BMC Bioinformatics

RESEARCH ARTICLE Open Access

An elastic-net logistic regression approach to generate classifiers and gene signatures for types of immune cells and T helper cell subsets

Arezo Torang[1], Paraag Gupta[1] and David J. Klinke II[1,2*]

Abstract

Background: Host immune response is coordinated by a variety of different specialized cell types that vary in time and location. While host immune response can be studied using conventional low-dimensional approaches, advances in transcriptomics analysis may provide a less biased view. Yet, leveraging transcriptomics data to identify immune cell subtypes presents challenges for extracting informative gene signatures hidden within a high dimensional transcriptomics space characterized by low sample numbers with noisy and missing values. To address these challenges, we explore using machine learning methods to select gene subsets and estimate gene coefficients simultaneously.

Results: Elastic-net logistic regression, a type of machine learning, was used to construct separate classifiers for ten different types of immune cell and for five T helper cell subsets. The resulting classifiers were then used to develop gene signatures that best discriminate among immune cell types and T helper cell subsets using RNA-seq datasets. We validated the approach using single-cell RNA-seq (scRNA-seq) datasets, which gave consistent results. In addition, we classified cell types that were previously unannotated. Finally, we benchmarked the proposed gene signatures against other existing gene signatures.

Conclusions: Developed classifiers can be used as priors in predicting the extent and functional orientation of the host immune response in diseases, such as cancer, where transcriptomic profiling of bulk tissue samples and single cells are routinely employed. Information that can provide insight into the mechanistic basis of disease and therapeutic response. The source code and documentation are available through GitHub: https://github.com/KlinkeLab/ImmClass2019.

Keywords: Gene signature, Machine learning, Elastic-net, In silico cytometry

Background

Host immune response is a coordinated complex system, consisting of different specialized innate and adaptive immune cells that vary dynamically and in different anatomical locations. As shown in Fig. 1, innate immune cells comprise myeloid cells, which include eosinophils, neutrophils, basophils, monocytes, and mast cells. Adaptive immune cells are mainly B lymphocytes and T lymphocytes that specifically recognize different antigens [1]. Linking innate with adaptive immunity are Natural Killer cells and antigen presenting cells, like macrophages and dendritic cells. Traditionally, unique cell markers have been used to characterize different immune cell subsets from heterogeneous cell mixtures using flow cytometry [2–4]. However, flow cytometry measures on the order of 10 parameters simultaneously and relies on prior knowledge for selecting relevant molecular markers, which could provide a biased view of the immune state within

*Correspondence: david.klinke@mail.wvu.edu
[1] Department of Chemical and Biomedical Engineering, West Virginia University, 1306 Evansdale Dr, 26506 Morgantown, WV, USA
[2] Department of Microbiology, Immunology, and Cell Biology, West Virginia University, 1 Medical Center Drive, 26506 Morgantown, WV, USA

a sample [5]. Recent advances in technology, like mass cytometry or multispectral imaging, have expanded the number of molecular markers, but the number of markers used for discriminating among cell types within a sample remains on the order of $10^{1.5}$.

In the recent years, quantifying tumor immune contexture using bulk transcriptomics or single-cell RNA sequencing data (scRNA-seq) has piqued the interest of the scientific community [6–10]. Advances in transcriptomics technology, like RNA sequencing, provide a much higher dimensional view of which genes are expressed in different immune cells (i.e., on the order of 10^3) [11]. Conceptually, inferring cell types from data using an expanded number of biologically relevant genes becomes more tolerant to non-specific noise and non-biological differences among samples and platforms. In practice, cell types can be identified using gene signatures, which are defined as sets of genes linked to common downstream functions or inductive networks that are co-regulated [12, 13], using approaches such as Gene Set Enrichment Analysis (GSEA) [12]. However, as microarray data can inflate detecting low abundance and noisy transcripts and scRNA-seq data can have a lower depth of sequencing, opportunities for refining methods to quantify the immune contexture using gene signatures still remain.

Leveraging transcriptomics data to identify immune cell types presents analytic challenges for extracting informative gene signatures hidden within a high dimensional transcriptomics space that is characterized by low sample numbers with noisy and missing values. Typically, the number of cell samples is in the range of hundreds or less, while the number of profiled genes is in the tens of thousands [14]. Yet, only a few number of genes are relevant for discriminating among immune cell subsets. Datasets with a large number of noisy and irrelevant genes decrease the accuracy and computing efficiency of machine learning algorithms, especially when the number of samples are very limited. Hence, feature selection algorithms may be used to reduce the number of redundant genes [15]. Using feature selection methods enable developing gene signatures in different biomedical fields of study [16]. There are many proposed feature selection methods that can select gene sets that enable classifying samples with high accuracy. In recent years, regularization methods have became more popular, which efficiently select features [17] and also control for overfitting [18]. As a machine learning tool, logistic regression is considered to be a powerful discriminative method [18]. However, logistic regression alone is not applicable for high-dimensional cell classification problems [19]. On the other hand, hybrid methods, like regularized logistic regression, have been successfully applied to high-dimensional problems [20]. Regularized logistic regression selects a small set of genes with the strongest effects on the cost function [17]. A regularized logistic regression can be also be applied with different regularization terms. The most popular regularized terms are LASSO, Ridge [21], and elastic-net [22], which impose the $l1$ norm, $l2$ norm, and linear combination of $l1$ norm and $l2$ norm regularization, respectively, to the cost function. It has been shown that, specifically in very high dimensional problems, elastic-net outperforms LASSO and Ridge [17, 22].

In this study, we focused on two-step regularized logistic regression techniques to develop immune cell signatures and immune cell and T helper cell classifiers using RNA-seq data for the cells highlighted in bold in Fig. 1. The first step of the process included a pre-filtering phase to select the optimal number of genes and implemented an elastic-net model as a regularization method for gene selection in generating the classifiers. The pre-filtering step reduced computational cost and increased final accuracy by selecting the most discriminative and relevant set of genes. Finally, we illustrate the value of the approach in annotating gene expression profiles obtained from single-cell RNA sequencing. The second step generated gene signatures for individual cell types using selected genes from first step and implemented a binary regularized logistic regression for each cell type against all other samples.

Results

We developed classifiers for subsets of immune cells and T helper cells separately with two main goals. First, we aimed to annotate RNA-seq data obtained from an enriched cell population with information as to the immune cell identity. Second, we developed gene signatures for different immune cells that could be used to quantify the prevalence from RNA-seq data obtained from a heterogeneous cell population. Prior to developing the classifiers, the data was pre-processed to remove genes that have low level of expression for most of samples (details can be found in Methods section) and normalized to increase the homogeneity in samples from different studies and to decrease dependency of expression estimates to transcript length and GC-content. Genes retained that had missing values for some of the samples were assigned a value of -1. Next, regularized logistic regression (elastic-net) was performed and the optimal number of genes and their coefficients were determined.

