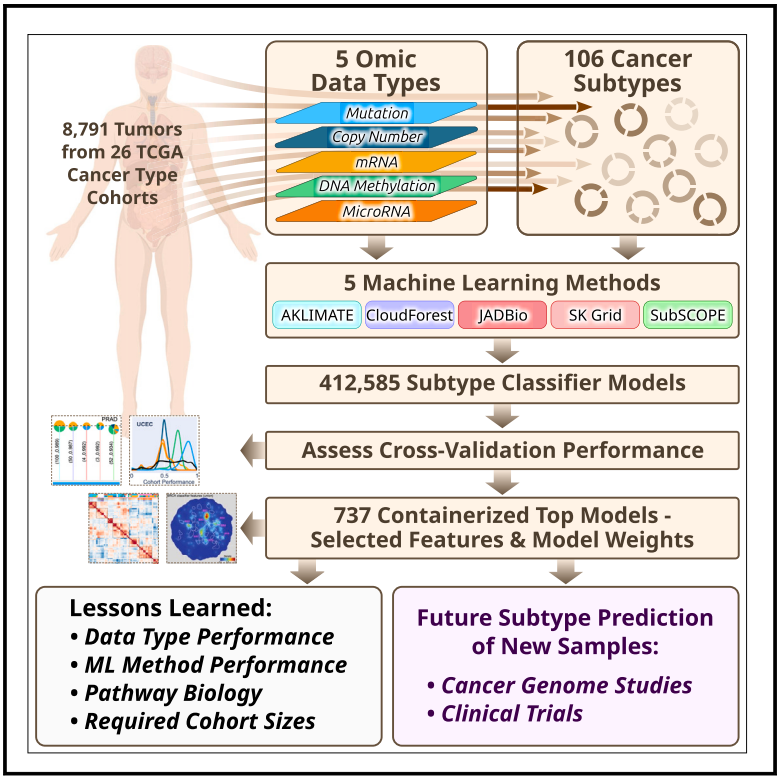
Cancer Cell

Article

Classification of non-TCGA cancer samples to TCGA molecular subtypes using compact feature sets

Graphical abstract



Highlights

- Five machine learning methods used to train classifiers for 106 TCGA cancer subtypes
- Classifier models based upon small numbers of features from five omic data types
- Insights into methods, data types, and cohort size for effective classification
- Public resource for classifying non-TCGA patient samples to TCGA cancer subtypes

Authors

Kyle Ellrott, Christopher K. Wong, Christina Yau, ..., Jean C. Zenklusen, Andrew D. Cherniack, Peter W. Laird

Correspondence

ellrott@ohsu.edu (K.E.), achernia@broadinstitute.org (A.D.C.), peter.laird@vai.org (P.W.L.)

In brief

Ellrott et al. provide a means to assign patient samples from clinical trials and other cancer genome studies to published TCGA molecular subtypes. Applying machine learning to data from five different molecular platforms for 8,791 TCGA tumor samples, they contribute a public resource of 737 top classifier models, which can form the foundation for clinical assay development.

Ellrott et al., 2025, Cancer Cell *43*, 1–18
February 10, 2025
https://doi.org/10.1016/j.ccell.2024.12.002

CellPress

Cancer Cell

CellPress
OPEN ACCESS

Article

Classification of non-TCGA cancer samples to TCGA molecular subtypes using compact feature sets

Kyle Ellrott,[1,25,*] Christopher K. Wong,[2,25] Christina Yau,[3,4,25] Mauro A.A. Castro,[5,25] Jordan A. Lee,[1,25] Brian J. Karlberg,[1,25] Jasleen K. Grewal,[6,19,25] Vincenzo Lagani,[7,8,20,25] Bahar Tercan,[9,25] Verena Friedl,[2] Toshinori Hinoue,[10] Vladislav Uzunangelov,[2,21] Lindsay Westlake,[11,12] Xavier Loinaz,[11] Ina Felau,[13] Peggy I. Wang,[13] Anab Kemal,[13] Samantha J. Caesar-Johnson,[13] Ilya Shmulevich,[9] Alexander J. Lazar,[14] Ioannis Tsamardinos,[7,15,16] Katherine A. Hoadley,[17] The Cancer Genome Atlas Analysis Network,[24] A. Gordon Robertson,[6,22] Theo A. Knijnenburg,[9,23] Christopher C. Benz,[4] Joshua M. Stuart,[2] Jean C. Zenklusen,[13] Andrew D. Cherniack,[11,12,18,26,*] and Peter W. Laird[10,26,27,*]

[1]Oregon Health and Science University, Portland, OR 97239, USA
[2]Biomolecular Engineering Department, School of Engineering, University of California, Santa Cruz, Santa Cruz, CA 95064, USA
[3]University of California, San Francisco, Department of Surgery, San Francisco, CA 94158, USA
[4]Buck Institute for Research on Aging, Novato, CA 94945, USA
[5]Bioinformatics and Systems Biology Laboratory, Federal University of Paraná, Curitiba, PR 81520-260, Brazil
[6]Canada's Michael Smith Genome Sciences Centre, BC Cancer, Vancouver, BC, Canada
[7]JADBio Gnosis DA, GR-700 13 Heraklion, Crete, Greece
[8]Institute of Chemical Biology, Ilia State University, Tbilisi 0162, Georgia
[9]Institute for Systems Biology, 401 Terry Avenue North, Seattle, WA 98109, USA
[10]Department of Epigenetics, Van Andel Institute, Grand Rapids, MI 49503, USA
[11]The Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA
[12]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA 02215, USA
[13]Center for Cancer Genomics, National Cancer Institute, Bethesda, MD 20892, USA
[14]Departments of Pathology & Genomic Medicine, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA
[15]Department of Computer Science, University of Crete, GR-700 13 Heraklion, Crete, Greece
[16]Institute of Applied and Computational Mathematics, Foundation for Research and Technology Hellas (FORTH), GR-700 13 Heraklion, Crete, Greece
[17]Department of Genetics, Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, NC 27519, USA
[18]Harvard Medical School, Boston, MA 02115, USA
[19]Present address: NVIDIA Corporation, Santa Clara, CA, USA
[20]Present address: Biological and Environmental Sciences and Engineering Division (BESE), SDAIA-KAUST Center of Excellence in Data Science and Artificial Intelligence, King Abdullah University of Science and Technology (KAUST), Thuwal, 23955, Saudi Arabia
[21]Present address: Predictive Sciences, Research & Early Development, Bristol Myers Squibb, Redwood City, CA 94063, USA
[22]Present address: Dxige Research Inc., Courtenay, BC, V9N 1C2; Canada
[23]Present address: Altius Institute for Biomedical Sciences, Seattle, WA 98121, USA
[24]Further details can be found in the supplemental information
[25]These authors contributed equally
[26]These authors contributed equally
[27]Lead contact
*Correspondence: ellrott@ohsu.edu (K.E.), achernia@broadinstitute.org (A.D.C.), peter.laird@vai.org (P.W.L.)
https://doi.org/10.1016/j.ccell.2024.12.002

SUMMARY

Molecular subtypes, such as defined by The Cancer Genome Atlas (TCGA), delineate a cancer's underlying biology, bringing hope to inform a patient's prognosis and treatment plan. However, most approaches used in the discovery of subtypes are not suitable for assigning subtype labels to new cancer specimens from other studies or clinical trials. Here, we address this barrier by applying five different machine learning approaches to multi-omic data from 8,791 TCGA tumor samples comprising 106 subtypes from 26 different cancer cohorts to build models based upon small numbers of features that can classify new samples into previously defined TCGA molecular subtypes—a step toward molecular subtype application in the clinic. We validate select classifiers using external datasets. Predictive performance and classifier-selected features yield insight into the different machine-learning approaches and genomic data platforms. For each cancer and data type we provide containerized versions of the top-performing models as a public resource.

Please cite this article in press as: Ellrott et al., Classification of non-TCGA cancer samples to TCGA molecular subtypes using compact feature sets, Cancer Cell (2024), https://doi.org/10.1016/j.ccell.2024.12.002

INTRODUCTION

Cancers have traditionally been classified by their originating organ or anatomic site, and within a cancer type by histological features, morphologic grade, and AJCC/UICC TNM stage.[1–3] Such cancer subtyping informs prognosis, guiding therapeutic approaches or surgical interventions. The separation of cancers by site of origin is substantiated by cancer genome studies, which find distinct genomic and transcriptomic biology associated with the main cancer types, reflective of their different tissue origins.[4–9] Large-scale cancer genome projects, such as The Cancer Genome Atlas (TCGA), also reveal previously unrecognized molecular heterogeneity and discrete subgroups within cancer types, providing an opportunity to enhance cancer classification by defining molecular subtypes within cancer types.

Molecular subtypes may eventually complement and even supersede traditional histopathological classifications.[10–16] However, clinical use of molecular subtyping for most cancers remains in its infancy. Clinical implementation of molecularly defined subtypes requires straightforward, accurate, and reproducible clinical assays that can place a new tumor into a previously defined molecular subtype classification scheme. Molecular subtypes are often initially identified using unsupervised or integrative clustering methods that produce results intrinsic to that specific dataset, and which do not carry over to patients from a different dataset. Gene expression signatures generated during subtype discovery may inform on the underlying differences in biology among the subtypes, but they have substantial feature redundancy, have not been tested by cross-validation or validation on other sample sets, are overfit for the subtype discovery dataset, and usually have low predictive power for samples from other studies.

In the work reported here, we start to bridge the gap between the discovery of molecular subtypes in an existing cancer cohort, and the application of these subtype labels for a newly diagnosed patient in the clinic. Using five different machine-learning (ML) methods, we trained classifier models that reduce feature redundancy and constrain or minimize feature numbers, while using cross-validation strategies to reduce overfitting, and to assess prediction performance. We produced 412,585 distinct classifier models, incorporating five different data types for 8,791 TCGA samples, comprising 26 different cancer cohorts and 106 subtypes. We created an online resource of 737 publicly available, containerized predictive models, representing the top models for each of the 26 cancer cohorts, five training algorithms, and data types. Our collection of classification models provides a rich resource of gene-based feature sets for the creation of compact cancer testing panels and kits to clinically subtype non-TCGA patient tumor samples.

RESULTS

TCGA tumor subtype definitions and classification model development

Tumor classification in the patient care setting generally starts with a known cancer type, informed by histopathology and anatomic location. We therefore took a cancer-type-centric approach to subtype classification. We retrieved the molecular subtypes reported by TCGA, which had defined cancer subtypes using appropriate data types and methods for each cohort (Figure 1; Table S1). For cancer cohorts with partially overlapping and related subtypes, we merged TCGA cohorts, resulting in 26 distinct cancer cohorts.

We assembled all genomic data for five data platforms (mutation, copy number, mRNA, DNA methylation, and miRNA) from PanCancer Atlas resources (gdc.cancer.gov/node/977). We used a gene-centric approach to facilitate the analysis of the biological significance of the selected features. To simplify future clinical assay development, we emphasized the selection of fewer features while retaining predictive performance.

We constructed subtype-balanced repeated cross-validation folds for all cohorts, and set these as training and test sets (Figure 2A). We applied five ML approaches: AKLIMATE,[17] CloudForest,[18] SK Grid,[19] JADBio,[20] and subSCOPE[8] (Figure 2B; Table S2). SK Grid and JADBio each employed a collection of embedded approaches, thus the true number of approaches implicitly tested in our study is far more than five. For AKLIMATE, CloudForest, SK Grid and JADBio, each cancer cohort was trained separately. In contrast, we trained subSCOPE's Neural Nets (NNs) on subtype data from all cancer cohorts simultaneously. We trained and tested all classifiers using the same cross-folds, and aggregated results into a single matrix. We generated performance statistics from the test cross-folds, and retained the classifier-selected features for further analysis. Because the word "accuracy" has a specific statistical meaning in the predictive model literature,[21] we avoided the colloquial usage of the term "accuracy," and preferred the term "performance," as defined by the overall weighted F1 score.[22]

Data types used to define subtypes influence prediction performance and classifier-selected features

The top models from each method demonstrated similar cohort-level performance (Figure S1). Differences in performance among subtypes within a cohort varied from 0.00 in ESCC and TGCT to 0.37 in COADREAD (Figure 3A). Four striking observations emerged from a comparison of prediction performances and selected features across cancer types. First, cancer cohorts with subtypes defined in the original TCGA publications by multi-omics or by histology often yielded highly accurate classifiers (Figure 3A). Second, for most cancer types, mRNAs predominated among features selected in the top models (Figure 3B; Tables S3, S4, and S5). Third, classifiers for cancer cohorts originally subtyped using mutation (SKCM) or DNA methylation (LGGGBM) often selected features from the corresponding data types for the top models (Figure 3B). Fourth, subtypes that were defined using summary statistics of genome-wide features such as mutational load, chromosomal instability, and CpG island methylator phenotype (CIMP) (e.g., GEA and COADREAD) are difficult to capture using individual gene-centric features, as used in our training, and thus the resulting classifiers achieved relatively low performance.

Models recapitulate PAM50 assignments in external validation tests

We investigated whether our BRCA subtype mRNA classifiers could accurately predict PAM50 label assignments[23] in two independent breast cancer cohorts. We used the METABRIC[24]

Please cite this article in press as: Ellrott et al., Classification of non-TCGA cancer samples to TCGA molecular subtypes using compact feature sets, Cancer Cell (2024), https://doi.org/10.1016/j.ccell.2024.12.002

CellPress
OPEN ACCESS

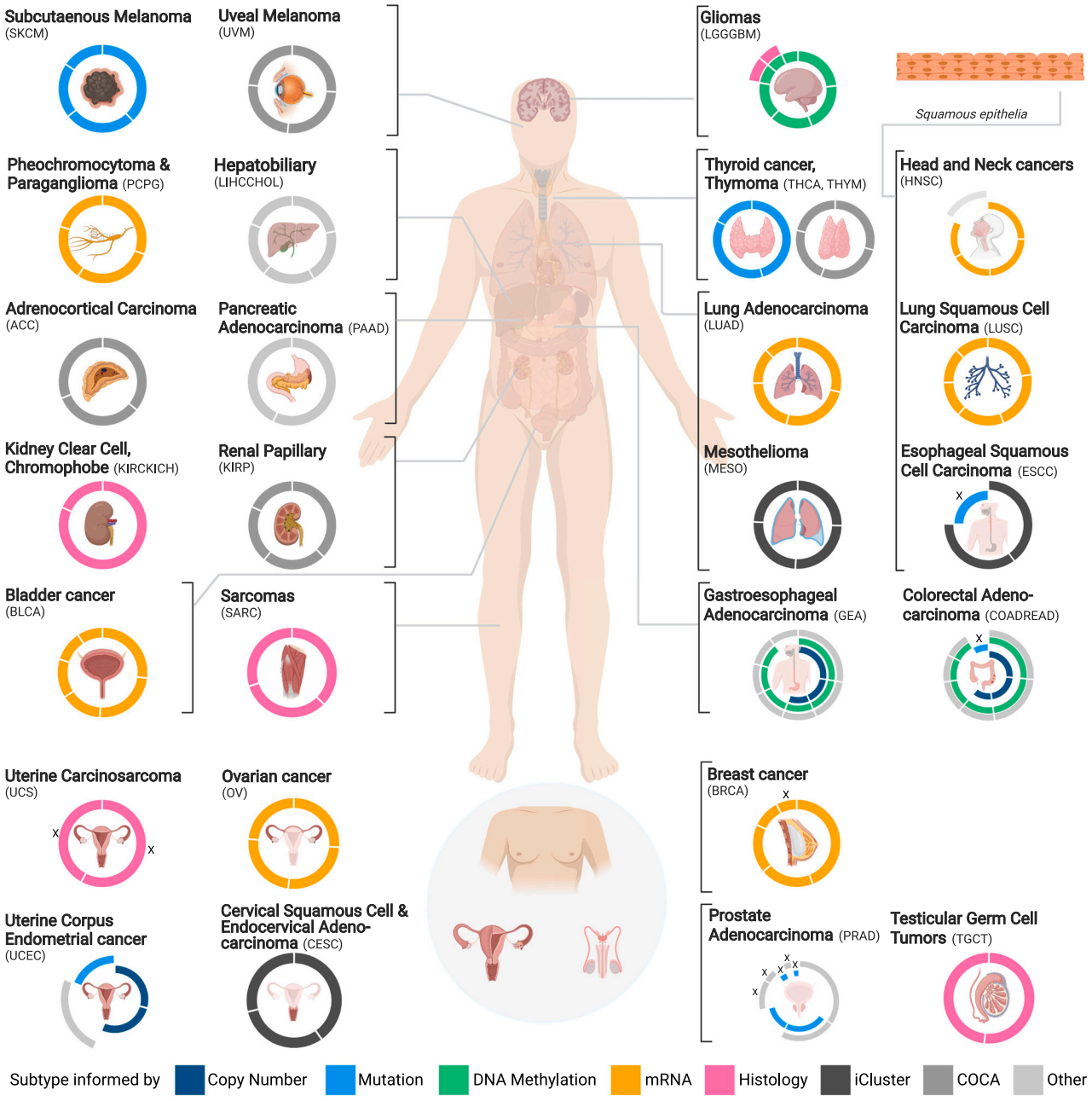


Figure 1. Cancer types and subtyping

An overview of the cancer cohorts and subtypes studied as part of this project, color-coordinated by the genomic data type(s) used to define the subtypes. For a given cancer type, subtypes are indicated with a ring around the corresponding inset organ view. Breaks in each ring distinguish the subtypes. In cases where the subtype is informed by more than one data type, concentric arcs are shown. Only subtypes with two or more samples are shown; "x"s mark small subtypes that were excluded from classifier development. See also Table S1.

and AURORA[25] breast cancer cohorts datasets, which offered challenges that we anticipate may be encountered in applying our models to other studies: METABRIC data were produced on a different mRNA platform—expression microarrays, while AURORA represents a small cohort with formalin-fixed, paraffin-embedded (FFPE) samples. We applied the SK Grid and AKLIMATE mRNA models to transformed METABRIC data, and obtained predictions of the PAM50 subtypes highly concordant with the PAM50 assignments in the original METABRIC publication (Figures 3C and 3D).[24] Samples with discordant classifier calls between METABRIC's PAM50 assignments and either SK Grid or AKLIMATE model predictions were more likely to have negative silhouette scores[26] ($p = 5.6 \times 10^{-15}$, one-sided Mann-Whitney Wilcoxon Test), representing samples with less robust PAM50 classification assignments (Figure 3C). The silhouette score represents a metric for

Please cite this article in press as: Ellrott et al., Classification of non-TCGA cancer samples to TCGA molecular subtypes using compact feature sets, Cancer Cell (2024), https://doi.org/10.1016/j.ccell.2024.12.002

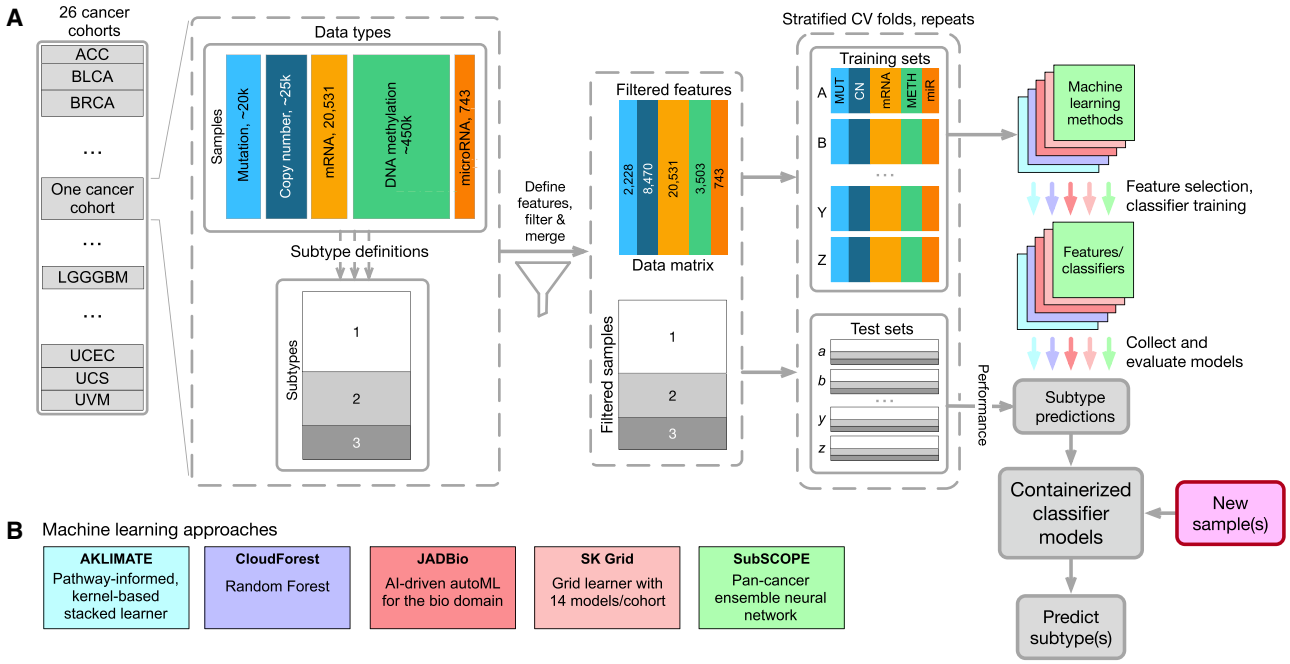


Figure 2. Process workflow

(A) Analysis workflow for classifier training and testing for an example cohort, using somatic mutations, copy number alterations, DNA methylation, and expression data for mRNAs and for miRNA mature strands. The (approximate) number of features for each data type in the original genomic data are provided, followed by the number in the filtered feature matrix (medians across all 26 cohorts).

