer mutation status more effectively than DNA methylation

We compared gene expression with DNA methylation as downstream readouts of the effects of cancer alterations. In these analyses, we considered both the 27K probe and 450K probe methylation datasets generated for the TCGA Pan-Cancer Atlas. As target genes, we used the same combined cancer-related gene set described in the "Selection of cancer-related genes" section. We used samples that had data for each of the data types being compared, including somatic mutation data to generate mutation labels. This process retained 7981 samples in the intersection of the expression, 27K methylation, 450K methylation, and mutation datasets, which we used for subsequent analyses. The most frequent missing data types were somatic mutation data (1114 samples) and 450K methylation data (1072 samples) (Fig. 3A).

For many genes, predictions are better than our baseline model where labels are permuted (values greater than 0 in the box plots), suggesting that there is considerable

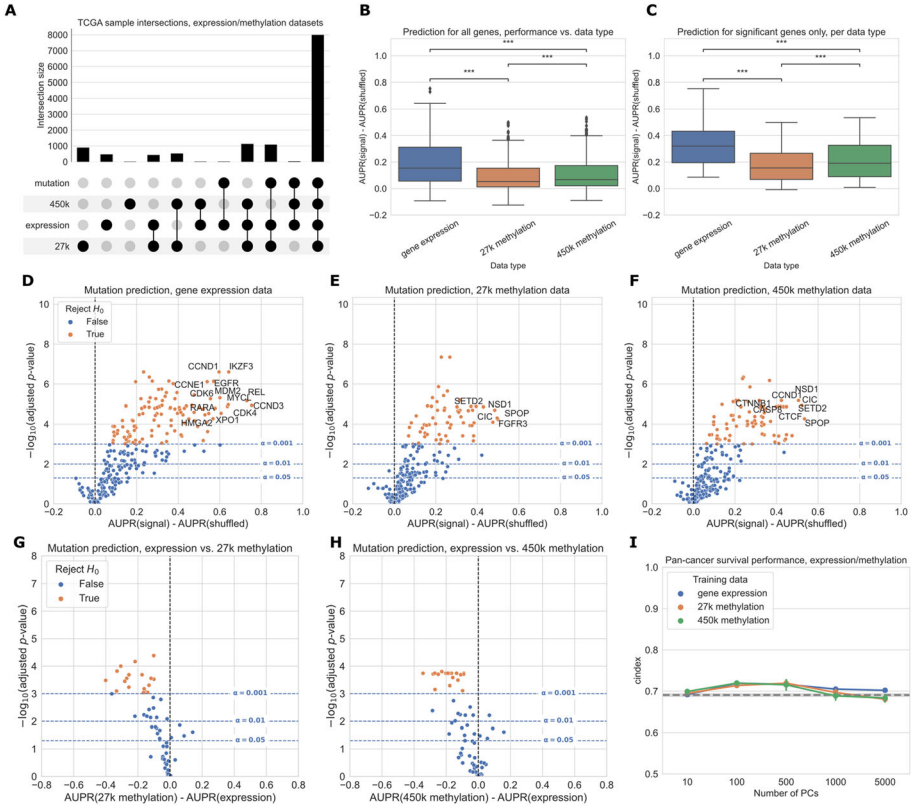

Fig. 3 A Count of overlapping samples between gene expression, 27K methylation, 450K methylation, and somatic mutation data used from TCGA. Only non-zero overlap counts are shown. Somatic mutation sample information is included because it is needed to generate the mutation presence/absence labels. **B** Predictive performance for genes in the cancer-related gene set, using each of the three data types as predictors. The gene expression predictor uses the top 8000 gene features by mean absolute deviation, and the methylation predictors use the top 5000 principal components as predictive features. Significance stars indicate results of Bonferroni-corrected pairwise Wilcoxon tests: $^{**}p < 0.01$, $^{***}p < 0.001$, ns: not statistically significant for a cutoff of $p = 0.05$. **C** Predictive performance for genes where at least one of the considered data types predicts mutation labels significantly better than the permuted baseline. **D**–**F** Predictive performance for each gene in the cancer-related gene set, for each data type, compared with a baseline model trained on permuted labels. **G**, **H** Direct comparison of performance using gene expression and each methylation dataset, for genes that perform significantly better than the baseline for both data types. Points (genes) to the left of y=0 perform better using gene expression-derived features, and points to the right perform better using methylation-derived features. **I** Pan-cancer survival prediction performance, quantified using c-index on the y-axis, for gene expression, 27K methylation, and 450K methylation. The x-axis shows results with varying numbers of principal components included for each data type. Models also included covariates for patient age, sample mutation burden, and sample cancer type; gray dotted line indicates mean performance for a covariate-only baseline model

predictive signal in both expression and methylation datasets across the cancer-related gene set (Fig. 3B). On aggregate across all genes, predictive performance is best overall for gene expression. Both before and after filtering for genes that exceed the significance threshold, gene expression with raw gene features provides a significant performance improvement relative to the 27K methylation and 450K methylation datasets (Fig. 3B, C). Results were similar with PCA-compressed gene expression features or raw CpG probes as predictors (Additional file 1: Fig. S5).

Considering each target gene in the cancer-related gene set individually, we observed that 113/272 genes significantly outperformed the permuted baseline using gene expression data, as compared to 62/272 genes for 27K methylation and 77/272 genes for

450K methylation (Fig. 3D–F, more information about specific genes in Additional file 1: Fig. S2). Some "well-predicted" genes that outperformed the permuted baseline tended to be similar between data types (Fig. 3D–F; genes in the top right of each plot). For example, *CIC* appears in the top right of all 3 plots, and *CCND1* appears in the top right of the gene expression and 450K methylation plots, suggesting that mutations in these genes have strong gene expression and DNA methylation signatures, and these signatures tend to be preserved across cancer types.

In addition to comparing mutation classifiers trained on different data types to the permuted baseline, we also compared classifiers trained on true labels directly to each other, for genes that performed significantly better than the baseline for both of the data types under consideration (Fig. 3G, H). We observed that 18/58 genes were significantly more predictable from expression data than 27K methylation data, and 21/69 genes were significantly more predictable from expression data than 450K methylation data. In both cases, no genes were significantly more predictable using the methylation data types. Still, we observed that some points were clustered around the origin, indicating that the data types appear to confer similar information about mutation status. That is, in these cases, matching the gene being studied with the "correct" data modality seems to be unimportant: mutation status has a strong signature which can be extracted from both expression and DNA methylation data roughly equally.

We additionally compared pan-cancer survival prediction performance using principal components derived from each data type; in general, results were comparable across the three data types (Fig. 3I). All data types outperformed the covariate-only baseline (see the "Methods" section) for lower numbers of PC features included, although performance was similar to the baseline for higher numbers of PCs. Confidence intervals between the best- and worst-performing data types overlap at most PC counts (with the exception of gene expression at 5000 PC features), suggesting that similarly to mutation prediction, the three data types tend to have comparable effectiveness for pan-cancer survival prediction.

Focusing on several selected genes of interest, we observed that relative classifier performance varies by gene (Fig. 4). Past work has indicated that mutations in *TP53* are highly predictable from gene expression data [11], and we observed that the methylation datasets provided similar predictive performance (Fig. 4A). Similarly, for *IDH1* both expression and methylation features result in similar performance, consistent with the previously observed role of IDH1 in regulating both DNA methylation and gene expression (Fig. 4D) [37]. Mutations in *KRAS* and *ERBB2* (*HER2*) were most predictable from gene expression data, and in both cases the methylation datasets significantly outperformed the baseline as well (Fig. 4B, E). Gene expression signatures of *ERBB2* alterations are historically well-studied in breast cancer [38], and samples with activating *ERBB2* mutations have recently been shown to share sensitivities to some small-molecule inhibitors across cancer types [39]. These observations are consistent with the pan-cancer *ERBB2* mutant-associated expression signature that we observed in this study. *NF1* mutations were also most predictable from gene expression data, although the gene expression-based *NF1* mutation classifier did not significantly outperform the baseline with permuted labels at a cutoff of $\alpha = 0.001$ (Fig. 4C). *SETD2* is an example of a gene that is more predictable from the methylation datasets than from gene expression, although gene expression with raw gene features significantly outperformed the

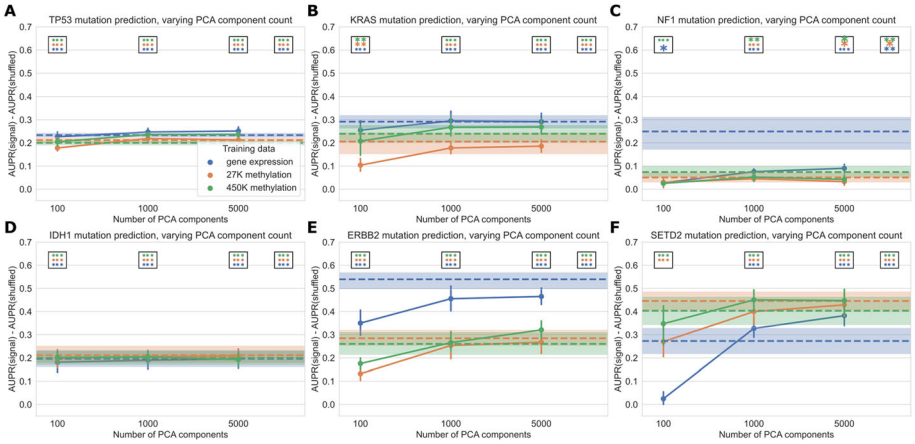

Fig. 4 Performance across varying PCA dimensions for specific genes of interest. Dotted lines represent results for "raw" features (8000 gene features for gene expression data and 8000 CpG probes for both methylation datasets, selected by largest mean absolute deviation). Error bars and shaded regions show bootstrapped 95% confidence intervals. Stars in boxes show statistical testing results compared with permuted baseline model; each box refers to the model using the number of PCA components it is over (far right box = models with raw features). ** $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, no stars: not statistically significant for a cutoff of $p = 0.05$

permuted baseline as well (Fig. 4F). *SETD2* is widely mutated across cancer types and affects H3K36 histone methylation most directly, but SETD2-mediated changes in H3K36 methylation have been linked to dysregulation of diverse cellular processes including DNA methylation and RNA splicing [23, 40].