Generating and validating an immune cell classifier

In developing the immune cell classifier, we determined the optimal number of genes in the classifier by varying the lambda value used in the regularized logistic regression of the training samples and assessing performance. To quantify the performance using different lambdas, a dataset was generated by combining True-Negative samples, which were created using a bootstrapping approach that randomly resampled associated genes

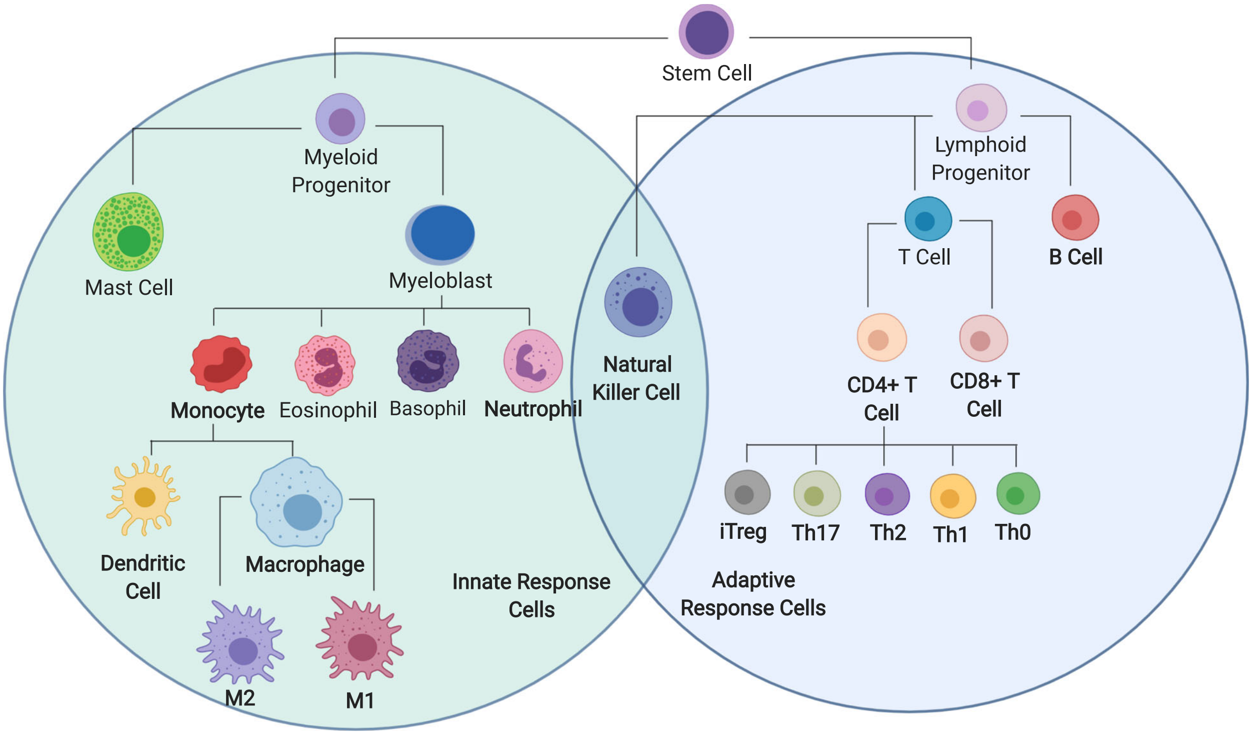

Fig. 1 Lineage tree representation of cells of the immune system. Immune cells are derived from hematopoietic stem cells (HSCs). HSCs differentiate into lymphoid and myeloid progenitors that further branch out to the more specific cell types associated with adaptive and innate immunity. This Figure indicates the main immune cell subsets and arrows are to show lineage relationships. Gene signatures were developed in this study for immune cells highlighted in bold

and their corresponding value from the testing datasets to create a synthetic dataset of similar size and complexity, with the original testing data, which were untouched during training and provided True-Positive samples. The accuracy of predicting the True-Positive samples were used to generate Receiver Operating Characteristic (ROC) curves (Fig. 2a). Performance using each lambda was quantified as the Area Under the ROC Curve (AUC).

The optimal lambda for immune cell classifier was the smallest value (i.e., highest number of genes) that maximized the AUC. Functionally, this lambda value represents the trade-off between retaining the highest number of informative genes (i.e., classifier signal) for developing the gene signature in the second step, while not adding non-informative genes (i.e., classifier noise). Consequently, we selected a lambda value of 1e-4 (452 genes) for the immune cell classifier, where the selected genes and their coefficients are shown in Additional file 1: Table S1.

To explore correlations between the weights of selected genes with their expression level, we generated heatmaps shown in Fig. 2, panels b and c. A high level of gene expression is reflected as a larger positive coefficient in a classifier model, while low or absent expression results in a negative coefficient. This is interpreted as, for example, if gene A is not in cell type 1, the presence of this gene in a sample decreases the probability for that sample to be cell type 1. For instance, E-cadherin (CDH1) was not detected in almost all monocyte samples and thus has a negative coefficient. Conversely, other genes are only expressed in certain cell types, which results in a high positive coefficient. For instance, CYP27B1, INHBA, IDO1, NUPR1, and UBD are only expressed by M1 macrophages and thus have high positive coefficients.

The differential expression among cell types suggests that the set of genes included in the classifier model may also be a good starting point for developing gene signatures, which is highlighted in Fig. 2d. Here, we focused on the expression of the 452 genes included in the classifier model and the correlations between samples clustered based on cell types. The off-diagonal entries in the correlation matrix are colored by euclidean distance with the color indicating similarity or dissimilarity using pink and blue, respectively. Color bars along the axes also highlight the cell types for the corresponding RNA-seq samples. As expected, RNA-seq samples from the same cell type were highly similar. More interestingly, correlation between different cell types can also be seen, like high similarity between CD4+ and CD8+ T cell samples, CD8+ T cell and NK cell samples, and monocyte and dendritic cell samples.