(B) The five ML approaches used in this study. MUT, mutation; CN, copy number; METH, DNA methylation. See also Table S2.

how similar that sample is to its assigned class, relative to the next-best class.[26]

The AKLIMATE model outputs a prediction probability for each subtype assignment for each tested sample. We used the difference in prediction probabilities for the top two subtype calls for each sample to estimate the AKLIMATE subtype prediction confidence for that sample. The difference would be expected to be small in cases where the sample subtype is not confidently predicted, because the top two predicted subtypes would have similar probabilities. We found that this estimate of subtype prediction confidence correlated very well with the silhouette scores obtained for the original METABRIC assignments, particularly for the luminal A (Spearman Rho = 0.69, $p = 5 \times 10^{-67}$) and basal-like (Spearman Rho = 0.60, $p = 10^{-12}$) subtypes (Figure 3E).

To assess the degree to which technical differences in the data production platform in other datasets may lower classifier performance, we investigated the performance of a rank-based machine-learning approach called multiclassPairs,[27] a multi-class version of the binary (two-class) k-top scoring pairs (kTSP) classifier,[28] which would be expected to be more tolerant of platform technology differences. This approach performed comparably to the other methods we tested with gene expression data (Figures S2A–S2D). While the rank-based multiclassPairs was never the top-performing method, its relative insensitivity to data platform may prove to be a benefit in some applications.

We investigated whether training classifiers on the subset of samples with high silhouette score subtype calls would produce higher-ranked models, or whether focusing on the subset of samples with low silhouette scores would instead yield better performance. We compared equally sized data subsets, enriching for either high silhouette scores, which we refer to as the "typical set," or enriching for low silhouette scores "atypical set," versus no enrichment for silhouette scores "full set" (Figure S2E). Remarkably, we found that building models using samples from the "full set," which includes both typical subtype core samples and ambiguously assigned samples, produced models with the highest concordance to the original METABRIC subtype assignments.

We investigated whether our models would perform well on a small cohort of FFPE-processed primary breast tumor samples from the AURORA study,[25] since FFPE processing is known to affect gene expression data characteristics.[29] We achieved a model performance similar to that obtained for METABRIC (Figures S2F–S2G).

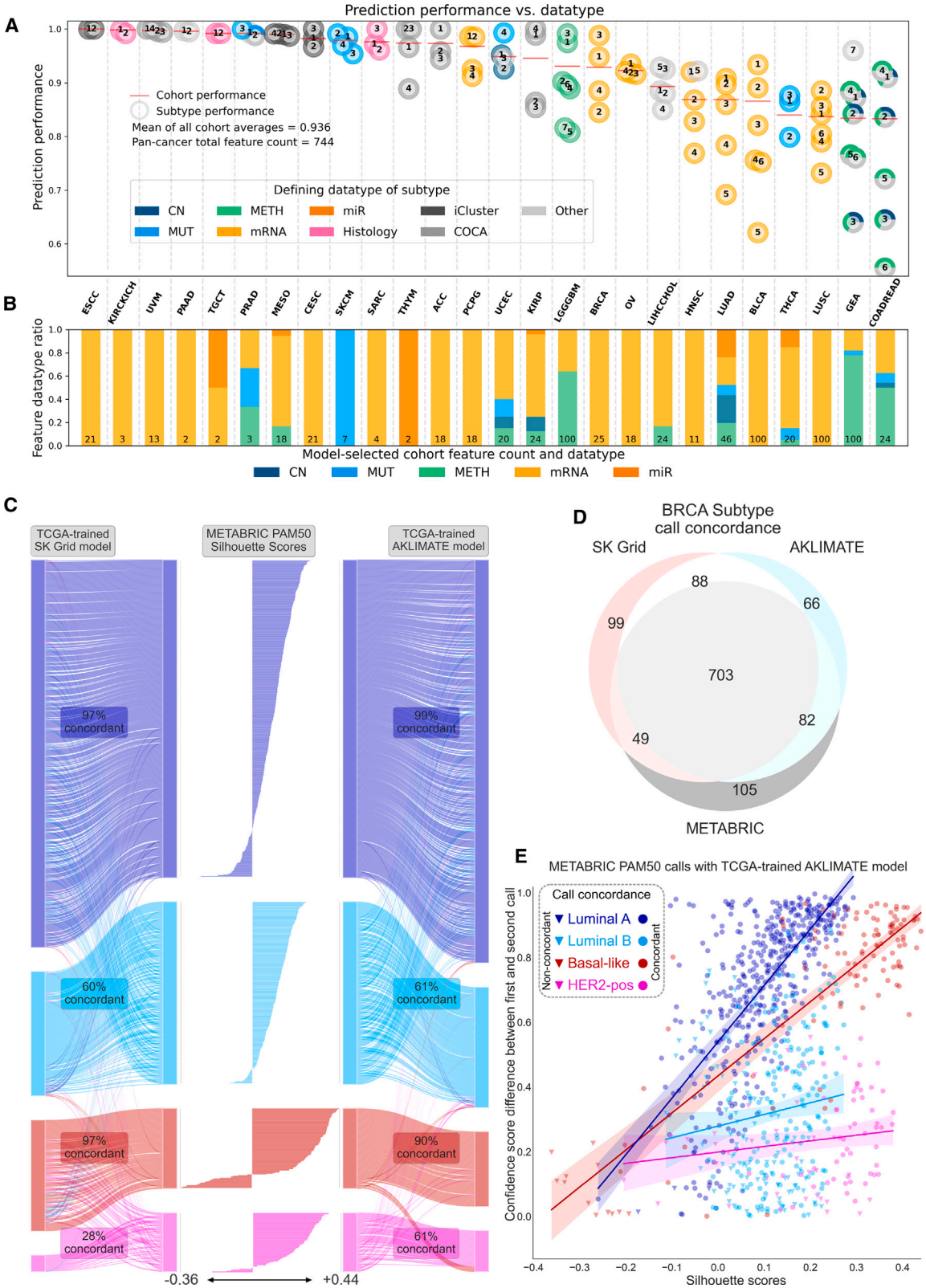
More input data types or numerous features do not drive model performance

To assess whether single data types can be used to efficiently categorize samples into subtypes, we compared classifier performance using all platforms jointly as data inputs with the performance obtained using each individual platform alone (Figures S3A and S3B). For half of the cancer cohorts, the best model built using a single data type as input achieved performance as good as models constructed using all data types jointly (Figure 4A; indicated by asterisks).

The five ML methods selected very different numbers of features, with some methods (e.g., CloudForest) constrained *a priori*, and others (e.g., JADBio) purposefully designed to select few

Please cite this article in press as: Ellrott et al., Classification of non-TCGA cancer samples to TCGA molecular subtypes using compact feature sets, Cancer Cell (2024), https://doi.org/10.1016/j.ccell.2024.12.002

CellPress
OPEN ACCESS



(legend on next page)

Please cite this article in press as: Ellrott et al., Classification of non-TCGA cancer samples to TCGA molecular subtypes using compact feature sets, Cancer Cell (2024), https://doi.org/10.1016/j.ccell.2024.12.002

features. To estimate which methods could achieve better performance per allowed feature, we determined the performance of each method, while constraining the number of features in the model outputs at varying feature set sizes, and calculated areas under the curve (AUC) for individual tumor types (Figure 4B). We refer to this aggregate measure across a range of feature set sizes as the ''cohort AUC.'' We found that this metric, which adjusts for the number of features, yielded similar results for the three methods analyzed (AKLIMATE, CloudForest, and JADBio), with JADBio often displaying slightly higher cohort AUC than the other two methods, reflecting a more efficient selection of features (Figure 4B). The shape of these curves revealed that performance rapidly plateaus at approximately 10 classifier features, suggesting parsimonious classifiers with very few features are sufficient to recapitulate the subtypes for most cohorts.

mRNA features predominate in top models for most cancer types

Models using mRNA feature inputs performed well on cancer cohorts with subtypes defined by multi-omics (Figures 4A and 4B). This suggests that subtypes defined by multiple data types or histology may represent distinct biologies that are relatively easy to capture at the transcriptome level. Classifiers developed using gene expression feature inputs significantly outperformed models derived using the next best single data types in 10 out of 26 cancer cohorts (Table S4). In the few tumor types for which features from data types other than mRNA were significantly more predictive, the best classifier-selected data type generally matched the one originally used to define the subtypes. For example, for LGGGBM and GEA, DNA methylation input features yielded significantly more accurate models than those that used gene expression features (Figures 4A and 4B; Table S4). SKCM subtypes were originally defined by mutation features, and models built with mutation input features significantly outperformed models using mRNA features (Figure 4B; Table S4).

Features shared across methods reflect known tumor biology

The ML literature suggests that for high-dimensional datasets it can be difficult for the same method to find a reproducible feature set.[30–32] Highly correlated features such as co-regulated genes can exacerbate this issue. Nonetheless, when two or more ML methods select the same feature, this feature likely provides a particularly strong signal on which to base subtype classification. When we compared the overlap in features selected by the best-performing models for each method, we found that models with fewer features tended to display a much stronger degree of feature overlap with other models. For example, we noted that SK Grid selected the smallest number of features (n = 9) for BRCA (Figure 5A), and all nine of these features were shared with at least one other BRCA model. Most features were selected by a single method, and some of these unique features may represent distinct members of a larger correlated set that captures the same biology. When the number of selected features is constrained, methods may be forced to choose arbitrarily among sets of equally good predictors. Newer feature-selection methods,[33] some of which are incorporated into JADBio, try to identify multiple, equally and optimally predictive subsets of features to address this issue.[20]

For each cohort's subtypes, we defined features that were selected by two or more methods as the *core set* of features. We found that core sets were enriched in biological themes. For example, for BRCA subtypes, all five methods selected the two mRNA features, *ESR1* and *FOXC1* (Figure 5A). Both are important in breast development and in breast cancer.[39–41] *ESR1* encodes the estrogen receptor alpha, an important biomarker in breast cancer classification, prognosis, and predicting treatment response. *FOXC1* encodes a chromatin remodeling factor that drives cell plasticity and partial EMT,[41] with expression strongly anti-correlated to the pioneer transcription factor *FOXA1* (selected by two methods), which interacts with *ESR1* to promote luminal cell fates.[42] Of the 38 mRNA features in the BRCA core set, 17 were genes represented in the original PAM50 breast cancer panel.[23]

For COADREAD subtypes, DNA methylation features dominated most ML approaches (Figure 5B), consistent with a strong influence of DNA methylation in the original subtype definitions. Many of the DNA methylation features have previously been reported to be hypermethylated in colorectal adenocarcinomas with a CpG island methylator phenotype (CIMP-high), or were identified as epigenetically silenced in previous studies (genes indicated in red font).[35,36] Notably, promoter methylation of *MLH1*, which is responsible for most sporadic cases of microsatellite instability (MSI), was selected by four out of five methods; similarly, *SFRP5*, a negative regulator of Wnt signaling, and frequently silenced by promoter methylation in CIMP-high colorectal cancer,[43] was selected by four methods.

For SKCM subtypes, the core feature set matched the somatic mutations (*NRAS*, *BRAF*, and *NF1*) used in the original subtype definitions (Figure 5C).[44] LGGGBM subtypes were originally defined by first splitting cases into *IDH1*/*IDH2* mutant versus wildtype cases, and then by further defining subtypes within each of these groups using DNA methylation profiles and

Figure 3. Overview of classifier performance metrics

(A) Classifier performance for each subtype in the 26 tumor type cohorts, representing the mean of the overall weighted F1 score of the most accurate model for predicting the subtypes within each tumor type (horizontal red bars). Subtype performance is plotted as round markers, numbered by subtype and colored by the data type used originally to define that subtype.

(B) Proportion of model-selected feature-set data types for the top model in each cohort. At the base of each stacked bar is the number of gene-based features utilized by a cohort's subtype classifier.

(C) Concordance of the original METABRIC PAM50 calls to SK Grid (left) and AKLIMATE (right) classifications. The central horizontal bars depict the silhouette scores for each sample.

(D) Venn diagram summarizing the union and intersection counts of samples across the METABRIC validation experiment.

(E) Comparison of sample silhouette scores vs. the difference in confidence score between the best and second-best sample prediction confidence calls for AKLIMATE; colored by subtype. Circles indicate samples with concordant calls and triangles indicate samples with discordant calls. The linear regression trend lines for each subtype, with associated 95% confidence intervals, are shown. See also Figures S1, S2, and Tables S3, S4, and S5.

Please cite this article in press as: Ellrott et al., Classification of non-TCGA cancer samples to TCGA molecular subtypes using compact feature sets, Cancer Cell (2024), https://doi.org/10.1016/j.ccell.2024.12.002

CellPress
OPEN ACCESS

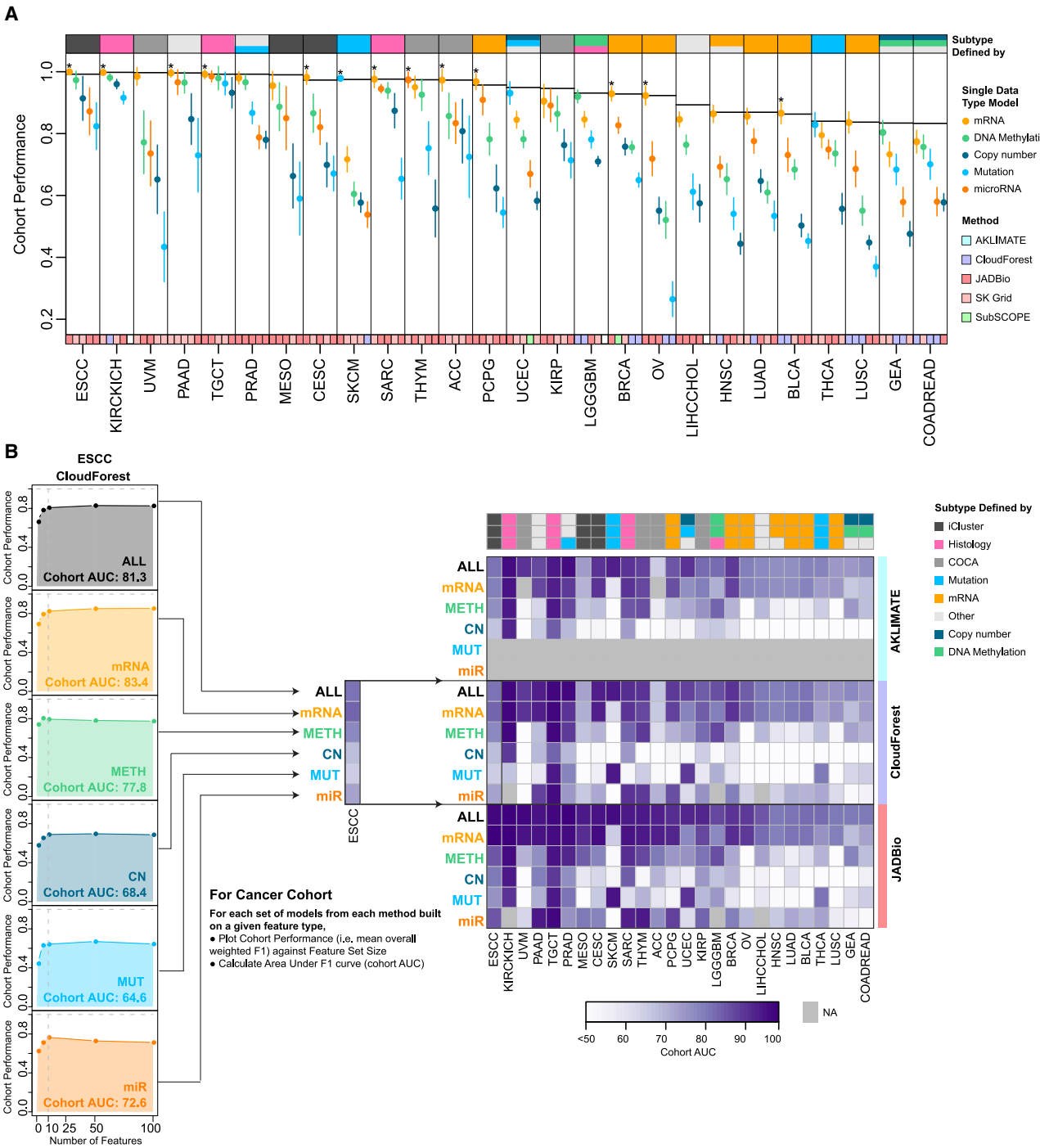


Figure 4. Performance of models using single data types vs. multi-omics

(A) The best-performing model for each data type is indicated by colored dot for each cohort, with vertical bars representing the range across subtypes. Asterisks denote cohorts where the top single data type model achieved performance equivalent to or better than the top multi-omics model performance, which is indicated by a horizontal black bar. The upper annotation track indicates the data type(s) originally used to define the subtypes. The bottom annotation bar indicates the method that produced the top models.

(B) Influence of feature set size on performance. For each cancer cohort, for each method and data type, a plot of cohort performance as a function of *a priori* defined feature set size produces an area under the curve (AUF1C). As an example, the curves for the ESCC cohort for CloudForest multi-omics and single data type models are shown on the left. A heatmap of the AUF1C values for multi-omics and single data type models is shown on the right. Above the heatmap, the upper annotation track indicates how the subtypes were originally defined. See also Figure S3 and Table S4.