Comparing six different readouts favors expression and DNA methylation

Next, we expanded our comparison to all five functional data modalities (six total readouts, since there are two DNA methylation platforms) available in the TCGA Pan-Cancer Atlas. As with previous experiments, we limited our comparison to the set of samples profiled for each readout, resulting in 5226 samples with data for all readouts. The data types with the most missing samples were RPPA data (2215 samples that were missing RPPA data) and 450K methylation (630 samples that were missing 450K methylation data) (Fig. 5A). Summarized over all genes in the cancer-related gene set, we observed that gene expression tended to produce better predictions than the other data types (Fig. 5B). This remained true when we looked only at the set of genes having at least one significant predictor (i.e., "well-predicted" genes) (Fig. 5C).

On the individual gene level, mutations in 33/217 genes were significantly predictable from RPPA data relative to the permuted baseline, compared to 25/217 genes from microRNA data and 2/217 genes from mutational signatures data (Fig. 5D–F). For the remaining data types on this smaller set of samples, 79/217 genes outperformed the baseline for gene expression data, 31/217 for 27k methylation, and 42/217 for 450k methylation. Compared to the methylation experiments (Fig. 3), we observed fewer "well-predicted" genes for the expression and methylation datasets here (likely due to the considerably smaller sample size) but relative performance was comparable (Additional file 1: Fig. S3). Direct comparisons between each added data type and gene expression data showed that for most "well-predicted" genes, RPPA, microRNA, and

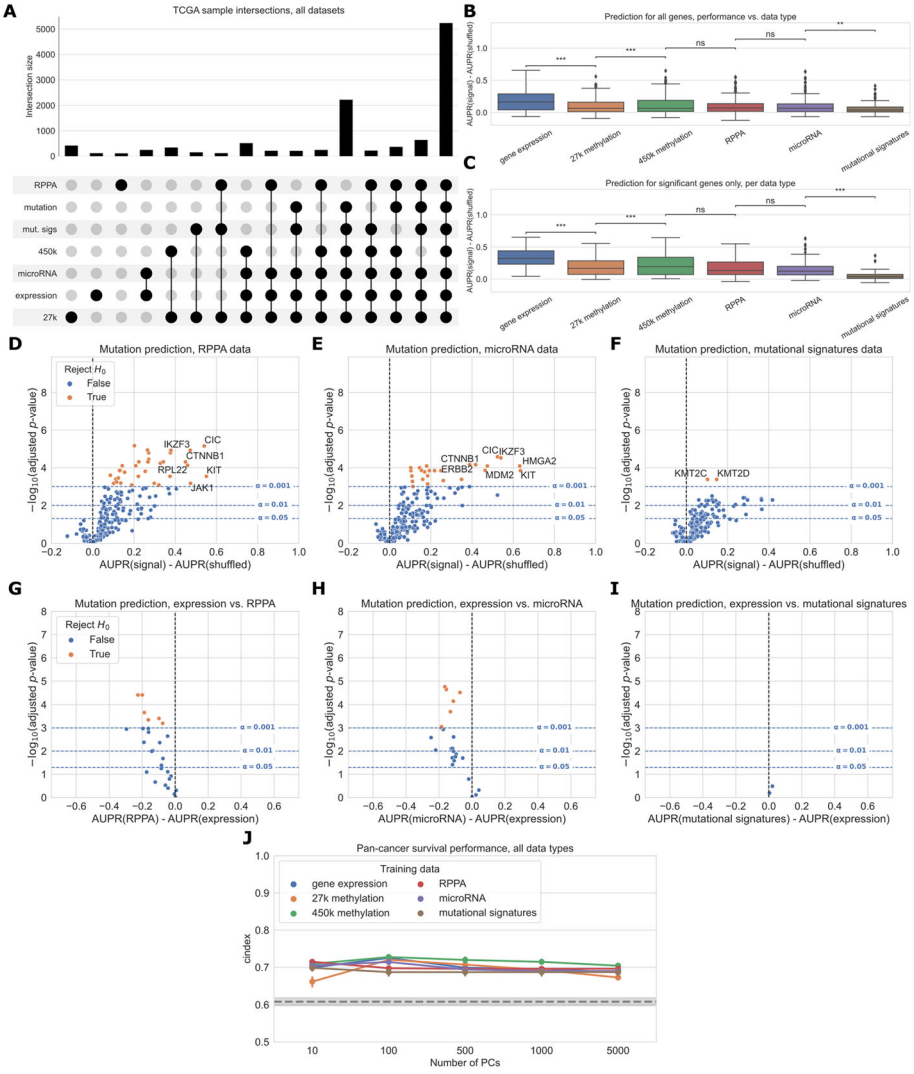

Fig. 5 A Overlap of TCGA samples between all data types used in mutation prediction comparisons. Only overlaps with more than 100 samples are shown. Somatic mutation sample information is included because it is needed to generate the mutation presence/absence labels. **B** Overall distribution of performance per data type across 217 genes from the cancer-related gene set. Each data point represents mean cross-validated AUPR difference, compared with a baseline model trained on permuted labels, for one gene; notches show bootstrapped 95% confidence intervals. Significance stars indicate results of Bonferroni-corrected pairwise Wilcoxon tests: $^{**}p < 0.01$, $^{***}p < 0.001$, ns: not statistically significant for a cutoff of $p = 0.05$. All pairwise tests were run, and corrected for, but only neighboring test results are shown. **C** Overall performance distribution per data type for genes where the permuted baseline model is significantly outperformed for one or more data types, resulting in a total of 39 genes. **D**–**F** Volcano-like plots showing predictive performance for each gene in the cancer-related gene set, in each of the added data types (RPPA, microRNA, mutational signatures). The *x*-axis shows the difference in mean AUPR compared with a baseline model trained on permuted labels, and the *y*-axis shows *p*-values for a paired t-test comparing cross-validated AUPR values within folds. **G**–**I** Direct comparison of performance using gene expression and each added data type, showing only genes that perform significantly better than the baseline model for both data types. Points (genes) to the left of *y*=0 perform better using gene expression-derived features, and points to the right perform better using the added data type (RPPA, microRNA, and mutational signatures respectively). **J** Pan-cancer survival prediction performance, quantified using c-index on the *y*-axis, for all data types. The *x*-axis shows results with varying numbers of principal components included for each data type. Models also included covariates for patient age, sample mutation burden, and sample cancer type; gray dotted line indicates mean performance for a covariate-only baseline model

mutational signatures data generally provide similar or worse performance compared to gene expression (Fig. 5G–I).

Performance using RPPA data (Fig. 5G) is notable because of its drastically smaller dimensionality than the other data types (190 proteins, compared to thousands of features for the expression and methylation data types). This suggests that each protein abundance measurement provides a high information content, although this is by design as the antibody probes used for the TCGA analysis were selected to cover established cancer-related pathways [41]. Similarly, the scope of the features captured by the mutational signatures data we used is limited to single-base substitution signatures; a broader spectrum of possible signatures is described in previous work [26, 42] including doublet-substitution signatures, small indel signatures, and signatures of structural variation, but these were not readily available for the TCGA exome sequencing data. The relatively poor predictive ability of mutational signatures likely stems from a combination of biological and technical factors, as there is no reason to expect that changes in somatic mutation patterns would be directly caused by most cancer driver mutations. Two exceptions are *KMT2C* and *KMT2D* (Fig. 5F), which may have a role in mediating DNA damage response [43].

As in the expression/methylation comparison, we also compared pan-cancer survival prediction performance between all six readouts, using the top principal components derived from each data type to ensure comparable information content (Fig. 5J). All six readouts performed comparably for this smaller set of samples, with slightly better performance across PC feature dimensions for the 450K methylation array. The covariate-only baseline predictor performed considerably worse than it did in the expression/methylation comparisons, with all -omics data types outperforming the baseline predictor at all PC numbers.