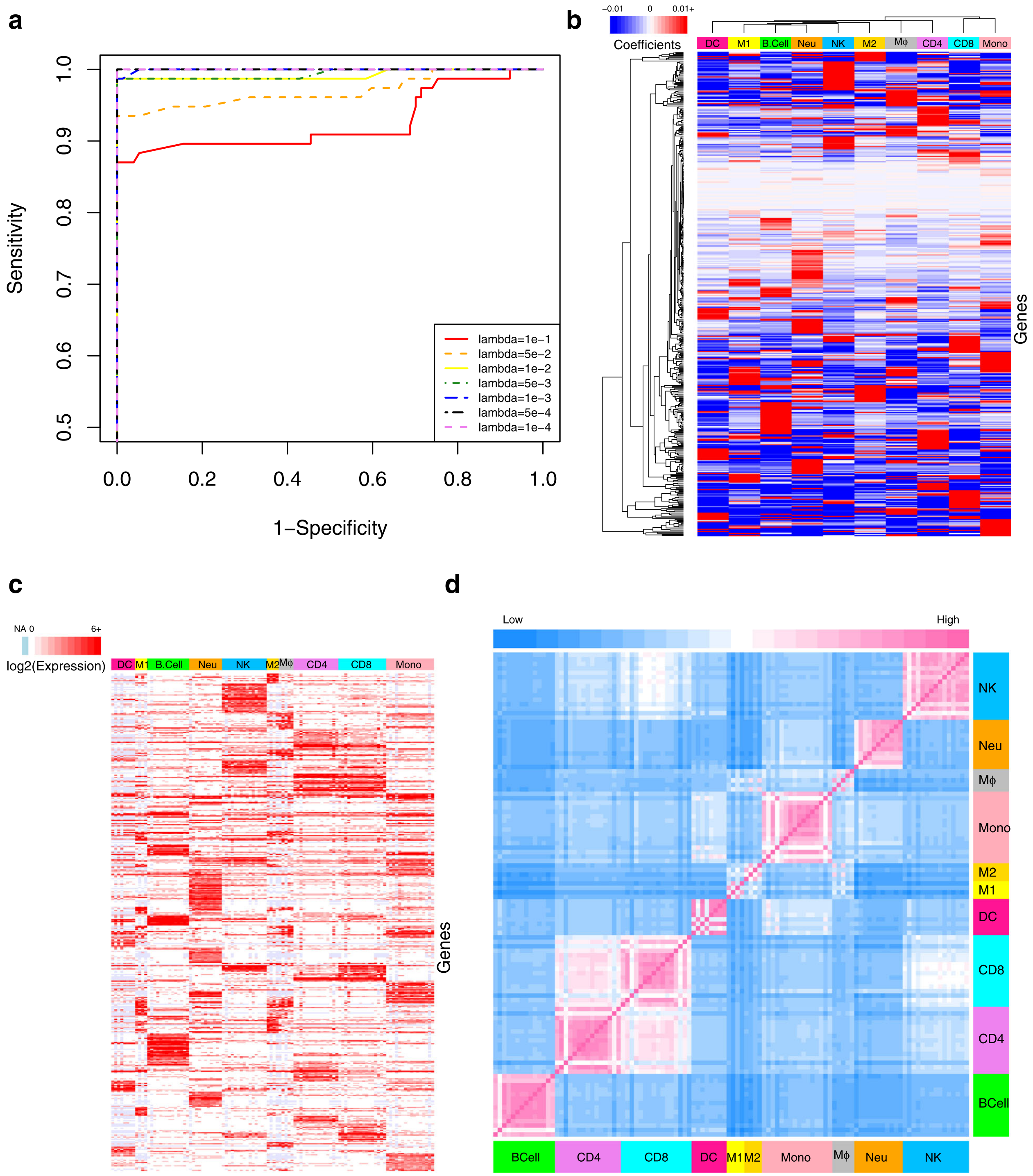

Fig. 2 Development of immune cell classifier and similarity heatmap. **a** ROC curve for the immune cell classifier was calculated using the indicated lambda values (shown in different colors and line styles) and 10-fold cross validation. The lambda value that maximized the AUC value was used for subsequent calculations. Elastic-net logistic regression was used to discriminate among ten immune cell types, where the value of the non-zero coefficients (panel **b**), expression levels (panel **c**), and similarity map (panel **d**) for the 452 genes included in the classifier are indicated by color bars for each panel. In panel **b**, blue to red color scheme indicates coefficients ranging from negative to positive values. Ordering of the genes is the same in panels **b** and **c**. In panel **c**, light blue indicates missing values and the intensity of red color (white/red color scale on the top-left) shows the log base 2 expression level. A color bar on top of this panel was used to separate samples of each cell type. Panel **d** illustrates the similarity between samples calculated using distance matrix based on same 452 genes. Color bars on the left and bottom sides are to separate samples of each cell type and the top color bar (light blue/pink color scale) shows the intensity of similarity or dissimilarity of samples

Collectively, these heatmaps illustrate that the selected genes are a highly condensed but are still a representative set of genes that include the main characteristics of the immune cell types. It is also notable to compare the clustering result of cell types based on their coefficients in the classifier shown in Fig. 2b with similarity matrix in Fig. 2d. Since in the classifier coefficients are forcing the model to separate biologically close cell types (like CD4+ T cell and CD8+ T cell), the clustering results suggest that the coefficient vectors are equally dissimilar (Fig. 2b). However, in the case of their expression values, their similarity remains (Fig. 2d).

Evaluating the immune cell classifier using scRNA-seq datasets

To evaluate the proposed classifier in immune cell classification, two publicly accessible datasets generated by scRNA-seq technology were used [23, 24]. The first dataset included malignant, immune, stromal and endothelial cells from 15 melanoma tissue samples [23]. We focused on the immune cell samples, which included 2761 annotated samples of T cells, B cells, M*phi* and NK cells, and 294 unresolved samples. The immune cells in this study were recovered by flow cytometry by gating on CD45 positive cells. Annotations were on the basis of expressed marker genes while unresolved samples were from the CD45-gate and classified as non-malignant based on inferred copy number variation (CNV) patterns (i.e., CNV score <0.04).

Following pre-processing to filter and normalize the samples similar to the training step, the trained elastic-net logistic regression model was used to classify cells into one of the different immune subsets based on the reported scRNA-seq data with the results summarized in Fig. 3a. The inner pie chart shows the prior cell annotations reported by [23] and the outer chart shows the corresponding cell annotation predictions by our proposed classifier. Considering T cells as either CD4+ T cell or CD8+ T cell, the overall similarity between annotations provided by [23] and our classifier prediction is 96.2%. The distribution in cells types contained within the unresolved samples seemed to be slightly different than the annotated samples as we predicted the unresolved samples to be mainly CD8+ T cells and B cells.

The only cell type with low similarity between our classifier predictions and prior annotations was NK cells, where we classified almost half of samples annotated previously as NK cells as CD8+ T cell. Discriminating between these two cell types is challenging as they share many of the genes related to cytotoxic effector function and can also be subclassified into subsets, like CD56bright and CD56dim NK subsets [25]. To explore this discrepancy, we compared all annotated samples based on their CD8 score and NK score provided by the classifier, as shown in Fig. 3b. Although the number of NK cell samples are relatively low, it seems that the NK samples consist of two groups of samples: one with a higher likelihood of being a NK cell and a second with almost equal likelihood for being either CD8+ T cell or NK cell. We applied principal component analysis (PCA) to identify genes associated with this difference and used Enrichr for gene set enrichment [26, 27]. Using gene sets associated with the Human Gene Atlas, the queried gene set was enriched for genes associated with CD56 NK cells, CD4+ T cell and CD8+ T cell. Collectively, the results suggests that the group of cells with similar score for NK and CD8 in the classifier model are Natural Killer T cells.

We also analyzed a second dataset that included 317 epithelial breast cancer cells, 175 immune cells and 23 non-carcinoma stromal cells, from 11 patients diagnosed with breast cancer [24]. We only considered samples annotated previously as immune cells, which were annotated as T cells, B cells, and myeloid samples by clustering the gene expression signatures using non-negative factorization. The scRNA-seq samples were similarly preprocessed and analyzed using the proposed classifier, with the results shown in Fig. 4. The inner pie chart shows the prior cell annotations reported by [24] and the outer chart shows the corresponding predicted cell annotation by our proposed classifier. Considering T cells as either CD4+ T cell or CD8+ T cell, 94.4% of reported T cells are predicted as the same cell type and other 5.6% is predicted to be DC or NK cells. However, for reported B cells and myeloid cells, we predicted relatively high portion of samples to be T cells (15.7% of B cells and 40% of myeloid cells). The rest of the myeloid samples were predicted to be macrophages or dendritic cells. Collectively, our proposed classifier agreed with many of the prior cell annotations and annotated many of the samples that were previously unresolved.

Developing a classifier for T helper cell subsets

To further apply this methodology to transcriptomic data, a separate classifier for distinguishing among T helper cells was developed using a similar approach to the immune cell classifier. We explored different values of the regression parameter lambda to find the optimal number of genes for this new dataset and visualized the performance of different lambdas by generating True-Negative samples using a bootstrapping approach whereby synthetic datasets were created by randomly resampling testing datasets. Original testing data that were completely untouched during training were used as True-Positive samples. The resulting True-Negative and True-Positive samples were used to generate ROC curves (Fig. 5a) and the AUC was used to score each lambda value. Generally, the lambda values for T helper cell classifier represents the trade-off between retaining genes and keeping the AUC