Please cite this article in press as: Ellrott et al., Classification of non-TCGA cancer samples to TCGA molecular subtypes using compact feature sets, Cancer Cell (2024), https://doi.org/10.1016/j.ccell.2024.12.002

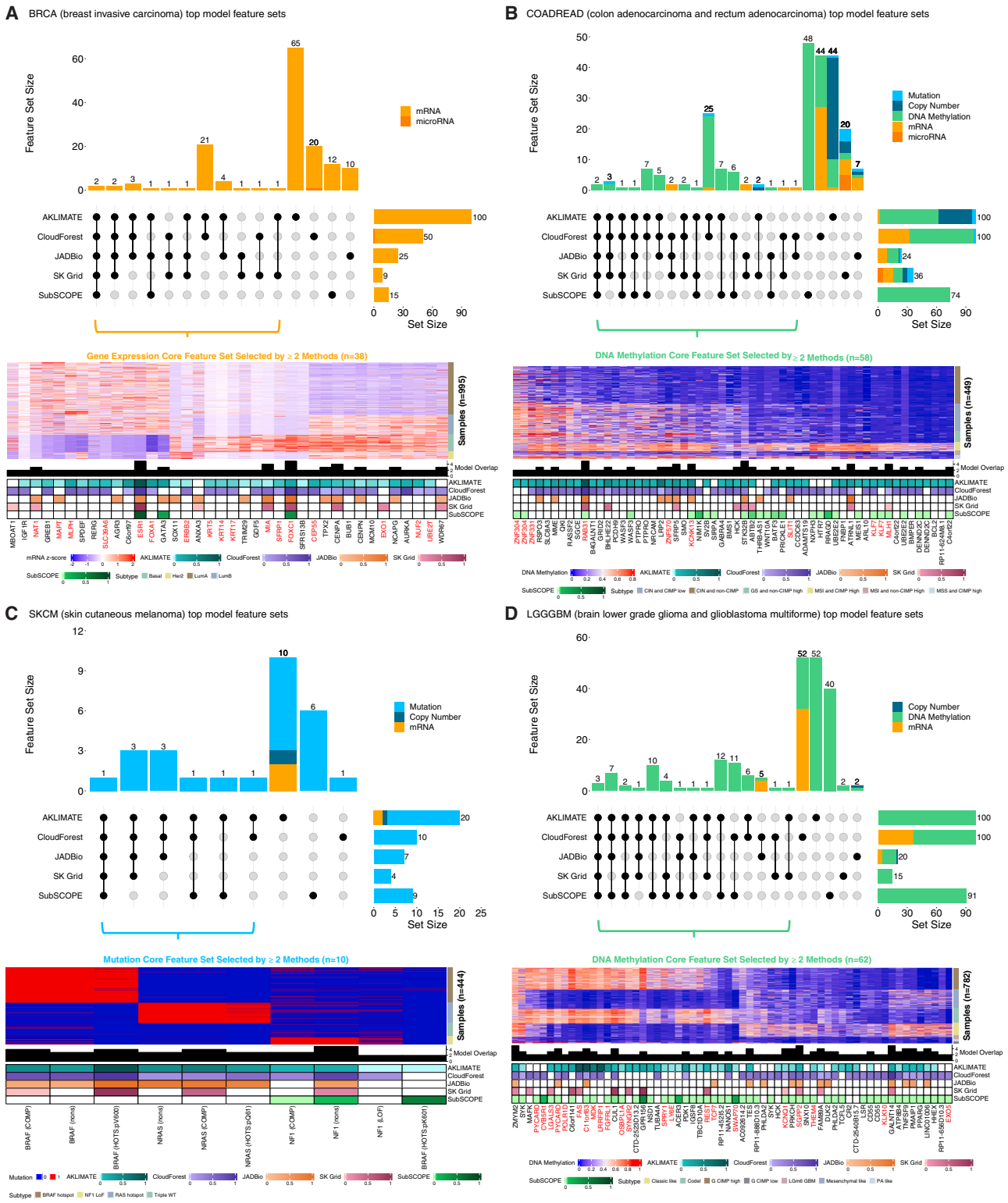


Figure 5. Feature sets of top performing models

(A–D) Analysis of the overlap in feature sets among top models for each of four cancer cohorts. (A) BRCA (expression clustering-based subtyping), (B) COADREAD (DNA methylation clustering-based subtyping), (C) SKCM (mutation clustering-based subtyping), and (D) LGGGBM (DNA methylation clustering-based subtyping). For each cancer cohort, we identified the top model for each method. Models were allowed to select up to 100 features, except JADBio, which limited its feature set to a maximum of 25. Overlap among the selected feature sets is represented in Upset plots.[34] A bar chart of model overlap indicates feature

(legend continued on next page)

histology, resulting in three *IDH1*/*IDH2*-mutant subtypes and four *IDH1*/*IDH2*-wildtype subtypes.[14,37] Interestingly, for LGGGBM, the classification methods overwhelmingly selected DNA methylation features (Figure 5D), and did not identify *IDH1*/*IDH2* mutations, perhaps because *IDH* mutation status spans multiple subtypes. Feature overlap among top models is shown for all cancer cohorts in Figure S4.

Classifier feature sets converge on common pathways

We investigated whether the core set of shared features selected by two or more models represent genes enriched for biological processes. We plotted the location of known cancer-associated genes from the COSMIC database on a two-dimensional projection of a comprehensive collection of pathways found in PathwayCommons[45] to serve as a reference landscape against which we could compare the features selected by classifiers (Figure 6A, left panel). Gene features selected by multiple classifiers for BRCA, LGGGBM, and COADREAD formed clusters within the projection, suggesting involvement in the same pathways (Figure 6A, right three panels). The features pooled from all TCGA cancer cohort subtype predictors also showed evidence of clustering (Figure S5A, center panel).

To both visualize and quantify the degree of pathway enrichment, we displayed genes and their pathway relationships as density clouds, such that functionally related genes form dense areas, which we refer to as "summits" (Figure 6B). We numbered summits from most to least dense and used gene set enrichment analysis to assign labels to the major landscape summits (Table S6). The COSMIC summits (Figure 6B, left panel, white outlines) revealed areas with known mutation associations to cancer biology. We mapped gene features selected in our classifier models onto the same two-dimensional pathway projection and identified summits representing classifier features for BRCA, LGGGBM, and COADREAD (Figure 6B, right three panels). Selected feature summits that coincided with COSMIC summits represent clusters of features chosen from pathways that have established roles in cancer biology (Table S7).

An analysis of the pooled set of features selected by subtype classifiers for all TCGA cancer cohorts revealed overlap with many known COSMIC cancer driver pathways (Figure S5B). For example, the densest aggregate TCGA-subtype classifier summit (T1) was enriched for genes from the TP53 pathway, and this summit strongly overlapped a COSMIC summit (C7) that was also enriched for TP53 (Figure S5B, right panel). On the other hand, feature summits that were far from any COSMIC summit represented cases in which genes were selected from pathways with less-established roles in cancer biology, but nevertheless contain genes that help distinguish TCGA cancer subtypes. For example, the T4 summit was enriched for biological oxidation, and T15 was enriched for organelle assembly, neither of which is represented in COSMIC (Figure S5B).

For BRCA, the second-densest summit (BR2, Figure 6B) included genes from the estrogen receptor (ER) signaling pathway, which overlaps with the COSMIC ER signaling summit (C2), representing a pathway that is well known for distinguishing luminal from basal breast cancers. In contrast, the major summit for BRCA subtyping (BR1) involved Rho/CDC42, which showed no overlap with any major COSMIC cluster, but nevertheless represents a set of cell proliferation features that distinguish aggressive basal-like from luminal tumors.

Cell-cycle-related summits were present in other tumor types that were either distinct from COSMIC (LG8 and LG12) or in close proximity to COSMIC summits but still non-overlapping (CR1). LGGGBM's major summits (LG1 and LG2) had clear enrichment for TP53-related signaling and overlapped the COSMIC TP53 summit (C7), illustrating the differential TP53 involvement in LGGGBM tumors, where tumors that were more GBM-like showed a lower frequency of TP53 alterations than other LGG subtypes. Features distinguishing COADREAD subtypes represented pathways involving Wnt signaling (CR4, Figure 6B), which partially overlapped with the COSMIC Wnt-signaling C4 and C19 summits. Wnt signaling is a key driver of most colorectal tumors, but the mechanisms by which this is achieved differ among COADREAD subtypes. For example, CIMP-high tumors display a reduced frequency of *APC* mutations,[36] but instead rely on epigenetic silencing of SFRP genes, which encode negative regulators of Wnt signaling.[43]

We hypothesized that if gene features selected by different models reflected related biology, they would be located closer to each other in network topology space than would be expected by chance. Using the same PathwayCommons compendium, we measured the pathway distance between the nearest neighbors of each method's gene feature lists to the gene feature lists of the other methods (Figure 6C). To control for overlapping genes that would have zero distances, we also compared the distances between second-nearest neighbors (Figure S5C). We found that the gene features selected by most methods for BRCA, LGGGBM, and COADREAD models were closer to one another than expected by chance (Figure 6C). Taken together, these findings suggest that equivalently predictive features for subtyping may result from the gene co-membership structure present in biological pathway space.

The subtype-specific expression signal for each gene can be represented by a scaled signal-to-noise ratio (SNR). Ranking the genes in each summit by decreasing SNR produces a "sail" shape (see Figure 6D). Commonly selected core features tended to cluster on the left side of the sails, indicating classifier features that had high SNRs, representing strong individual predictors. Some features with lower SNR were also selected, and these likely provide orthogonal classification information when used in conjunction with other features. We analyzed enrichment of pathway hallmark gene sets with selected feature lists for

membership in the top models of the five methods. The set of features selected in two or more models for each cancer cohort is designated as the "core" feature set for that cancer cohort. The heatmaps represent a hierarchical clustering analysis of the core feature measurements for the main selected data type for all cohort samples. Sample rows in heatmaps are organized by subtype. The method annotation panels indicate min-max normalized feature importance values, with 1 indicating the most important feature (the entire model feature set was normalized regardless of platform). Gene symbols (heatmap columns) are colored red to indicate membership in the corresponding annotation list; PAM50 membership (BRCA), DNA methylation literature support (COADREAD and LGGGBM cohorts).[35,36–38] See also Figure S4.

Please cite this article in press as: Ellrott et al., Classification of non-TCGA cancer samples to TCGA molecular subtypes using compact feature sets, Cancer Cell (2024), https://doi.org/10.1016/j.ccell.2024.12.002

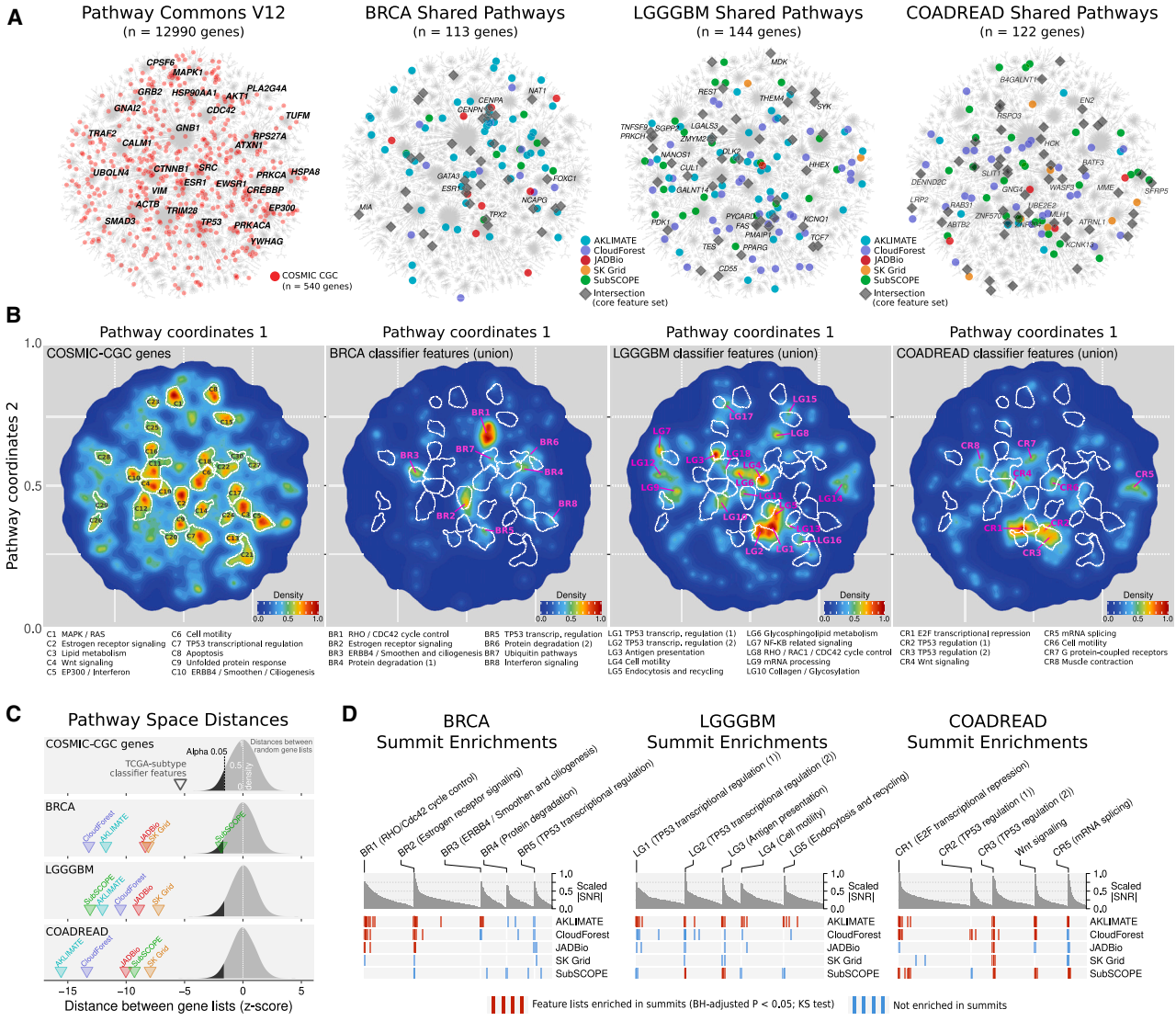


Figure 6. Pathways and biology of classifier features

(A) Pathway space representation of PathwayCommons V12 (gray background graph).[45] Left panel: pathway space location of cancer-associated genes from the COSMIC-CGC database (release v95, tier 1 collection) (red circles). Labels represent the top 30 "hub" genes in the graph. Right panels: pathway space locations of BRCA, LGGGBM, and COADREAD classifier feature lists (colored circles). Dark diamonds indicate intersections in ≥2 ML methods; text labels indicate intersections in ≥3 machine-learning methods.

(B) Density of selected genes in the pathway space depicted in (A). Summits represent dense collections of genes, and are numbered from the most to the least dense. Left panel: COSMIC-CGC summits. Right panels: single cohort summits. White outlines indicate the locations of COSMIC-CGC summits in the left panel.

(C) Distances between classifier feature lists in pathway space. The x axis shows the average shortest-path distance to nearest neighbors between gene lists. Top panel: Distance from TCGA-subtype classifier features to COSMIC-CGC genes. Lower panels: Distance from the classifier feature list of one method to the gene lists of the other methods, expressed as z-scores, using the distribution of random gene list distances.

(D) Enrichment analysis of BRCA, LGGGBM, and COADREAD summits. Genes in each summit were ranked by a signal-to-noise ratio (SNR) metric. Tracks below the summits show the distribution of feature lists from the top-performing models. Feature lists enriched in summits are indicated in red. See also Figure S5 and Tables S6 and S7.

BRCA, LGGGBM, and COADREAD subtype classifiers, again noting the clustering of selected core features toward the left side of the sails, indicating features with a high SNR (Figure S5D). Taken together, these results indicate that multiple independent ML subtype classifier methods tend to select features that have high SNR, and equivalently predictive features tend to reside close together in biological pathway space.

Determinants of classification performance identified by meta-analysis

We undertook a comprehensive meta-analysis to identify specific characteristics of the data and ML classifiers that lead to better or worse cancer subtype classification performance. We collected 55 meta-features (Table S8) that describe the data and subtyping tasks for the 26 tumor types. We distinguished

between meta-features representing sample or feature characteristics supplied as training inputs, versus those identified during classifier training.

We investigated the similarities among the meta-features and their relationships to subtyping performance by comparing the 55 meta-features to each other, calculating correlations across the 26 tumor types, and clustering the meta-features based on their correlation coefficients (Figures S6A and S6B). We identified seven clusters of meta-features that were mutually correlated to each other, three of which had significant influence on classifier performance (Figures 7A and S6; Table S8). We assigned descriptive labels to each of these *meta-feature groups* (MFGs) inferred from their meta-feature composition.

The largest group of meta-features was MFG1 (subtype cohesiveness), which includes features that reflect the separation of the subtype classes in the data used for classification, such as the silhouette scores for all platforms. MFG2 has distinct meta-features with high positive association with overall weighted F1 score, including a meta-feature representing the percent of the cohort falling into the rarest class, and another reflecting the percent of continuous features used in the input.

The strongest anticorrelation with overall weighted F1 scores was observed for MFG7 (cohort and trained model complexity), which includes meta-features such as the number of principal components that captured 70% of the gene expression variance, the entropy of the subtype class variable (used here as a measure for uniformity of the subtype sizes), the number of samples in the rarest subtype class, the number of features selected by a model, and the overall number of subtype classes.

Finally, four meta-feature groups lacked meta-features with strong correlations (or anti-correlations) to overall weighted F1 score; the features in these groups included the number of input features, the variance of the features in gene expression space, and the balance in the number of input features per data type.

How many samples are needed to train classifiers?

We used our large resource of trained models for 26 different cancer cohorts of various size and complexity to estimate the number of training samples necessary to accurately classify subtypes in a given tumor type. The extrapolated prediction performance as a function of training set size can be affected by the classifier model used, the dataset provided as features, and the classification label fidelity. This issue has been addressed in statistics and ML communities[46] and, more recently, in genomics settings.[47,48] We approached this problem by fitting a power-law function, the "learning curve,"[49] to extrapolate the behavior of a model to larger sample sizes. We sub-sampled the original input data and repeated the training of classifiers to determine the classification performance with fewer samples as inputs. We found that the same general trend held across all of the cancer cohorts (Figure 7B).

At the cohort level, providing approximately 150 samples for training appears to be adequate to approach maximum model performance. For most cancers, a larger cohort than this did not appreciably improve the subtype classifier accuracy. For example, if a low overall weighted F1 score of 0.70 is obtained with only 50 samples, it is highly likely that collecting two to three times more samples will fail to obtain an overall weighted F1 score exceeding 0.80, only marginally improving performance. Thus, in these situations, different data types (e.g., proteomics) or different subtype definitions should be considered, rather than simply adding more samples of the same data type.

Remarkably, the curves of performance versus sample size for all cohorts fit roughly the same shape (Figure 7B), suggesting that a non-linear regression fit to the curves could predict future subtyping performance. We considered 87 different possible fitting functions and found that a Burr Type XII distribution yielded the closest approximation of the curves in Figure 7B for the 15 cohorts with at least 250 samples (Figure 7C). For a prospectively accruing cancer cohort, we found that 70 samples were sufficient to extrapolate a reliable curve and derive an estimate of classification performance. To illustrate how performance can be extrapolated, we applied the function to estimate subtype classification performance on the adrenocortical carcinoma (ACC) cohort, for which only 76 samples were available, (Figure 7D). We estimate that doubling the sample size for ACC would result in an increase of performance from the current 0.88 range (similar to LGGGBM performance) up to 0.92 (similar to BRCA).

