Dohmen *et al. Genome Biology* (2022) 23:123
https://doi.org/10.1186/s13059-022-02683-1

Genome Biology

METHOD

Open Access

Identifying tumor cells at the single-cell level using machine learning

Jan Dohmen[1], Artem Baranovskii[2,3], Jonathan Ronen[1], Bora Uyar[1], Vedran Franke[1*] and Altuna Akalin[1*]

*Correspondence: vedran.franke@mdc-berlin.de; altuna.akalin@mdc-berlin.de

[1] Bioinformatics and Omics Data Science Platform, Berlin Institute For Medical Systems Biology, Max Delbrück Center for Molecular Medicine in the Helmholtz Association (MDC), Hannoversche Str.28, 10115, Berlin, Germany
Full list of author information is available at the end of the article

Abstract

Tumors are complex tissues of cancerous cells surrounded by a heterogeneous cellular microenvironment with which they interact. Single-cell sequencing enables molecular characterization of single cells within the tumor. However, cell annotation—the assignment of cell type or cell state to each sequenced cell—is a challenge, especially identifying tumor cells within single-cell or spatial sequencing experiments. Here, we propose ikarus, a machine learning pipeline aimed at distinguishing tumor cells from normal cells at the single-cell level. We test ikarus on multiple single-cell datasets, showing that it achieves high sensitivity and specificity in multiple experimental contexts.

Keywords: Single-cell genomics, Machine learning, Cell type classification, Cancer

Background

Cancer is a disease that stems from the disruption of cellular state. Through genetic perturbations, tumor cells attain cellular states that give them proliferative advantage over the surrounding normal tissue [1]. The inherent variability of this process has hampered efforts to find highly effective common therapies, thereby ushering the need for precision medicine [2]. The scale of single-cell experiments is poised to revolutionize personalized medicine by effective characterization of the complete heterogeneity within a tumor for each individual patient [3, 4].

Recent expansion of single-cell sequencing technologies has exponentially increased the scale of knowledge attainable through a single biological experiment [5]. The information contained within a single high-throughput single-cell experiment enables not only characterization of variable stable states (i.e., cell types, and cell states) but also functional annotation of individual cells, such as prediction of the differentiation potential, susceptibility to perturbations, and inference of cell–cell interactions [6].

As with all new technologies, high-throughput single-cell sequencing also created new computational challenges [7]. A problem in single-cell data analysis is cell annotation—assignment of a particular cell type or a cell state to each sequenced cell. The size of

the generated datasets made manual annotation approaches utterly unfeasible, while the peculiarities of data generation prompted the development of novel innovative classification methods [8–13]. This is especially apparent in datasets stemming from cancer tissues, where the variability in the transcriptomic states does not conform to classically defined cell types. An outstanding question is whether there exist transcriptomic commonalities between cancer cells originating from different cancers, and whether it is possible to create a model which would discriminate between cancer cells and the surrounding tissue across different cancer types, and datasets.

High-throughput single-cell technologies provide unprecedented precision of characterization of biological systems, with all the technical and biological influences being evident in the data. In cancer biology, this heterogeneity of data composition presents a particular problem, because it is very hard to enumerate, and correct for, all of the technical and biological variables which are giving rise to the measured variability [14]. For example, cell dissociation produces artifacts which mimic MAP kinase pathway activation [15], while it is impossible to know the exact environment influencing each cell. Cells might be in gradients of oxygen availability, under different physical constraints, or influenced by multiple varying signaling molecules. This variability presents a challenge for developing machine learning models, because the data coming from different conditions will have different underlying distributions, meaning that methods trained on one dataset will not generalize to other datasets [16].

Currently, there are three types of methods for mitigating distributional differences between single-cell RNA sequencing datasets: (1) manifold matching methods that try to find commonalities between low dimensional representations of multiple datasets and align them into one space [17]; (2) domain adaptation deep learning tools that try to model (explicitly or implicitly) the batch effects through the latent space embeddings [18–24]; and gene set based classifiers that use learned marker genes and robust statistics to transfer knowledge between datasets [8].

An issue recurrently arising in machine learning in biology is how to determine the generalization boundaries of trained models (i.e., on which datasets the model will fail). It is not evident whether the learned model will perform equally well on the data profiled using different sequencing technologies (i.e., Drop-Seq, 10X, CEL-seq or Fluidigm C1), produced by different laboratories, or originating from a different biological source (i.e., same cell types, coming from different human individuals). Because the sources of the heterogeneity are frequently unknown, the models need to be explicitly tested for robustness on datasets corresponding to different biological conditions and profiled using different technologies [25]. Therefore, special care needs to be taken that the learned associations really are between variables of interest and are not confounded by the properties of the data. Both manifold matching and domain adaptation methods follow a tradeoff between the removal of unwanted variance, while preserving biological heterogeneity [25]. Gene signature based methods lie on the opposite part of the tradeoff spectrum, whereby the gene lists represent a strong inductive bias about a biological property (cell type). If the gene lists are carefully tested, then the methods achieve a markedly low false positive rate.

We set out to answer a simple question: "Is it possible to make a classifier that would correctly differentiate tumor cells from normal cells in multiple cancer types?". We have

built ikarus, a stepwise machine learning pipeline for tumor cell classification. ikarus consists of two steps: (1) discovery of a comprehensive tumor cell signature in the form of a gene set by consolidation of multiple expertly annotated single-cell datasets and (2) training of a robust logistic regression classifier for stringent discrimination of tumor and normal cells followed by a network-based propagation of cell labels using a custom built cell–cell network [26]. With the goal of developing a robust, sensitive, and reproducible in silico tumor cell sorter, we have tested ikarus on multiple single-cell datasets of various cancer types, obtained using different sequencing technologies, to ascertain that it achieves high sensitivity and specificity in multiple experimental contexts. We have strictly adhered to machine learning best practices to avoid contamination of results by information leakage from training into testing process.

Results

Identification of a robust marker gene set

Cell type annotations in any particular experiment are inherently noisy. This is partly due to the properties of single-cell data, such as the different number of detected genes in each cell, the influence of sample processing, and our limited knowledge of biomarkers that are necessary for comprehensive annotation of cell types and cell states. We hypothesized that we can find robust markers of cellular states by comparing multiple independent annotated datasets from diverse origins.

We have employed a two-step procedure to find tumor-specific gene markers. First, using differential expression analysis, we selected genes that are either enriched or depleted in cancer cells per dataset (see 9). To obtain the final gene list, we took an intersection of the gene sets from each of the datasets (Fig. 1A). We have applied a standard cross validation approach for gene set selection, where datasets were either used as training, validation, and test sets. For cross validation, we have used the two lung cancer datasets from Laughney [27] and Lambrechts [28], a colorectal cancer from [29], neuroblastoma dataset from Kildisiute [30], and a head and neck cancer datasets from [31]. For each pair of datasets, we have selected the gene signature and trained the logistic classifier. The resulting classifier accuracy was validated on the datasets that were not used for training (Additional File 3: Cross validation results). As the performance metric, we have chosen a minimal balanced accuracy on the validation set (measuring the worst performance of the classifier on the validation set). The cross validation procedure showed that the gene signature selection using multiple datasets increases the generalization performance of the classifier (Fig. 1C). The best performing classifier combined the colorectal cancer dataset from Lee et al. [29] with the lung cancer from Laughney et al. [27], and achieved a minimal balanced accuracy of 0.97 on the validation data. The performance of the best performing gene set was tested on the hepatocellular carcinoma [32] (balanced accuracy of 0.93), and the lung carcinoid dataset [33] (balanced accuracy of 0.99).

The resulting tumor gene signature contained 162 genes that were significantly enriched in cancer cells across multiple datasets (Additional File 2: Gene Signatures). The resulting set of genes showed high specificity for cancer cells, from the head and neck cancer samples [31] (Additional File 1: Fig. S1A). This result indicates that the gene