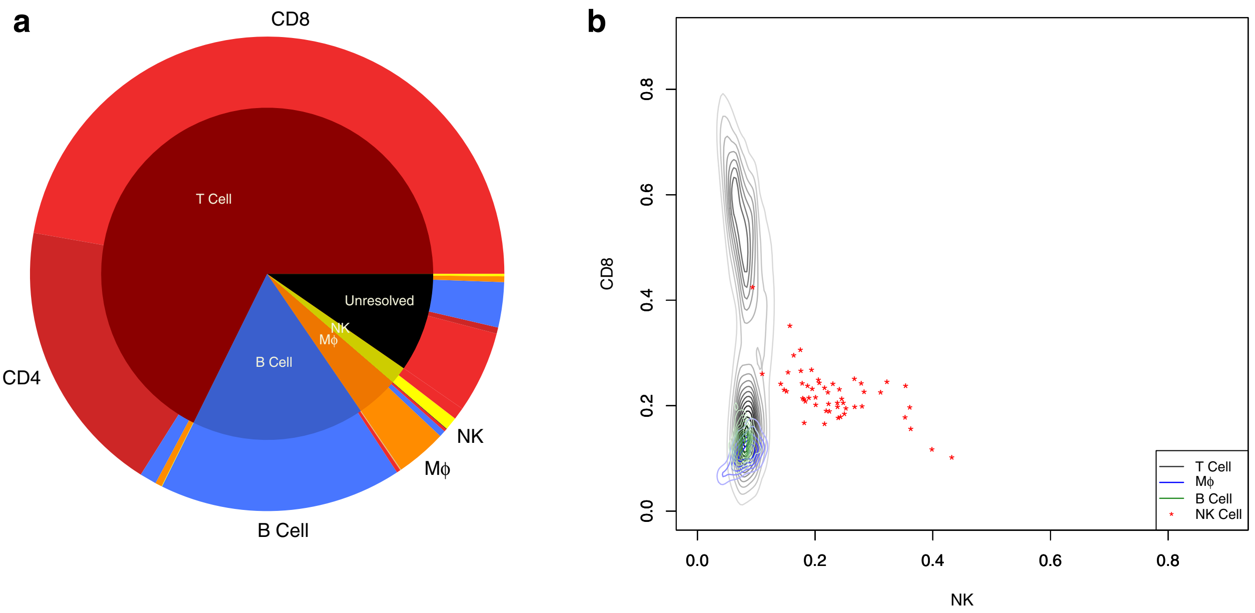

Fig. 3 Immune cell annotation prediction based on scRNA-seq data against prior annotations reported in melanoma dataset. **a** The inner pie chart summarizes the cell annotations reported by Tirosh et al. [23] and includes 298 unannotated CD45-positive non-malignant cells (labeled as Unresolved) isolated from melanoma tissue samples. Unannotated samples were acquired following gating for CD45+ single cells and classified as non-malignant based on inferred copy number variation patterns. Using gene expression values reported for each scRNA-seq sample, a new cell annotation was determined based on the closest match with the alternative cell signatures determined using elastic-net logistic regression, which are summarized in outer pie chart. **b** The contour plot for the likelihood of a sample to be either an NK cell or CD8+ T cell based on gene expression stratified by cells previously annotated by [23] to be T cells, macrophages, B cells, or NK cells

high. However, there appeared to be an inflection point at a lambda value of 0.05 whereby adding additional genes, by increasing lambda, reduced the AUC. Consequently, we selected a lambda value equal to 0.05 (72 genes) for the T helper classifier. The selected genes and their coefficients are listed in Additional file 1: Table S1. The gene list was refined subsequently by developing a gene signature.

Similar to the immune cell classifier, the coefficients of the selected genes for the T helper cell classifier correlated with their expression levels, as seen by comparing the heatmaps shown in Fig. 5, panels b and c. For instance, FUT7 has been expressed in almost all T helper cell samples except for iTreg that result in a negative coefficient for this cell type. In addition, there are sets of genes for each cell type that have large coefficients only for certain T helper cell subsets, like ALPK1, TBX21, IL12RB2, IFNG, RNF157 for Th1 that have low expression in other cells. As illustrated in Fig. 5d, the genes included in the classifier don't all uniquely associate with a single subset but collectively enable discriminating among T helper cell subsets. Interestingly, the T helper subsets stratified into two subgroups where naive T helper cells (Th0) and inducible T regulatory (iTreg) cells were more similar than effector type 1 (Th1), type 2 (Th2), and type 17 (Th17) T helper cells. Similar to the immune cell classifier, we also noted that the clustering of the classifier coefficients is different from what similarity matrix shows in Fig. 5d because the classifier coefficients aim to create a "classifying distance" among closely related cell types.

Finally by comparing the results of immune cell classifier with that of the T helper classifier, the intensity of differences among cell types can be seen in Figs. 2c and 5c. In the first figure you can find completely distinct set of genes in each cell type. Meanwhile, the gene sets in the second figure are not as distinct which could be due to the low number of samples or high biological similarity between T helper cell types.

Application of the classifiers

Clinical success of immune checkpoint inhibitors (ICI) for treating cancer coupled with technological advances in assaying the transcriptional signatures in individual cells, like scRNA-seq, has invigorated interest in characterizing the immune contexture within complex tissue microenvironments, like cancer. However as illustrated by the cell annotations reported by [24], identifying immune cell types from noisy scRNA-seq signatures using less biased methods remains an unsolved problem. To address this problem, we applied our newly developed classifiers to characterize the immune contexture in melanoma and explored differences in immune contexture that associate with immune checkpoint response. Of note, some patients

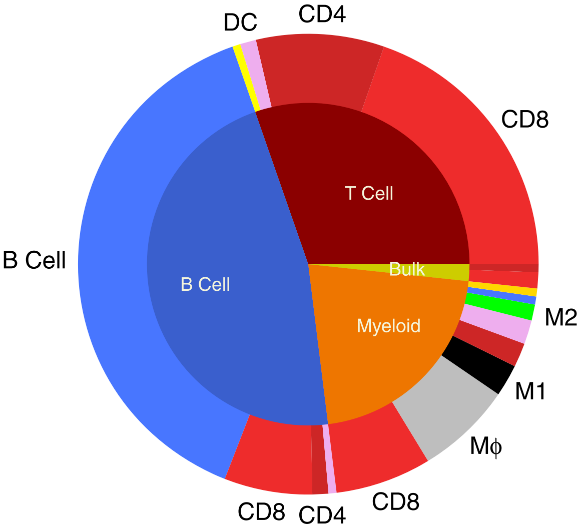

Fig. 4 Immune cell annotation prediction against prior annotations reported in breast cancer scRNA-seq dataset. The inner pie chart summarizes the cell annotations reported by Chung et al. [24], which annotated scRNA-seq results by clustering by gene ontology terms using likelihood ratio test. Using the gene expression profile reported for each scRNA-seq sample, a new cell annotation was determined based on the closest match with the alternative cell signatures determined using elastic-net logistic regression, which is summarized in the outer pie chart

with melanoma respond to ICIs durably but many others show resistance [28]. Specifically, we annotated immune cells in the melanoma scRNA-seq datasets [23, 29] using our classifiers separately for each patient sample and ordered samples based on the treatment response, with the results shown in Fig. 6a, b. We used the percentage of cell type in each tumor sample as it was more informative and meaningful than using absolute cell numbers. It is notable that untreated and NoInfo samples likely include both ICI-resistant and ICI-sensitive tumors.

In comparing samples from resistant tumors to untreated tumors, we found interestingly that there are samples with high prevalence of NK in untreated tumors (Mel53, Mel81, and Mel82) while no samples in resistant tumors have a high prevalence of NK cells. The mentioned untreated tumors also have no or very low number of Th2 cells in their populations. In addition, untreated tumors have a more uniform distribution of immune cell types in contrast to ICI-resistant ones, which could reflect a therapeutic bias in immune cell prevalence in the tumor microenvironment due to ICI treatment.

Next, we combined the annotation data from both classifiers and applied PCA and clustering analysis, as shown in Fig. 6, panels c and d. Using scrambled data to determine principal components and their associated eigenvalues that are not generated by random chance (i.e., a negative control), we kept the first and second principal components that capture 68% and 21% of the total variance, respectively, and neglected other components that fell below the negative control of 8.4%. As it shown in Fig. 6c, resistant samples mainly located in lowest value of second principal component (PC2). Upon closer inspection of the cell loadings within the eigenvectors, the low values of PC2 correspond to a low prevalence of Mϕ or high percentage of B cells. In addition, based on the first principal component (PC1), resistant samples have either the lowest values of PC1 (Mel74, Mel75, Mel58, Mel 78), which correspond to higher than average prevalence of CD8+ T cells, or the highest values of PC1 (Mel60, Mel72, Mel94), which show a higher than average prevalence of B cells.