DISCUSSION

The discovery of molecular subtypes for all major cancer types has been one of the most influential outcomes of TCGA, contributing to the high impact of the landmark TCGA publications.[50–55] However, the complex methods and high-dimensional data used to identify these subtypes produced results specific to the TCGA datasets, and did not result in classifiers that could be implemented for other samples in clinical trials or research studies. To expand the utility of TCGA subtypes, we used ML approaches to train subtype classifiers that require only small sets of gene-centric features. ML algorithms have proven effective for classification problems using large-scale and heterogeneous datasets,[30–32] including the prediction of cancer types and clinical outcomes, as well as for the identification of the tumor tissue-of-origin using genomic, epigenomic or transcriptomic data.[5–9,56,57–63] We used five different ML approaches to produce 412,585 distinct subtype classifier models incorporating five different data types, and 100 stratified 5-fold train-test partitions of 8,791 TCGA samples, comprising 26 different cancer cohorts and 106 subtypes.

Our analysis delivered important resources, useful tools and insights into biology. First, we created an online resource of 737 publicly available, containerized predictive models with top-ranked performance, representing the top models for each of the 26 cancer cohorts, data types, and training algorithms. We provide an easy-to-run Docker container for each of these models, along with their selected features, mean overall weighted F1, and the lowest prediction score that provides 95% prediction accuracy for a new sample (https://github.com/NCICCGPO/gdan-tmp-models). These models provide a foundation for clinical assay development. Second, we trained classifiers based upon each of the five individual data types from TCGA. These single-platform predictors expand the application of TCGA subtyping to studies that employ different data types than used for the original TCGA subtype discovery. Third, we showed that external cancer datasets that use different molecular assay platforms or include FFPE samples can be readily

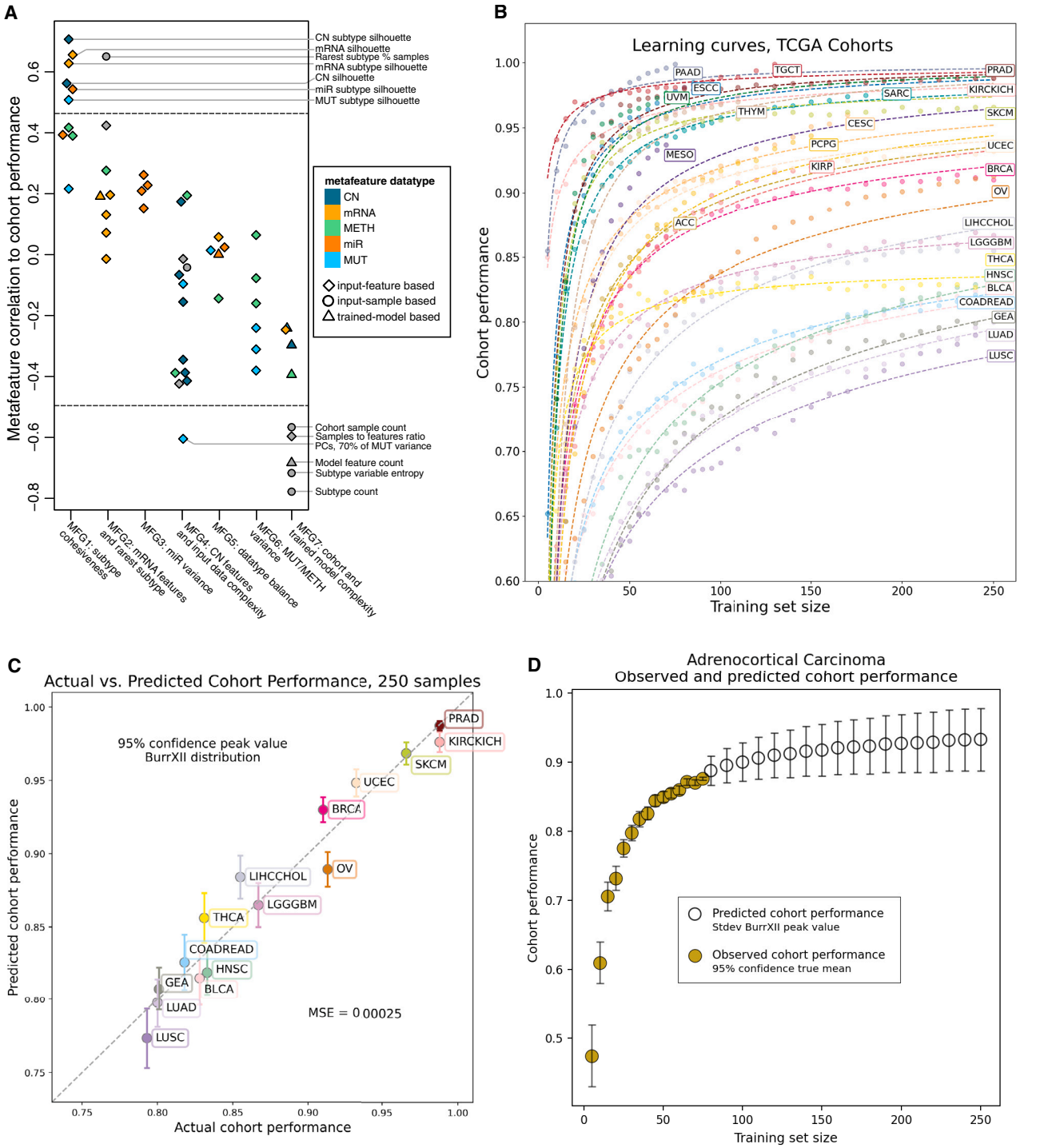


Figure 7. Factors influencing accurate subtype classification

(A) Correlation between meta-features and subtype classifier performance in 26 TCGA cohorts. Meta-features are grouped into seven meta-feature groups (MFG1-7) by hierarchical clustering. Two horizontal dashed lines mark significance thresholds (FDR-corrected Spearman correlation p values ≤ 0.05). PCs, principal components.

(B) Learning curves for 26 TCGA cohorts. Cohort performance as a function of sample size; each cohort was randomly sampled 100 times at each sample size-increment, and predictive accuracy was averaged across the sub-samplings.

(C) Predicted vs. actual cohort performance for the 15 cohorts with at least 250 tumor samples. The predicted performance at a sample size of 250 was estimated from the power-law curve fit to the sample-size range of 35 through 70 for each cohort.

(D) Representative extension of power-law accuracy prediction for a smaller TCGA cohort (adrenocortical carcinoma, 76 total samples). See also Figure S6 and Table S8.

Please cite this article in press as: Ellrott et al., Classification of non-TCGA cancer samples to TCGA molecular subtypes using compact feature sets, Cancer Cell (2024), https://doi.org/10.1016/j.ccell.2024.12.002

CellPress
OPEN ACCESS

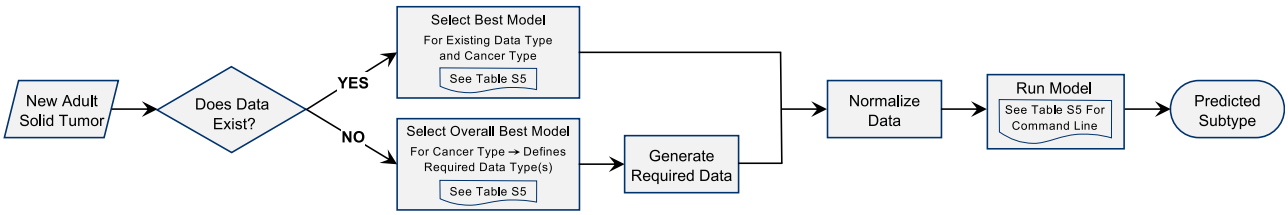


Figure 8. Guide to selection of the best model for a new sample

Decision chart to guide the selection of data and model to assign a TCGA subtype label to a non-TCGA patient sample. If genomic data exists for a new sample as indicated in the upper branch, then the best-performing model can be selected from Table S5 for the existing data type. In the lower branch the best overall model in Table S5 will dictate the type of data required for that model.

transformed to provide accurate subtype prediction using our classifier models. Fourth, we gained insight into the relative strengths of different data platforms across cancer types, both for original subtype discovery and prospective subtype prediction. Fifth, even when different models and data types didn't select identical features, they tended to select genes from similar or related pathways, underscoring the redundancy of informative features chosen by different models, and revealing biological distinctions between molecular subtypes. Sixth, we demonstrated that 70 samples are generally sufficient to estimate the ultimate classification accuracy for a prospectively accruing cancer cohort.

For most research applications, the classifier models can be applied directly to samples from other studies, even when only one data type is available, after appropriate data transformation to match the range and distribution in the TCGA cohort. We provide a simple flow diagram to guide the user to select the best appropriate model to classify a new sample (Figure 8). The derived feature sets and trained models also provide a valuable starting point for clinical assay development. These clinical assays may rely on platforms using different technology than TCGA, such as real-time PCR or bead-based hybridization assays, which will require further adjustment of features or model weights.

We anticipate that these tumor subtype classifiers will be useful in prospective clinical cancer research and practice, possibly helping to realize the as-yet unfulfilled promise of translating genomic findings to the clinic. By providing a set of classifiers that are amenable to practical implementations, our work unlocks the biology of TCGA cancer subtypes for research and clinical applications.

Cancer subtype-specific biological differences were frequently best defined by their mRNA features. Regardless of whether the training input included all data types or just mRNA features, the most-accurate models were often dominated by mRNA features. This may reflect, in part, the gene-centric nature of mRNA features. The copy number, DNA methylation, and miRNA features are more indirectly linked to individual gene function.

For mRNA-based predictors in ESCC, MESO, and ACC, JADBio achieved higher classification accuracy than the two methods that include random forest approaches (CloudForest and SK Grid). This suggests that non-redundant feature selection, not just ranking features by importance, can improve classification performance using more-parsimonious feature sets. For example, when a set of collinear, correlated features carry the same predictive component for a classifier, this may force a model to distribute importance among the features. Selection of features that offer some additional orthogonal information, on the other hand, might choose only one feature from a strongly correlated group and filter out the rest as redundant, resulting in more-accurate models with more-parsimonious feature sets.[45]

We expected the robustness of cluster-based subtypes to influence our ability to train models. We assessed this with the METABRIC breast cancer dataset. We found that the silhouette score was correlated with the difference in classification confidence between the best and second-best subtype call. Given this finding, one might hypothesize that training models with poorly clustered samples would increase classification error. We investigated the effect of using equally sized subsets of samples in training, enriched for well-clustered, subtype-prototypical TCGA samples, or enriched for poorly clustered samples. We found that enriching for either prototypical samples, or for poorly clustered samples did not improve classification performance. Models trained using the same number of samples representing the full range of silhouette scores appeared to yield the best-performing classifiers. An important conclusion from this analysis is that a full and diverse set of samples should be used in training to maximize sample size, even if the dataset contains samples that are ambiguously assigned to a subtype.

Limitations of the study

Even the most accurate classifier will be limited by the validity of the original subtype definitions. When we trained our classifiers, we assumed that all major subtypes for a particular cancer cohort were represented in the TCGA dataset. In practice, new and undocumented subtypes may be encountered in non-TCGA datasets. Thus, it may be beneficial to assign an "unknown" label to a new sample in a situation where no existing subtype has a strong-enough class prediction score.

Genome-wide features such as chromosomal instability and microsatellite instability are not captured well by the gene-centric feature sets that we used to train our classifiers. This may explain the poorer performance of all of our classifier methods with cancers like GEA and COADREAD that are characterized by such genome-wide disturbances.

We did not integrate the predictions of our various ML methods to create an "ensemble" method that might achieve higher classification accuracy. Combining the predictions of individual methods would increase the number of features needed to assign a subtype. To assess the tradeoff between parsimony and accuracy, we assessed a simple ensemble that took as input the class prediction scores of all of the methods and calculated a subtype assignment by averaging these scores for four cancer types. We

found that the ensemble predictions outperformed the best individual methods in two of these cohorts, albeit by narrow margins. Thus, if one could relax the requirement of using a small feature set for clinical application, improved subtype predictions may be possible for some cancer cohorts using model ensembles.

The recent explosion in deep-learning techniques popularized in large language models (LLMs) raises the possibility of improving biological classification tasks. For example, a recently published approach called Geneformer uses an attention-aware transformer pretrained on millions of single-cell transcriptomes to improve several molecular biology prediction tasks.[62] The common theme of transformer-based models is their capacity to continue to improve performance as more data are added to the training set. Modern LLMs have billions of parameters that can be used to model the intricate relationships within complex datasets, but usually require over a petabyte of data to train. Currently, our dataset of 8,791 samples does not provide enough power to benefit those more complex machine learning methods and we sought parsimonious feature sets, which are a poor fit for an LLM. Nevertheless, our results may inform approaches to fine-tuning of large models.

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Peter W. Laird (peter.laird@vai.org).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- Gene expression, microRNA, DNA methylation, somatic mutations, and copy number data have been deposited at the National Cancer Institute Genomic Data Commons (GDC) Publication Page and are publicly available as of the date of publication (see: https://gdc.cancer.gov/about-data/publications/CCG-TMP-2022). Files are listed in the key resources table and can be found on the paper's publication page at the GDC. All other data reported in this paper will be shared by the lead contact upon request.
- All original code and containerized pre-trained ML models have been deposited at GitHub and are publicly available as of the date of publication (see https://github.com/NCICCGPO/gdan-tmp-models). Also included are publicly available tutorials on applying these models on new datasets. Repositories are listed in the key resources table. All other original code reported in this paper will be shared by the lead contact upon request.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

CONSORTIA

Kyle Ellrott, Rehan Akbani, Victor H. Apolonio, Rameen Beroukhim, Bradley M. Broom, Christopher C Benz, Samantha J. Caesar-Johnson, Mauro A. A. Castro, Vinicius S. Chagas, Paulos Charonyktakis, Kami E. Chiotti, John A. Demchok, Esther Drill, Ina Felau, Martin L. Ferguson, Verena Friedl, Galen F Gao, Gad Getz, Jasleen K. Grewal, D. Neil Hayes, Toshinori Hinoue, Katherine A. Hoadley, Stephanie H. Hoyt, Steven J.M. Jones, Zhenlin Ju, Brian J Karlberg, Anab Kemal, Taek-Kyun Kim, Theo A. Knijnenburg, Vincenzo Lagani, Avantika Lal, Alexander J. Lazar, Jordan A. Lee, Xavier Loinaz, Eve Lowenstein, Akinyemi I. Ojesina, Daniele Ramazzotti, Lewis R. Roberts, A. Gordon Robertson, Whijae Roh, Andre Schultz, Hui Shen, Ronglai Shen, Ilya Shmulevich, Paul T. Spellman, Chip Stewart, Adam Struck, Joshua M. Stuart, Roy Tarnuzzer, Bahar Tercan, Ioannis Tsamardinos, Vladislav Uzunangelov, Chen Wang, Peggy I. Wang, Zhining Wang, Lindsay Westlake, Christopher K. Wong, Liming Yang, Christina Yau, Jean C. Zenklusen, Andrew D. Cherniack, Peter W. Laird.

ACKNOWLEDGMENTS

The authors would like to acknowledge the support of the National Cancer Institute. This work was funded through NIH/NCI grants U24CA264029 to A.D.C., U24CA264023 to P.W.L., U24CA264007 to K.E., U24CA264021 to K.A.H., and U24CA264009 to J.M.S. and C.C.B. We would also like to thank Dr. Ana Robles for helpful comments, edits, and suggestions.

AUTHOR CONTRIBUTIONS

Conceptualization, K.E., I.S., T.A.K., C.C.B., J.M.S., J.C.Z., A.D.C., and P.W.L.; data curation, K.E., C.K.W., M.A.A.C., J.A.L., B.J.K., V.F., T.H., V.U., L.W., and T.A.K.; formal analysis, K.E., C.K.W., C.Y., M.A.A.C., J.A.L., B.J.K., J.K.G., V.L., V.F., V.U., I.T., T.A.K., and J.M.S.; funding acquisition, K.E., I.S., K.A.H., C.C.B., J.M.S., J.C.Z., A.D.C., and P.W.L.; investigation, K.E., C.K.W., C.Y., M.A.A.C., J.A.L., B.J.K., J.K.G., V.L., B.T., V.F., V.U., and T.A.K.; methodology, K.E., C.K.W., M.A.A.C., J.A.L., J.K.G., V.L., B.T., V.U., I.T., T.A.K., J.M.S., A.D.C., and P.W.L.; project administration, K.E., I.F., P.I.W., A.K., S.J.C.-J., J.C.Z., I.S., T.A.K., C.C.B., J.M.S., A.D.C., and P.W.L.; software, K.E., C.K.W., C.Y., M.A.A.C., J.A.L., B.J.K., J.K.G., V.L., B.T., V.F., V.U., X.L., T.A.K., and J.M.S.; supervision, K.E., A.K., I.F., S.J.C.-J., J.C.Z., I.T., K.A.H., T.A.K., C.C.B., J.M.S., A.D.C., and P.W.L.; validation, K.E., C.K.W., M.A.A.C., J.A.L., B.J.K., J.K.G., B.T., V.U., I.S., T.A.K., and J.M.S.; visualization, K.E., C.K.W., C.Y., M.A.A.C., J.A.L., B.J.K., J.K.G., V.F., V.U., X.L., A.G.R., T.A.K., C.C.B., J.M.S., A.D.C., and P.W.L.; writing – original draft, K.E., C.K.W., C.Y., M.A.A.C., J.A.L., J.K.G., V.L., T.A.K., C.C.B., J.M.S., A.D.C., and P.W.L.; writing – review & editing, K.E., C.K.W., C.Y., M.A.A.C., J.A.L., B.J.K., J.K.G., V.L., B.T., V.F., P.I.W., A.J.L., I.T., K.A.H., A.G.R., T.A.K., C.C.B., J.M.S., J.C.Z., A.D.C., and P.W.L.

DECLARATION OF INTERESTS

A.D.C. receives research support from Bayer and consults for KaryoVerse. W.R. is currently working at Pfizer. I.T., P.C., and V.L. are or were directly or indirectly affiliated with JADBio—Gnosis DA, S.A., which offers the JADBio service commercially. V.F. is an employee and stock option owner of Bluestar Genomics Inc. L.R.R. has received grants from Bayer, Boston Scientific, Exact Sciences, FujiFilm Medical Sciences, Gilead Sciences, GlycoTest, RedHill, and Target PharmaSolutions and serves in a consulting/advising role for AstraZeneca, Bayer, Eisai, Exact Sciences, Gilead Sciences, Global Life Science Consulting, GRAIL, LLC, Hepion, MedEd Design, Medscape, Novartis Venture Fund, QED, RedHill, and The Lynx Group. A.J.L. has consulting relationships with AbbVie, Astra-Zeneca, Bayer, Bio-AI Health, BMS, Caris, Decipher, Foghorn Therapeutics, GRAIL, GSK, Illumina, Invitae/Archer DX, Iterion Therapeutics, Merck, Novartis, Nucleai, Paige, Pfizer, Regeneron, Roche/Genentech, SpringWorks, Tempus, and ThermoFisher. W.R. and G.G. are co-inventors of a patent application related to lung adenocarcinoma expression subtypes (U.S. Provisional Patent Application No.: 63/293,349). P.W.L. serves on the Scientific Advisory Boards of FOXO Technologies, Inc., and Tagomics, LLC. J.M.S. is a stock owner of Nantomics Inc. V.U. is an employee and stock owner of Bristol Myers Squibb.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.ccell.2024.12.002.