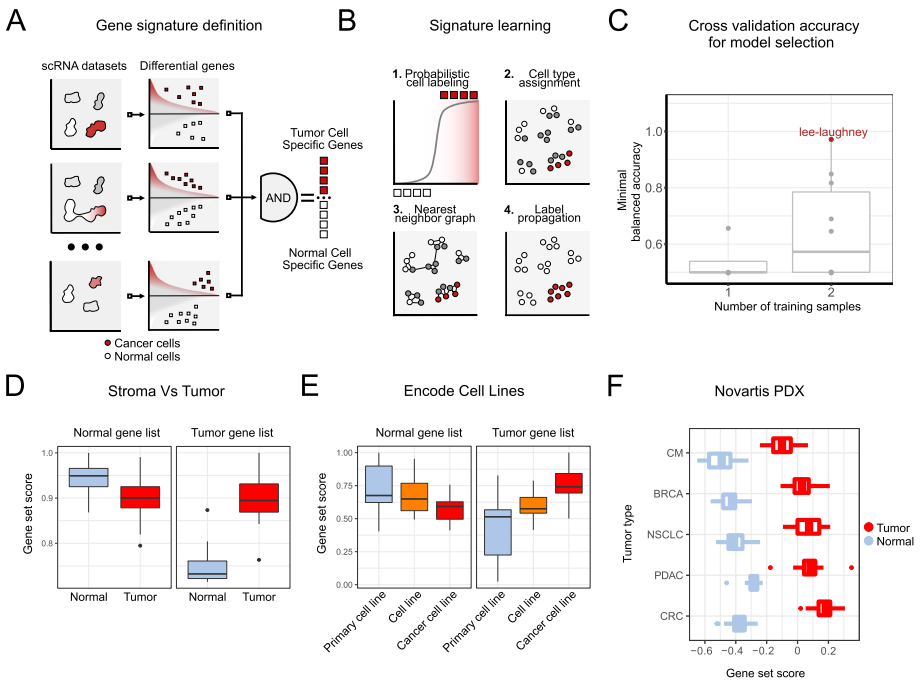

Fig. 1 Integration of multiple datasets enables robust extraction of informative gene sets. **A**, **B** ikarus workflow. ikarus is a two-step procedure for classifying cells. In the first step, integration of multiple expert labeled datasets enables the extraction of robust gene markers. The gene markers are then used in a composite classifier consisting of logistic regression and network propagation. **C** Comparison of cross validation accuracy for signature derivation and model selection. Minimal balanced accuracy on the validation set was chosen as the metric of choice (i.e., worse performance on the test set). Models trained on just one dataset achieved lower balanced accuracy than models trained on two datasets (p value given by the two sided Wilcoxon test is 0.063). The combination of colorectal cancer from Lee et al. [29] and lung cancer from Laughney et al. [27] achieved the highest minimal balanced accuracy of 0.97. **D** Comparison of gene signature scores in laser microdissected gastric cancer data [34]. The normal gene list shows lower signature scores in cancer samples (p value 0.052, $N=8$, Mood's median test), when compared to the cancer-associated normal tissue. The tumor gene signature is significantly higher for cancer samples than the normal tissue (p value 0.003, $N=8$, Mood's median test). **E** Primary cells and cancer cell lines have significantly different gene signature distributions. The normal-cell gene signature shows a gradual reduction in gene signature score distribution when compared in primary cells, cell lines, and tumor cell lines. The gene signature shows the complete opposite effect. Cancer cell lines have the higher gene signature score distribution, followed by cell lines, and primary cells. Distributions were compared using pairwise Wilcoxon tests with BH-FDR correction. All adjusted p values were lower than 0.01. **F** Patient-derived xenografts (PDX) show significantly higher tumor gene signature score, than the normal gene signature score. The same pattern is observed in multiple cancer types. Normal and tumor signature distributions were compared using Wilcoxon tests, for each cancer type, followed by BH-FDR correction. All adjusted p values were lower than 0.01

set contains information relevant for discriminating tumor cells from non-tumor cells in multiple different tumor types.

The same procedure was applied to the healthy cell types. We extracted genes enriched in each cell type, when compared to the tumor cells. The resulting gene set was then merged between multiple datasets. This "normal" cell gene signature contains both cell type specific markers and genes which are specifically depleted in the tumor cells (Additional File 1: Fig. S1B).

To validate the specificity of the novel tumor specific gene set, we have analyzed a gastric cancer dataset [34], where multiple areas of cancer and cancer-associated normal

cells were separated using laser-capture microdissection (LCM) and profiled using RNA sequencing. Using normal and tumor gene signatures that were identified by ikarus, we have scored the tumor and the associated normal cells. As expected, dissected sections coming from the cancerous lesions had significantly higher median tumor score than the surrounding normal tissue (Fig. 1D, right panel). In line with the latter, normal tissue scored higher than cancerous lesions when the normal gene signature was scored (Fig. 1D, left panel).

As another line of evidence, we have downloaded the expression data for primary, normal, and cancer cell lines from the ENCODE database [35, 36] (see 9). Tumor signature scores were on average highest in cancer cell lines, diminished in normal stable cell lines, reaching its lowest average in primary cells (Fig. 1E, left panel). When scoring using the normal (non-cancer) cell signature, an opposite trend was observed, i.e., score was highest in primary cells, intermediate in normal stable cells and the lowest in cancer cell lines (Fig. 1E, right panel).

Further, we tested the discriminatory power of the normal and tumor gene lists in multiple cancer types. To this end, we have used the patient-derived xenograft (PDX) samples from five cancer types provided by [37] and all of the cancer cell lines provided by the cancer cell line encyclopedia (CCLE) [38]. The tumor signature score was significantly higher than the normal signature score in all PDX cancer types (Fig. 1F) and all cancer cell lines screened in CCLE (Additional File 1: Fig. S1C). Surprisingly, the tumor signature list produced significantly reduced scores for cell lines stemming from blood-related cancers (LAML, CLL, LCML, MM, DLBC).

Accurate delineation of cancer cells

In the first step of classification, ikarus derives the tumor and normal gene set scores. The tumor and normal gene set scores are then used in a logistic regression classifier, to delineate cells with high probability of being tumorous or normal. The classification step is followed by the propagation of the cancer/normal label through a custom based cell–cell network (Fig. 1B). The cell–cell network is derived from the same gene sets that are used for robust scoring. By using only tumor or cell type specific genes, the resulting network separates communities that represent either tumor or normal cell states.

Figure 2A shows the performance of ikarus classification on all of the validation and test datasets. ikarus achieves an average balanced accuracy of 0.98 which is substantially higher than other classical machine learning methods. In addition to the standard machine learning methods (SVM, random forest, and logistic regression), we have compared ikarus to the top ranking tailored cell type classifiers, as evaluated in the recent comparison of methods for cell classification [8]: SingleCellNet [9], ACTINN [39], ItClust [10], scCaps [40], scHPL [12], CellAssign [41] from scvi-tools [42], and scMRMA [43]. We would like to emphasize that for the published methods, we have used the default hyperparameter settings from the corresponding descriptions or provided tutorials.

We have chosen balanced accuracy as a measure of performance because of the large imbalance of classes. The datasets contained, on average, 7 times more normal cells than annotated cancer cells (Additional File 2: Datasets). To give an unbiased view on the performance, Fig. 2B shows the area under the receiver operating characteristic (AUROC) distribution for the different datasets. ikarus also achieves a higher average AUROC

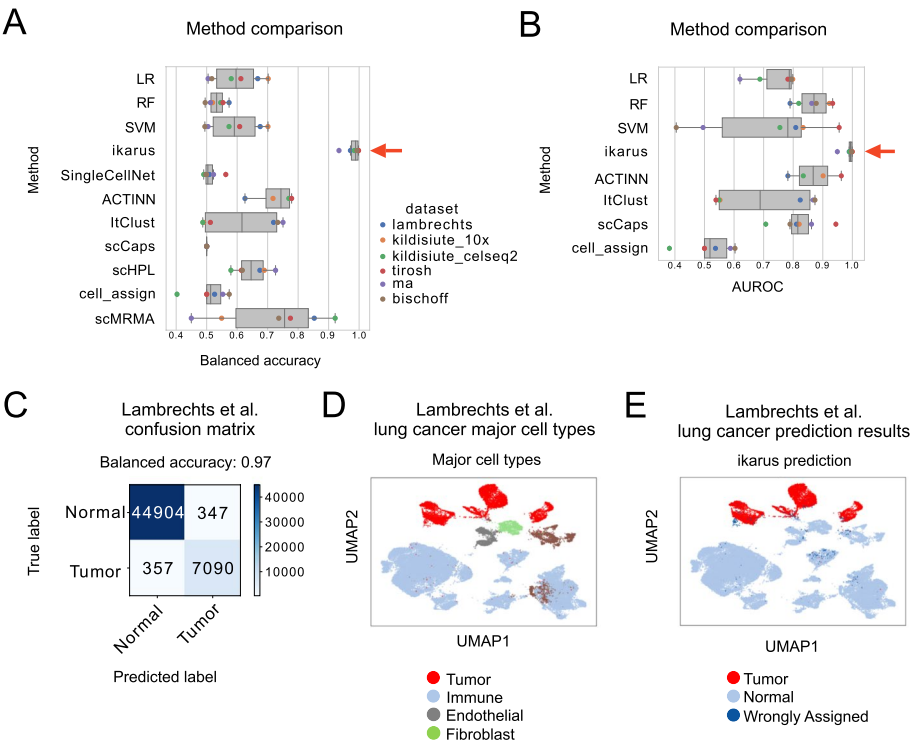

Fig. 2 ikarus accurately classifies tumor and normal cells. **A** Balanced accuracy for classification of tumor and normal cells, for each of the test datasets. Red arrow highlights performance of ikarus classifier. ikarus achieved a significantly higher performance than competing methods (AUC distributions were compared using ANOVA with post hoc pairwise comparison. *p* values were adjusted using BH-FDR. All adjusted *p* values were lower than 0.01). **B** Area under receiver operating characteristic (AUROC) performance for each classifier. AUROC could only be calculated for methods outputting class probabilities. **C** Confusion matrix showing the performance of ikarus classifier on the Lambrechts lung cancer dataset. **D** Cell type annotation of the Lambrechts et al. dataset. **E** Lambrechts dataset labeled by ikarus classifier

than other methods. Having a high AUROC value and low balanced accuracy is an indication of class imbalance. The classification error of the classical machine learning methods, having high AUROC and low balanced accuracy, is not uniformly distributed—they struggle with a high false positive rate.

We have additionally compared all methods with datasets subsampled to include an equal number of normal and tumor cells. ikarus showed a higher median balanced accuracy in discriminating tumor cells from normal cells (Additional File 1: Fig. S2A). During the comparison of subsampled datasets, we have noticed an increase in variance of ikarus results. This is because the subsampling reduces the connectivity of the cell–cell network which is used for network propagation.

We have also tested the performance of different classification methods by scaling down the input genes, from all profiled genes, to the tumor and normal gene signatures. The reduction of input to only normal and tumor gene signatures surprisingly increased the performance of all classifiers, indicating that the signatures contain information for proper discrimination between tumor and normal cells (Additional File 1: Fig. S2B).

Figure 2C shows the classification accuracy for the Lambrechts lung cancer dataset [28]. The Lambrechts dataset was not used for training nor gene signature definition. Figure 2D and E show the classification accuracy overlaid on UMAP [44] embeddings

of the Lambrechts lung cancer dataset [28]. ikarus correctly classifies normal cells, irrespective of the underlying cell types. The erroneous classifications are equally distributed between false positives and false negatives. (UMAPs for other validation and test datasets are reported in Additional File 1: Fig. S3 A-E).