In hierarchical clustering, the optimal number of clusters was selected based on calculation of different cluster indices using the NbClust R package [30] which mainly identified two or three clusters as the optimal number. In considering three groupings of the hierarchical clustering results shown in Fig. 6d, seven out of eight ICI-resistant samples clustered in first two clusters while the third cluster mainly contained untreated samples. The comparison of results from PCA and clustering analyses shows that the first cluster contained samples with extreme low value of PC1 which itself divided into two groups; one with extreme low value of PC2 and the other with higher amount of PC2. The second cluster located in highest amount of PC1 and lowest amount of PC2. All remained samples were clustered as third group, which were predominantly untreated samples. The difference in clustering suggests dissimilarities between ICI-resistant and untreated samples and the possibility of having ICI-sensitive tumors in untreated samples. D

Developing gene signatures

While classifiers are helpful for annotating scRNA-seq data as the transcriptomic signature corresponds to a

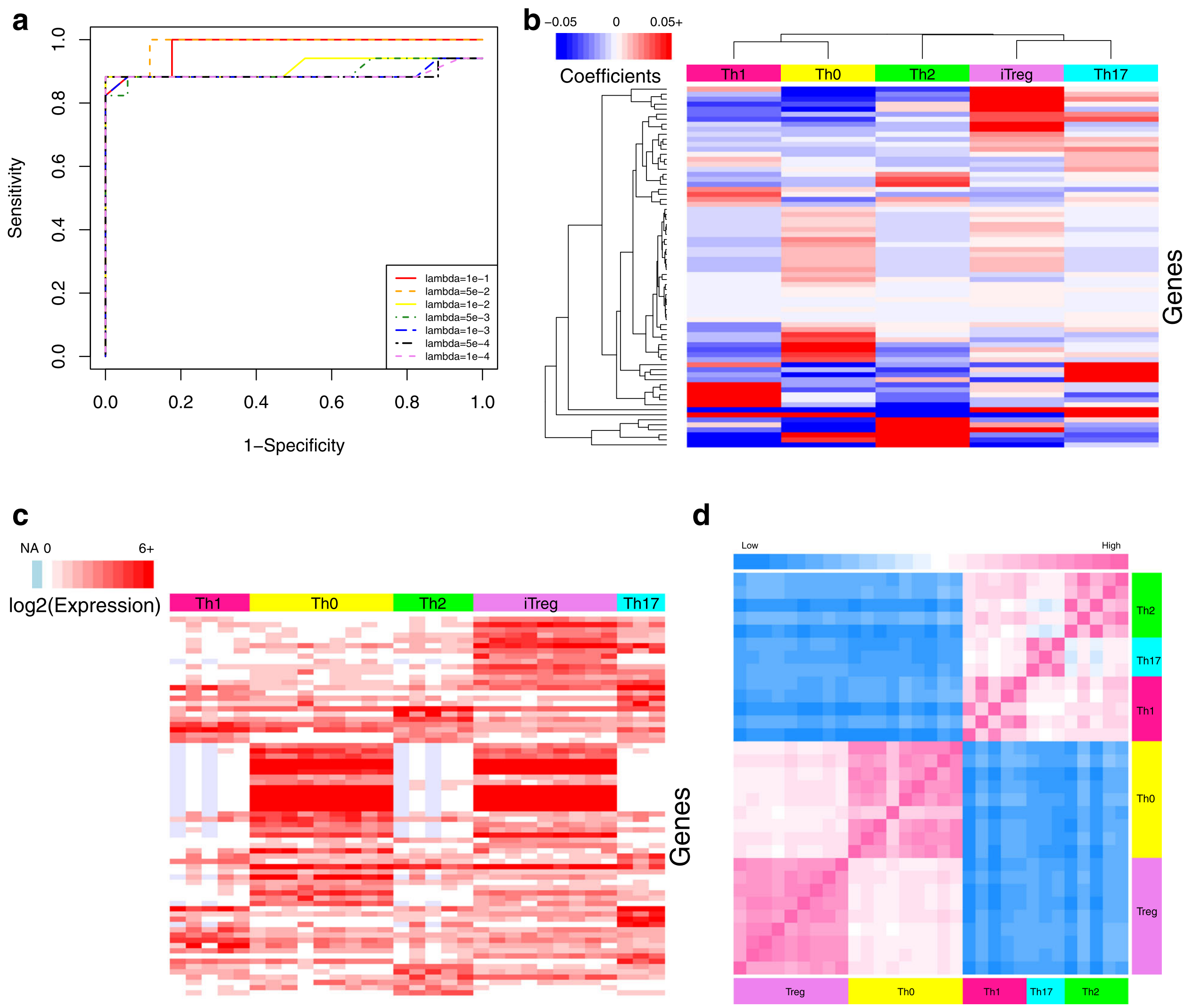

Fig. 5 Development of T helper cell classifier and similarity heatmaps **a** ROC curve for the T helper cell classifier was calculated using the indicated lambda values (shown in different colors and line styles) and 10-fold cross validation. The lambda value that maximized the AUC value was used for subsequent calculations. Elastic-net logistic regression to discriminate among five T helper cell types, where the value of the non-zero coefficients (panel **b**), expression levels (panel **c**), and similarity map (panel **d**) for the 72 genes included in the classifier are indicated by color bars for each panel. In panel **b**, blue to red color scheme indicates coefficients ranging from negative to positive values. Ordering of the genes is the same in panels **b** and **c**. In panel **c**, light blue indicates missing values and the intensity of red color (white/red color scale on the top-left) indicates the log base 2 expression level. A color bar on top of this panel was used to separate samples of each cell type. Panel **d** illustrates the similarity between samples calculated using an euclidean distance matrix based on the same 72 genes, where the color indicates the distance (pink: high similarity/low distance; blue: low similarity/high distance). Color bar on the top/side of the heatmap indicates the cell type of origin

single cell, gene signatures are commonly used to determine the prevalence of immune cell subsets within transcriptomic profiles of bulk tissue samples using deconvolution methods, called in silico cytometry [31]. Leveraging the classifier results, we generated corresponding gene signatures using binary elastic-net logistic regression. Specifically, classifier genes with non-zero coefficients were used as initial features of the models, which were then regressed to the same training and testing datasets as used for developing the classifiers. Lambda values were selected for each immune and T helper cell subset based on similar method of lambda selection for classifiers and their values and corresponding AUC are shown in Additional file 2: Table S2. Finally, all generated signatures are summarized in Additional file 3: Table S3.

We visualized the expression levels of the remaining set of genes, which at least occur in one gene signature, in Fig. 7. The expression of genes retained in immune cell signatures (Fig. 7a) and T helper cell signatures (Fig. 7b) were clustered by similarity in expression (rows) and by similarity in sample (columns). For both immune and T helper cell subsets, samples of