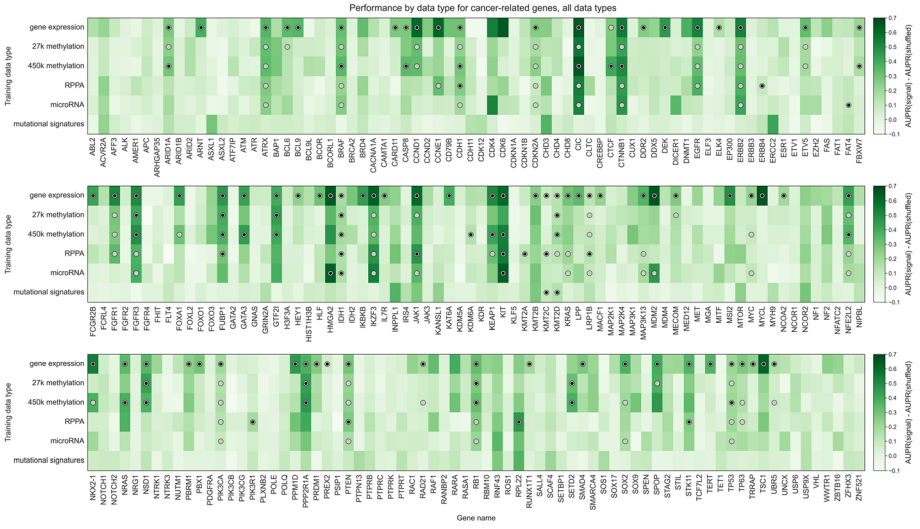

Fig. 6 Heatmap displaying predictive performance for mutations in each of the 217 genes from the cancer-related gene set, across all six TCGA data modalities. Each cell quantifies performance for a target gene, using predictive features derived from a particular data type. Gray shaded dots indicate that the given data type provides significantly better predictions than the permuted baseline for the given gene; black inner dots indicate the same and additionally that the given data type provides statistically equivalent performance to the data type with the best average performance (determined by pairwise *t*-tests across data types with FDR correction)

When we constructed a heatmap depicting predictive performance for each gene across data types, we found that many genes tended to be well-predicted by more than one data type (Fig. 6). Of the 86 genes that are well-predicted using at least one data type (gray circles in Fig. 6), 52/86 (60.5%) are well-predicted by multiple data types, meaning that multiple -omics readouts contain a detectable signature of presence/absence of a mutation in the corresponding gene. Of the remaining 34 genes, 28/34 (82.4%) are well-predicted by gene expression alone. This supports our observation that in a surprising number of cases, choosing the "correct" data modality is unimportant for driver genes with strong functional signatures, although gene expression may be the best "default" choice as it tends to be a strong predictor in the majority of cases. Exceptions included *ERBB4*, *KMT2A*, *PIK3R1*, and *RPL22* (only well-predicted using RPPA data); *FAT4* (only well-predicted using microRNA data); and *KDM6A* (only well-predicted using 450K methylation data).

Simple multi-omics integration provides little performance benefit

We also trained "multi-omics" classifiers to predict mutations in six well-studied and widely mutated driver genes across various cancer types: *EGFR*, *IDH1*, *KRAS*, *PIK3CA*, *SETD2*, and *TP53*. Each of these genes is well-predicted from several data types in our earlier experiments (Fig. 6), consistent with having strong pan-cancer driver effects. For the multi-omics classifiers, we considered all pairwise combinations of the top three performing individual data types (gene expression, 27K methylation, and 450K methylation), in addition to a model using all three data types. We trained a classifier for multiple data types by concatenating features from the individual data types, then fitting the same elastic net logistic regression model as we used for the single-omics models. Here, we show results using the top 5000 principal components from each data type as predictive features, to ensure that feature count and scale is comparable among data types; results for raw features are shown in Additional file 1: Fig. S6. We additionally ran the same experiments using a 3-layer fully-connected neural network for classification, with principal components as input, and results are shown in Additional file 1: Fig. S7. In general, we found predictions using elastic net logistic regression to be more robust across cross-validation folds and hyperparameter choices than predictions using the neural network, although the neural network provided a slight performance improvement using multiple -omics types for some genes.

For each of the six target genes, we observed comparable performance between the best single-omics classifier (blue boxes in Fig. 7A) and the best multi-omics classifier (orange boxes in Fig. 7A). Across all classifiers and data types, we found varied patterns based on the target gene. For *IDH1* and *TP53*, performance is relatively consistent regardless of data type(s), suggesting that baseline performance is high and there is little room for improvement as data is added (Fig. 7C, G). The *TP53* classifier using raw features showed a statistically significant improvement when multiple data types were integrated, although the difference in mean performance was relatively small (Additional file 1: Fig. S6, p=0.0078). For *EGFR*, *KRAS*, and *PIK3CA*, combining gene expression with methylation data results in statistically equivalent or worse performance to gene expression alone; classifiers trained only on methylation data generally do not perform

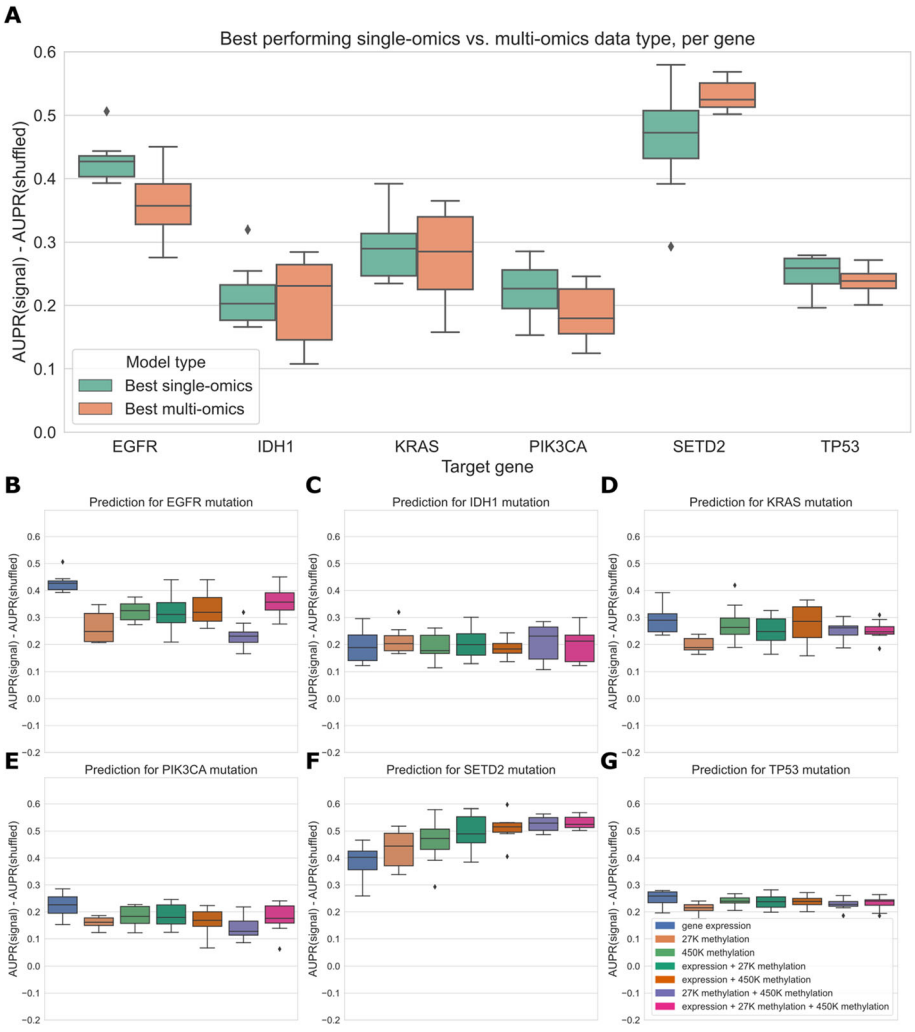

Fig. 7 A Comparing the best-performing model (i.e., highest mean AUPR relative to permuted baseline) trained on a single data type against the best "multi-omics" model for each target gene. None of the differences between single-omics and multi-omics models were statistically significant using paired-sample Wilcoxon tests across cross-validation folds, for a threshold of 0.05. **B**–**G** Classifier performance, relative to baseline with permuted labels, for mutation prediction models trained on various combinations of data types. Each panel shows performance for one of the six target genes; box plots show performance distribution over 8 evaluation sets (4 cross-validation folds × 2 replicates).

as well as those that integrate expression data (Fig. 7B, D, E). Previously, we saw that the best classifiers for *SETD2* used methylation data alone (Fig. 6). When we added multiple data types to our *SETD2* classifier, we did observe an increase in performance (Fig. 7F), although this improvement was not statistically significant in a paired-sample t-test for α=0.05 (p=0.078). Overall, we observed that combining data types in a relatively simple manner, by concatenating features from each individual data type, provided little or no improvement in predictive ability over the best individual data type. This supports our earlier observations of the redundancy of gene expression and methylation data as functional readouts, since our multi-omics classifiers are not in general able to extract gains in predictive performance as more data types are added for this set of cancer drivers.