In order to test the robustness of ikarus across different single-cell sequencing technologies, we applied ikarus on a dataset of neuroblastoma tumors sequenced by either 10X genomics or CEL-Seq2 protocols by Kildisiute et al. [30]. ikarus achieved a high classification accuracy (balanced accuracy of 0.98) on all datasets, irrespective of the profiling technology (Figures S3B and S3C). The false positive rate we observed in the test datasets (1–3%) can be partly attributable to occasional erroneous labeling of cells by the authors of the corresponding studies. The lack of a perfectly labeled single-cell tumor sequencing dataset makes it difficult to quantify the actual rate of false positive predictions by our method. One possible strategy to remedy this issue is to test our method on a dataset that is presumably free of tumor cells, in other words, a healthy tissue sample. To ascertain the actual false positive rate for tumor cell classification, we have tested ikarus on the single-cell data from peripheral blood of a healthy individual [45], where all cells are expected to be non-tumorous. ikarus labeled all cells as non-tumorous (Additional File 1: Fig. S3F).

Because the datasets used for training and testing consist predominantly of carcinomas, we decided to test ikarus performance on a synovial sarcoma sample [46]. On sarcoma samples, ikarus achieved a reduced balanced accuracy of 0.51, which was primarily driven by a high false negative rate—ikarus missed sarcoma tumor cells (Additional File 1: Fig. S3G).

Further, we were interested in how the accuracy of the classification changes in regards to the size and structure of the gene set. First, we have conducted an ablation study, where we removed from 20 to 80% of randomly selected genes from the list (Fig. 3A). The removal of up to 40% of the genes from the list leads to a ~12% (from 99 to 87%) drop in median accuracy. If 80% of the gene list is removed, the classification becomes random (median accuracy tends to ~50%).

Next, we explored how much the accuracy of the classification depends on individual genes. To test this, we sequentially removed each individual gene from the set and repeated the classification. For 160 out of the 162 genes, there was no observable change in the classification accuracy on the test datasets (Fig. 3B). The accuracy on the Lambrechts lung cancer [28] dataset was, however, particularly sensitive to the omittance of two genes: serum amyloid A (SAA1) and fibrinogen beta chain (FGB) (Fig. 3C). Each gene is a marker for a tumor specific cell cluster in the Lambrechts dataset (Fig. 3C, D), and their removal influences the classification probability of cells constituting that particular cluster. Such dependence was not observed for other test datasets (Additional File 2: Effects of SAA1 and FGB).

Properties of the tumor gene signature

Having observed the high accuracy performance of ikarus based on the detected tumor gene signature, we ventured forth to obtain a deeper characterization of the functional

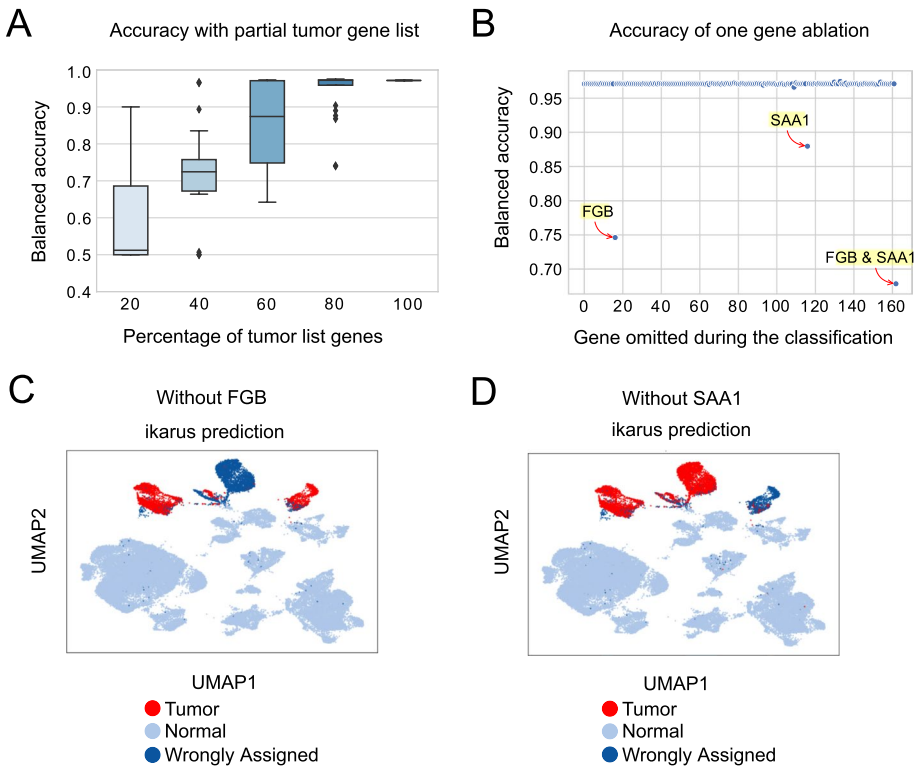

Fig. 3 ikarus performance under perturbation. **A** Performance of ikarus classifier, on the Lambrechts' validation set, when 20 to 80% of the input list is removed. ikarus performance is significantly reduced when the 60% of the gene list is removed (p value < 0.01, as determined by a one sample t-test, with BH-FDR correction). **B** Performance of ikarus classifier, on the Lambrechts' validation set, upon single gene removal from the gene list. One hundred sixty out of the 162 genes were inconsequential for the classification. FGB (fibrinogen) and SAA1 (serum amyloid alpha) ablation significantly reduced the classification accuracy, but only on the Lambrechts et al. lung cancer dataset. **C**, **D** FGB and SAA1 are respectively strong markers for one of the cell state clusters in the lung cancer dataset

content of these genes. Specifically, their involvement in the development of cancer and their roles in the prognosis for the patients.

Firstly, we were interested whether the genes within the tumor gene signature conform into expression modules, or whether their expression distribution is independent. We calculated the pairwise Pearson correlation between the genes from the signature for all datasets. To our surprise, the correlation between the genes was largely zero (Fig. 4A). We found only a single module (containing 34 genes) that was robustly present in all datasets (Additional File 1: Fig. S4A). Genes in this module are annotated as belonging to the cell cycle. We further inspected whether the classification accuracy depends on these cell cycle-related genes. The removal of the 34 cell cycle-related genes did not affect the classification accuracy (results not shown).

The tumor gene signature had, to our surprise, little overlap with established cancer-related gene sets. When compared with the gene sets annotated in the CancerSEA database of cancer functional states [47], our tumor gene list had zero or very few overlaps with most CancerSEA gene sets, except for the cell cycle genes, which shared only 9 genes with our tumor gene list (Fig. 4B). Co-expression analysis, using SEEK [48], again showed that the tumor gene signature is partially related to cell cycle and

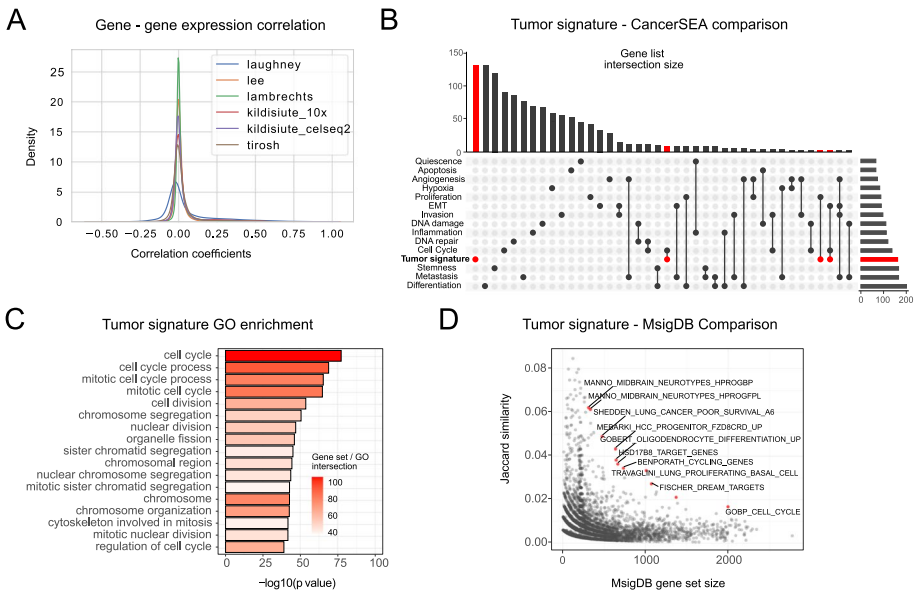

Fig. 4 Properties of the tumor gene signature. **A** Tumor gene signature co-expression analysis in tumor cells. Co-expression is measured as Pearson correlation between all pairs of genes from the gene list. **B** Tumor gene signature shows little overlap with known cancer-associated gene sets. The tumor gene signature was compared to the CancerSEA database. Out of the 162 genes, 134 showed no overlap with any of the gene lists. 9 genes overlapped with the cell cycle gene list. Only intersections of size 3 and more are shown. **C** Results of the gene co-expression analysis using SEEK. Top 150 co-expressed genes identified by SEEK were used as an input for GO enrichment analysis. **D** Tumor gene signature shows limited overlap with most of the gene sets from MsigDB. The most enriched gene set corresponds to the cell cycle

DNA replication (Fig. 4C). In addition, we saw no overlap with the cancer hallmarks from MSIGDB [49] (Additional File 1: Fig. S4B). When compared to the complete MSIGDB database, the tumor gene signature preferentially overlapped with the cell cycle hallmark (Fig. 4D). Gene ontology (GO) analysis using gprofiler2 [50] showed an enrichment of terms exclusively related to cell cycle and mitosis (Additional File 1: Fig. S4C). We have tested the GO and SEEK enrichment after the removal of the cell cycle module. The analysis did not result in any statistically enriched terms.