Received: August 16, 2024
Revised: August 26, 2024
Accepted: December 5, 2024
Published: January 2, 2025

REFERENCES

1. Kumar, V., Abbas, A.K., and Aster, J.C. (2021). Robbins & Cotran Pathologic Basis of Disease, 10th Edition (Elsevier), pp. 1–1392.
2. Kleihues, P., and Sobin, L.H. (2000). World Health Organization classification of tumors. Cancer *88*, 2887. https://doi.org/10.1002/1097-0142(20000615)88:12<2887::aid-cncr32>3.0.co;2-f.
3. Sobin, L.H., Gospodarowicz, M.K., and Wittekind, C. (2011). TNM Classification of Malignant Tumours, 7th Edition (John Wiley & Sons), pp. 1–336.
4. Hoadley, K.A., Yau, C., Hinoue, T., Wolf, D.M., Lazar, A.J., Drill, E., Shen, R., Taylor, A.M., Cherniack, A.D., Thorsson, V., et al. (2018). Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. Cell *173*, 291–304.e6. https://doi.org/10.1016/j.cell.2018.03.022.
5. Penson, A., Camacho, N., Zheng, Y., Varghese, A.M., Al-Ahmadie, H., Razavi, P., Chandarlapaty, S., Vallejo, C.E., Vakiani, E., Gilewski, T., et al. (2020). Development of Genome-Derived Tumor Type Prediction to Inform Clinical Cancer Care. JAMA Oncol. *6*, 84–91. https://doi.org/10.1001/jamaoncol.2019.3985.
6. Jiao, W., Atwal, G., Polak, P., Karlic, R., Cuppen, E., PCAWG Tumor Subtypes and Clinical Translation Working Group, Danyi, A., de Ridder, J., van Herpen, C., Lolkema, M.P., et al. (2020). A deep learning system accurately classifies primary and metastatic cancers using passenger mutation patterns. Nat. Commun. *11*, 728. https://doi.org/10.1038/s41467-019-13825-8.
7. Salvadores, M., Mas-Ponte, D., and Supek, F. (2019). Passenger mutations accurately classify human tumors. PLoS Comput. Biol. *15*, e1006953. https://doi.org/10.1371/journal.pcbi.1006953.
8. Grewal, J.K., Tessier-Cloutier, B., Jones, M., Gakkhar, S., Ma, Y., Moore, R., Mungall, A.J., Zhao, Y., Taylor, M.D., Gelmon, K., et al. (2019). Application of a Neural Network Whole Transcriptome-Based Pan-Cancer Method for Diagnosis of Primary and Metastatic Cancers. JAMA Netw. Open *2*, e192597. https://doi.org/10.1001/jamanetworkopen.2019.2597.
9. Lu, M.Y., Chen, T.Y., Williamson, D.F.K., Zhao, M., Shady, M., Lipkova, J., and Mahmood, F. (2021). AI-based pathology predicts origins for cancers of unknown primary. Nature *594*, 106–110. https://doi.org/10.1038/s41586-021-03512-4.
10. Zehir, A., Benayed, R., Shah, R.H., Syed, A., Middha, S., Kim, H.R., Srinivasan, P., Gao, J., Chakravarty, D., Devlin, S.M., et al. (2017). Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. Nat. Med. *23*, 703–713. https://doi.org/10.1038/nm.4333.
11. Guinney, J., Dienstmann, R., Wang, X., de Reyniès, A., Schlicker, A., Soneson, C., Marisa, L., Roepman, P., Nyamundanda, G., Angelino, P., et al. (2015). The consensus molecular subtypes of colorectal cancer. Nat. Med. *21*, 1350–1356. https://doi.org/10.1038/nm.3967.
12. Bailey, P., Chang, D.K., Nones, K., Johns, A.L., Patch, A.M., Gingras, M.C., Miller, D.K., Christ, A.N., Bruxner, T.J.C., Quinn, M.C., et al. (2016). Genomic analyses identify molecular subtypes of pancreatic cancer. Nature *531*, 47–52. https://doi.org/10.1038/nature16965.
13. Rouzier, R., Perou, C.M., Symmans, W.F., Ibrahim, N., Cristofanilli, M., Anderson, K., Hess, K.R., Stec, J., Ayers, M., Wagner, P., et al. (2005). Breast cancer molecular subtypes respond differently to preoperative chemotherapy. Clin. Cancer Res. *11*, 5678–5685. https://doi.org/10.1158/1078-0432.CCR-04-2421.
14. Ceccarelli, M., Barthel, F.P., Malta, T.M., Sabedot, T.S., Salama, S.R., Murray, B.A., Morozova, O., Newton, Y., Radenbaugh, A., Pagnotta, S.M., et al. (2016). Molecular Profiling Reveals Biologically Discrete Subsets and Pathways of Progression in Diffuse Glioma. Cell *164*, 550–563. https://doi.org/10.1016/j.cell.2015.12.028.
15. Bagaev, A., Kotlov, N., Nomie, K., Svekolkin, V., Gafurov, A., Isaeva, O., Osokin, N., Kozlov, I., Frenkel, F., Gancharova, O., et al. (2021). Conserved pan-cancer microenvironment subtypes predict response to immunotherapy. Cancer Cell *39*, 845–865.e7. https://doi.org/10.1016/j.ccell.2021.04.014.
16. Harris, T.J.R., and McCormick, F. (2010). The molecular pathology of cancer. Nat. Rev. Clin. Oncol. *7*, 251–265. https://doi.org/10.1038/nrclinonc.2010.41.
17. Uzunangelov, V., Wong, C.K., and Stuart, J.M. (2021). Accurate cancer phenotype prediction with AKLIMATE, a stacked kernel learner integrating multimodal genomic data and pathway knowledge. PLoS Comput. Biol. *17*, e1008878. https://doi.org/10.1371/journal.pcbi.1008878.
18. Bressler, R., Kreisberg, R.B., Bernard, B., Niederhuber, J.E., Vockley, J.G., Shmulevich, I., and Knijnenburg, T.A. (2015). CloudForest: A Scalable and Efficient Random Forest Implementation for Biological Data. PLoS One *10*, e0144820. https://doi.org/10.1371/journal.pone.0144820.
19. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., and Thirion, B. (2011). Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. *12*, 2825–2830.
20. Tsamardinos, I., Charonyktakis, P., Papoutsoglou, G., Borboudakis, G., Lakiotaki, K., Zenklusen, J.C., Juhl, H., Chatzaki, E., and Lagani, V. (2022). Just Add Data: Automated Predictive Modeling for Knowledge Discovery and Feature Selection. npj Precis. Oncol. *6*, 38. https://doi.org/10.1038/s41698-022-00274-8.
21. Kitchenham, B.A., Pickard, L.M., MacDonell, S.G., and Shepperd, M.J. (2001). What accuracy statistics really measure. IEE Proc. Softw. *148*, 81–85. https://doi.org/10.1049/ip-sen:20010506.
22. Dice, L.R. (1945). Measures of the amount of ecologic association between species. Ecology *26*, 297–302. https://doi.org/10.2307/1932409.
23. Parker, J.S., Mullins, M., Cheang, M.C.U., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., Hu, Z., et al. (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. J. Clin. Oncol. *27*, 1160–1167. https://doi.org/10.1200/JCO.2008.18.1370.
24. Curtis, C., Shah, S.P., Chin, S.F., Turashvili, G., Rueda, O.M., Dunning, M.J., Speed, D., Lynch, A.G., Samarajiwa, S., Yuan, Y., et al. (2012). The

genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature *486*, 346–352. https://doi.org/10.1038/nature10983.

25. Garcia-Recio, S., Hinoue, T., Wheeler, G.L., Kelly, B.J., Garrido-Castro, A.C., Pascual, T., De Cubas, A.A., Xia, Y., Felsheim, B.M., McClure, M.B., et al. (2023). Multiomics in primary and metastatic breast tumors from the AURORA US network finds microenvironment and epigenetic drivers of metastasis. Nat. Cancer *4*, 128–147. https://doi.org/10.1038/s43018-022-00491-x.
26. Rousseeuw, P.J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. *20*, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7.
27. Marzouka, N.A.D., and Eriksson, P. (2021). multiclassPairs: an R package to train multiclass pair-based classifier. Bioinformatics *37*, 3043–3044. https://doi.org/10.1093/bioinformatics/btab088.
28. Tan, A.C., Naiman, D.Q., Xu, L., Winslow, R.L., and Geman, D. (2005). Simple decision rules for classifying human cancers from gene expression profiles. Bioinformatics *21*, 3896–3904. https://doi.org/10.1093/bioinformatics/bti631.
29. Esteve-Codina, A., Arpi, O., Martinez-García, M., Pineda, E., Mallo, M., Gut, M., Carrato, C., Rovira, A., Lopez, R., Tortosa, A., et al. (2017). A Comparison of RNA-Seq Results from Paired Formalin-Fixed Paraffin-Embedded and Fresh-Frozen Glioblastoma Tissue Samples. PLoS One *12*, e0170632. https://doi.org/10.1371/journal.pone.0170632.
30. Kotsiantis, S.B. (2007). Supervised Machine Learning: A Review of Classification Techniques. In Emerging Artificial Intelligence Applications in Computer Engineering, I. Maglogiannis, K. Karpouzis, B.A. Wallace, and J. Soldatos, eds. (IOS Press), pp. 3–24.
31. Kourou, K., Exarchos, T.P., Exarchos, K.P., Karamouzis, M.V., and Fotiadis, D.I. (2015). Machine learning applications in cancer prognosis and prediction. Comput. Struct. Biotechnol. J. *13*, 8–17. https://doi.org/10.1016/j.csbj.2014.11.005.
32. Bzdok, D., Krzywinski, M., and Altman, N. (2017). Points of Significance: Machine learning: a primer. Nat. Methods *14*, 1119–1120. https://doi.org/10.1038/nmeth.4526.
33. Lagani, V., Athineou, G., Farcomeni, A., Tsagris, M., and Tsamardinos, I. (2017). Feature Selection with the R Package MXM: Discovering Statistically Equivalent Feature Subsets. J. Stat. Softw. *80*, 1–25.
34. Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R., and Pfister, H. (2014). UpSet: Visualization of Intersecting Sets. IEEE Trans Vis Comput Graph *20*, 1983–1992. https://doi.org/10.1109/TVCG.2014.2346248.
35. Hinoue, T., Weisenberger, D.J., Lange, C.P.E., Shen, H., Byun, H.M., Van Den Berg, D., Malik, S., Pan, F., Noushmehr, H., van Dijk, C.M., et al. (2012). Genome-scale analysis of aberrant DNA methylation in colorectal cancer. Genome Res. *22*, 271–282. https://doi.org/10.1101/gr.117523.110.
36. Liu, Y., Sethi, N.S., Hinoue, T., Schneider, B.G., Cherniack, A.D., Sanchez-Vega, F., Seoane, J.A., Farshidfar, F., Bowlby, R., Islam, M., et al. (2018). Comparative Molecular Analysis of Gastrointestinal Adenocarcinomas. Cancer Cell *33*, 721–735.e8. https://doi.org/10.1016/j.ccell.2018.03.010.
37. Noushmehr, H., Weisenberger, D.J., Diefes, K., Phillips, H.S., Pujara, K., Berman, B.P., Pan, F., Pelloski, C.E., Sulman, E.P., Bhat, K.P., et al. (2010). Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma. Cancer Cell *17*, 510–522. https://doi.org/10.1016/j.ccr.2010.03.017.
38. Knijnenburg, T.A., Wang, L., Zimmermann, M.T., Chambwe, N., Gao, G.F., Cherniack, A.D., Fan, H., Shen, H., Way, G.P., Greene, C.S., et al. (2018). Genomic and Molecular Landscape of DNA Damage Repair Deficiency across The Cancer Genome Atlas. Cell Rep. *23*, 239–254.e6. https://doi.org/10.1016/j.celrep.2018.03.076.
39. Porras, L., Ismail, H., and Mader, S. (2021). Positive Regulation of Estrogen Receptor Alpha in Breast Tumorigenesis. Cells *10*, 2966. https://doi.org/10.3390/cells10112966.
40. Elian, F.A., Yan, E., and Walter, M.A. (2018). FOXC1, the new player in the cancer sandbox. Oncotarget *9*, 8165–8178. https://doi.org/10.18632/oncotarget.22742.
41. Ray, T., Ryusaki, T., and Ray, P.S. (2021). Therapeutically Targeting Cancers That Overexpress FOXC1: A Transcriptional Driver of Cell Plasticity, Partial EMT, and Cancer Metastasis. Front. Oncol. *11*, 721959. https://doi.org/10.3389/fonc.2021.721959.
42. Wu, Y., Li, Z., Wedn, A.M., Casey, A.N., Brown, D., Rao, S.V., Omarjee, S., Hooda, J., Carroll, J.S., Gertz, J., et al. (2023). FOXA1 Reprogramming Dictates Retinoid X Receptor Response in ESR1-Mutant Breast Cancer. Mol. Cancer Res. *21*, 591–604. https://doi.org/10.1158/1541-7786.MCR-22-0516.
43. Suzuki, H., Watkins, D.N., Jair, K.W., Schuebel, K.E., Markowitz, S.D., Chen, W.D., Pretlow, T.P., Yang, B., Akiyama, Y., Van Engeland, M., et al. (2004). Epigenetic inactivation of SFRP genes allows constitutive WNT signaling in colorectal cancer. Nat. Genet. *36*, 417–422.
44. Cancer Genome Atlas Network (2015). Genomic Classification of Cutaneous Melanoma. Cell *161*, 1681–1696. https://doi.org/10.1016/j.cell.2015.05.044.
45. Rodchenkov, I., Babur, O., Luna, A., Aksoy, B.A., Wong, J.V., Fong, D., Franz, M., Siper, M.C., Cheung, M., Wrana, M., et al. (2020). Pathway Commons 2019 Update: integration, analysis and exploration of pathway data. Nucleic Acids Res. *48*, D489–D497. https://doi.org/10.1093/nar/gkz946.
46. Anderson, J. (1983). The Architecture of Cognition (Harvard University Press), p. 314.
47. Figueroa, R.L., Zeng-Treitler, Q., Kandula, S., and Ngo, L.H. (2012). Predicting sample size required for classification performance. BMC Med. Inform. Decis. Mak. *12*, 8. https://doi.org/10.1186/1472-6947-12-8.
48. Mukherjee, S., Tamayo, P., Rogers, S., Rifkin, R., Engle, A., Campbell, C., Golub, T.R., and Mesirov, J.P. (2003). Estimating dataset size requirements for classifying DNA microarray data. J. Comput. Biol. *10*, 119–142. https://doi.org/10.1089/106652703321825928.
49. Cortes, C., Jackel, L.D., Solla, S.A., Vapnik, V., and Denker, J.S. (1993). Learning Curves: Asymptotic Values and Rate of Convergence. In NIPS'93: Proceedings of the 6th International Conference on Neural Information Processing Systems, J. Cowan, G. Tesauro, and J. Alspector, eds. (Morgan Kaufmann Publishers Inc), pp. 327–334.
50. Cancer Genome Atlas Research Network (2014). Comprehensive molecular characterization of gastric adenocarcinoma. Nature *513*, 202–209. https://doi.org/10.1038/nature13480.
51. Cancer Genome Atlas Network (2012). Comprehensive molecular portraits of human breast tumours. Nature *490*, 61–70. https://doi.org/10.1038/nature11412.
52. Cancer Genome Atlas Research Network (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature *455*, 1061–1068. https://doi.org/10.1038/nature07385.
53. Cancer Genome Atlas Research Network (2011). Integrated genomic analyses of ovarian carcinoma. Nature *474*, 609–615. https://doi.org/10.1038/nature10166.
54. Cancer Genome Atlas Network (2012). Comprehensive molecular characterization of human colon and rectal cancer. Nature *487*, 330–337. https://doi.org/10.1038/nature11252.
55. Cancer Genome Atlas Research Network (2012). Comprehensive genomic characterization of squamous cell lung cancers. Nature *489*, 519–525. https://doi.org/10.1038/nature11404.
56. Darmofal, M., Suman, S., Atwal, G., Toomey, M., Chen, J.F., Chang, J.C., Vakiani, E., Varghese, A.M., Balakrishnan Rema, A., Syed, A., et al. (2024). Deep-Learning Model for Tumor-Type Prediction Using Targeted Clinical Genomic Sequencing Data. Cancer Discov. *14*, 1064–1081. https://doi.org/10.1158/2159-8290.CD-23-0996.
57. Amrane, M., Oukid, S., Gagaoua, I., and Ensari, T. (2018). Breast Cancer Classification Using Machine Learning. Proceedings of 2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting

(EBBT), Istanbul, 18-19 April 2018, 1–4. https://doi.org/10.1109/EBBT.2018.8391453.

58. Tan, A.C., and Gilbert, D. (2003). Ensemble machine learning on gene expression data for cancer classification. Appl. Bioinformatics *2*, S75–S83.

59. Alharbi, F., and Vakanski, A. (2023). Machine Learning Methods for Cancer Classification Using Gene Expression Data: A Review. Bioengineering (Basel) *10*, 173. https://doi.org/10.3390/bioengineering10020173.

60. Gao, F., Wang, W., Tan, M., Zhu, L., Zhang, Y., Fessler, E., Vermeulen, L., and Wang, X. (2019). DeepCC: a novel deep learning-based framework for cancer molecular subtype classification. Oncogenesis *8*, 44. https://doi.org/10.1038/s41389-019-0157-8.

61. Chen, R., Yang, L., Goodison, S., and Sun, Y. (2020). Deep-learning approach to identifying cancer subtypes using high-dimensional genomic data. Bioinformatics *36*, 1476–1483. https://doi.org/10.1093/bioinformatics/btz769.

62. Theodoris, C.V., Xiao, L., Chopra, A., Chaffin, M.D., Al Sayed, Z.R., Hill, M.C., Mantineo, H., Brydon, E.M., Zeng, Z., Liu, X.S., and Ellinor, P.T. (2023). Transfer learning enables predictions in network biology. Nature *618*, 616–624. https://doi.org/10.1038/s41586-023-06139-9.

63. Liu, E.T., and Mockus, S.M. (2020). Tumor Origins Through Genomic Profiles. JAMA Oncol. *6*, 33–34. https://doi.org/10.1001/jamaoncol.2019.3981.

64. Shen, R., Olshen, A.B., and Ladanyi, M. (2009). Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. Bioinformatics *25*, 2906–2912. https://doi.org/10.1093/bioinformatics/btp543.

65. Cancer Genome Atlas Research Network, Brat, D.J., Verhaak, R.G.W., Aldape, K.D., Yung, W.K.A., Salama, S.R., Cooper, L.A.D., Rheinbay, E., Miller, C.R., Vitucci, M., et al. (2015). Comprehensive, Integrative Genomic Analysis of Diffuse Lower-Grade Gliomas. N. Engl. J. Med. *372*, 2481–2498. https://doi.org/10.1056/NEJMoa1402121.

66. Farshidfar, F., Zheng, S., Gingras, M.C., Newton, Y., Shih, J., Robertson, A.G., Hinoue, T., Hoadley, K.A., Gibb, E.A., Roszik, J., et al. (2017). Integrative Genomic Analysis of Cholangiocarcinoma Identifies Distinct IDH-Mutant Molecular Profiles. Cell Rep. *18*, 2780–2794. https://doi.org/10.1016/j.celrep.2017.02.033.

67. The Cancer Genome Atlas Research Network (2017). Integrated genomic characterization of oesophageal carcinoma. Nature *541*, 169–175. https://doi.org/10.1038/nature20805.