Discussion

We carried out a large-scale comparison of data types in the TCGA Pan-Cancer Atlas as functional readouts of genetic alterations in cancer, integrating results across cancer types and across driver genes. Overall, we found that gene expression captures signatures of mutation state most effectively in general, relative to a baseline model, but we saw that for many genes other data types could be equally effective at predicting mutation presence or absence. For pan-cancer survival prediction, we found that the functional readouts tended to be similarly effective, outperforming a simple baseline using age and sample mutation burden in most cases. Our multi-omics modeling experiment indicated that the mutation state information captured by gene expression and DNA methylation is highly redundant, as added data types resulted in no gain or modest gains in classifier performance.

Comparing mutation status prediction using raw and PCA compressed expression and DNA methylation data, we observed that feature extraction using PCA provided no benefit compared to using raw gene or CpG probe features. Other studies using DNA methylation array data have found that nonlinear dimension reduction methods, such as variational autoencoders and capsule networks, can be effective for extracting predictive features [44, 45]. The latter approach is especially interesting because capsule networks and "capsule-like methods" can be constrained to extract features that align with known biology (i.e., that correspond to known disease pathways or CpG site annotations). This can improve model interpretability as well as predictive performance. Similar methods have been applied to extract biologically informed features from gene expression data (see, for instance, [46, 47]). A more comprehensive study of dimension reduction methods in the context of mutation prediction, including the features selected by these methods and their biological relevance and interpretation, would be a beneficial area of future work. In addition to methods for extracting features, another aspect of the study that could be explored further is methods for labeling samples as mutated or not more efficiently. Although the mutation calls we used from MC3 represent the consensus of multiple algorithms aggregated through a standard pipeline, benchmarking other methods for identifying mutated samples could improve the utility of our method, such as calling mutations directly from RNA-seq data to avoid the need for paired samples [48, 49].

In contrast to many other studies demonstrating the benefits of integrating multiple -omics data types for various cancer-related prediction problems [50–54], we found that combining multiple data types to predict mutation status was generally not effective for this problem. The method we used to integrate different data types by concatenating feature sets is sometimes referred to as "early" data integration (discussed in more detail in [55, 56]). It is possible that more sophisticated data integration methods, such as "intermediate" integration methods that learn a set of features jointly across datasets, would produce improved predictions. We do not interpret our results as concrete evidence that multi-omics integration is not effective for this problem; rather, we see them as an indication that this is a challenging data integration problem for which further investigation is needed. We also present this problem as a set of benchmark tasks on which multi-omics integration methods can be evaluated. In addition to the methodological questions, the issue of data integration also has implications for the underlying biology: a more nuanced understanding of when different data readouts provide

redundant information, and when they can contribute unique information about cancer pathology and development, could have many translational applications.

One limitation of the current study is that, for the mutation prediction problem, we only evaluated classifiers that were trained on pan-cancer data. Considering every possible combination of target gene and TCGA cancer type (85 target genes × 33 cancer types × 6 data types) would have drastically increased the computational load and presented a large multiple testing burden. Alternatively, choosing only a subset of gene/cancer type combinations to study would have biased our results toward known driver gene/cancer type relationships, which we aimed to avoid. In future work, it would be interesting to identify classifiers that perform well in a certain cancer type but not in the pan-cancer context and to compare these instances across different cancer types. As a motivating example, other studies have shown that activating mutations in Ras isoforms (*HRAS*, *KRAS*, *NRAS*) tend to have similar effects to one another in thyroid cancer, producing similar gene expression signatures [15]. In multiple myeloma, however, activating *KRAS* and *NRAS* mutations produce distinct expression signatures, necessitating separate classifiers [57]. A high-throughput computational pipeline to identify cases where functional signatures of a particular cancer driver are either concordant or discordant between cancer types could identify opportunities for context-specific protein function prediction, improve biomarker identification, and suggest cases where drugs targeting specific alterations might produce discordant results in different cancer types.

As with any study relying on observational, cross-sectional data such as the TCGA Pan-Cancer Atlas, the conclusions that we can draw are limited by the data. In particular, for any of our "well-predicted" genes (i.e., genes that, when mutated, have strong signatures in one or more data types), we cannot definitively distinguish correlation from causation. To directly assess the effects of particular mutations on various data modalities, some studies use cell line data from sources such as the Cancer Cell Line Encyclopedia (CCLE) [58]. While this approach could help to isolate the causal effect of a given mutation on a given cell line, cell lines are sometimes an imperfect match for the cancers they are derived from [59]. We are also limited in that we cannot assign timing or clonal status to mutations, or fully characterize intratumor heterogeneity, with certainty from the bulk sequencing data generated by TCGA (although some features of tumor mutational processes over time can be estimated from bulk data, e.g., [60]). As methods for generating large longitudinal datasets at single-cell resolution mature and scale, we will need to revise the way we think about cellular function and dysregulation in cancer cells, as dynamic and adaptive processes rather than a single representative snapshot of a tumor.

Conclusions

Based on our results, for studies focused on the functional consequences of cancer mutations, we recommend that researchers of cancers prioritize downstream readouts based on the gene or genes of interest (Fig. 6). On balance, prediction of mutation status is best in general using gene expression data, and prediction of patient survival is similar for all data types in the study. However, the finding that for many genes, multiple functional profiles contain much of the same information will be useful for some study designs, given varying cost and stability of different readouts. In addition to gene

expression, results using DNA methylation and RPPA measurements as predictive features were promising, especially considering the substantially lower dimensionality of the RPPA dataset compared to other data types. It is important to note that the specific technologies chosen by TCGA, and the tradeoffs inherent in such a high-throughput study, could influence the comparison: it is possible that, for instance, another technology for measuring DNA methylation (such as bisulfite sequencing) or another technique for measuring protein abundance (such as mass spectrometry-based proteomics) could change performance for those data types. Future technology advances, in both quality and quantity of data, are likely to improve our understanding of the full picture of functional consequences of mutations in cancer cells.

Methods

Mutation data download and preprocessing

To generate binary mutated/non-mutated gene labels for our machine learning model, we used mutation calls for TCGA samples from MC3 [61] and copy number threshold calls from GISTIC2.0 [62]. MC3 mutation calls were downloaded from the Genomic Data Commons (GDC) of the National Cancer Institute, at https://gdc.cancer.gov/about-data/publications/pancanatlas. Copy number threshold calls are from an older version of the GDC data and are available here: https://figshare.com/articles/dataset/TCGA_PanCanAtlas_Copy_Number_Data/6144122. We removed hypermutated samples (defined as five or more standard deviations above the mean non-silent somatic mutation count) from our dataset to reduce the number of false positives (i.e., non-driver mutations). After this filtering, 9074 TCGA samples with mutation and copy number data remained. Any sample with a non-silent somatic variant in the target gene was included in the positive set. We also included copy number gains in the target gene for oncogenes and copy number losses in the target gene for tumor suppressor genes in the positive set; all remaining samples were considered negative for mutation in the target gene.

Omics data download and preprocessing

RNA sequencing, 27K and 450K methylation array, microRNA, and RPPA datasets for TCGA samples were all downloaded from GDC, at the same link provided above. Mutational signatures' information for TCGA samples with whole-exome sequencing data was downloaded from the International Cancer Genome Consortium (ICGC) data portal, at https://dcc.icgc.org/releases/PCAWG/mutational_signatures/Signatures_in_Samples/SP_Signatures_in_Samples. For our experiments, we used only the "single base signature" (SBS) mutational signatures, generated in [26]. In general, before training classifiers or extracting PCA components from all of the data types, we standardized (took z-scores of) each column/feature of all data types. For the RNA sequencing dataset, we generally used only the top 8000 gene features by mean absolute deviation as predictors in our single-omics models, except where specified otherwise. For the RPPA, microRNA, and mutational signatures datasets, all columns/features were used.

To remove missing values from the methylation datasets, we removed the 10 samples with the most missing values, then performed mean imputation for probes with 1 or 2 values missing. All probes with missing values remaining after sample filtering and

imputation were dropped from the analysis. This left us with 20,040 CpG probes in the 27K methylation dataset and 370,961 CpG probes in the 450K methylation dataset. For experiments where "raw" methylation data was used, we used the top 100,000 probes in the 450K dataset by mean absolute deviation for computational efficiency, and we used all of the 20,040 probes in the 27K dataset. For experiments where "compressed" methylation data was used, we used principal component analysis (PCA), as implemented in the scikit-learn Python library [63], to extract the top 5000 principal components from the methylation datasets. We initially applied the beta-mixture quantile normalization (BMIQ) method [64] to correct for variability in signal intensity between type I and type II probes, but we observed that this had no effect on our results. We report uncorrected results in the main paper for simplicity.

Construction of a set of cancer genes

To get a comprehensive picture of classification performance across a wide variety of cancer-related genes, we integrated several curated gene sets from the literature into a single "merged" cancer gene set. The individual gene sets we integrated were from Vogelstein et al. [34] (all genes from Table S2A), Bailey et al. [35] (only genes annotated as "pan-cancer" drivers in Table S1), and the COSMIC Cancer Gene Census [36] (all Tier 1 genes annotated as "somatic"). In addition, the COSMIC CGC dataset contains 3 possible "roles in cancer" for each gene: oncogene, TSG, and fusion gene; for this analysis we dropped genes that are annotated only as fusion genes (i.e., no oncogene or TSG annotation). These filters resulted in a starting dataset of 511 cancer-related genes, which we reduced further for each experiment based on the number of mutated (i.e., positively labeled) samples as described in the next section.