The enrichment of cell cycle and DNA replication-related functional terms in our tumor gene set (Additional File 2: Enrichment analysis of tumor gene signature) led us to hypothesize that the novel gene set differentiates promptly cycling cells. To test this hypothesis, we inspected the correlation of the tumor gene set scores with the growth rates detected in Patient Derived Xenograft (PDX) samples from [37] and the doubling times of the cancer cell lines from CCLE [38]. Unexpectedly, there was no correlation between the tumor signature score and the PDX growth rates in any of the reported cancer types (Additional File 1: Fig. S4C). Repetition of the analysis on the cell line doubling times from CCLE again revealed the same lack of correlation (Additional File 1: Fig. S4C).

We were interested in whether the tumor cell signature is predictive of survival outcomes in cancer. From the protein atlas database (http://www.proteinatlas.org) [51], we extracted genes predictive of survival in one or more cancers. The overlap of tumor gene signature with the extracted gene set showed that more than 75% of

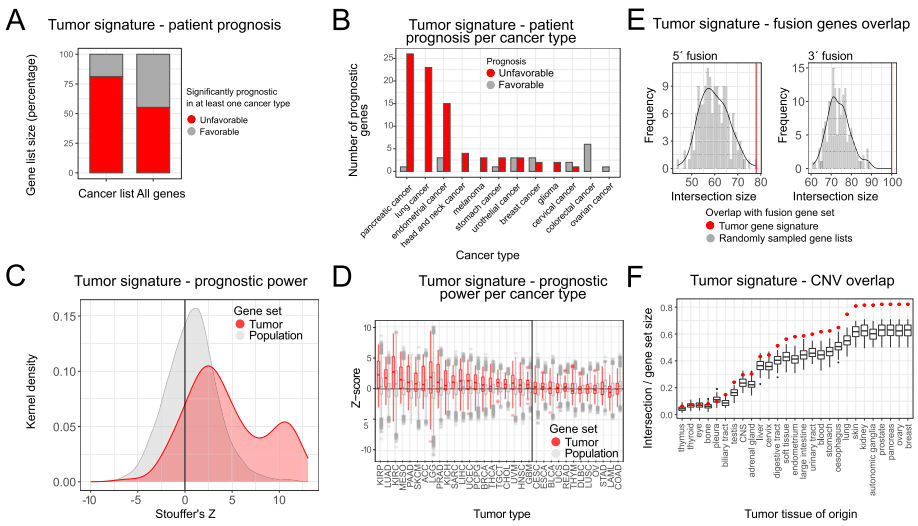

Fig. 5 Tumor gene signature is predictive of survival and associated with copy number variations. **A** Tumor gene signature genes are more likely to be significantly prognostic for patient survival in at least one cancer type (enrichment *p* value was lower than 0.01, as determined by a chi-square test). Data was extracted from the protein atlas. **B** The association of tumor gene signature genes with poor survival outcomes is evident in multiple cancer types. *p* values for within cancer comparisons are reported in Additional File 3: Statistics. **C** Data from [52] show that the tumor gene signature genes have much higher Stoufer's *z* (association with poor survival outcomes) than rest of protein coding genes (*p* value was lower than 0.01, based on the Wilcoxon rank-sum test). **D** Increased association of the tumor signature genes with negative survival is apparent in 23 out of 33 cancer types (the BH-FDR adjusted *p* value was lower than 0.01 as determined by Wilcoxon rank-sum test). **E** Genes from the tumor gene signature are more likely to participate in 3′ or 5′ fusions. When compared to sets of randomly drawn genes, the tumor signature genes had a significantly higher probability of participating in genomic fusions. The red vertical line depicts the overlap of the tumor signature list with the corresponding gene fusion list (the BH-FDR adjusted *p* value was lower than 0.01, as determined by the resampling test). **F** Genes from the tumor gene signature are frequently found in amplified or deleted genomic regions. We have measured the percentage of the gene list which overlaps with the known CNV regions for each cell line in the CCLE dataset. The tumor gene signature was found to be enriched in 22/27 cancer types (BH-FDR adjusted empirical *p* value was lower than 0.01 based on resampling tests). Background distributions were derived from expression matched randomly sampled gene lists

tumor signature genes are predictive of unfavorable prognosis in at least one cancer type (Fig. 5A). Interestingly, when stratified by cancer type, tumor signature genes reported to be unfavorable are overrepresented among 5 cancer types: liver, renal, pancreatic, lung, and endometrial cancers (Fig. 5B). An analogous analysis was done with data taken from [52], where the authors systematically calculated the risk predictive status for all genes in TCGA cancer types. The analysis showed that the cancer specific genes have a significantly higher Stoufer's *Z* value (a measure of how significantly the gene expression predicts the risk status in any cancer) than the rest of the annotated gene set (Fig. 5C). Furthermore, the same trend was observed in 21 out of 33 profiled cancer types (Welch two sample *t*-test, Bonferroni adjusted p value < 0.05) (Fig. 5D).

Next, we wanted to explore how often the genes from the tumor gene list participate in genomic rearrangements, particularly gene fusions, which are frequent drivers of oncogenic events in multiple cancer types. We downloaded the known cancer gene fusions from the ChiTaRS [53] database and inspected the overlap with the novel cancer defining gene list. To establish enrichment, we compared the overlap with a background

consisting of random gene sets. Genes from the tumor gene signature have a significantly higher probability of participating in both 3′ and 5′ fusions, than a random set of genes (Fig. 5E).

Gain and loss of DNA content is a ubiquitous property of tumor cells. Copy number variation (CNV) profiles that arise from genomic rearrangements create unique genomic signatures that can be used for characterization and discrimination of different tumor types [54]. We wondered how prevalent genes from the tumor gene signature are in the known CNV regions. To this end, we have compared the intersection of the tumor signature list with the CNV data from CCLE. We compared the tumor gene list intersection to a background distribution constructed by randomly sampling expression matched gene sets. The tumor gene list had a significantly higher overlap with the known CNV regions in the majority of profiled cancer types (Fig. 5F) irrespective of CNV frequency.

Multi-omics analysis reduces the false positive rate of classification

Characterization of biological systems from multiple viewpoints often produces synergistic insights into the underlying biology. We wondered whether the classification accuracy of ikarus algorithm can be improved by using multi-omics measurements. To this end, we have used inferCNV [55] to extract copy number variations (CNVs) from the single-cell RNA sequencing data. InferCNV is a Bayesian method, which agglomerates the expression signal of genomically adjointed genes to ascertain whether there is a gain or loss of a certain larger genomic segment. We have used inferCNV to call copy number variations in all samples used in the manuscript.

Firstly, we wondered whether the copy number variations could be used as universal markers for discriminating between normal and cancer cells. We trained random forest classifiers to discriminate between the expert labeled normal and tumor cells. One classifier was trained on each sample. Each of the classifiers was tested on all samples. As expected, when evaluated on the sample which was used for the training, each random forest classifier correctly discriminated between the cancer and tumor cells (Fig. 6A). The classifiers, however, did not generalize to other cancer types—they all suffered from a high false positive rate. We tried to improve the generalization of the classification by training on multiple datasets. We trained a random forest classifier on joint Lee et al. and Laughney et al. data and tested on all other datasets. Using multiple datasets for training did not improve the results of the classification on out of sample cells (Fig. 6B).

We then wondered whether the CNV calls could be used in conjunction with the gene expression data to improve ikarus classification of tumor and normal cells. We looked at the average CNV value and the variance of CNV values in cells, which were misclassified by ikarus in data from Lee et al. and Laughney et al. Both the average CNV value and the variance of CNVs were significantly higher in cancer cells, which were misclassified as normal cells (Fig. 6C). This indicated that by integrating the CNV scores with the gene expression classifier, we might increase the classification accuracy.

We have added an additional proofreading step into the classification procedure. We trained a logistic classifier on inferred CNVs, with ikarus predicted cell type labels as the dependent variable. Cells which obtained highly probable discordant class labels from the CNV classifier had their labels flipped. Using the proofreading step, the average balanced accuracy stayed the same for all of the samples. We have however noticed a

sudden drop in the false positive rate, with a marginal increase in the false negative rate (Fig. 6D, Additional File 2: Results).

Discussion

We have implemented a two-step approach for solving a problem that is perceived as simple: discriminating tumor cells from normal cells. In the first step, ikarus pipeline integrates multiple expert labeled datasets to extract gene sets which discriminate tumor cells from normal cells. In the second step, ikarus uses a robust gene set scoring along with adaptive network propagation for cell classification. By using robust gene set scoring and network propagation, we have mitigated two common problems in single-cell analysis: the influence of batch effects on sample comparison and parameter optimization during clustering.

The effect of technical differences between single-cell datasets is usually resolved using integration methods. Single-cell integration methods require extensive tuning of sets of parameters, most of which have non-intuitive effects on the results. Moreover, the accuracy of the resulting integrations cannot be trivially evaluated without extensive usage of biological priors. Gene set scoring methods are robust, because they use "within sample" rank based scores instead of direct comparison of measured expression values between different samples. The only technical variable that influences gene set scoring is the percentage of genes from the gene list which are detected in each cell. We have however extensively tested the influence of the number of genes on the classification accuracy.