68. Shen, H., Shih, J., Hollern, D.P., Wang, L., Bowlby, R., Tickoo, S.K., Thorsson, V., Mungall, A.J., Newton, Y., Hegde, A.M., et al. (2018). Integrated Molecular Characterization of Testicular Germ Cell Tumors. Cell Rep. *23*, 3392–3406. https://doi.org/10.1016/j.celrep.2018.05.039.

69. Goldhirsch, A., Winer, E.P., Coates, A.S., Gelber, R.D., Piccart-Gebhart, M., Thürlimann, B., and Senn, H.J.; Panel members (2013). Personalizing the treatment of women with early breast cancer: highlights of the St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2013. Ann. Oncol. *24*, 2206–2223. https://doi.org/10.1093/annonc/mdt303.

70. Nielsen, T., Wallden, B., Schaper, C., Ferree, S., Liu, S., Gao, D., Barry, G., Dowidar, N., Maysuria, M., and Storhoff, J. (2014). Analytical validation of the PAM50-based Prosigna Breast Cancer Prognostic Gene Signature Assay and nCounter Analysis System using formalin-fixed paraffin-embedded breast tumor specimens. BMC Cancer *14*, 177. https://doi.org/10.1186/1471-2407-14-177.

71. Weigelt, B., Mackay, A., A'Hern, R., Natrajan, R., Tan, D.S.P., Dowsett, M., Ashworth, A., and Reis-Filho, J.S. (2010). Breast cancer molecular profiling with single sample predictors: a retrospective analysis. Lancet Oncol. *11*, 339–349. https://doi.org/10.1016/S1470-2045(10)70008-5.

72. Roh, W., Geffen, Y., Cha, H., Miller, M., Anand, S., Kim, J., Heiman, D.I., Gainor, J.F., Laird, P.W., Cherniack, A.D., et al. (2022). High-Resolution Profiling of Lung Adenocarcinoma Identifies Expression Subtypes with Specific Biomarkers and Clinically Relevant Vulnerabilities. Cancer Res. *82*, 3917–3931. https://doi.org/10.1158/0008-5472.CAN-22-0432.

73. Kamoun, A., de Reyniès, A., Allory, Y., Sjödahl, G., Robertson, A.G., Seiler, R., Hoadley, K.A., Groeneveld, C.S., Al-Ahmadie, H., Choi, W., et al. (2020). A Consensus Molecular Classification of Muscle-invasive Bladder Cancer. Eur. Urol. *77*, 420–433. https://doi.org/10.1016/j.eururo.2019.09.006.

74. Chu, A., Robertson, G., Brooks, D., Mungall, A.J., Birol, I., Coope, R., Ma, Y., Jones, S., and Marra, M.A. (2016). Large-scale profiling of microRNAs for The Cancer Genome Atlas. Nucleic Acids Res. *44*, e3. https://doi.org/10.1093/nar/gkv808.

75. Chiu, H.S., Martínez, M.R., Bansal, M., Subramanian, A., Golub, T.R., Yang, X., Sumazin, P., and Califano, A. (2017). High-throughput validation of ceRNA regulatory networks. BMC Genom. *18*, 418. https://doi.org/10.1186/s12864-017-3790-7.

76. Mullokandov, G., Baccarini, A., Ruzo, A., Jayaprakash, A.D., Tung, N., Israelow, B., Evans, M.J., Sachidanandam, R., and Brown, B.D. (2012). High-throughput assessment of microRNA activity and function using microRNA sensor and decoy libraries. Nat. Methods *9*, 840–846. https://doi.org/10.1038/nmeth.2078.

77. Zhou, W., Triche, T.J., Jr., Laird, P.W., and Shen, H. (2018). SeSAMe: reducing artifactual detection of DNA methylation by Infinium BeadChips in genomic deletions. Nucleic Acids Res. *46*, e123. https://doi.org/10.1093/nar/gky691.

78. Taylor, A.M., Shih, J., Ha, G., Gao, G.F., Zhang, X., Berger, A.C., Schumacher, S.E., Wang, C., Hu, H., Liu, J., et al. (2018). Genomic and Functional Approaches to Understanding Cancer Aneuploidy. Cancer Cell *33*, 676–689.e3. https://doi.org/10.1016/j.ccell.2018.03.007.

79. Ellrott, K., Bailey, M.H., Saksena, G., Covington, K.R., Kandoth, C., Stewart, C., Hess, J., Ma, S., Chiotti, K.E., McLellan, M., et al. (2018). Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines. Cell Syst. *6*, 271–281.e7. https://doi.org/10.1016/j.cels.2018.03.002.

80. Bailey, M.H., Tokheim, C., Porta-Pardo, E., Sengupta, S., Bertrand, D., Weerasinghe, A., Colaprico, A., Wendl, M.C., Kim, J., Reardon, B., et al. (2018). Comprehensive Characterization of Cancer Driver Genes and Mutations. Cell *173*, 371–385.e18. https://doi.org/10.1016/j.cell.2018.02.060.

81. Sanchez-Vega, F., Mina, M., Armenia, J., Chatila, W.K., Luna, A., La, K.C., Dimitriadoy, S., Liu, D.L., Kantheti, H.S., Saghafinia, S., et al. (2018). Oncogenic Signaling Pathways in The Cancer Genome Atlas. Cell *173*, 321–337.e10. https://doi.org/10.1016/j.cell.2018.03.035.

82. Chang, M.T., Asthana, S., Gao, S.P., Lee, B.H., Chapman, J.S., Kandoth, C., Gao, J., Socci, N.D., Solit, D.B., Olshen, A.B., et al. (2016). Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity. Nat. Biotechnol. *34*, 155–163. https://doi.org/10.1038/nbt.3391.

83. Roh, Y., Heo, G., and Whang, S.E. (2019). A Survey on Data Collection for Machine Learning: a Big Data – AI Integration Perspective. Preprint at arXiv. https://doi.org/10.48550/arXiv.1811.03402.

84. Shrikumar, A., Greenside, P., and Kundaje, A. (2019). Learning Important Features Through Propagating Activation Differences. Preprint at arXiv. https://doi.org/10.48550/arXiv.1704.02685.

85. Borboudakis, G., and Tsamardinos, I. (2019). Forward-Backward Selection with Early Dropping. J. Mach. Learn. Res. *20*, 1–39.

86. Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. J. Roy. Stat. Soc. B *58*, 267–288.

87. Tsamardinos, I., Greasidou, E., and Borboudakis, G. (2018). Bootstrapping the out-of-sample predictions for efficient and accurate cross-validation. Mach. Learn. *107*, 1895–1922. https://doi.org/10.1007/s10994-018-5714-4.

88. Dunning, M., Lynch, A., and Eldridge, M. (2015). illuminaHumanv3.db: Illumina HumanHT12v3 annotation data (chip illuminaHumanv3) R package version 1.26.0. https://bioconductor.org/packages/release/data/annotation/html/illuminaHumanv3.db.html.

89. Sondka, Z., Bamford, S., Cole, C.G., Ward, S.A., Dunham, I., and Forbes, S.A. (2018). The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. Nat. Rev. Cancer *18*, 696–705. https://doi.org/10.1038/s41568-018-0060-1.

90. Pau, G., Fuchs, F., Sklyar, O., Boutros, M., and Huber, W. (2010). EBImage–an R package for image processing with applications to cellular phenotypes. Bioinformatics *26*, 979–981. https://doi.org/10.1093/bioinformatics/btq046.

91. Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J.P., and Tamayo, P. (2015). The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst. *1*, 417–425. https://doi.org/10.1016/j.cels.2015.12.004.

92. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., and Mesirov, J.P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. USA *102*, 15545–15550. https://doi.org/10.1073/pnas.0506580102.

93. Wilson, D.J. (2019). The harmonic mean p-value for combining dependent tests. Proc. Natl. Acad. Sci. USA *116*, 1195–1200. https://doi.org/10.1073/pnas.1814092116.

94. Ripley, B.,Venables, W. Class*: Functions for Classification*. R package, version 7.3-19. Doi:10.32614/CRAN.package.class.

95. Korotkevich, G., Sukhov, V., Budin, N., Shpak, B., Artyomov, M.N., and Sergushichev, A. (2021). Fast gene set enrichment analysis. https://www.biorxiv.org/content/10.1101/060012v3.

Please cite this article in press as: Ellrott et al., Classification of non-TCGA cancer samples to TCGA molecular subtypes using compact feature sets, Cancer Cell (2024), https://doi.org/10.1016/j.ccell.2024.12.002

CellPress
OPEN ACCESS

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Sample data and CV folds	This paper	TMP_20230209.tar.gz
Model prediction performance metrics	This paper	big_results_matrix.tsv.gz
Feature lists as model-by-feature matrix	This paper	collected_features_matrix.tsv.gz
Feature importance scores	This paper	Feature importance scores
Feature lists	This paper	feature_lists.tsv.gz
Model-by-feature matrix for top models per method/cohort combination	This paper	collected_features_matrix_top_models_lte_100.tsv.gz
Model-by-gene matrix for top model per cohort (excluding genes from CNVR features)	This paper	collected_genes_matrix_top_models_lte_100_exclude_CNVR_cohort_level.tsv.gz
Model-by-gene matrix for top models per method/cohort combination (excluding genes from CNVR features)	This paper	collected_genes_matrix_top_models_lte_100_exclude_CNVR.tsv.gz
Model prediction performance metrics for top models per method/cohort combination	This paper	top_performing_models_lte_100_features.tsv.gz
Model prediction performance metrics for top model per cohort	This paper	very_top_performing_models.tsv.gz
Pathway data	This paper	pathway_commons_filtered_by_source_no_chem_edges.sif.gz
Sample prediction and feature list reporting file format	This paper	tmp_model_output_files_format.md
Model scoring script	This paper	tmp_prediction_file_to_performance_metrics.py
Pathway for generating pruned version (line below)	This paper	pathway_commons_v12_main_component_20211122.RData
Pathway for Figure 6	This paper	pathway_commons_v12_pruned_for_landscape_analysis_20211122.RData
Quantile rescaling script, for Figure 3D	This paper	quantile_rescale.py
METABRIC subtype silhouette scores, for Figure 3D	This paper	METABRIC_PAM50_silhouettes_no_Normal-Like_2022-03-15.tsv
AKLIMATE subtype predictions on rescaled METABRIC samples, for Figures 3C and 3D	This paper	aklimate_predict_metabric_brca.tsv
SK Grid subtype predictions on rescaled METABRIC samples, for Figure 3C	This paper	metabric_adaboost.tsv
JADBio model data associated with Docker Image	This paper	models_jadbio.tar.gz
CloudForest model data associated with Docker image	This paper	models_cf..tar.gz
AKLIMATE importance scores for Figure 5	This paper	aklimate_feature_importance_scores_20200807.tar.gz
Figure 5 input data; COADREAD methylation analysis	This paper	20220425_TMP_DNA_methylation_features_analysis_COAD.tsv
Figure 5 input data: LGGGBM methylation analysis	This paper	20220425_TMP_DNA_methylation_features_analysis_LGGGBM.tsv
Figure 5 input data; BRCA PAM50 membership	This paper	brca_pam50_hits.tsv
Figure 5 input data; JADBIO feature importance	This paper	jadbio_ft_importances_f1.tar.gz
Figure 5 input data; top models mapping of classifier name to feature set name	This paper	modelID_performance2importance.json
Figure 5 input data; all models mapping of classifier name to feature set name	This paper	modelID_performance2importance_ALLCOHORTS.json

(*Continued on next page*)

Please cite this article in press as: Ellrott et al., Classification of non-TCGA cancer samples to TCGA molecular subtypes using compact feature sets, Cancer Cell (2024), https://doi.org/10.1016/j.ccell.2024.12.002

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Sub-sampling experiment raw results, 100 predictions at each sampling size for each cohort	This paper	TMP_sub-sampling_experiment.tgz
Result files from SK Grid	This paper	skgrid_results_20210708.tar.gz
Result files from AKLIMATE	This paper	aklimate_predictions_and_features_20200630.tar.gz
Result files from subSCOPE	This paper	subSCOPE_results.tar.gz
Result files from the Gnosis platform	This paper	gnosis-results.tar.gz
Result files from CloudForest	This paper	Cloud_Forest_v12_sample_predictions.zip
User renaming of features to TMP nomenclature	This paper	ft_name_convert.tar.gz
Top model info (name, parameters, ft list) - Docker	This paper	model_info.json
METABRIC expression data	Curtis et al.[24]	https://ega-archive.org/studies/EGAS00000000083
AURORA expression data	Gene Expression Omnibus	Gene expression GSE212375; PAM50 - PMC9886551 Table S2
Software and algorithms		
subSCOPE docker Image	This paper	syn30993770; subscope.tar.gz
AKLIMATE docker Image	Uzunangelov et al.[17]	syn29659459; aklimate.tar.gz
JADBio docker Image	Tsamardinos et al.[20]	syn31114207; jadbio.tar.gz
SK Grid docker Image	This paper	syn29658355; sk_grid.tar.gz
CloudForest docker Image	Bressler et al.[18]	syn30267068; cloudforest.tar.gz
subSCOPE Workflow	This paper	https://github.com/NCICCGPO/gdan-tmp-models
AKLIMATE Workflow	This paper	https://github.com/NCICCGPO/gdan-tmp-models
JADBio Workflow	This paper	https://github.com/NCICCGPO/gdan-tmp-models
SK Grid Workflow	This paper	https://github.com/NCICCGPO/gdan-tmp-models
CloudForest Workflow	This paper	https://github.com/NCICCGPO/gdan-tmp-models
Consensus prediction calling	This paper	https://github.com/NCICCGPO/gdan-tmp-models
PathwaySpace	This paper	https://github.com/sysbiolab/PathwaySpace
Other		
Public data files at the NCI Publication page	This paper	https://gdc.cancer.gov/about-data/publications/CCG-TMP-2022

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

In this study, conducted by the Tumor Molecular Pathology (TMP) Analysis Working Group of the Genomic Data Analysis Network (GDAN), we used the Cancer Genome Atlas (TCGA) Research Network tumor and matched normal samples with informed consent under their local Institutional Review Boards.

METHOD DETAILS

Cohort and cancer subtype definition

For most cohorts, we adopted the cancer cohort abbreviations and subtype classifications reported by TCGA cancer landmark studies (Table S1; Figure 1). The approaches used in these studies for subtype discovery included unsupervised clustering or expert-defined decision trees, or a combination of both. Clustering-based subtypes were inferred from individual data types, merged single-platform cluster assignments or multi-platform integrative clustering (iCluster).[64] Decision tree groupings were based on specific somatic alterations (e.g., mutations or fusions) or other tumor characteristics (e.g., MSI or viral infection).

For gliomas, IDH1/2 mutational status has been shown to be a better indication of molecular similarity than tumor grade,[14,65] and so we merged Glioblastomas and Low-Grade Gliomas into LGGGBM.

In the TCGA Hepatocellular Carcinoma and Cholangiocarcinoma cohorts, some tumors were shown to molecularly cluster with the other tumor type.[66] Given that precise anatomical boundaries can be ambiguous for large tumors in these cancer types, we merged these two into a single cohort: LIHCCHOL.

Gastrointestinal tumors were regrouped into tumor types that had shared or dissimilar molecular characteristics.[36,67] For ESCA we separated the molecularly distinct squamous tumors as an ESCC cohort, and we merged the remaining ESCA esophageal

Please cite this article in press as: Ellrott et al., Classification of non-TCGA cancer samples to TCGA molecular subtypes using compact feature sets, Cancer Cell (2024), https://doi.org/10.1016/j.ccell.2024.12.002

adenocarcinomas (EAC) with all STAD tumors to form a GEA cohort (gastroesophageal adenocarcinomas), consistent with the TCGA interpretation of GI cancer subtype relationships.[36,67] The COAD and READ tumors were merged into a COADREAD group (Colon and Rectal adenocarcinomas).

We were not able to identify molecular subtypes for TCGA Kidney Clear Cell (KIRC) and Chromophobe (KICH) tumors, and so merged these into a single KIRCKICH cohort for the purposes of constructing a two-class classifier of histological subtype.

For four cohorts (HNSC, ESCC, UCEC, GEA), subtypes were based on a combination of clustering and decision tree analysis. We used histological classifications for two tumor types (TGCT, SARC), given that molecular classification for TGCT did not differ meaningfully from histologic classification,[68] and, for SARC, no molecular subtypes sufficiently covered all histological subtypes.

Before classifier training we removed subtypes that had too few samples for cross-fold validation (Table S1). For the BRCA cohort, we eliminated the Normal-like subtype, as it is not included in the St. Gallen Guidelines for Intrinsic Subtype,[69] and is not included in the PROSIGNA clinical assay.[70] As well, a study of micro-dissected breast cancers with more than 90% tumor cellularity had no samples classified as Normal-like.[71]

With the exceptions noted above, for each cancer cohort we retrieved molecular subtypes from prior TCGA publications, which had defined them using varying approaches and data types (summarized in Table S1). For LUAD, we used recent molecular subtype definitions.[72] For BLCA we used consensus MIBC subtypes.[73] The methods that had been used for subtyping included unsupervised clustering, or expert-defined decision trees, or a combination of both. Where appropriate, clustering-based subtypes were inferred from single platforms (gene expression, mutation, or DNA methylation); other subtypes were based on multi-platform integrative clustering (iCluster) or on methods that merged single-platform cluster assignments (Clustering of Cluster Assignments, COCA, or Similarity Network Fusion, SNF). Decision tree groupings were based on specific somatic alterations (e.g., mutations or fusions) or other tumor characteristics (e.g., MSI or viral infection). For four cohorts (HNSC, ESCC, UCEC, GEA), subtypes were based on a combination of clustering and decision tree analysis.

Altogether, we worked with data from 8,791 TCGA samples from 26 cancer cohorts. Each TCGA sample had five genomic data types available. Each sample had been assigned to one of 106 molecular subtypes. Subtypes were ordered and numbered within cancer cohort by declining sample size prior to data filtering. Cohort and subtype labels were considered as the 'ground truth' for our classifier model development.

Classifier model development

Given the molecular subtypes in each cohort, we then generated subtype-balanced repeated cross-validation folds, and set these as training and test sets. To develop classifier models capable of assigning a new sample to a previously defined subtype, we tested five ML approaches: AKLIMATE, CloudForest, SK Grid, JADBio, and subSCOPE (Figure 2). We note that SK Grid and JADBio each employed a collection of embedded approaches (see STAR Methods). Therefore, the true number of approaches implicitly tested in our study is far more than five. For AKLIMATE, CloudForest, SK Grid and JADBio, each cohort was trained separately. In contrast, we trained subSCOPE's Neural Nets (NNs) on subtype data from all cancer cohorts simultaneously. For all cohorts, we assembled multi-platform genomic data from PanCancer Atlas resources (gdc.cancer.gov/node/977) (see STAR Methods). We trained and tested all classifiers using the same cross-folds, and aggregated results into a single matrix. We generated performance statistics from the test cross-folds, and retained the classifier-selected features for further analysis.

Dataset creation

The data for this study was created by aggregating molecular profiles found in the NCI's Genomic Data Commons data system. In that system, all results are separated by sample and data type. We organized the data by creating files that combine all samples and data types into a single matrix, with one for each of the 26 tumor type cohorts included in the study (Figure 2A). Since machine-learning methods can generate complex models with poor interpretability, we used a gene-centric approach, emphasizing the selection of fewer features, while retaining predictive performance, to facilitate the analysis of the biological significance of the selected features. We divided each cohort's samples into training and test sets, using a 5-fold cross-validation, stratified across the retained subtype labels. For each tumor type, we generated 100 repeats, i.e., 100 ways in which the samples were divided over 5-folds. The header of each cross-validation column is of the format Rx:Fy, where x denotes the repeat (from 1 to 100) and y denotes the fold (from 1 to 5). In this column, values of 0 indicate training samples, while values of 1 indicate test samples.