Comparing data modalities

We made three main comparisons in this study: one between different sets of genes using only expression data, one comparing expression and DNA methylation data types, and one comparing all data types. This choice in comparisons was mainly due to sample size limitations, as running a single comparison using all data types would force us to use only samples that are profiled for every data type, which would discard a large number of samples that lack profiling on only one or a few data types. Thus, for each of the three comparisons, we used the intersection of TCGA samples having measurements for all of the datasets being compared in that experiment. This resulted in three distinct sets of samples: 9074 samples shared between {expression, mutation} data, 7981 samples shared between {expression, mutation, 27K methylation, 450K methylation}, and 5226 samples shared between {expression, mutation, 27K methylation, 450K methylation, RPPA, microRNA, mutational signatures}. When we dropped samples between experiments as progressively more data types were added, we observed that the dropped samples had approximately the same cancer type proportions as the dataset as a whole. In other words, samples that were profiled for one data type but not another did not tend to come exclusively from one or a few cancer types. Exceptions included acute myeloid leukemia (LAML) which had no samples profiled in the RPPA data, and ovarian cancer (OV) which had only 8 samples with 450K methylation data. More

detailed information on cancer type proportions profiled for each data type is provided in Additional file 1: Fig. S1 and Additional file 2.

For each target gene, in order to ensure that the training dataset was reasonably balanced (i.e., that there would be enough mutated samples to train an effective classifier), we included only cancer types with at least 15 mutated samples and at least 5% mutated samples, which we refer to here as "valid" cancer types. After applying these filters, the number of valid cancer types remaining for each gene varied based on the set of samples used: more data types resulted in fewer shared samples, and fewer samples generally meant fewer valid cancer types. In some cases, this resulted in genes with no valid cancer types, which we dropped from the analysis. Out of the 511 genes from the "merged" cancer gene set described in the previous section, for the analysis using {expression, mutation} data, we retained 268 target genes; for the {expression, mutation, 27k methylation, 450k methylation} analysis, we retained 272 genes; and for the analysis using all data types, we retained 217 genes.

We additionally explored mutation prediction from gene expression alone using three gene sets of equal size: the cancer-related genes from the merged dataset described previously, a set of frequently mutated genes in TCGA, and a set of random genes with mutations profiled by MC3. To match the size of the merged cancer gene set, we took the 268 most frequently mutated genes in TCGA as quantified by MC3, all of which had at least one valid cancer type. For the random gene set, we first filtered to the set of all genes with one or more valid cancer types by the same criteria (15 total samples mutated and at least 5% of samples mutated), then sampled 268 of the resulting 1348 genes uniformly at random. Based on the results of the gene expression experiments, we used the merged cancer-related gene set for all subsequent experiments comparing -omics data types.

Training classifiers to detect cancer mutations

We trained logistic regression classifiers to predict whether or not a given sample had a mutational event in a given target gene using data from various -omics datasets as explanatory variables. Our model was trained on -omics data (X) to predict mutation presence or absence (y) in a target gene. To control for varying mutation burden per sample and to adjust for potential cancer type-specific expression patterns, we included one-hot encoded cancer type and $\log_{10}$(sample mutation count) in the model as covariates. Since our -omics datasets tend to have many dimensions and comparatively few samples, we used an elastic net penalty to prevent overfitting [65] in line with the approach used in Way et al. [9] and Way et al. [13]. Elastic net logistic regression finds the feature weights $\hat{w} \in \mathbb{R}^p$ solving the following optimization problem:

$$\hat{w} = \text{argmin}_w \ell(X, y; w) + \alpha\lambda||w||_1 + \frac{1}{2}\alpha(1-\lambda)||w||_2$$

where $i \in \{1, \ldots, n\}$ denotes a sample in the dataset, $X_i \in \mathbb{R}^p$ denotes features (omics measurements) from the given sample, $y_i \in \{0, 1\}$ denotes the label (mutation presence/absence) for the given sample, and $\ell(\cdot)$ denotes the negative log-likelihood of the observed data given a particular choice of feature weights, i.e.

$$\ell(X, y; w) = -\sum_{i=1}^{n} y_i \log\left(\frac{1}{1+e^{-w^\top X_i}}\right) + (1-y_i) \log\left(1-\frac{1}{1+e^{-w^\top X_i}}\right)$$

This optimization problem leaves two hyperparameters to select: α (controlling the tradeoff between the data log-likelihood and the penalty on large feature weight values) and λ (controlling the tradeoff between the L1 penalty and L2 penalty on the weight values). Although the elastic net optimization problem does not have a closed form solution, the loss function is convex, and iterative optimization algorithms are commonly used for finding reasonable solutions. For fixed values of α and λ, we solved for $\hat{w}$ using stochastic gradient descent, as implemented in scikit-learn's SGDClassifier method.

Given weight values $\hat{w}$, it is straightforward to predict the probability of a positive label (mutation in the target gene) $P(y^* = 1 \mid X^*; \hat{w})$ for a test sample X^*:

$$P(y^* = 1 \mid X^*; \hat{w}) = \frac{1}{1+e^{-\hat{w}^\top X^*}}$$

and the probability of no mutation in the target gene, $P(y^* = 0 \mid X^*; \hat{w})$, is given by (1 – the above quantity).

For each target gene, we evaluated model performance using two replicates of 4-fold cross-validation, where train and test splits were stratified by cancer type and sample type. That is, each training set/test set combination had equal proportions of each cancer type (BRCA, SKCM, COAD, etc.) and each sample type (primary tumor, recurrent tumor, etc.). To choose the elastic net hyperparameters, we used 3-fold nested cross-validation, with a grid search over the following hyperparameter ranges: λ = [0.0, 0.05, 0.1, 0.3, 0.5, 0.7, 0.9, 1.0] and α = [0.0001, 0.001, 0.01, 0.1, 1, 10]. Using the grid search results, for each evaluation fold, we selected the set of hyperparameters with the optimal area under the precision-recall curve (AUPR), averaged over the three inner folds.

Evaluating mutation prediction classifiers

Area under the receiver-operator curve (AUROC) [66] and area under the precision-recall curve (AUPR) [67] are metrics that are frequently used to quantify classification performance for a continuous or probabilistic output, such as that provided by logistic regression. These metrics summarize performance across a variety of binary label thresholds, rather than requiring choice of a single threshold to determine positive or negative predictions. In the main text, we report results using AUPR, summarized using average precision. AUPR has been shown to distinguish between models more accurately than AUROC when there are few positively labeled samples [68, 69]. As an additional correction for imbalanced labels, in many of the results in the main text, we report the difference in AUPR between a classifier fit to true mutation labels and a classifier fit to data where the mutation labels are randomly permuted. In cases where mutation labels are highly imbalanced (very few mutated samples and many non-mutated samples), a classifier with permuted labels may perform well simply by chance, e.g., by predicting the negative/non-mutated class for most samples. To maintain the same label balance for the classifiers with permuted labels as the classifiers with the true labels, we permuted labels separately in the train and test sets for each cross-validation split. Additionally, to maintain the same label proportions within each cancer type after permuting the labels, we permuted labels independently for each cancer type.

Recall that for each target gene and each -omics dataset, we ran two replicates of 4-fold cross-validation, for a total of eight performance results. To make a statistical comparison between two models using these performance distributions, we used paired-sample *t*-tests, where performance measurements derived from the same cross-validation fold are considered paired measurements. We used this approach to compare a model trained on true labels with a model trained on permuted labels (addressing the question, "for the given gene using the given data type, can we predict mutation status better than random"), and to compare a model trained on data type A with a model trained on data type B (addressing the question, "for the given gene, can we make more effective mutation status predictions using data type A or data type B").

We corrected for multiple tests using a Benjamini-Hochberg false discovery rate correction. For experiments where we chose a binary threshold for accepting/rejecting H_0, we set a conservative corrected threshold of $p = 0.001$; we were able to estimate the number of false positives by examining genes with better performance for permuted mutation labels than true labels. We chose this threshold to ensure that none of the observed false positive genes were considered significant, since we would never expect permuting labels to improve performance. However, our results were not sensitive to the choice of this threshold, and we display cutoffs of $p = 0.05$ and $p = 0.01$ in many of our plots as well.

Survival prediction using -omics datasets

As a complementary comparison to mutation prediction, we constructed predictors of patient survival using the clinical data available from the GDC, in the TCGA-CDR-SupplementalTableS1.xlsx file. Following the methods described in [70], as the clinical endpoint we used overall survival (OS), except in nine cancer types with few deaths observed where we used progression-free intervals (PFI) as the clinical endpoint (BRCA, DLBC, LGG, PCPG, PRAD, READ, TGCT, THCA, and THYM). For prediction, we used Cox regression as implemented in the scikit-survival Python package [71], with patient age at diagnosis and $\log_{10}$(sample mutation count) included as covariates, as well as a one-hot encoded variable for cancer type. To ensure that the per-feature information content was comparable between -omics data types, we preprocessed the -omics datasets using PCA and extracted the top k principal components; in the case where the number of features in the original dataset was less than k, we used all available PCs (that is, we set $k = \min(p, k)$ where p is the number of features in the unprocessed dataset). We plot results over multiple values of k: $k \in \{10, 100, 500, 1000, 5000\}$.