A common step in single-cell analysis is aggregation of cells into clusters, which are then used for cell type annotation. Clustering is, however, a procedure with an inherently high number of parametric options. It is extremely hard, if not impossible, to choose a set of parameters that would produce the same level of accuracy (same cell types) on different datasets, which often necessitates manual intervention to deduce the best clustering resolution. Because cell types and cell states form highly connected modules within the cell–cell similarity graph, we have therefore opted to replace clustering with network propagation. Network propagation is a procedure where the uncertainty of cell annotation can be reduced by integrating the annotation score of each individual cell with the scores of its nearest neighbors. Network propagation represents a parameterless alternative to clustering, while retaining the same level of sensitivity for cell annotation.

By exploring a multi-omics approach, we have tried to increase the accuracy of the normal–tumor cell discrimination. Using inferred CNVs, we have shown that the information from copy number variations does not generalize across different cancer types. The inclusion of the copy number variation as a proofreading step reduced the false positive rate of the classifiers. It is still an open question, though, by how much would single-cell multi-omic measurements improve the classification (for example, by concurrently measuring mutations, CNVs, chromatin accessibility, and expression in the same single cells). Currently, such methods are either in their infancy, and the required data is not available, or have a very limited profiling range (profile only a handful of loci).

ikarus is currently constrained by the reliance on well annotated single-cell datasets. For both the gene set definition and testing, we rely on expert provided cell annotation. This requirement has limited our training and testing capabilities to the handful of profiled, and annotated cancer types. We have determined that ikarus produces accurate

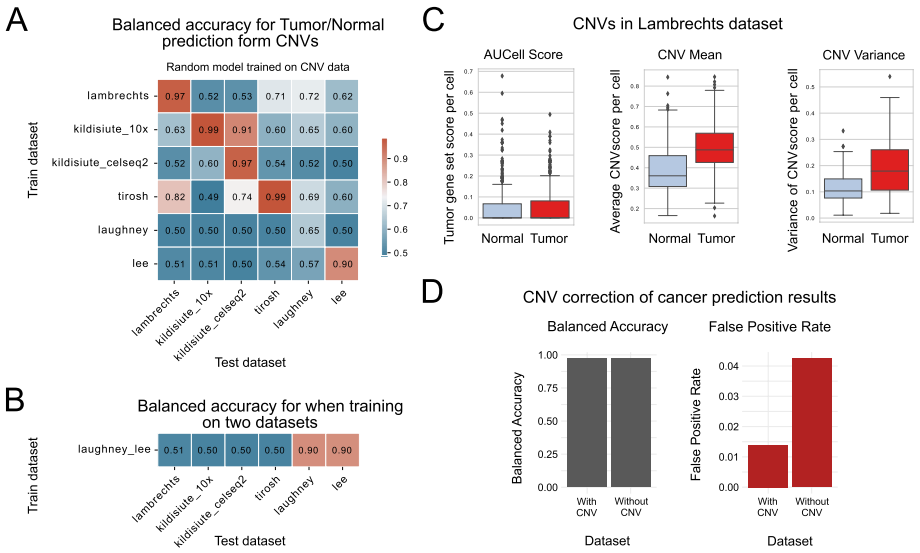

Fig. 6 Multi-omics decreases the false positive rate. **A** Balanced accuracy of random forest classifiers trained on each of the tumor types. The classifiers have excellent performance on the same samples they were trained on or on similar tumors (such as the Kildisiute et al. neuroblastoma sample), while they fail to generalize to other tumor types. **B** Training on multiple tumor types does not improve the generalization of the classifiers. **C** Cells misclassified by ikarus can be discriminated based on the average CNV and variance of CNV values. Tumor cells misclassified as normal cells have significantly higher values of both the average CNV and the variance of CNV, than the corresponding normal cells. **D** Integration of the CNV proofreading decreases the false positive rate from 4% to 1%, with the same average balanced accuracy

classifications in epithelial tumors, and the neuroblastoma; however, it showed reduced accuracy in classifying cells from synovial sarcoma. This implies that multiple trained models will be required for comprehensive discrimination of all cancer types. The exponentially increasing number of single-cell datasets will enable us to increase both the number of training datasets, as well as to test ikarus on currently unavailable tumor types, for instance, soft tissue sarcomas. Moreover, the increasing quality of single-cell datasets, most importantly, increasing gene coverage, will also increase the utility of ikarus as a gene set based classification method.

Conclusion

By integrating multiple datasets, we have derived a tumor signature gene list which is surprisingly refractory to annotation. The gene list contains a sub-module ($n = 34$) that encompasses genes involved in the cell cycle. All of the other genes, however, showed little modularity and a lack of enrichment in any single annotation category. Interestingly, the genes were highly expressed in all of the available PDX and CCLE cancer models. The ablation studies showed that the classifier was robust to the removal of any one of the genes. The low co-expression, combined with the lack of sensitivity to the gene removal indicates that the tumor signature genes provide mutually synergistic information towards the classification accuracy.

ikarus classifier, however, is not limited to tumor cell detection. It can be used to detect any cellular state, such as cell types. The only requirements are that the cellular state is present in at least two independent experiments, which are expertly annotated.

Automatic, parameterless discrimination between tumor cells and the surrounding tumor-associated tissue has multifactorial utility. Tumor cells can be streamlined into algorithms for neoepitope prediction, thereby enabling direct, clinically relevant insights. Furthermore, the increasing availability of multi-omics measurements would enable automatic genetic characterization of tumor subpopulations and the subpopulation-based recommendation of best therapeutic course of action. Application of automatic tumor classification on spatial sequencing datasets enables direct annotation of histological samples, thereby facilitating automated digital pathology.

The current scale of development in single-cell biology (on both the technological and computational levels) shows promise for quantitative characterization of the complete tumor heterogeneity, for each individual. However, before the personalized medicine approach can be readily adopted, every step in the data analysis needs to be completely automated, with robust performance guarantees. ikarus pipeline represents one step towards the implementation of personalized cancer therapy.

Methods

ikarus workflow

The presented ikarus pipeline consists of two major steps. In the first step, ikarus uses multiple expert labeled datasets to define gene signatures and builds a cell type specific classifier. In the second step, based on the constructed gene sets and classifier, unknown cells of interest are scored. The classifier's scores are then propagated through a custom cell-cell network, which eventually leads to the cell annotations. While the first step is optional, as users can provide their own gene lists, the latter steps are mandatory to make a prediction. For making predictions ikarus' API is modeled as the scikit-learn workflow, which means 1. load data, 2. initialize a model, 3. fit the model, and 4. make the actual predictions on unknown data. In general, annotated data objects are used as data format (AnnData, https://anndata.readthedocs.io). Each individual step is described in more detail in corresponding subsections below.

Defining gene signature lists

The count matrices of the input AnnData objects are expected to be normalized to the total number of reads per cell and log transformed with a base of 2 and a pseudocount of 1. In addition to that, each AnnData object must contain for each cell the corresponding cell type in the observation section, possibly in multiple columns for multiple hierarchical cell type levels (e.g., tumor and normal cells, or tumor, epithelial and immune cells).

It is important to take care that the input data is not scaled and that it contains the complete set of profiled genes and not a preselected set (such as highly variable genes).

Then, for each gene in the input dataset, a *t*-test with overestimated variance is used to compute an approximation of log 2-fold change between two cell type classes, one upregulated and one downregulated class. Those classes are provided by the user and should be chosen in accordance with the considered columns of the AnnData observation section. Users can either perform only one comparison (e.g., tumor vs. normal

cells) but also multiple (e.g., tumor vs. epithelial cells and tumor vs. immune cells). This is done independently for each dataset.

For each gene and for each pair-wise comparison, the associated log 2 fold changes (if $p_{adj} < 0.1$, neglected otherwise) of the different input datasets are averaged. According to these average values the genes are then sorted, highest to lowest and a user-defined number of top genes is selected (for our analyses, we used the top 300 genes). The final list of upregulated genes is derived by taking either the intersection or the union of selected genes across all of the comparisons.

The whole procedure is performed once for the case that the class of interest (e.g., tumor) is upregulated (here we take the intersection of selected genes across all of the comparisons) and once for the case that this class is downregulated (here we take the union of selected genes across all of the comparisons). That way, we obtain two final gene sets. One set representing genes enriched in the class of interest (i.e., enriched in tumor cells), and a set depleted in the class of interest (depleted in tumor cells).

A combination of Lee, Laughney, Lambrechts, Tirosh, and Kildisiute datasets was used to conduct a cross validation analysis. For each pair of datasets, gene signature selection was performed, followed by training of the logistic classifier. The resulting classifier accuracy was validated on the three datasets which were not used for training. The accuracy of the top performing classifiers was furthermore tested on the hepatocellular carcinoma (Ma) and carcinoid datasets (Bischoff). Cross validation results can be found in the Additional File 3: Cross Validation Results. As the performance metric, a minimal balanced accuracy on the validation set was chosen (i.e., what is the worst performance of the classifier on the validation set).

For comparison, classifiers were also trained on gene lists extracted from each of the datasets.

Cell scoring using gene sets

Both the tumor and normal gene sets were used to score each of the cells in each of the experiments using AUCell [56]. As input to AUCell, we provide the gene expression matrices that were normalized to the total number of sequenced reads per cell and subsequently transformed using the $\log2(x+1)$ function. AUCell requires that the dataset contains at least 80% of the genes from the input gene set.

We have noticed that the AUCell scores do not behave properly in some of the bulk sequencing datasets. Namely, samples which had similar transcriptomes sometimes had widely different AUCell scores. The user is encouraged to use different gene set scoring algorithms like ssGSEA.

Logistic classifier training

A logistic classifier was trained on the combined Lee et al. and Laughney et al. datasets. Scores of normal and tumor gene signatures were used as the input and the tumor/normal class assignment as the target variables.