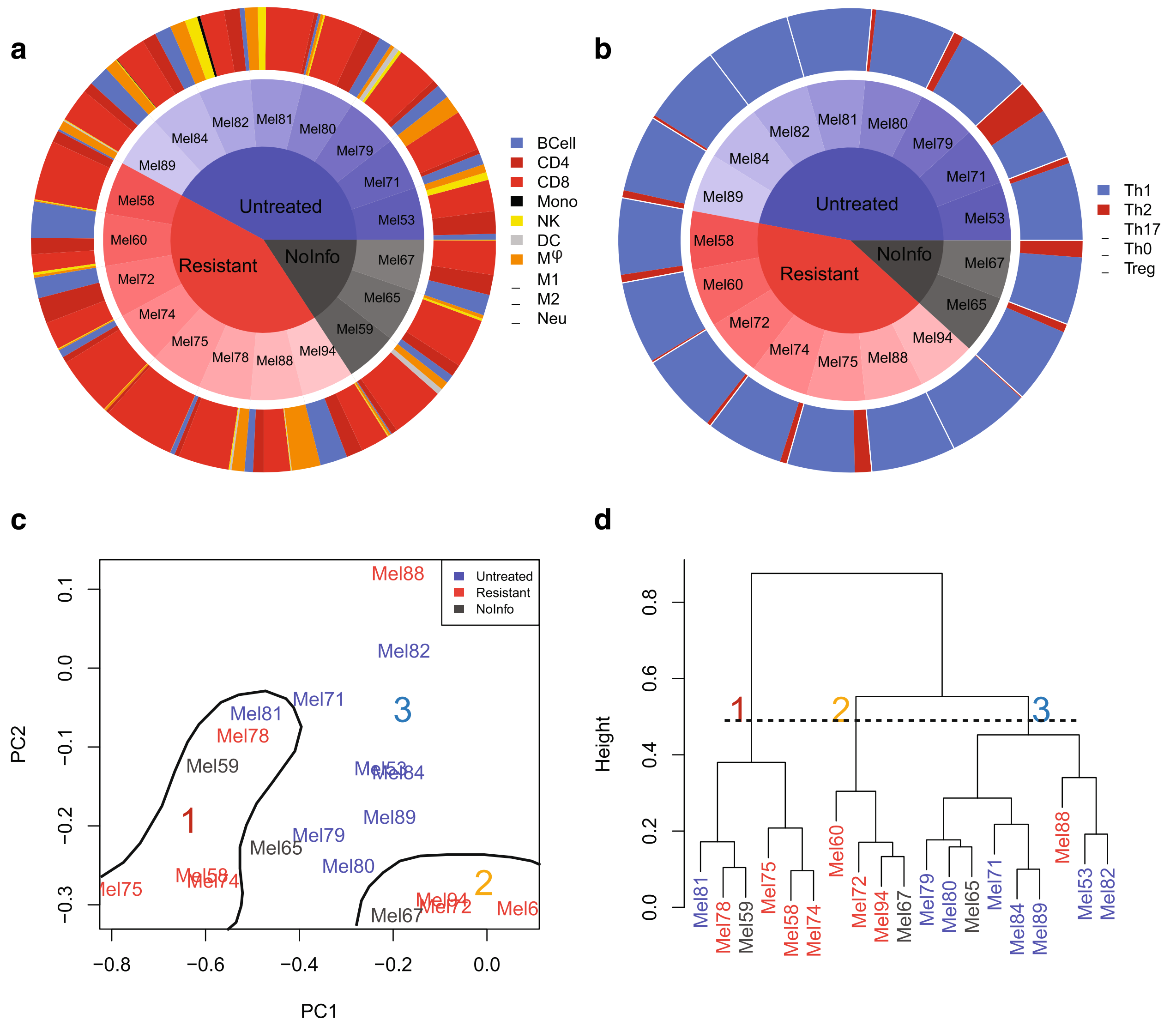

Fig. 6 Annotation of scRNA-seq results from melanoma dataset stratified by patient treatment status. Treatment status of patients diagnosed with melanoma was stratified based on their response to ICIs ([23, 29]). **a** The distribution in immune cell annotations and **b** T helper cell annotations based on scRNA-seq data were separated into samples obtained from ICI-resistant tumors, untreated tumors, and tumors reported in melanoma data without information about treatment status. Distributions are shown based on the percentage of all immune cells measured for each patient. Cell annotations were based on immune cell classifier and T helper cell classifier results. **c** PCA analysis was applied to the data obtained from both classifiers and the results for the first and second principal components were plotted. Red, blue, and grey colors indicate resistant, untreated and NoInfo (samples that have no information about their treatment status in the reference works) tumors, respectively. **d** Samples were hierarchically clustered based on the percentages of the nine immune cells and five T helper cells and same coloring applied to show tumor types

same cell type were mainly clustered together. The only exception is for macrophages (Mϕ and M2) which can be attributed to high biological similarity and a low number of technical replicates for these cell types.

In general, the gene sets generated from the logistic regression model performed well with far fewer requisite genes in the testing set, a desirable result for a gene set intended to be used for immunophenotyping. In Fig. 8, the results of the benchmarking are shown separated by comparative gene set. Both the CIBERSORT and Single-Cell derived gene sets contain an average of 64 and 135 genes, respectively, while the logistic regression gene set contains an average of just 19. The new logistic regression gene set performed comparably to the existing contemporary gene sets and far exceeded the performance of the manually curated gene set used previously [6]. The benchmarking results indicate that the logistic regression gene sets are an improvement in efficacy over compact gene sets, such as those that are manually annotated or

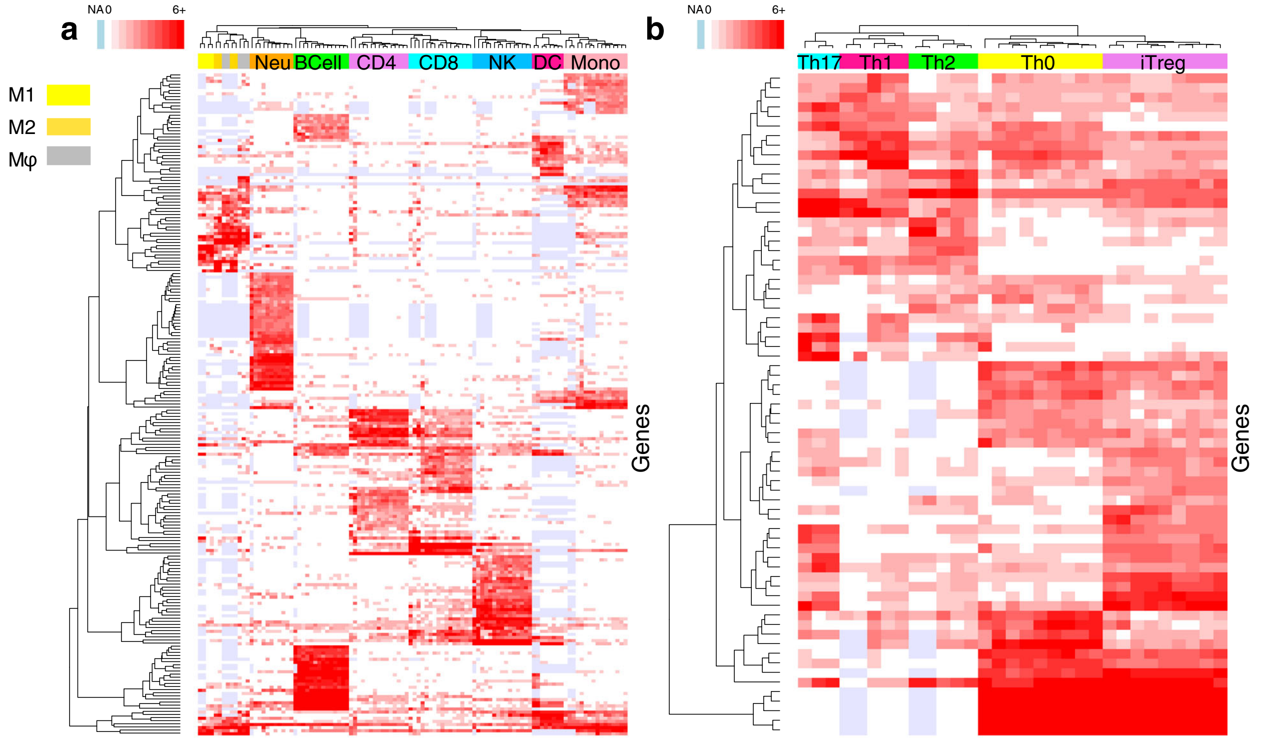

Fig. 7 Heatmaps of the expression levels for the final list of genes created by gene signatures. The expression of genes retained in immune cell signatures (panel **a**) and T helper cell signatures (panel **b**) were clustered by similarity in expression levels (rows) and by similarity in samples (columns). The color bar at the top indicates the samples cell type. Light blue shows missing values and the intensity of red color (white/red color scale on the top-left color bar) indicates the log base 2 expression level in both panels

hand-picked. Meanwhile, the logistic regression gene sets also demonstrate an optimization of broader gene sets that contain too many genes for deep specificity when used in further analysis. The inclusion of too many genes in a set can dilute the real data across a constant level of noise, while including too few lacks the power to draw conclusions with high confidence. The logistic regression gene sets demonstrate a balance of these two issues through its highly refined selection of genes that can be fine-tuned using its lambda parameter.