To model pan-cancer survival (results shown in main paper), we used the elastic net Cox regression implementation in scikit-survival (i.e., the CoxnetSurvivalAnalysis method). To select hyperparameters for the elastic net Cox regression model, we performed a grid search over λ = [0.0, 0.05, 0.1, 0.3, 0.5, 0.7, 0.9, 1.0] and α = [0, 1e−5, 1e−4, 5e−4, 0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1, 10, 100, 1000].

We measured survival prediction performance using the censored concordance index (c-index) [72], which quantifies agreement between the order of survival time predictions and true outcomes for a held-out dataset; higher c-index values indicate more accurate survival prediction performance. Similar to the mutation prediction experiments, we calculated c-index values on held-out subsets of the data for two replicates

of 4-fold cross-validation, resulting in eight performance measurements for each model. As a baseline, we constructed survival models using only non-omics covariates. Covariates included patient age at diagnosis, $\log_{10}$(sample mutation count), and a one-hot encoded variable for sample cancer type.

Multi-omics mutation prediction experiments

To predict mutation presence or absence in cancer genes using multiple data types simultaneously, we concatenated individual datasets into a large feature matrix, then used the same elastic net logistic regression method described previously. For this task, we considered only the gene expression, 27K methylation, and 450K methylation datasets. We used only these data types to limit the number of multi-omics combinations; the expression and methylation datasets resulted in the best overall performance across the single-omics experiments, so we limited combinations to those datasets. In the main text, we report results using the top 5000 principal components for each dataset, which ensures that most variance is captured (approximately 95–98% of variance for each data type). In Additional file 1: Fig. S6, we also report results using "raw" features: for gene expression, we used all 15,639 genes available in our RNA sequencing dataset, and for the 27K and 450K methylation datasets, we used the top 20,000 CpG probes by mean absolute deviation.

To construct the multi-omics models, we considered each of the pairwise combinations of the datasets listed above, as well as a combination of all 3 datasets. When combining multiple datasets, we concatenated along the column axis and included covariates for cancer type and sample mutation burden as before. For all multi-omics experiments, we used only the samples from TCGA with data for all three data types (i.e., the same 7981 samples used in the single-omics experiments comparing expression and methylation data types). We considered a limited subset of well-performing genes from the merged cancer gene set as target genes, including *EGFR*, *IDH1*, *KRAS*, *PIK3CA*, *SETD2*, and *TP53*. We selected these genes because we had previously observed that they have good predictive performance and because they represent a combination of alterations that have strong gene expression signatures (*KRAS*, *EGFR*, *IDH1*, *TP53*) and strong DNA methylation signatures (*IDH1*, *SETD2*, *TP53*).

For the experiments predicting mutation status using a 3-layer fully connected neural network, described in the "Results" section and Additional file 1: Fig. S7, we used the top 5000 principal components as input for each data type. We selected hyperparameters for each of the 8 outer cross-validation splits using a single inner train/validation split and a random search over 20 hyperparameter combinations. The hyperparameter ranges that we sampled from in the random search were as follows: learning_rate [0.1, 0.01, 0.001, 5e−4, 1e−4], h1_size [1000, 500, 250], dropout [0.5, 0.75, 0.9], and weight_decay [0, 0.1, 1, 100]. Here, h1_size refers to the size of the first hidden layer, and the size of the second hidden layer was always set to h1_size/2. As in the elastic net grid search, we chose the combination of hyperparameters with the best AUPR on the validation set and retrained the model with those hyperparameters to make predictions on the test set. We trained our networks with the Adam optimizer [73], with ReLU activation after the hidden layers and sigmoid activation to make predictions, and using binary cross-entropy as the loss function as implemented in the

PyTorch BCEWithLogitsLoss function, through the skorch library which provides interoperability between PyTorch and scikit-learn.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-022-02705-y.

Additional file 1. All supplementary figures and supplementary information.

Additional file 2. Number of samples from each TCGA cancer type that are "dropped" as more data types are added to the analysis. The "base" column indicates the number of samples that are present per cancer type in the final intersection of all data types (i.e. each sample counted in the last column has data for each of the 7 data types, including gene expression (not listed here) and somatic mutations).

Additional file 3. Review history.

Acknowledgements

We would like to thank Alexandra Lee, Ariel Hippen, Ben Heil, Milton Pividori, and Natalie Davidson for reviewing the software associated with this work and providing insightful feedback. This research was supported in part by the University of Pittsburgh Center for Research Computing through the resources provided. Figure 1 (the schematic of the background and evaluation pipeline) was created using BioRender.com.

Review history

The review history is available as Additional file 3.

Peer review information

Wenjing She was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Authors' contributions

JC: conceptualization, methodology, software, visualization, writing—original draft, writing—review and editing. BCC: methodology, writing—review and editing. MC: methodology, writing—review and editing. CSG: conceptualization, funding acquisition, methodology, supervision, writing—review and editing. The authors read and approved the final manuscript.

Funding

This work was supported by grants from the National Institutes of Health's National Human Genome Research Institute (NHGRI) under award R01 HG010067 to CSG and the National Institutes of Health's National Cancer Institute (NCI) under awards R01 CA237170 to CSG, R01 CA216265 to BCC, and R01 CA253976 to BCC. The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Availability of data and materials

The datasets analyzed during this study were previously published as part of the TCGA Pan-Cancer Atlas project and are publicly available from the NIH NCI Genomic Data Commons (GDC) [74]. The mutational signatures dataset was downloaded from the ICGC Data Portal [75]. Scripts used to download and preprocess the datasets for this study are available at https://github.com/greenelab/mpmp/tree/master/00_download_data [76].

All analyses were implemented in the Python programming language and are available at Zenodo [77] and in the following GitHub repository: https://github.com/greenelab/mpmp [76] under the open-source BSD 3-clause license. Scripts to download large data files from GDC and other sources are located in the 00_download_data directory. Scripts to run experiments comparing data modalities used individually are located in the 02_classify_mutations directory, scripts to run multi-omics experiments are located in the 05_classify_mutations_multimodal directory, and scripts to run survival prediction experiments are located in the 06_predict_survival directory. The Python environment was managed using conda, and directions for setting up the environment can be found in the README.md file. Most analyses were run on the HTC CPU cluster at the University of Pittsburgh, except the neural networks which were trained and evaluated on the PMACS LPC GPU cluster at the University of Pennsylvania; scripts for training classifiers both locally for a single gene and on a Slurm cluster to reproduce the analysis of many genes in parallel are provided in the linked GitHub repo. This manuscript was written using Manubot [78] and is available on GitHub at https://github.com/greenelab/mpmp-manuscript under the CC0-1.0 license [79] and at Zenodo [80].

As a data resource, coefficients and hyperparameter choices for final models fit using all data types are available on Figshare: coefficients are available at https://doi.org/10.6084/m9.figshare.19576012 [81] and hyperparameters are at https://doi.org/10.6084/m9.figshare.19576048 [82]. File format/entries are described in the supplementary material in Additional file 1.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Genomics and Computational Biology Graduate Group, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. [2]Department of Epidemiology, Geisel School of Medicine, Dartmouth College, Lebanon, NH, USA. [3]Department of Molecular and Systems Biology, Geisel School of Medicine, Dartmouth College, Lebanon, NH, USA. [4]Department of Computational and Systems Biology, School of Medicine, University of Pittsburgh, Pittsburgh, PA, USA. [5]Department of Biochemistry and Molecular Genetics, University of Colorado School of Medicine, Aurora, CO, USA. [6]Center for Health AI, University of Colorado School of Medicine, Aurora, CO, USA.