Cell annotation using network propagation

ikarus implements the cell annotation as an iterative two-step process of cell type assignment and label propagation. In each iteration, we assign labels to cells with a decreasing

stringency threshold, which are then propagated to their nearest neighbors. Firstly, the labels are assigned to the most probable cells, based on a robust stringency threshold. Cells below the stringency threshold have their LR probabilities masked to zero. The stringency threshold is defined based on the order statistic of the gene set score difference between the two classes of interest. In the first iteration, it is the 90% percentile of the (tumor–normal) gene set score difference. The label propagation is then obtained by computing the dot product of neighborhood connectivities and LR class probability estimates. Annotations are derived from the propagated class probabilities. Within each iteration step, the stringency threshold is reduced using exponential decay:

$$N(t) = N_0 e^{-\lambda t}$$

where

N_0 is the starting stringency threshold;

t is an iteration step;

λ is an exponential decay constant.

The cell–cell graph, used for label propagation, is constructed using the normal and tumor gene signatures according to [57], as adopted in [58].

The label propagation procedure is repeated until less than 0.1% of cell annotations change.

CNV correction

Classification improvement using copy number variations

To improve the classification results, we used inferred CNV scores as an additional source of information. InferCNV [55] was used to compute CNV scores. A cutoff value of 0.1 was chosen for gene selection. CNV prediction was performed via HMM (Hidden Markov Models). For tumor sub-clustering the parameters were kept default (hclust = 'ward.D2', tumor_subcluster_pval = 0.05), though tumor_subcluster_partition_method was set to 'qnorm' as this is claimed to be reasonable faster than 'random_trees' No prior information on distinct clusters was provided.

In a self-supervised fashion, we used the current ikarus cell annotations as pseudo-labels to train an additional logistic regression model (LR). The LR takes as its input per cell inferred CNV values and predicts the cell annotations. The LR itself is trained on all cells from the validation dataset, e.g., Lambrechts et al., Kildisiute et al., Puram et al. This model was then used to make predictions on the same dataset assuming that logistic regression should not overfit on this task. The outcome is then considered as the final corrected ikarus prediction.

Gene set characterization

Gene set activity in cell lines and PDX models

Tumor and normal gene signature scores were calculated for bulk RNA sequencing data from laser microdissected data from gastric cancer, ENCODE cell lines, CCLE cell lines, and PDX data. Because AUCell was developed for single-cell RNA sequencing data, the signature scores were calculated using ssGSEA as implemented in the escape Bioconductor package [59]. The tumor gene signature scores were compared to the cell line doubling times and PDX growth rates that were provided as annotations to the datasets.

ENCODE cell lines dataset was further stratified into three groups: primary cell line, cell line, and cancer cell line. The stratification was done manually based on the annotation provided by ENCODE.

Comparison with published gene sets

The tumor signature gene set assembled in this study was characterized by comparison with publicly available gene sets provided by multiple public resources. We considered gene sets of various provenances, e.g., all Homo sapiens gene sets published by MsigDB [49], cancer-specific gene sets that represent distinct functional tumor states (CancerSEA) [47], novel gene lists covering previously unidentified members of cellular signaling pathways [60].

Gene sets as provided by MsigDB ($n = 31120$) were assessed via the interface provided by the R package msigdbr version 7.2.1. Next, intersections and unions of the cancer gene set (this study) with every human gene set from that release (version 7.4) of MsigDB, as well as the members of the intersections and original sizes of query gene sets, were computed.

CancerSEA resource provides a collection of functional cancer gene sets derived from a multitude of single-cell studies, thereby supplying a single-cell level scope to the cancer functional hallmarks. For characterization of the cancer gene set assembled in this study, we downloaded 14 gene sets $n_{min} = 66$, $n_{max} = 201$ from the CancerSEA resource (http://biocc.hrbmu.edu.cn/CancerSEA/goDownload) representative of distinct cancer functional states and intersected it with our gene set. The results of this analysis were presented with the UpSet plot framework [61] implemented in the ComplexHeatmap R package [62], mode intersect.

To account for recent advances in the annotation of cellular signaling pathways, we downloaded a novel collection of gene sets composed of genes previously unmapped to any signal transduction pathway. Namely, we acquired 11 gene sets of various sizes $n_{min} = 10$, $n_{max} = 164$) and intersected with the tumor signature gene set of this study. The visualization of this analysis was similar, i.e., the intersections were presented with an UpSet plot implemented in ComplexHeatmap R package, mode intersect.

Gene fusions

Data on human gene fusions were downloaded from the ChiTaRS resource as was provided on August 16, 2019 (http://chitars.md.biu.ac.il/index.html) [53]. First, we constructed a background distribution from randomly selected sets of genes that were expression-matched to the tumor signature gene set (this study). Every random gene set was intersected with fused genes from the database and the resulting intersection sizes were used to fill a background distribution. Lastly, the tumor signature gene set from this study was intersected with the list of fused genes to compare with the background distribution. This analysis was done separately on 5′- and 3′-fused genes.

Co-expression analysis

To investigate genes that are co-expressed with the tumor signature gene list across many datasets, we took advantage of web-based platform SEEK (https://seek.princeton.edu/seek/) [48]. We queried the tumor gene list to the SEEK search engine and downloaded a SEEK-generated ranked list of co-expressed genes. For further analyses we used top 150 genes from the ranked list.

Gene ontology (GO) analysis

The GO analyses throughout this study were done using the framework provided by gprofiler2 R package [50]. From the default run settings, p values threshold was changed to 10e−4 and correction_method option set to g_SCS.

Gene set sensitivity testing

To measure ikarus' robustness on the extracted tumor gene list, we performed the following analyses:

Gene set size

Using the Lambrechts et al. lung cancer dataset, we iteratively computed ikarus' balanced accuracies ablating a random section of the tumor gene list in a cumulative stepwise manner before scoring and prediction steps. Namely, we randomly removed 20, 40, 60, and 80% percent of the tumor gene list. Every ablation percentage was itself reiterated 25 times. The predictions were not CNV-corrected.

Single gene ablation

Further, we investigated the prediction value of individual genes in the gene list. We employed a similar procedure as before, but in contrast to ablating the whole sections of the list, we removed an individual gene from the list per iteration before computing ikarus' balanced accuracies.

Gene set prognostic power analysis

To infer the prognostic power of the generated gene set, we referred to prognostic data available in the TCGA. Namely, we downloaded a dataset of Cox-proportional hazard model z-scores that were generated for every gene expression feature across all available tumor types [43]. The distributions of the gene set z-scores were compared to the distribution of all gene expression z-scores (population) in every cancer type individually. Additionally, the cited research provided estimates of Stoufer's z-scores per gene expression feature. This metric represents a normalized prognostic average over all cancer types available in the dataset. Here, the same procedure was used; we compared the distribution of Stoufer's Zs in the gene set to the distribution of all gene expression features.

CNV analysis

To investigate the overrepresentation of copy number amplifications among the genes in the extracted tumor gene list, we referred to the COSMIC database. Namely, we downloaded a complete COSMIC collection of copy number alterations and stratified it by tumor tissue of origin ($n = 27$). Next, we iteratively intersected the tumor gene list from

this study with significantly amplified genes (denoted as "gain" in the COSMIC table) over tumor tissues of origin. As a random control, we prepared similarly sized random gene sets ($n = 162$) that were expression-matched to the original tumor gene list. Expression matching was done on Laughney et al., Lee et al., and Lambrechts et al. datasets independently. In total, 150 random gene sets were generated, 50 sets per expression dataset.

Comparison to reference methods

The methods used as reference for ikarus were installed and used to the best of our knowledge. We used default hyperparameter settings from the corresponding description or provided tutorials. We did not perform any kind of hyperparameter optimization. We would like to point out that the published classification methods have many tunable parameters, and tuning the parameters might significantly increase their performance. Both CellAssign and scMRMA assume a marker gene list for the target cell type prediction. We provided here the tumor and normal gene signatures generated with ikarus.

Dataset subsampling

To ascertain whether the classification methods accuracy can be improved by balancing the classes, we randomly subsampled 1000 tumor and 1000 normal cells 100 times for each of the following datasets used for validation, Lambrechts et al., Kildisiute et al., and Puram et al., and evaluated the classifier performance on the datasets.

Statistical testing

Statistical tests performed, groups in comparison, and sample sizes are summarized in Supplementary Additional File 3: Statistical Comparisons. In cases of multiple testing, p values were adjusted using Benjamini-Hochberg (FDR) method [63]. For situations where tests were not applicable, random background distributions were simulated against which the probabilities of observing an event under question were estimated. Testing approaches of such kind are reported as "empirical" in Supplementary Additional File 3: Statistical Comparisons. If the adjusted p value was lower than 0.01, it was reported as statistically significant.

For comparing the distribution of non-tumor cells from ikarus' misclassifications for the Lambrechts et al. lung cancer dataset with the actual distribution of cell types, we performed pairwise for each of the misclassified groups 2×2 fisher exact test (Additional File 2: Lambrechts misclassification).

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-022-02683-1.

Additional file 1. Supplementary figures 1–4.

Additional file 2. Tables with data description and results.

Additional file 3. Results of statistical analysis.

Additional file 4. Review history.

Acknowledgements

We would like to sincerely thank Florian Uhlitz for a very constructive discussion and critical reading of the manuscript.

Review history
Review history is available as Additional file 4.

Peer review information
Stephanie McClelland was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Authors' contributions
AA conceptualized and planned the project. JD and AB jointly executed all of the computational analyses. BU, VF, and AA wrote the manuscript, with comments from JD and AB. JR critically edited the manuscript. BU, VF, and AA jointly supervised the work. VF led the day-to-day supervision and made sure the project was completed in a pre-defined timeline. All authors read, edited, and confirmed the final manuscript. AA acquired funding and supervised the project.