Constructing single-feature matrices

mRNA gene expression feature matrix

Batch-corrected mRNA matrices were obtained from the TCGA PanCancer Atlas Projects (https://gdc.cancer.gov/about-data/publications/pancanatlas).[4] The batch-correction adjusted for sequencer type, sequencing centers (the University of North Carolina [UNC] and the British Columbia Cancer Agency [BCCA]), and a plate effect observed in PRAD. Briefly, genes were adjusted using a novel algorithm called EB++; a variant of the Empirical Bayes/ComBat algorithm. Genes with mostly zero reads or with residual batch effects (~10% of genes) were removed from the adjusted samples and replaced with NAs. First, PRAD batches 312 and 320 were adjusted to the rest of the PRAD batches, then UNC Illumina GAII sequenced samples (UCEC, COAD, READ) were adjusted to the UNC Illumina HiSeq data. BCCA Illumina GAII-sequenced samples (LAML, STAD, ESCA) were also adjusted to Illumina HiSeq

Please cite this article in press as: Ellrott et al., Classification of non-TCGA cancer samples to TCGA molecular subtypes using compact feature sets, Cancer Cell (2024), https://doi.org/10.1016/j.ccell.2024.12.002

generated data. Most of the GBM gene expression data were generated using older microarray technology, which sometimes yielded values between 0 and 1. The log2 transformation of those values and further batch effects adjustment resulted in some negative numbers in the LGGGBM expression data.

miRNA expression feature matrix

To generate miRNA-seq data for ~11 thousand samples across TCGA projects we had used two library construction protocols: poly(A) selection ('MultiMACS') and total RNA ('Direct'),[74] taking care to use only one protocol per project (see https://gdc.cancer.gov/node/977, PanCanAtlas_miRNA_sample_information_list.txt). Initially we used Illumina GAII sequencers, and then HiSeq sequencers. We annotated aligned miRNA reads with miRBase v16, which contained 1212 mature strands.

Batch correction of miRNAseq mature strand RPM data addressed library protocols and sequencers. We batch-corrected normalized abundance (i.e., reads per million, RPM) data for 743 expressed mature strands (https://gdc.cancer.gov/node/977). These included approximately 650 mature strands that were sufficiently highly expressed to behave well in batch correction; we removed weakly abundant mature strands, because such strands are unlikely to be biologically influential.[75,76] We then added non-batch-corrected RPM profiles for approximately ~100 mature strands that were known to be important in some cancers but which we had removed following batch correction. For three classifier projects we combined two TCGA cohorts: KIRCKICH, LIHCCHOL, and LGGGBM. No miRNAs were available for GBM, so we excluded miRNAs from work with the combined LGGGBM cohort. For the LIHCCHOL and KIRCKICH combined cohorts, we anticipated that, despite batch correction, there could be residual batch effects in the miRNA-seq data due to the different library protocols, so we used no miRNA data in generating classifiers for LIHCCHOL and KIRCKICH cancer cohorts.

DNA methylation feature matrix

We used the pre-processed DNA methylation β value matrices generated for the TCGA PanCancer Atlas analysis projects,[4] obtained from the PanCanAtlas Publications page (https://gdc.cancer.gov/about-data/publications/pancanatlas). Data for ten cancer types, including BRCA, COADREAD, GEA, KIRCKICH, KIRP, LGGGBM, LUAD, LUSC, OV, and UCEC, were generated using two generations of Infinium arrays, the older HumanMethylation27 (HM27) and the newer HumanMethylation450 (HM450: 396,065 features). Therefore, we used the merged HM27-HM450 data matrix containing 22,601 features shared between the HM27 and HM450 platforms. For the other sixteen cancer types, we analyzed the HM450 data matrix, which included 396,065 features.

For DNA hypermethylation features, we selected CpG sites that acquired cancer-associated hypermethylation. We first examined the DNA methylation profiles from histologically normal tissues to identify features that lacked tissue-specific DNA methylation. For the merged HM27-HM450 data, we used 1,064 normal tissue samples from 22 different tissue types, which included BLCA (n = 19), BRCA (n = 112), CESC (n = 3), COADREAD (n = 81), ESCC (n = 2), GEA (n = 38), LGGGBM (n = 2), HNSC (n = 50), KIRCKICH (n = 344), KIRP (n = 48), LIHCCHOL (n = 59), LUAD (n = 52), LUSC (n = 67), OV (n = 12), PAAD (n = 10), PCPG (n = 3), PRAD (n = 50), SARC (n = 4), SKCM (n = 2), THCA (n = 56), THYM (n = 2), and UCEC (n = 46). The HM450 data included 720 normal tissue samples from 21 different tissue types consisting of BLCA (n = 19), BRCA (n = 86), CESC (n = 3), COADREAD (n = 41), ESCC (n = 2), GEA (n = 13), LGGGBM (n = 2), HNSC (n = 50), KIRCKICH (n = 157), KIRP (n = 43), LIHCCHOL (n = 59), LUAD (n = 30), LUSC (n = 41), PAAD (n = 10), PCPG (n = 3), PRAD (n = 50), SARC (n = 4), SKCM (n = 2), THCA (n = 56), THYM (n = 2), and UCEC (n = 45). We also analyzed HM450 data from three whole blood samples from the *sesameData* R/Bioconductor package.[77] We then selected 10,114 and 100,405 features methylated at < 10% frequency in any tissue type using a β value of >0.2 to define positive DNA methylation in the merged HM27-HM450 and HM450 datasets, respectively. Finally, in each tumor type, we identified features that were methylated at a β value of ≥0.3 in more than 2% of tumors, yielding DNA methylation feature metrics that vary in the numbers of features across tumor types.

For Loss-of-function (LOF) features, we started with the DNA hypermethylation matrices generated above that we binarized using a β value of ≥0.3 to define positive DNA methylation and <0.3 to specify lack of methylation. We identified 889 and 8,797 features in the merged HM27-HM450 and HM450 platforms, respectively, located within CpG island promoter regions (1,500-bp flanking regions upstream and downstream of Transcription Start Sites) associated with 587 genes covered in the mutation feature matrix. We created LOF features for DNA hypermethylation by extracting features overlapping these promoter features on the mutated genes from the binarized feature matrices. If there were multiple features associated with the same gene, a sample identified as methylated at more than half the features for the corresponding gene was also labeled as methylated at the gene level.

Copy number feature matrix

Copy number features were derived from TCGA PanCancer Atlas gene-level thresholded copy number data from GISTIC 2.0.[78] All gene level deletions (−1 or −2) were all set to −1, and all gene-level amplifications (1 or 2) were set to +1. Then, in each tumor type, genes within a cytoband were compressed so that those with identical copy number profiles across all tumors were represented by one gene. If possible, known oncogenes or tumor suppressors were chosen as the representative genes.

Mutation feature matrix

Mutation features were derived from the PanCanAtlas MC3 publication mc3.v0.2.8.PUBLIC.LAML_PATCH.maf.oncokb.txt.[79] We generated four types of mutation features, which covered 587 genes and 470 hotspots. These included non-silent mutations, hotspot mutations, loss-of-function mutations, and composites. Non-silent mutation features were created from somatic variants and coded with the prefix B:MUTA:nons. A variant was considered non-silent if the "variant classification" field was "Missense_Mutation", "Nonsense_Mutation", "Frame_Shift_Del", "Splice_Site", "Frame_Shift_Ins", "In_Frame_Del", "In_Frame_Ins", "Translation_Start_Site", "Nonstop_Mutation", or "Splice_Region". We collected the list of genes from the driver mutations published by the TCGA PanCanAtlas group[80] and from the non-silent catalog.[81] Hotspot mutation features were coded with the prefix B:MUTA:HOTS and represent the presence of variants resulting in amino acid changes that occur in protein mutation hotspots.[82] Loss-of-function

mutation features were coded with the prefix B:MUTA:LOF, and represent events in which either 1) a non-silent mutation occurred alongside a deletion in the same gene in the same sample or 2) a promoter hypermethylation event was observed in the gene. For these calls, we used the binarized versions for both the methylation and copy number data. Finally, we created composite mutation features to indicate an event in which any impactful variant occurred in a sample; we encoded these with the prefix B:MUTA:COMP. These features represent cases in which any of the 3 previous types, non-silent or hotspot or loss-of-function were recorded.

Filtering out missing values

We applied a two-step approach to identify and remove samples and features containing missing values, performing this separately for each tumor type's single data type tables. These single data-type feature matrices contained measurement data from the five major genomic platforms.

(1) mRNA gene expression [continuous values]
(2) miRNA mature strand expression [continuous values] (as noted, for KIRCKICH, LIHCCHOL and LGGGBM cohorts we used no miRNA data)
(3) DNA methylation beta values [continuous values between 0 and 1]
(4) Copy number data [ternary; −1, 0, 1]
(5) Non-silent Mutations [binarized]
(6) Hotspot Mutations [binarized]
(7) "Loss-of-function" Mutations [binarized]
(8) Composite Mutations [binarized]

In a first step, we iteratively removed samples and features with more than 20% missing data until no such samples, nor such features remained. Specifically, we first removed all samples with more than 20% missing data. Then, we removed all features with more than 20% missing data across samples. We continued this process until all remaining samples and features had less than 20% missing data. We tried two versions of this: one in which we first removed samples, and one in which we first removed features. In the end, we chose the approach that minimized the product of the number of removed samples and features for each cancer cohort. Finally, in a second step, we removed all features with missing values, leading to data tables that had no missing values.

Merging of single data type matrices

We combined the eight data tables after filtering out missing values, as described above, and excluding samples that we had removed in one or more of the eight filtered data tables. In other words, the samples in the resulting combined matrix were the intersection of the samples in the eight filtered data tables. Automated diagnostic plots as well as automated and manual spot-checking ensured that these large numerical tables were correctly merged. The samples (i.e., the rows in the final tables) were TCGA samples, identified by 12-character TCGA case ID barcodes (e.g., TCGA-02-0001). The columns were the features. Each feature name was structured to contain the 1) variable type, i.e., binary/categorical/numerical; 2) molecular platform i.e., MUTA for mutation, METH for methylation, MIR for microRNA mature strands, GEXP for messenger RNA, CNVR for copy number variation; 3) additional molecular annotations, such as HOTS for mutational hotspots; 4) gene name; and 5) additional molecular label.

An example feature ID is "B:MUTA:HOTS:PIK3R1:pR348:". This is a binary mutation hotspot call for a mutation at AA position 348 in PIK3R1. In the combined data table, the first column contains the samples, and the second column contains the subtypes, which were used as labels for classification. Subsequent columns contain the features.

Building a cohort of machine learning methods

We employed five distinct feature selection and/or model fitting pipelines. These pipelines include 1) CloudForest, led by ISB; 2) AKLIMATE, led by UCSC; 3) subSCOPE, led by the BC Cancer's Genome Sciences Center; 4) SK Grid, led by OHSU; and 5) JADBio, led by JADBio Gnosis DA S.A. These algorithms were built upon different machine-learning classification philosophies and employed diverse feature selection approaches (including filtering, wrapping, embedding, and/or prior knowledge) either before or during their model fitting. Some of these methods searched for sparse sets of classifier features (e.g., JADBio, SK Grid) while others attempted to capture discriminating features rich in established biology and pathway knowledge (e.g., AKLIMATE, subSCOPE) — all aiming to maximize cancer subtype prediction performance. This report considers, in depth, performance metrics of these different classification algorithms. For the translational setting of classifying a new patient sample into a previously defined TCGA subtype, our results provide guidance for matching methods with feature sets to predict subtypes of a particular cancer type.

Applying the CloudForest method

CloudForest is a Random Forest (RF) package written in Go, which is particularly well suited for large, heterogeneous, genomics and biomedical datasets.[18] CloudForest was run as a standard RF classification model with 50,000 trees, a minimum leaf size of 5, balanced bagging, and default parameters for other options. The pipeline is implemented as an RF workflow with feature reduction steps. Specifically, first an RF is trained using all features. Then the 1,000 best features are selected and a second RF model is trained on the same samples using only the best 1,000 best features. This process is repeated for the best 100, 50, 10, 5 and 1 features. This leads to 7 trained RF models (all features, 1000, 100, 50, 10, 5 features, and 1 best feature). The model training was done using the training folds; classification performance was reported on the held-out test sets. Feature importance was measured using Gini

impurity. Summarized feature importance scores were averaged over the folds and repeats. CloudForest experiments were performed for each of the 26 tumor types separately, and within each tumor type, using six different feature sets separately: 1) binary mutation calls, 2) ternary copy number data, 3) continuous gene expression data, 4) continuous DNA methylation data, 5) continuous miRNA expression data, and 6) the previous five datasets combined.

Applying the AKLIMATE method

Algorithm for Kernel Learning with Integrative Modules of Approximating Tree Ensembles (AKLIMATE) is a kernel-based stacked learner that is informed by biological pathway gene memberships.[17] Based on a Multiple Kernel Learning (MKL) approach, AKLIMATE utilizes a set of combined kernels, each of which represents a different aspect of sample-sample similarities. First, samples are evaluated using random forest models. The outputs of these predictions are translated into distance matrices, which are then passed into an Elastic Net MKL system. The component random forest models used by the MKL are trained on feature sets derived from prior biological knowledge. The prior knowledge used for these elements includes biological pathway information. Using this biological information-based modeling, the AKLIMATE model has been previously used to predict microsatellite instability in endometrial and colon cancers, survival in breast cancer, and shRNA knockdown viability in cancer cell lines.[17]

In the work reported here, we used AKLIMATE as a feature-scoring tool. The pipeline begins by training an AKLIMATE model. AKLIMATE is given the sample data for the training set as well as input "feature sets". The feature importance scores are extracted from the trained AKLIMATE model. Many thousands of features may be assigned a non-zero importance score in the AKLIMATE model. To get a smaller model with n input features, the n most important features are used to subset the training data. This subset of training data is used to train a random forest classifier (using the *Ranger R* package). Finally, the test set of sample data that was held out at the beginning is used to assess the sample subtype prediction performance of the random forest classification model.

The AKLIMATE pipeline used a compendium consisting of ~17,000 feature sets collected from several sources including: MSigDB, genesigDB, PathwayCommons, KEGG, Reactome, PID, and genomic position neighborhood.

In the work described here, AKLIMATE used 4 feature s: GEXP, CNVR, METH, and MUTA. In pilot experiments, the AKLIMATE pipeline found that miRNA features introduced noise in to the data, resulting in reduced classification performance. Consequently, miRNA features were used in no AKLIMATE models in this study.

The copy number features in the common, combined dataset were a "compressed" representation of the original, full copy number data of the TCGA samples. The full copy number data have many correlated features that can be attributed to copy number changes for genes that are located on the same physical region of the chromosome. In an effort to reduce the redundancy in the data, the copy number features were organized into highly correlated groups and a representative of the group was chosen to be included in the "compressed" copy number data. Since AKLIMATE leverages the biological prior knowledge contained with sample features, the compression likely had a negative effect on the ability of AKLIMATE to take full advantage of the copy number data. The CNVR compression causes reduction of copy number representation in pathway space, since the selected feature within a correlated copy number group is randomly selected (from a pathway point of view).

Applying the subSCOPE method

subSCOPE is a deep neural network-based system. Its training was markedly different than the other ML approaches used, on two major points: its training set and how important features were identified. Unlike other systems that worked on one tumor cohort at a time, subSCOPE was trained on the entire cancer cohort set jointly, learning how to identify all subtypes across all cancers at the same time. Traditionally, deep neural networks continue to improve as more data are added to the training set, while other methods tend to converge.[83] We took this approach to expand the total training set size, so trained for a problem with 8,791 samples, rather than a series of problems with approximately 100–500 samples each. Also, by training all problems at the same time, the neural network would have the opportunity to identify common patterns across multiple cancer types. Unlike other methods, subSCOPE was first trained on all input features. Once the model was trained, a second system, we used *DeepLift*[84] for feature importance calculations.

Applying the SK grid method

The SK Grid system is designed as an 'off the shelf' machine-learning pipeline that utilizes methods available in the popular Python package 'Scikit-Learn'.[19] Feature selection is done with Recursive Feature Elimination (RFE) and Forward-Backward Early Dropping (FBED)[85] to generate feature sets of mixed and single TCGA data types. Feature selection for each cancer cohort was 1) run independently across all molecular features (gene expression, copy number variation, miRNA, methylation, mutation status), 2) run independently for each of the five molecular feature types to produce separate feature sets, and 3) a concatenated set of the second step of independent runs.

Each of these selected feature sets was input to a set of 14 SK Grid classifiers to identify the most accurate combination of feature set and classifier. Default hyperparameter settings were used for each classifier. The classifiers utilized were: Adaboost, Bernoulli Naive Bayes, Decision Tree, Extra Trees, Gaussian Naive Bayes, Gaussian Process, K Nearest Neighbors, Logistic Regression, Multi-layer Perceptron, Multinomial Naive Bayes, Passive Aggressive, Random Forest, Stochastic Gradient Descent, and Support Vector Machine.

Applying the JADBio method

JADBio is an AutoML system developed by JADBio Gnosis DA S.A.,[20] operating as a Software as a Service through a web interface (http://jadbio.com), as well as an API. Once the user starts an analysis, a knowledge-based decision support system (namely, the Algorithm and Hyper-Parameter Space selection, AHPS, system) selects the appropriate algorithms and hyper-parameter value combinations to try with the specific data, depending on the dataset characteristics (most importantly, sample size and number

Please cite this article in press as: Ellrott et al., Classification of non-TCGA cancer samples to TCGA molecular subtypes using compact feature sets, Cancer Cell (2024), https://doi.org/10.1016/j.ccell.2024.12.002

of features), as the preferences specified by the user. Among the user inputs is the desired level of tuning effort, which trades off between computational time and thoroughness in exploring the hyper-parameter space. Notably, the AHPS selects the analysis protocol as well, i.e., whether to use a hold-out approach or (repeated) cross validation, along with the number of folds and repetitions.

Machine learning modeling methods include Random Forest, other Decision Trees, Support Vector Machines, and Generalized Linear Models. Feature selection methods include the Statistical Equivalent Signatures (SES)[33] and Lasso[86] algorithms.

Once the space of possible configurations (i.e., machine-learning pipelines) is defined, the best configuration of pre-processing methods, feature selection/modeling algorithms, and respective hyper-parameters is selected through a grid search. As part of the performance analysis, JADBio calculates conservative, unbiased estimates of best model's performance on unseen data, using a specialized bootstrapping algorithm on the training data.[87] In the context of the analyses presented in this work, the pipeline used the JADBio default settings, with a few modifications. First, the system utilized the same protocol (repeated cross validation) and folds employed by the other groups, for the sake of comparability. While executing the repeated cross validation procedure, JADBio employed several techniques for speeding up computations while preserving the quality of the results; for example, the system can automatically decide to avoid further cross-validation repetitions if improvements in performance are deemed improbable. The tuning effort was set to ''extensive'' for all analyses, except for BLCA, COADREAD, LGGGBM, LUSC (''preliminary''), and GEA, HNSC, LUAD, OV, SKCM, THCA, and UCEC (''normal''). In each analysis, the best configuration was identified from a balanced accuracy metric. Finally, predictive performances were computed as described in the section below, for ensuring the comparability of our results with the those from the other machine-learning systems.