Discussion

Recent developments in RNA sequencing enable a high fidelity view of the transcriptomic landscape associated with host immune response. Despite considerable progress in parsing this landscape using gene signatures, gaps remain in developing unbiased signatures for individual immune cell types from healthy donors using high dimensional RNA-seq data. Here, we developed two classifiers - one for immune cell subsets and one for T helper cell subsets - using elastic-net logistic regression with cross validation. The features of these classifiers were used as a starting point for generating gene signatures that captured with fifteen binary elastic-net logistic regression models the most relevant gene sets to distinguish among different immune cell types without including too much noise.

Gene signatures in previous studies have been developed and used mainly as a base for deconvoluting the tumor microenvironment to find the presence of immune cells from bulk RNA measures. Therefore, as the first step, determining cell-specific gene signatures critically influences the results of deconvolution methods [32]. Newman et al. defined gene signatures for immune cells using two-sided unequal variances t-test as base matrix for CIBERSORT [8]. In another study, Li et al. in developing TIMER, generated gene signatures for six immune cell types with selecting genes with expression levels that have a negative correlation with tumor purity [9]. More recently, Racle et al. developed a deconvolution tool based on RNA-seq data (EPIC) by pre-selecting genes based on ranking by fold change and then selected genes by manually curating and comparing the expression levels in blood and tumor microenvironment [10]. Finally, quanTIseq (the most recently developed tool for deconvolution) was developed for RNA-seq data based on the gene signatures generated by quantizing the expression levels into different bins and selecting high quantized genes for each cell type that have low or medium expression in other cell types [7]. Although all methods obtained high accuracy

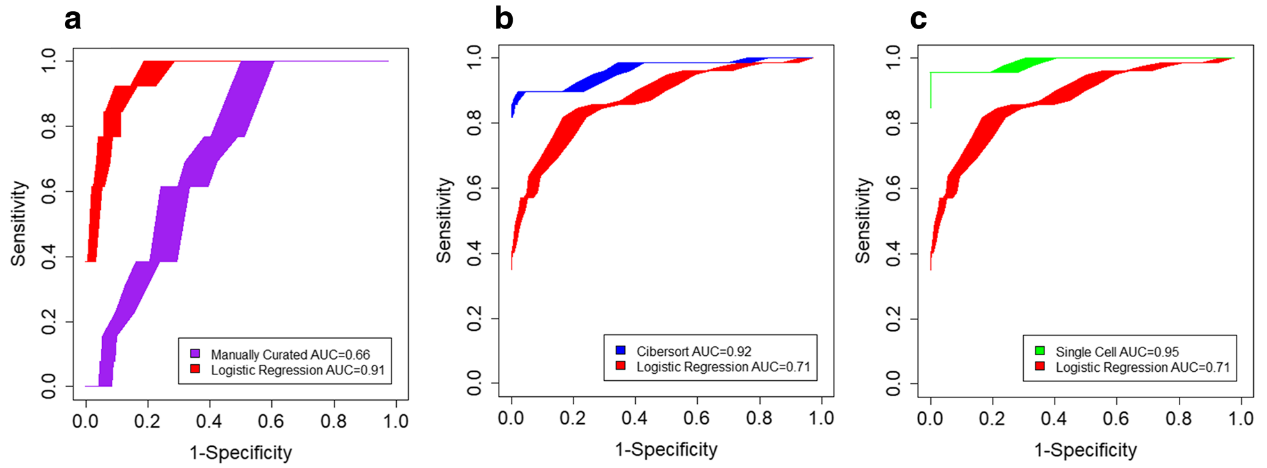

Fig. 8 Benchmarking ROC performance curves. ROC curves to illustrate relative performance between logistic regression gene set and the manually curated (Panel **a**), CIBERSORT (Panel **b**), and single cell gene sets (Panel **c**). The logistic regression gene set's performance is shown in red. Shaded regions are 95% confidence intervals about the average ROC curve simulated from 1000 iterations

based on their developed signatures, a more rigorous and unbiased gene signature developed by RNA-seq data and precise feature selection methods can further improve the accuracy and validate the process for downstream analyses.

In addition, to identify cell types based on their transcriptome, clustering techniques have been used in many studies [33, 34]. However, there are high variability levels of gene expression even in samples from the same cell type. Moreover, transcriptomics data has high dimensions (tens of thousands) and this is too complicated for clustering techniques as only few number of genes are discriminative. To overcome these problems some studies used supervised machine learning methods like Support Vector Machine (SVM) [35, 36]. However, to the best of our knowledge, this paper is the first to apply two-step regularized logistic regression on RNA-seq transcriptomic of immune cells. This method increases the chance to capture the most discriminative set of genes for each cell type based on the power of an elastic-net [22]. In addition, using a two-step elastic net logistic regression enabled eliminating the most irrelevant genes while keeping the highest number of possible significant genes in the first step and more deeply selecting among them in the second step to generate robust gene signatures for immune cells.

Moreover, contemporary methods have only considered a limited number of immune cell types, and specifically T helper subsets as individual cell types have been neglected [23, 24, 29] in comprehensive studies. Therefore, the other novel aspect of this study is the separation of models for immune cells and T helper cells and development of gene signatures for a large number of immune cell types (fifteen different immune cell types) including different T helper cell subsets. The ability to identify a greater number of immune cell types enables studying immune system in different diseases in more depth. As we used publicly available RNA-seq datasets for immune cells and T helper cells, we acknowledge that our developed classifiers and gene signatures may be still constrained by the limited number of samples specifically for T helper cells. As more data describing the transcriptome of immune cells will become accessible, one can update the classifiers and gene signatures. Despite the limited number of samples used in the approach, the developed classifiers can even be applied to completely untouched and large datasets [23, 24] that have been generated using scRNA-Seq technology which creates noisier data.

Conclusions

Here, we developed an immune cell classifier and classifier for T helper cell subsets along with gene signatures to distinguish among fifteen different immune cell types. Elastic-net logistic regression was used to generate classifiers with 10-fold cross-validation after normalizing and filtering two separate RNA-seq datasets that were generated using defined homogeneous cell populations. Subsequently, we generated gene signatures using a second step of binary regularized logistic regression applied to the RNA-seq data using previously selected classifier genes. As an external validation, the resulting classifiers accurately identified the type of immune cells in scRNA-seq datasets. Our classifiers and gene signatures can be considered for different downstream applications. First, the classifiers may be used to detect the type of immune cells in under explored bulk tissue samples profiled using RNA-seq and to verify the identity of immune cells annotated with low confidence. Second, the gene signatures could be used to study tumor micro-environments and the inter-dependence of immune response with cancer cell phenotypes, which is emerging to be an important clinical question.