Funding
Open Access funding enabled and organized by Projekt DEAL. Akalin lab is supported by Berlin Institute of Health: Digital Health Accelerator in connection with this project. In addition, we acknowledge funding from the German Federal Ministry of Education and Research (BMBF) as part of the RNA Bioinformatics Center of the German Network for Bioinformatics Infrastructure (de.NBI) [031 A538C RBC (de.NBI)].

Availability of data and materials
Code
ikarus is a python package that can be found on the following link:
https://github.com/BIMSBbioinfo/ikarus
Code for reproducing the figures can be found on the following link:
https://github.com/BIMSBbioinfo/ikarus%2D%2D-auxiliary
Zenodo repository of the software libraries is available here [64].
Both repositories are available under the MIT license.
Datasets
Single-cell RNA-seq data:
Gene expression values for single-cell RNA-seq experiments are available through the corresponding publications. If not explicitly declared otherwise, the 10X genomics protocol was used for scRNA-seq.
Laughney et al. provide a lung adenocarcinoma (primary tumors and metastases) dataset that include 40505 cells coming from 17 patients. For our purpose, 1091 cells are considered tumorous, and 39,414 are normal.
63,689 cells from 23 colorectal cancer patients are coming from Lee et al. 16,248 cells are considered tumorous and 47,441 are normal. After our cross validation analysis, these two datasets serve as input for model building.
A non-small-cell lung cancer dataset is coming from Lambrechts et al. It considers 52,698 cells, of which 7447 are tumorous and 45,251 are normal. This dataset is used for model testing.
Puram et al. published 5578 single cells from 18 head and neck squamous cell carcinoma patients. They performed Fluorescence-activated cell sorting for scRNA-seq. 2215 cells are tumorous, 3363 cells are normal. We use this dataset for model testing.
Kildisiute et al. published a neuroblastoma cell atlas. We used 6442 cells (10X) from 5 patients (1766 tumorous, 4676 normal) and 13281 cells (CEL-seq2) from 16 patients (1630 tumorous, 11651 normal) as two distinct datasets for model testing. Ma et al. made a hepatocellular carcinoma dataset available with 17,164 tumor and 39,557 non-tumor cells. It was used as another test set. We also used a lung carcinoid dataset by Bischoff et al. for testing. It includes 8097 tumor and 55,230 non-tumor cells. Jerby-Arnon et al. published a synovial sarcoma dataset that we used for testing. It includes 8323 tumor and 851 non-tumor cells. A comprehensive description of the datasets can be found in the Additional File 2: Datasets.
For both input datasets, Laughney et al. lung adenocarcinoma and Lee et al. colorectal cancer, we considered a refined annotation for tumorous cells. Based on gene sets from MSigDB (v7.1) [49] hallmark collection HALLMARK_E2F_TARGETS, HALLMARK_G2M_CHECKPOINT, HALLMARK_MYC_TARGETS_V1, HALLMARK_MYC_TARGETS_V2, HALLMARK_P53_PATHWAY, HALLMARK_MITOTIC_SPINDLE, HALLMARK_HYPOXIA, HALLMARK_ANGIOGENESIS, and HALLMARK_GLYCOLYSIS, we scored each cell. If the average over all considered hallmark gene list scores (in the range 0–1) exceeds a reasonable threshold (0.45 for Laughney et al., 0.35 for Lee et al.), the cell is considered tumorous. Thresholds are chosen to minimize the amount of false positives with respect to the initial annotation of normal and tumor cell sources. The distribution of normal and tumor cell sources obtained from the initial annotation is provided in Additional File 2: Datasets.
ENCODE cell line dataset:
Gene expression values for primary cells, cell lines, and cancer cell lines were downloaded in batch from the ENCODE portal with the following query: Assay title: "polyA plus RNA-seq"; Status: "released"; Perturbation: "not perturbed"; Organism: "Homo sapiens"; Biosample classification: "cell line", "primary cell"; Genome assembly: "GRCh38". Identifiers corresponding to the acquired data totaling 860 files are provided in the supplementary materials (Additional File 2: Encode IDs). Downloaded expression tables were merged and standardized by a custom R script prepare.data_encode.R to a combined gene expression matrix that includes all input data, HGNC symbol gene annotation and cell annotations. For gene expression quantification log2(TPM) with a pseudocount of 1 was used. Based on those components, an AnnData object is created which is then provided as an input to the ikarus package. Cancer cell line annotation was done manually and is provided in Additional File 2: Enrichment analysis of tumor gene signature.
Microdissection dataset:
The gastric cancer microdissection dataset comprises laser capture microdissected (LCM) stromal and cancer regions collected from a patient cohort ($n = 8$) totaling 16 samples. Microdissected tissue for each sample was pooled together before library preparation to account for the absence of replicates. Gene expression quantification of stromal and cancer samples, as provided by the authors of the study [34] in the form of raw counts, was first standardized to ikarus format and then used as an input to ikarus pipeline.
Databases:

- Human protein atlas (https://www.proteinatlas.org/humanproteome/pathology) [51]
- Prognostic genes [52]
- Gene fusion (ChiTaRs) [53]
- SEEK (co-expression database) (https://seek.princeton.edu/seek/) [48]
- g:Profiler (https://biit.cs.ut.ee/gprofiler/) [50]
- CancerSEA (http://biocc.hrbmu.edu.cn/CancerSEA/home.jsp) [47]
- MsigDB (GO, Hallmark gene sets) [49]
- Atlas of co-essential modules [60]
- DepMap Achilles scores (https://depmap.org/portal/download/) [65]
- COSMIC (cancer.sanger.ac.uk) [66]

Declarations

Ethics approval and consent to participate

Ethics approval is not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Bioinformatics and Omics Data Science Platform, Berlin Institute For Medical Systems Biology, Max Delbrück Center for Molecular Medicine in the Helmholtz Association (MDC), Hannoversche Str.28, 10115, Berlin, Germany. [2]Non-coding RNAs and Mechanisms of Cytoplasmic Gene Regulation Lab, Berlin Institute for Medical Systems Biology, Hannoversche Str. 28, 10115 Berlin, Germany. [3]Free University Berlin, Kaiserswerther Str. 16-18, 14195 Berlin, Germany.

Received: 23 November 2021 Accepted: 6 May 2022
Published online: 30 May 2022

References

1. Turajlic S, Sottoriva A, Graham T, Swanton C. Resolving genetic heterogeneity in cancer. Nat Rev Genet. 2019;20:404–16.
2. Moscow JA, Fojo T, Schilsky RL. The evidence framework for precision cancer medicine. Nat Rev Clin Oncol. 2018;15:183–92.
3. Bassiouni R, Gibbs LD, Craig DW, Carpten JD, McEachron TA. Applicability of spatial transcriptional profiling to cancer research. Mol Cell. 2021;81:1631–9.
4. Nath A, Bild AH. Leveraging Single-cell approaches in cancer precision medicine. Trends Cancer Res. 2021;7:359–72.
5. Svensson V, Vento-Tormo R, Teichmann SA. Exponential scaling of single-cell RNA-seq in the past decade. Nat Protoc. 2018;13:599–604.
6. Luecken MD, Theis FJ. Current best practices in single-cell RNA-seq analysis: a tutorial. Mol Syst Biol. 2019;15:e8746.
7. Lähnemann D, Köster J, Szczurek E, McCarthy DJ, Hicks SC, Robinson MD, et al. Eleven grand challenges in single-cell data science. Genome Biol. 2020;21:31.
8. Abdelaal T, Michielsen L, Cats D, Hoogduin D, Mei H, Reinders MJT, et al. A comparison of automatic cell identification methods for single-cell RNA sequencing data. Genome Biol. 2019;20:194.
9. Tan Y, Cahan P. SingleCellNet: A computational tool to classify single cell RNA-Seq data across platforms and across species. Cell Syst. 2019;9:207–213.e2.
10. Hu J, Li X, Hu G, Lyu Y, Susztak K, Li M. Iterative transfer learning with neural network for clustering and cell type classification in single-cell RNA-seq analysis. Nat Mach Intell. 2020;2:607–18.
11. Andreatta M, Corria-Osorio J, Müller S, Cubas R, Coukos G, Carmona SJ. Interpretation of T cell states from single-cell transcriptomics data using reference atlases. Nat Commun. 2021;12:2965.
12. Michielsen L, Reinders MJT, Mahfouz A. Hierarchical progressive learning of cell identities in single-cell data. Nat Commun. 2021;12:2799.
13. Ranjan B, Schmidt F, Sun W, Park J, Honardoost MA, Tan J, et al. scConsensus: combining supervised and unsupervised clustering for cell type identification in single-cell RNA sequencing data. BMC Bioinformatics. 2021;22:186.
14. Grabski IN, Irizarry RA. A probabilistic gene expression barcode for annotation of cell-types from single cell RNA-seq data. bioRxiv. 2020:2020.01.05.895441. https://doi.org/10.1101/2020.01.05.895441.
15. van den Brink SC, Sage F, Vértesy Á, Spanjaard B, Peterson-Maduro J, Baron CS, et al. Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations. Nat Methods. 2017;14:935–6.
16. Tran HTN, Ang KS, Chevrier M, Zhang X, Lee NYS, Goh M, et al. A benchmark of batch-effect correction methods for single-cell RNA sequencing data. Genome Biol. 2020;21:12.
17. Argelaguet R, Cuomo ASE, Stegle O, Marioni JC. Computational principles and challenges in single-cell data integration. Nat Biotechnol. 2021. https://doi.org/10.1038/s41587-021-00895-7.
18. Lopez R, Regier J, Cole MB, Jordan MI, Yosef N. Deep generative modeling for single-cell transcriptomics. Nat Methods. 2018;15:1053–8.
19. Brbić M, Zitnik M, Wang S, Pisco AO, Altman RB, Darmanis S, et al. MARS: discovering novel cell types across heterogeneous single-cell experiments. Nat Methods. 2020;17:1200–6.
20. Li X, Wang K, Lyu Y, Pan H, Zhang J, Stambolian D, et al. Deep learning enables accurate clustering with batch effect removal in single-cell RNA-seq analysis. Nat Commun. 2020;11:2338.