Evaluation of prediction performance

Project researchers developing the ML pipelines submitted prediction result files in a pre-defined format (see Publication Page: https://gdc.cancer.gov/about-data/publications/CCG-TMP-2022). One set of files described the individual sample subtype predictions for each repeat-fold. Another set of files described the feature list used for each model. Some models generated features; and those were recorded in the form of a tab-separated file.

The sample prediction files were used to estimate the performance of each model. The following performance metrics were computed on the cohort and/or subtype levels with the scikit-learn Python module[19]: accuracy, balanced accuracy, weighted F1 score, precision, and recall. These results are available on the Publication Page (See: https://gdc.cancer.gov/about-data/publications/CCG-TMP-2022).

At the cancer cohort level, we defined 'cohort performance' as the overall weighted F1 score for all samples within a cohort. This metric balances precision (positive predictive value) and recall (sensitivity).[22] We separately defined 'subtype performance' as the within-subtype weighted F1 score.

The top-performing model was determined by both a model's cohort-level weighted F1 score and the number of model features. For the work reported here, the evaluation considered only models that used at most 100 features (JADBio intentionally limited its feature set to a maximum of 25). In the case of ties, the model with the lower standard deviation in cohort-level weighted F1 was considered superior. The top models for each method/cohort pair and the single top model for each cohort are provided on the Publication Page.

Table S5 lists 737 models, representing the best-performing models for each of the 26 cancer cohorts, data types, and model training algorithms. The Table can be sorted to select the best model for the appropriate cancer and data type. Each model lists the selected features, the mean overall weighted F1 obtained from cross-validation, and the bash command to run the model. A Docker container for these 737 models can be found on GitHub at https://github.com/NCICCGPO/gdan-tmp-models. Each model outputs a prediction score, reflecting the classification confidence for the model. The lowest prediction score that provided 95% accuracy in cross-validation testing is listed for comparison to the score obtained for a newly classified sample. The 95% accuracy threshold was computed over the test predictions in all cross validation repeats except for the CloudForest models. CloudForest uses resubstitution predictions. Some models had low prediction performance, with the prediction accuracy never reaching the 95% correct threshold at even the highest prediction score. Such models have a value of "#N/A" for this threshold. Some subSCOPEmodels have a value of "#N/A" for the mean overall weighted F1. This is attributable to subSCOPE models being trained as pan-cancer predictors, in contrast with all other methods being trained on single cancer cohorts.

Containerization of the top models

Ready-to-run prediction algorithms were created for every top model. These included parameter files and runner code that would open configuration and input files, and then execute the prediction algorithm. For each cancer type and data modality a top model container was generated that included both 'all feature' and 'single feature' options; for example, 'gene expression only' models.

The runner code and model configuration files were compiled into Docker images, so that all runtime dependencies would be containerized. This containerization allows models to run on a variety of target systems and to be incorporated easily into cloud-based prediction systems. In addition, tool wrappers were written in the Common Workflow Language (CWL), to allow consistent invocation patterns. Finally, a unified workflow was created that tied all of the tools into a single invocation.

The docker images were uploaded to the Synapse Docker Image Registry, and were also compiled into an archived tar-ball that was uploaded to the GDC (tool wrappers and the full integrated pipeline were published to https://github.com/NCICCGPO/gdan-tmp-models). The repository also includes utility code and example documentation.

Applying models to external datasets

The pre-trained models were validated independently with two external datasets: METABRIC[24,51] and AURORA.[25] To evaluate the models, we used 939 fresh-frozen METABRIC primary tumor samples and 28 formalin-fixed paraffin-embedded (FFPE) AURORA primary tumor samples. Both external data contained PAM50 annotations and underwent the same analysis pipeline as outlined below.

A gene-by-gene numerical transformation was used to convert the expression values to the RNA-seq-based quantification that we had used for TCGA samples. Each gene was transformed independently with a quantile rescaling procedure, using the array of values across the two cohorts as the source (METABRIC or AURORA) and destination (TCGA BRCA) distributions. The quantile transform code is made available as part of the containerized model toolkit on GitHub.

The quantile rescaling method was applied to external data for each trained model separately, since each model used a different subset of the full TCGA BRCA sample cohort. The Normal-like samples were retained in this step; all samples were used in the rescaling. The trained models were applied to their corresponding rescaled external data, resulting in BRCA subtype predictions for each sample. The collective subtype calls from all models for a sample was consolidated into a single subtype call per sample using a majority vote approach, where ties were broken based on highest mean model confidence (consensus method). The percent prediction concordance for a set of predictions was calculated as the proportion of samples in which the predicted subtype matches the PAM50 annotation. Samples annotated with the Claudin and Normal-like subtype were present in the AURORA and/or METABRIC cohorts. Since the Normal-like and Claudin samples were excluded from all BRCA model training, these samples were excluded from this analysis.

Subtype calling patterns in external datasets

This set of experiments was intended to measure the correlation of subtype silhouette scores of the training sample set with the prediction performance of ML models. The distribution of silhouette scores in the training set as well as the sample set size within each subtype was preserved. The models were trained on TCGA BRCA samples and tested on the METABRIC discovery sample set.[24,51] The TCGA BRCA sample set (n = 995) consists of 535 BRCA_1 (LumA), 205 BRCA_2 (LumB), 175 BRCA_3 (Basal), and 80 BRCA_4 (Her2) samples. The METABRIC discovery sample set (n = 997) contained 466 LumA, 268 LumB, 118 Basal, 87 Her2, and 58 Normal-like samples. The inference of PAM50 subtypes for the TCGA expression-based clusters was established in previous work.[51] The METABRIC expression data were assayed using the Illumina HumanHT12v3 microarray platform. To ensure compatibility with the models trained on TCGA's RNA-seq data, the Illumina probe ID's were mapped to TCGA Feature IDs using the updated probe-to-gene annotation available from the IlluminaHumanv3.db package.[88] In cases where multiple Illumina probes mapped to the same TCGA Feature ID, we used the median expression value. 19,215 TCGA feature IDs were present in both the TCGA BRCA dataset and the METABRIC dataset. The classifier model training was performed using TCGA BRCA training data in the overlapping feature set.

Sample silhouette scores were calculated separately for the TCGA BRCA sample set and the METABRIC sample set. Both expression matrices were reduced to represent the 50-gene PAM50 gene set.[23] The silhouette score[26] was computed for each sample in the two matrices, based on the subtype assignment. The TCGA subtype assignment was used for the TCGA BRCA samples (BRCA_1, BRCA_2, BRCA_3, BRCA_4). The METABRIC PAM50 assignment was used for the METABRIC samples (LumA, LumB, Basal, Her2, Normal-like). A percentile rank was computed for each sample within its subtype. The samples with the highest silhouette width value in their subtypes were given the highest percentile ranks (i.e., 100%).

The silhouette score was calculated by generating an expression matrix constructed using a subset of ENTREZ ID mapped probes that covered the genes used in the PAM50 signature, resulting in an expression matrix of 50 genes. This matrix was then passed to the sklearn.metrics.silhouette_score function to compute the silhouette width.

Generating model training sample sets

To assess whether discordant model predictions were attributable, at least in part, to ambiguity in the assignment of the PAM50 subtype labels in that cohort, we investigated whether training classifiers on the subset of samples with high silhouette score subtype calls would produce higher-ranked models, or whether focusing on the subset of samples with low silhouette scores would instead yield better performance. Since this approach would reduce sample size, we compared *equally sized* data subsets, enriching for either high silhouette scores, which we refer to as the "typical set", or enriching for low silhouette scores "atypical set", versus no enrichment for silhouette scores "full set". Random sampling was used to generate 30 ML model training sets from each of three subsets of the TCGA-BRCA cohort based on the sample silhouette width ranking (see Figure S1B). The "full set" represents 30 equally sized random samplings of the full cohort. The "typical set" represents the half of the cohort with the highest per-sample silhouette width scores. The "atypical set" represents the half of the cohort with the lowest silhouette width scores. Each training set was created to have an identical number of samples of each subtype, which were: BRCA_1 (215), BRCA_2 (83), BRCA_3 (71), and BRCA_4 (33). Each training set was used to train a separate AKLIMATE model that used 100 gene expression features. The models from the "full set" had marginally higher prediction performance than both the "typical set" (p = 0.046, one-sided Mann-Whitney Wilcoxon test with Bonferroni multiple test correction) and the "atypical set" (p = 0.053).

Performance and feature set size relationship

Three of the five methods (AKLIMATE, CloudForest, JADBio) provided mixed and single-data type models with increasing numbers of features selected. Selected feature set sizes ranged from 5 to 100 for AKLIMATE models, 1 to 100 for CloudForest models, and 1 to 25 for JADBio models. For each set of models within a given cancer cohort built on a given data type, the cohort performance was

plotted (i.e., mean overall weighted F1) against feature set size up to 100 features. For JADBio, the performance was extrapolated to 100 features as the cohort performance of JADBio models with the largest number of features. Assuming cohort performance of 0 at 0 features, the area under the F1 curve (AUF1C) was calculated using the composite trapezoid rule. The AUF1C has a maximum possible value of 99.5 (which is achieved if all models with $\geq$ 1 features have cohort performance of 1). The AUF1C values for mixed and single-data type models from each method within each cancer cohort were visualized using the *heatmap.plus R* package.

Training set power analysis

To quantify the effect of cohort size on classifier predictive performance, we selected the SK Grid model and feature set with the greatest performance for each cohort. Within each cohort, we repeated 100 samplings without replacement at sample-size steps of five to 100 in increments of five, then from 100 to 250 total samples in increments of ten. For each sampling, the cohort's top SK Grid model was trained and tested with 10-fold cross-validation, using overall weighted F1 scoring. At each sampling size, the 100 predictions were averaged to generate a sample size/accuracy 'learning curve'.

We fit a three-parameter inverse power-law to each learning curve. For making predictions, we fit the inverse power-law function to the 100 individual (unaveraged) cohort performance score-sets in the sample-size range of 35–70. We then projected 100 discrete predictions of overall weighted F1 score to a sample size of 250. For each of the 15 cohorts with at least 250 samples, we compared this projected distribution to the actual observed score. A systematic search of the SciPy library of statistical distributions and their characteristics yielded the Burr Type XII curve as the best-fit predictor. For each of the 11 cohorts with less than 250 total samples, we used this curve and approach to project a prediction of cohort performance at a cohort size of 250 samples.

Aggregating TCGA subtype features

The TCGA subtype features were aggregated by cohort and gene symbol. For each feature (n = 3706) that mapped to a gene symbol (n = 3241), Table S3 shows the number of ML methods that detected that feature in a given cohort. This could be from 0 to 5 methods. For the pathway analyses assessing single cohorts, all features that mapped to the pathway space were depicted in Figure 6 (BRCA genes n = 113; LGGGBM genes n = 144; COADREAD genes n = 122). For the pathway analyses assessing all cohorts, we used a core feature set of n = 742 genes that were detected by at least two ML methods in at least one cohort; from these 742 'core' features, 533 were annotated in the pathway space depicted in Figure S4. Because CNVR could include large chromosome segments that contained multiple genes, CNVR features were excluded from the gene aggregation provided in Table S3.

Generating pathway maps

Analysis showed that the ML methods selected large sets features that did not overlap with each other. Given this pattern, we asked whether such distinct features shared an underlying common biology. The redundancy of biological networks can create equivalencies in the data (e.g., high gene expression correlation) between genes in the same pathway. Thus, it is possible that the features selected by methods shared more commonality than a direct comparison of feature sets might suggest. For this analysis, we created a comprehensive pathway map to both visualize and quantify the degree to which selected features were related to one another (both within and across methods). To this end, gene-gene interactions were collected from expert-curated data sources (PathwayCommons)[45] using the following pathway sources: Reactome, NCI-PID, HPRD, PANTHER, CORUM, KEGG, INOH, NetPath, Pathbank, and InnateDB. This resulted in 13,589 genes interconnected by 434,371 interactions, which is available for download from the GDC Publication Page (See: https://gdc.cancer.gov/about-data/publications/CCG-TMP-2022).

Visualizing cancer pathways

A single visual overview of these pathways was created using a geodesic projection onto the plane (Figure S4A), using edge betweenness centrality to guide the layout process. The layout attempts to arrange communities around hub vertices on the resulting two-dimensional pathway space. The final map is formed by the main connected component, represented by 12,990 vertices (genes) and 359,540 edges (interactions), and is available at the GDC Publication Page. In order to generate a compact image, edges of low betweenness were pruned using forward selection and backward elimination, and retaining edges of high centrality in the main connected component. The pruned graph is available at the GDC Publication Page. To serve as a guideline for understanding the major organizational patterns on the map, we indicated the location of 540 genes from COSMIC-CGC database (release v95, tier 1 collection[89] onto the map. This revealed clusters of genes at several dozen subregions (Figure S4A, left panel). To quantify and further aid the visualization of subregion that contained related genes, we used a heat diffusion strategy based on a Weibull decay function, which projected a binary signal around selected genes (read ''hot spots'' in high density areas). Watershed segmentation[90] was used to map ridges and valleys in the landscape topography, and we refer to the major visible areas as ''summits''. We numbered summits, starting with the summit with the highest heat density. A gene set enrichment analysis was used to identify pathways with overrepresented gene membership in summits, and a thematic name was assigned to each summit (Tables S6 and S7). The COSMIC summits represented many major cancer-associated pathways, of which the top 30 were annotated (Figure S4B, left panel) to allow assessing the pathway implications of the feature lists selected by the ML methods. For example, for the TCGA-subtype classification analysis, 533 core features were plotted on the pathway map (Figure S4A, center panel) and subsequently used for heat diffusion analysis to reveal TCGA-subtype specific summits (Figure S4B, right panel, indicates the top 30 densest TCGA subtype summits).

Assessing pathway distance

Features that were selected by one classifier method were compared to features from another method by measuring the distance between genes represented in their feature sets within the pathway map. If the genes of a method were close in pathway space to genes of another method, this would support the idea that a method chose features that were equivalently predictive to those chosen by another method, and that shared underlying biology gives rise to the predictive equivalence of features. The genes for a particular method were collected from its predictive features. Some of these features were distinct, while others were shared between methods. We calculated the *shortest-path distance to the 1st nearest neighbors* (DNN1) between gene lists (inter-method distances) (Figure 6C). In order to avoid including 0-length pathway distances for genes shared between methods, we also calculated the *shortest-path distance to the 2nd nearest neighbors* (DNN2) (Figure S4C). The distances of nearest neighbors were compared to random controls in which the same number of genes were randomly selected for a method.

Summit enrichment analysis

To compare the important features selected by the five ML methods, we implemented a gene set enrichment visualization that annotates sparse classifier feature lists with gene sets. For each set of genes in a summit, we assigned to each gene a scaled |SNR| score, as follows. Let A and B be two disjoint sample sets in a cohort C, such that A represents a given subtype and B all the other subtypes. Let $\bar{x}_1, s_1$ and $\bar{x}_2, s_2$ denote the mean and standard deviation of the expression levels of a gene g i n sets A and B, respectively. The SNR of gene g is given by $p_g = (\bar{x}_1 - \bar{x}_2)/(s_1 + s_2)$, and is represented by a vector $v_g = (p_1, p_2, \ldots, p_n)$, where p_i denotes the SNR in ith subtype. A scaled |SNR| was calculated by $u_g = \Sigma|v_g|$, and then scaled by $\hat{u_g} = u_g/max(u)$. Sorting a summit's genes by descending scaled |SNR| gave each summit a sail-like profile. Given a sparse list of classifier features, we assessed the enrichment of the features in a gene set using a Kolmogorov-Smirnov (KS) test, which indicated whether the distribution of these features was aligned with the sorted distribution of the genes in that summit.

Hallmark enrichment analysis

We used the same gene set enrichment visualization approach as was described in 'summit enrichment analysis' (above) to assess whether MSigDB Hallmark gene sets[91] were enriched with classifier features. For this analysis, we noted that some machine-learning methods returned feature sets that were very sparse in the Hallmark gene set collection. Our first goal was to associate as many classifier features as practical with Hallmarks. Our second goal was to ensure that the gene set represented the "feature space" assessed by the machine-learning methods. To address both issues, we used nearest neighbor classification (NNC) to assign to gene sets the classifier features that had not been annotated in Hallmarks. The NNC classifier was trained on the expression data of the genes annotated in Hallmarks, with genes representing data points and gene sets representing class labels. The NNC classifier then assigned a class label to each classifier feature not annotated in Hallmarks, using the Euclidean distance to the nearest neighbor points in the training data. If there was more than one nearest neighbor, a majority vote was used; ties were broken at random.

Subtype level gene set enrichment analysis

We assessed the enrichment of gene sets in a cohort's subtype using the Gene Set Enrichment Analysis (GSEA).[92] Briefly, GSEA assigns a normalized enrichment score (NES) to the distribution of a gene set across a rank-ordered list of genes (called a 'phenotype'). For each subtype, the phenotype was defined by ranking genes by a signal-to-noise ratio (SNR) metric, comparing samples in that subtype to all other samples in the cohort. We assessed the enrichment of the Hallmark gene set collection,[91] assigning to each Hallmark a NES score and a p-value. Because each cohort has multiple subtypes, more than one NES score and p-value were assigned to a given gene set in a cohort. In order to identify Hallmark gene sets that were highly ranked in a cohort, we aggregated the NES scores and p-values for each Hallmark in a cohort as follows. We aggregated the NES scores by the sum of the |NES| scores. The p-values were aggregated by the harmonic mean p-value (HMP)[93] and then adjusted by the Benjamini-Hochberg (BH) method from the *p.adjust*() function in *R* 4.1.0.

Meta-feature analysis

We collected 55 different *meta-features* (Table S8) that described aspects of the data and subtyping tasks for each of the 26 tumor types that either were supplied as *inputs* or were identified during *training* by the classifiers. These included input-related meta-features such as the number of samples within each cancer cohort, the size of the rarest subtype and the silhouette scores for each data platform. They also included meta-features identified during training, such as the number of features selected for classification by the best method and percent of DNA methylation features among selected features of the top model. We compared all the meta-features to each other as well as to the classifier performance in subtyping (overall weighted F1 score), and clustered the meta-features based on the similarities across the 26 tumor types.

QUANTIFICATION AND STATISTICAL ANALYSIS

The enrichment analysis was executed in *R* 4.1.0 and *RStudio* 1.4.1106. We used nearest neighbor classification (NNC) from the R *class* package version 7.3–19[94] for associating Hallmark features and ensuring that the gene sets were in the appropriate feature space. For GSEA we used the *fgsea* R/*Bioconductor* package version 1.18.0.[95] The PathwaySpace software to produce the "landscape summits" can be found at https://github.com/sysbiolab/PathwaySpace. Power analysis and performance scoring was performed with Python3 and Scikit-learn.

Please cite this article in press as: Ellrott et al., Classification of non-TCGA cancer samples to TCGA molecular subtypes using compact feature sets, Cancer Cell (2024), https://doi.org/10.1016/j.ccell.2024.12.002

CellPress
OPEN ACCESS

ADDITIONAL RESOURCES

Containerized pre-trained machine learning models: https://www.synapse.org/#!Synapse:syn29568296.

Containerized pre-trained machine-learning model workflows, and tutorials on how to apply these models to a user's own datasets: https://github.com/NCICCGPO/gdan-tmp-models.

National Cancer Institute Genomic Data Commons Publication Page that contains intermediate analysis results and underlying data shown in figures: https://gdc.cancer.gov/about-data/publications/CCG-TMP-2022.