Received: 15 November 2021 Accepted: 14 June 2022
Published online: 27 June 2022

References

1. Sanchez-Vega F, Mina M, Armenia J, Chatila WK, Luna A, La KC, et al. Oncogenic Signaling Pathways in The Cancer Genome Atlas. Cell. 2018;173:321–337.e10.
2. Smith JC, Sheltzer JM. Systematic identification of mutations and copy number alterations associated with cancer patient prognosis. eLife. 2018;7. https://doi.org/10.7554/eLife.39217.
3. Hofree M, Carter H, Kreisberg JF, Bandyopadhyay S, Mischel PS, Friend S, et al. Challenges in identifying cancer genes by analysis of exome sequencing data. Nat Commun. 2016;7(1). https://doi.org/10.1038/ncomms12096.
4. Tokheim CJ, Papadopoulos N, Kinzler KW, Vogelstein B, Karchin R. Evaluating the evaluation of cancer driver genes. Proc Natl Acad Sci U S A. 2016;113(50):14330–5. https://doi.org/10.1073/pnas.1616440113.
5. Zhao S, Liu J, Nanga P, Liu Y, Cicek AE, Knoblauch N, et al. Detailed modeling of positive selection improves detection of cancer driver genes. Nat Commun. 2019;10(1). https://doi.org/10.1038/s41467-019-11284-9.
6. Nussinov R, Jang H, Tsai C-J, Cheng F. Review: Precision medicine and driver mutations: computational methods, functional assays and conformational principles for interpreting cancer drivers. PLoS Comput Biol. 2019;15(3):e1006658. https://doi.org/10.1371/journal.pcbi.1006658.
7. Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. 2013;45(10):1113–20. https://doi.org/10.1038/ng.2764.
8. Guinney J, Ferté C, Dry J, McEwen R, Manceau G, Kao K, et al. Modeling RAS phenotype in colorectal cancer uncovers novel molecular traits of RAS dependency and improves prediction of response to targeted agents in patients. Clin Cancer Res. 2013;20(1):265–72. https://doi.org/10.1158/1078-0432.CCR-13-1943.
9. Way GP, Sanchez-Vega F, La K, Armenia J, Chatila WK, Luna A, et al. Machine learning detects Pan-cancer Ras pathway activation in The Cancer Genome Atlas. Cell Rep. 2018;23:172–180.e3.
10. Li X, Li S, Wang Y, Zhang S, Wong K-C. Identification of pan-cancer Ras pathway activation with deep learning. Brief Bioinform. 2020;22(4). https://doi.org/10.1093/bib/bbaa258.
11. Knijnenburg TA, Wang L, Zimmermann MT, Chambwe N, Gao GF, Cherniack AD, et al. Genomic and molecular landscape of DNA damage repair deficiency across The Cancer Genome Atlas. Cell Rep. 2018;23:239–254.e6.
12. Kang J, Lee A, Lee YS. Prediction of PIK3CA mutations from cancer gene expression data. PLoS One. 2020;15(11): e0241514. https://doi.org/10.1371/journal.pone.0241514.
13. Way GP, Zietz M, Rubinetti V, Himmelstein DS, Greene CS. Compressing gene expression data using multiple latent space dimensionalities learns complementary biological representations. Genome Biol. 2020;21(1):109. https://doi.org/10.1186/s13059-020-02021-3.
14. Grzadkowski MR, Holly HD, Somers J, Demir E. Systematic interrogation of mutation groupings reveals divergent downstream expression programs within key cancer genes. BMC Bioinformatics. 2021;22(1):1–34. https://doi.org/10.1186/s12859-021-04147-y.
15. Haan D, Tao R, Friedl V, Anastopoulos IN, Wong CK, Weinstein AS, et al. Using transcriptional signatures to find cancer drivers with LURE. In: Biocomputing 2020: World Scientific; 2019. https://doi.org/10.1142/9789811215636_0031.
16. Banerjee S, Simonetti FL, Detrois KE, Kaphle A, Mitra R, Nagial R, Söding J. Tejaas: reverse regression increases power for detecting trans-eQTLs. Genome Biol. 2021;22(1):1–6. https://doi.org/10.1186/s13059-021-02361-8.
17. Cieślik M, Chinnaiyan AM. Cancer transcriptome profiling at the juncture of clinical translation. Nat Rev Genet. 2017; 19(2):93–109. https://doi.org/10.1038/nrg.2017.96.
18. Noushmehr H, Weisenberger DJ, Diefes K, Phillips HS, Pujara K, Berman BP, et al. Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma. Cancer Cell. 2010;17(5):510–22. https://doi.org/10.1016/j.ccr.2010.03.017.
19. Christensen BC, Smith AA, Zheng S, Koestler DC, Houseman EA, Marsit CJ, et al. DNA methylation, isocitrate dehydrogenase mutation, and survival in glioma. J Natl Cancer Inst. 2011;103(2):143–53. https://doi.org/10.1093/jnci/djq497.
20. Yan H, Parsons DW, Jin G, McLendon R, Rasheed BA, Yuan W, et al. *IDH1* and *IDH2* mutations in gliomas. N Engl J Med. 2009;360(8):765–73. https://doi.org/10.1056/NEJMoa0808710.
21. Lu C, Ward PS, Kapoor GS, Rohle D, Turcan S, Abdel-Wahab O, et al. IDH mutation impairs histone demethylation and results in a block to cell differentiation. Nature. 2012;483(7390):474–8. https://doi.org/10.1038/nature10860.
22. Huang Y, Rao A. Connections between TET proteins and aberrant DNA modification in cancer. Trends Genet. 2014; 30(10):464–74. https://doi.org/10.1016/j.tig.2014.07.005.
23. Fahey CC, Davis IJ. SETting the stage for cancer development: SETD2 and the consequences of lost methylation. Cold Spring Harb Perspect Med. 2017;7(5):a026468. https://doi.org/10.1101/cshperspect.a026468.
24. Helleday T, Eshtad S, Nik-Zainal S. Mechanisms underlying mutational signatures in human cancers. Nat Rev Genet. 2014;15(9):585–98. https://doi.org/10.1038/nrg3729.
25. Nusinow DP, Szpyt J, Ghandi M, Rose CM, McDonald ER III, Kalocsay M, et al. Quantitative proteomics of the cancer cell line encyclopedia. Cell. 2020;180:387–402.e16.

26. Alexandrov LB, Kim J, Haradhvala NJ, Huang MN, Tian Ng AW, et al. The repertoire of mutational signatures in human cancer. Nature. 2020;578(7793):94–101. https://doi.org/10.1038/s41586-020-1943-3.
27. Chen Y-C, Gotea V, Margolin G, Elnitski L. Significant associations between driver gene mutations and DNA methylation alterations across many cancer types. PLoS Comput Biol. 2017;13(11):e1005840. https://doi.org/10.1371/journal.pcbi.1005840.
28. Youn A, Kim KI, Rabadan R, Tycko B, Shen Y, Wang S. A pan-cancer analysis of driver gene mutations, DNA methylation and gene expressions reveals that chromatin remodeling is a major mechanism inducing global changes in cancer epigenomes. BMC Med Genomics. 2018;11(1):98. https://doi.org/10.1186/s12920-018-0425-z.
29. Heim D, Montavon G, Hufnagl P, Müller K-R, Klauschen F. Computational analysis reveals histotype-dependent molecular profile and actionable mutation effects across cancers. Genome Med. 2018;10(1):83. https://doi.org/10.1186/s13073-018-0591-9.
30. Louhimo R, Hautaniemi S. CNAmet: an R package for integrating copy number, methylation and expression data. Bioinformatics. 2011;27(6):887–8. https://doi.org/10.1093/bioinformatics/btr019.
31. Jia P, Zhao Z. Impacts of somatic mutations on gene expression: an association perspective. Brief Bioinform. 2016;18:413–25. https://doi.org/10.1093/bib/bbw037.
32. Vaske CJ, Benz SC, Sanborn JZ, Earl D, Szeto C, Zhu J, et al. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. Bioinformatics. 2010;26(12):i237–45. https://doi.org/10.1093/bioinformatics/btq182.
33. Ding J, McConechy MK, Horlings HM, Ha G, Chun Chan F, Funnell T, et al. Systematic analysis of somatic mutations impacting gene expression in 12 tumour types. Nat Commun. 2015;6(1):8554. https://doi.org/10.1038/ncomms9554.
34. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Kinzler KW. Cancer genome landscapes. Science. 2013;339(6127):1546–58. https://doi.org/10.1126/science.1235122.
35. Bailey MH, Tokheim C, Porta-Pardo E, Sengupta S, Bertrand D, Weerasinghe A, et al. Comprehensive characterization of cancer driver genes and mutations. Cell. 2018;173:371–385.e18.
36. Sondka Z, Bamford S, Cole CG, Ward SA, Dunham I, Forbes SA. The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. Nat Rev Cancer. 2018;18(11):696–705. https://doi.org/10.1038/s41568-018-0060-1.
37. Bai H, Harmancı AS, Erson-Omay EZ, Li J, Coşkun S, Simon M, et al. Integrated genomic characterization of IDH1-mutant glioma malignant progression. Nat Genet. 2015;48(1):59–66. https://doi.org/10.1038/ng.3457.
38. Bertucci F, Borie N, Ginestier C, Groulet A, Charafe-Jauffret E, Adélaïde J, et al. Identification and validation of an ERBB2 gene expression signature in breast cancers. Oncogene. 2004;23(14):2564–75. https://doi.org/10.1038/sj.onc.1207361.
39. Robichaux JP, Elamin YY, Vijayan RSK, Nilsson MB, Hu L, He J, et al. Pan-cancer landscape and analysis of ERBB2 mutations identifies poziotinib as a clinically active inhibitor and enhancer of T-DM1 activity. Cancer Cell. 2019;36:444–457.e7.
40. McDaniel SL, Strahl BD. Shaping the cellular landscape with Set2/SETD2 methylation. Cell Mol Life Sci. 2017;74(18):3317–34. https://doi.org/10.1007/s00018-017-2517-x.
41. Li J, Lu Y, Akbani R, Ju Z, Roebuck PL, Liu W, et al. TCPA: a resource for cancer functional proteomics data. Nat Methods. 2013;10(11):1046–7. https://doi.org/10.1038/nmeth.2650.
42. Li Y, Roberts ND, Wala JA, Shapira O, Schumacher SE, et al. Patterns of somatic structural variation in human cancer genomes. Nature. 2020;578(7793):112–21. https://doi.org/10.1038/s41586-019-1913-9.
43. Chang A, Liu L, Ashby JM, Wu D, Chen Y, O'Neill SS, et al. Recruitment of KMT2C/MLL3 to DNA damage sites mediates DNA damage responses and regulates PARP inhibitor sensitivity in cancer. Cancer Res. 2021;81(12):3358–73. https://doi.org/10.1158/0008-5472.CAN-21-0688.
44. Levy JJ, Titus AJ, Petersen CL, Chen Y, Salas LA, Christensen BC. MethylNet: an automated and modular deep learning approach for DNA methylation analysis. BMC Bioinformatics. 2020;21(1):108. https://doi.org/10.1186/s12859-020-3443-8.
45. Levy JJ, Chen Y, Azizgolshani N, Petersen CL, Titus AJ, Moen EL, et al. MethylSPWNet and MethylCapsNet: biologically motivated organization of DNAm neural networks, inspired by capsule networks. NPJ Syst Biol Appl. 2021;7(1):33. https://doi.org/10.1038/s41540-021-00193-7.
46. Kuenzi BM, Park J, Fong SH, Sanchez KS, Lee J, Kreisberg JF, et al. Predicting drug response and synergy using a deep learning model of human cancer cells. Cancer Cell. 2020;38:672–684.e6.
47. Omar M, Mulder L, Coady T, Zanettini C, Imada EL, Dinalankara W, et al. Biological constraints can improve prediction in precision oncology. bioRxiv. 2021. https://doi.org/10.1101/2021.05.25.445604.
48. Piskol R, Ramaswami G, Li J. Reliable identification of genomic variants from RNA-Seq data. Am J Hum Genet. 2013;93(4):641–51. https://doi.org/10.1016/j.ajhg.2013.08.008.
49. Coudray A, Battenhouse AM, Bucher P, Iyer VR. Detection and benchmarking of somatic mutations in cancer genomes using RNA-seq data. PeerJ. 2018;6:e5362. https://doi.org/10.7717/peerj.5362.
50. Kristensen VN, Lingjærde OC, Russnes HG, Vollan HKM, Frigessi A, Børresen-Dale A-L. Principles and methods of integrative genomic analyses in cancer. Nat Rev Cancer. 2014;14(5):299–313. https://doi.org/10.1038/nrc3721.
51. Kim D, Li R, Dudek SM, Ritchie MD. Predicting censored survival data based on the interactions between meta-dimensional omics data in breast cancer. J Biomed Inform. 2015;56:220–8. https://doi.org/10.1016/j.jbi.2015.05.019.
52. Cappelli E, Felici G, Weitschek E. Combining DNA methylation and RNA sequencing data of cancer for supervised knowledge extraction. BioData Mining. 2018;11(1):22. https://doi.org/10.1186/s13040-018-0184-6.
53. Sharifi-Noghabi H, Zolotareva O, Collins CC, Ester M. MOLI: multi-omics late integration with deep neural networks for drug response prediction. Bioinformatics. 2019;35(14):i501–9. https://doi.org/10.1093/bioinformatics/btz318.
54. Uzunangelov V, Wong CK, Stuart JM. Accurate cancer phenotype prediction with AKLIMATE, a stacked kernel learner integrating multimodal genomic data and pathway knowledge. PLoS Comput Biol. 2021;17(4):e1008878. https://doi.org/10.1371/journal.pcbi.1008878.
55. Gligorijević V, Pržulj N. Methods for biological data integration: perspectives and challenges. J R Soc Interface. 2015;12(112):20150571. https://doi.org/10.1098/rsif.2015.0571.
56. Zitnik M, Nguyen F, Wang B, Leskovec J, Goldenberg A, Hoffman MM. Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities. Inform Fusion. 2019;50:71–91. https://doi.org/10.1016/j.inffus.2018.09.012.