21. Zhou X, Chai H, Zeng Y, Zhao H, Yang Y. scAdapt: virtual adversarial domain adaptation network for single cell RNA-seq data classification across platforms and species. Brief Bioinform. 2021;22. https://doi.org/10.1093/bib/bbab281.
22. Ge S, Wang H, Alavi A, Xing E, Bar-Joseph Z. Supervised adversarial alignment of single-cell RNA-seq data. J Comput Biol. 2021;28:501–13.
23. Chen L, He Q, Zhai Y, Deng M. Single-cell RNA-seq data semi-supervised clustering and annotation via structural regularized domain adaptation. Bioinformatics. 2021;37:775–84.
24. Kimmel JC, Kelley DR. Semisupervised adversarial neural networks for single-cell classification. Genome Res. 2021;31:1781–93.
25. Luecken MD, Büttner M, Chaichoompu K, Danese A, Interlandi M, Mueller MF, et al. Benchmarking atlas-level data integration in single-cell genomics. Nat Methods. 2022;19:41–50.
26. Ronen J, Akalin A. netSmooth: Network-smoothing based imputation for single cell RNA-seq. F1000Res. 2018;7:8.
27. Laughney AM, Hu J, Campbell NR, Bakhoum SF, Setty M, Lavallée V-P, et al. Regenerative lineages and immune-mediated pruning in lung cancer metastasis. Nat Med. 2020;26:259–69.
28. Lambrechts D, Wauters E, Boeckx B, Aibar S, Nittner D, Burton O, et al. Phenotype molding of stromal cells in the lung tumor microenvironment. Nat Med. 2018;24:1277–89.
29. Lee H-O, Hong Y, Etlioglu HE, Cho YB, Pomella V, Van den Bosch B, et al. Lineage-dependent gene expression programs influence the immune landscape of colorectal cancer. Nat Genet. 2020;52:594–603.
30. Kildisiute G, Kholosy WM, Young MD, Roberts K, Elmentaite R, van Hooff SR, et al. Tumor to normal single-cell mRNA comparisons reveal a pan-neuroblastoma cancer cell. Sci Adv. 2021;7. https://doi.org/10.1126/sciadv.abd3311.
31. Puram SV, Tirosh I, Parikh AS, Patel AP, Yizhak K, Gillespie S, et al. Single-cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer. Cell. 2017;171:1611–1624.e24.
32. Ma L, Wang L, Khatib SA, Chang C-W, Heinrich S, Dominguez DA, et al. Single-cell atlas of tumor cell evolution in response to therapy in hepatocellular carcinoma and intrahepatic cholangiocarcinoma. J Hepatol. 2021;75:1397–408.
33. Bischoff P, Trinks A, Wiederspahn J, Obermayer B, Pett JP, Jurmeister P, et al. The single-cell transcriptional landscape of lung carcinoid tumors. Int J Cancer. 2022. https://doi.org/10.1002/ijc.33995.
34. Grunberg N, Pevsner-Fischer M, Goshen-Lago T, Diment J, Stein Y, Lavon H, et al. Cancer-associated fibroblasts promote aggressive gastric cancer phenotypes via heat shock factor 1-mediated secretion of extracellular vesicles. Cancer Res. 2021;81:1639–53.
35. Kilpinen H, Waszak SM, Gschwind AR, Raghav SK, Witwicki RM, Orioli A, et al. Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. Science. 2013;489:57–74.
36. Davis CA, Hitz BC, Sloan CA, Chan ET, Davidson JM, Gabdank I, et al. The Encyclopedia of DNA elements (ENCODE): data portal update. Nucleic Acids Res. 2018;46:D794–801.
37. Gao H, Korn JM, Ferretti S, Monahan JE, Wang Y, Singh M, et al. High-throughput screening using patient-derived tumor xenografts to predict clinical trial drug response. Nat Med. 2015;21:1318–25.
38. Ghandi M, Huang FW, Jané-Valbuena J, Kryukov GV, Lo CC, McDonald ER 3rd, et al. Next-generation characterization of the Cancer Cell Line Encyclopedia. Nature. 2019;569:503–8.
39. Ma F, Pellegrini M. ACTINN: automated identification of cell types in single cell RNA sequencing. Bioinformatics. 2020;36:533–8.
40. Wang L, Nie R, Yu Z, Xin R, Zheng C, Zhang Z, et al. An interpretable deep-learning architecture of capsule networks for identifying cell-type gene expression programs from single-cell RNA-sequencing data. Nat Mach Intell. 2020;2:693–703.
41. Zhang AW, O'Flanagan C, Chavez EA, Lim JLP, Ceglia N, McPherson A, et al. Probabilistic cell-type assignment of single-cell RNA-seq for tumor microenvironment profiling. Nat Methods. 2019;16:1007–15.
42. Gayoso A, Lopez R, Xing G, Boyeau P, Wu K, Jayasuriya M, et al. scvi-tools: a library for deep probabilistic analysis of single-cell omics data. bioRxiv. 2021:2021.04.28.441833. https://doi.org/10.1101/2021.04.28.441833.
43. Li J, Sheng Q, Shyr Y, Liu Q. scMRMA: single cell multiresolution marker-based annotation. Nucleic Acids Res. 2022;50:e7.
44. McInnes L, Healy J, Melville J. UMAP: Uniform manifold approximation and projection for dimension reduction. arXiv [stat.ML]. 2020; Available: http://arxiv.org/abs/1802.03426.
45. Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, et al. Massively parallel digital transcriptional profiling of single cells. Nat Commun. 2017;8:14049.
46. Jerby-Arnon L, Neftel C, Shore ME, Weisman HR, Mathewson ND, McBride MJ, et al. Opposing immune and genetic mechanisms shape oncogenic programs in synovial sarcoma. Nat Med. 2021;27:289–300.
47. Yuan H, Yan M, Zhang G, Liu W, Deng C, Liao G, et al. CancerSEA: a cancer single-cell state atlas. Nucleic Acids Res. 2019;47:D900–8.
48. Zhu Q, Wong AK, Krishnan A, Aure MR, Tadych A, Zhang R, et al. Targeted exploration and analysis of large cross-platform human transcriptomic compendia. Nat Methods. 2015;12:211–4 3 p following 214.
49. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst. 2015;1:417–25.
50. Raudvere U, Kolberg L, Kuzmin I, Arak T, Adler P, Peterson H, et al. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). Nucleic Acids Res. 2019;47:W191–8.
51. Uhlen M, Zhang C, Lee S, Sjöstedt E, Fagerberg L, Bidkhori G, et al. A pathology atlas of the human cancer transcriptome. Science. 2017;357. https://doi.org/10.1126/science.aan2507.
52. Smith JC, Sheltzer JM. Genome-wide identification and analysis of prognostic features in human cancers. bioRxiv. 2021:2021.06.01.446243. https://doi.org/10.1101/2021.06.01.446243.
53. Gorohovski A, Tagore S, Palande V, Malka A, Raviv-Shay D, Frenkel-Morgenstern M. ChiTaRS-3.1-the enhanced chimeric transcripts and RNA-seq database matched with protein-protein interactions. Nucleic Acids Res. 2017;45:D790–5.
54. Shlien A, Malkin D. Copy number variations and cancer. Genome Med. 2009;1:62.

55. Tickle T, Tirosh I, Georgescu C, Brown M, Haas B. inferCNV of the Trinity CTAT Project. Klarman Cell Observatory, Broad Institute of MIT and Harvard, Cambridge, MA. USA. Available: https://github.com/broadinstitute/infercnv.
56. Aibar S, González-Blas CB, Moerman T, Huynh-Thu VA, Imrichova H, Hulselmans G, et al. SCENIC: single-cell regulatory network inference and clustering. Nat Methods. 2017;14:1083–6.
57. Coifman RR, Lafon S, Lee AB, Maggioni M, Nadler B, Warner F, et al. Geometric diffusions as a tool for harmonic analysis and structure definition of data: diffusion maps. Proc Natl Acad Sci U S A. 2005;102:7426–31.
58. Haghverdi L, Büttner M, Alexander Wolf F, Buettner F, Theis FJ. Diffusion pseudotime robustly reconstructs lineage branching. Nature Methods. 2016:845–8. https://doi.org/10.1038/nmeth.3971.
59. Borcherding N, Andrews J. escape: Easy single cell analysis platform for enrichment. 2021.
60. Wainberg M, Kamber RA, Balsubramani A, Meyers RM, Sinnott-Armstrong N, Hornburg D, et al. A genome-wide atlas of co-essential modules assigns function to uncharacterized genes. Nat Genet. 2021;53:638–49.
61. Lex A, Gehlenborg N, Strobelt H, Vuillemot R, Pfister H. UpSet: visualization of intersecting sets. IEEE Trans Vis Comput Graph. 2014;20:1983–92.
62. Gu Z, Eils R, Schlesner M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics. 2016;32:2847–9.
63. Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. J R Stat Soc. 1995;57:289–300.
64. Dohmen J, Baranovskii A, Ronen J, Uyar B, Franke V, Akalin A. Tumor cell classification at the single cell level. Zenodo. 2022. https://doi.org/10.1101/2021.10.15.463909.
65. Tsherniak A, Vazquez F, Montgomery PG, Weir BA, Kryukov G, Cowley GS, et al. Defining a cancer dependency map. Cell. 2017;170:564–576.e16.
66. Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, et al. COSMIC: the catalogue of somatic mutations in cancer. Nucleic Acids Res. 2019;47:D941–7.

Publisher's Note