57. Lin Y-HT, Way GP, Barwick BG, Mariano MC, Marcoulis M, Ferguson ID, et al. Integrated phosphoproteomics and transcriptional classifiers reveal hidden RAS signaling dynamics in multiple myeloma. Blood Adv. 2019;3(21):3214–27. https://doi.org/10.1182/bloodadvances.2019000303.
58. Ghandi M, Huang FW, Jané-Valbuena J, Kryukov GV, Lo CC, McDonald ER III, et al. Next-generation characterization of the Cancer Cell Line Encyclopedia. Nature. 2019;569(7757):503–8. https://doi.org/10.1038/s41586-019-1186-3.
59. Yu K, Chen B, Aran D, Charalel J, Yau C, Wolf DM, et al. Comprehensive transcriptomic analysis of cell lines as models of primary tumors across 22 tumor types. Nat Commun. 2019;10(1):3574. https://doi.org/10.1038/s41467-019-11415-2.
60. McGranahan N, Favero F, de Bruin EC, Birkbak NJ, Szallasi Z, Swanton C. Clonal status of actionable driver events and the timing of mutational processes in cancer evolution. Sci Transl Med. 2015;7(283):283ra54. https://doi.org/10.1126/scitranslmed.aaa1408.
61. Ellrott K, Bailey MH, Saksena G, Covington KR, Kandoth C, Stewart C, et al. Scalable open science approach for mutation calling of tumor exomes using multiple genomic pipelines. Cell Systems. 2018;6:271–281.e7.
62. Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R, Getz G. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol. 2011;12(4):R41. https://doi.org/10.1186/gb-2011-12-4-r41.
63. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Machine Learn Res. 2011;12:2825–30.
64. Teschendorff AE, Marabita F, Lechner M, Bartlett T, Tegner J, Gomez-Cabrero D, et al. A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. Bioinformatics. 2012;29(2):189–96. https://doi.org/10.1093/bioinformatics/bts680.
65. Zou H, Hastie T. Regularization and variable selection via the elastic net. J Royal Statistical Soc B. 2005;67(2):301–20. https://doi.org/10.1111/j.1467-9868.2005.00503.x.
66. Fawcett T. An introduction to ROC analysis. Pattern Recognition Letters. 2006;27(8):861–74. https://doi.org/10.1016/j.patrec.2005.10.010.
67. Raghavan V, Bollmann P, Jung GS. A critical investigation of recall and precision as measures of retrieval system performance. ACM Trans Inf Syst. 1989;7(3):205–29. https://doi.org/10.1145/65943.65945.
68. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One. 2015;10(3):e0118432. https://doi.org/10.1371/journal.pone.0118432.
69. Cao C, Chicco D, Hoffman MM. The MCC-F1 curve: a performance evaluation technique for binary classification. arXiv; 2020.
70. Smith JC, Sheltzer JM. Genome-wide identification and analysis of prognostic features in human cancers. Cell Rep. 2022;38(13):110569. https://doi.org/10.1016/j.celrep.2022.110569.
71. Pölsterl S. scikit-survival: a library for time-to-event analysis built on top of scikit-learn. J Machine Learn Res. 2020;21:1–6.
72. Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med. 1996;15(4):361–87. https://doi.org/10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4.
73. Kingma DP, Ba J. Adam: a method for stochastic optimization. arXiv; 2017.
74. The Cancer Genome Atlas (TCGA) Research Network. TCGA Pan-Cancer Atlas. Webpage. 2022. https://gdc.cancer.gov/about-data/publications/pancanatlas. Accessed 7 Dec 2020.
75. ICGC-TCGA Pan-Cancer Analysis Mutational Signatures Working Group. ICGC/TCGA Mutational Signatures. Webpage. 2022. https://dcc.icgc.org/releases/PCAWG/mutational_signatures. Accessed 5 Mar 2021.
76. Crawford J, Christensen BC, Chikina M, Greene CS. Code for 'Widespread redundancy in -omics profiles of cancer mutation states'. GitHub. 2022. https://github.com/greenelab/mpmp.
77. Crawford J, Christensen BC, Chikina M, Greene CS. Code for 'Widespread redundancy in -omics profiles of cancer mutation states'. Zenodo. 2022. https://doi.org/10.5281/zenodo.6590133.
78. Himmelstein DS, Rubinetti V, Slochower DR, Hu D, Malladi VS, Greene CS, et al. Open collaborative writing with Manubot. PLoS Comput Biol. 2019;15(6):e1007128. https://doi.org/10.1371/journal.pcbi.1007128.
79. Crawford J, Christensen BC, Chikina M, Greene CS. Widespread redundancy in -omics profiles of cancer mutation states. GitHub. 2022. https://github.com/greenelab/mpmp-manuscript.
80. Crawford J, Christensen BC, Chikina M, Greene CS Widespread redundancy in -omics profiles of cancer mutation states. Zenodo. 2022. https://doi.org/10.5281/zenodo.6638831.
81. Crawford J, Christensen BC, Chikina M, Greene CS. Widespread redundancy in -omics profiles of cancer mutation states. figshare. 2022. https://doi.org/10.6084/m9.figshare.19576012.
82. Crawford J, Christensen BC, Chikina M, Greene CS. Widespread redundancy in -omics profiles of cancer mutation states. figshare. 2022. https://doi.org/10.6084/m9.figshare.19576048.

Publisher's Note