Methods

Data acquisition

RNA-seq datasets for 15 different immune cell types including T helper cells, were obtained from ten different studies [37–46], which were publicly accessible via the *Gene Expression Omnibus* [47]. The list of samples is provided as Additional file 4: Table S4. The cell types were divided into two groups: immune cells that include B cells, CD4+ and CD8+ T cells, monocytes (Mono), neutrophils (Neu), natural killer (NK) cells, dendritic cells (DC), macrophage (Mϕ), classically (M1) and alternatively (M2) activated macrophages, and the T helper cells that include Th1, Th2, Th17, Th0, and Regulatory T cells (Treg). The goal was to train the gene selection model on immune cell types, and CD4+ T cell subsets (T helper cells), separately. If these two groups of cells are analyzed together, many of the genes that potentially could be used to discriminate among T helper cell subsets might be eliminated as they overlap with genes associated with CD4+ T cells.

In short, a total of 233 samples were downloaded and divided into two sets of 185 and 48 samples, for immune cells and T helper cells, respectively. Moreover, immune cell samples were further divided into 108 training and 77 testing samples. Training and testing numbers for T helper samples were 31 and 17, respectively. Training and testing data include samples from all studies. For a verification dataset, scRNA-seq data derived from CD45+ cell samples obtained from breast cancer [24] and melanoma [23] were used with GEO accession numbers of GSE75688 and GSE72056, respectively.

Data normalization

The expression estimates provided by the individual studies were used, regardless of the underlying experimental and data processing methods (Additional file 4: Table S4). For developing individual gene signatures and cell classification models, we did not use raw data due to sample heterogeneity such as different experimental methods and data processing techniques used by different studies as well as differences across biological sources. Rather, we applied a multistep normalization process before training models. To eliminate obvious insignificant genes from our data, for immune cell samples, genes with expression values higher than or equal to five counts, in at least five samples were kept, otherwise, they were eliminated from the study. However, for T helper samples, due to fewer number of samples, four samples with values higher than or equal to five counts were enough to be considered in the study. After first step of filtering, the main normalization step was used to decrease dependency of expression estimates to transcript length and GC-content [48, 49]. For all four sets of samples, including training and testing samples for immune cells and for T helper cells, expression estimates were normalized separately by applying *withinLaneNormalization* and *betweenLaneNormalization* functions from EDASeq package [50] in the R programming language (R 3.5.3), to remove GC-content biases and between-lane differences in count distributions [50]. After normalization, the second step of filtration, which was similar to the first step, was applied to eliminate genes with insignificant expression.

Missing values

In contrast to previous studies that only considered intersection genes [51] and to avoid deleting discriminative genes, we kept genes with high expression as much as possible. However, for most of genes, values for some samples were not reported. Hence, to deal with these missing values, we used an imputation method [52] and instead of mean imputation we set a dummy constant since mean imputation in this case is not meaningful and can increase error. Specifically, we generated a training set for each group of cell types, by duplicating the original training set 100 times and randomly eliminating ten percent of expression values. We next set -1 for all these missing values (both original missing values and those we eliminated) as a dummy constant because all values are positive and it is easier for the system to identify these values as noise. This approach makes the system learn to neglect a specific value (-1) and treat it like noise, instead of learning it as a feature of the samples.

Classifier training and testing

Considering the few number of training samples in comparison with the high dimensions (15453 genes in immune cell samples and 9146 genes in the T helper samples) and to avoid both over fitting the model and adding noise to the prediction model, we used regularization with logistic regression to decrease the total number of genes and select the most discriminative set of genes. To perform gene selection, we trained a lasso-ridge logistic regression (elastic-net) model, which automatically sets the coefficients of a large number of genes to zero and prunes the number of genes as features of the classifier. We cross-validated the model by implementing cv.glmnet function with nfold=10 from glmnet package [21] in R programming language, using training sets for both groups of cell types. We normalized the gene expression values using a log2 transform over training sets to decrease the range of values that can affect the performance of the model (log2(counts+1)). In order to find the optimal number of genes, we tried seven different lambdas and tested the results over the testing samples (*cv.glmnet(family="multinomial", alpha=0.93, thresh=1e-07, lambda=c(0.1, 0.05, 0.01, 0.005, 0.001, 0.0005, 0.0001), type.multinomial="grouped", nfolds=10)*). To select the optimal value for lambda, True-Negative samples were generated using a bootstrapping approach that randomly samples testing datasets to create a synthetic dataset with similar size and complexity but without underlying biological correlation, then we generated ROC curves and considered original testing datasets as True-Positive samples.

Developing gene signatures

Genes selected by the classifier models were used as initial sets to build gene signatures. In this case, we trained a new binary elastic-net model for each cell type by considering a certain cell type as one class and all other cell types as another class. The training and testing samples used to build gene signatures were the training and testing samples used in developing the classifiers with the difference being that they only contained the selected genes. Similar steps including dealing with missing values, applying log2 and visualization by ROC to select optimal number of genes were applied for each cell type. This two-step gene selection approach has the advantage that it eliminates a large number of undiscriminating genes at the first and finally select few number of genes for each cell type.

Benchmarking

Fisher exact testing was used for each gene set to characterize true and systematically scrambled data as a measure of performance of the gene set as a means of distinguishing between cell subtypes. In order to establish negative control values for determining specificity, a bootstrapping approach was used [53], where data was scrambled by randomly resampling with replacement expression values by gene as well as by patient to create a synthetic dataset with a similar size and complexity of the original dataset. The threshold for expression binarization for Fisher exact testing was selected based on gene expression histograms of the data to separate the measured expression from background noise levels, with 2.48 being used as the threshold (after log2 normalization). One-thousand iterations (N_{boot}) were processed and compiled in order to produce ROC curves with 95% confidence intervals shaded about the averaged ROC curve for each gene set's performance. A bootstrapping approach for generating a negative control sample is appropriate when a sufficiently large bootstrap sample (i.e., $N_{boot} \geq 1000$) and the original dataset is sufficiently diverse (i.e., $N_{data} \geq 30$) [54]. The tested gene sets were the logistic regression gene set, the CIBERSORT gene set [8], the single cell gene set [29], and the manually curated gene set that had been used previously [6].

Additional files

Additional file 1: Table S1. Coefficients of immune cell classifier and T helper cell classifier. Coefficients of immune cell classifier were located in the first sheet and coefficients of T helper cells were located in the second sheet. (XLSX 102 kb)

Additional file 2: Table S2. Lambda selection by AUC values. Lambdas with corresponding calculated AUC. The final column shows the selected lambdas. (XLSX 79 kb)

Additional file 3: Table S3. Genes in developed gene signature for immune and T helper cells. Yellow boxes show genes with negative impact in possibility of being related cell type. (XLSX 14 kb)

Additional file 4: Table S4. Data information used in training models. The second sheet shows the names that were used in creating the datasets. (XLSX 78 kb)

Abbreviations

AUC: Area under the ROC curve; CNV: Copy number variation; ICI: Immune checkpoint inhibitor; PCA: Principal component analysis; ROC: Receiver-operator curves; scRNA-seq: Single-cell RNA-seq; SVM: Support vector machine

Authors' contributions

Designed study: AT and DK; performed analyses and interpreted results: AT, PG, and DK; and drafted initial manuscript: AT, PG, and DK. All authors edited and approved the final version of the manuscript.

Funding

This work was supported by the National Science Foundation (NSF) (CBET-1644932 to DJK) and the National Cancer Institute (NCI) (R01CA193473 to DJK). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NCI, the National Institutes of Health, or the National Science Foundation. The funding bodies played no roles in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Availability of data and materials

The datasets supporting the conclusions of this article are available in Gene Expression Omnibus repository [https://www.ncbi.nlm.nih.gov] with the following GEO accession numbers: GSE60424, GSE64655, GSE36952, GSE84697, GSE74246, GSE70106, GSE55536, GSE71645, GSE66261, GSE96538, GSE75688, GSE72056. R scripts used in the analyses can be found on GitHub [https://github.com/KlinkeLab/ImmClass2019].

Ethics approval and consent to participate

The results described in this manuscript consist of secondary analyses of existing data and was determined by the West Virginia University IRB to qualify for an exemption from human subject research under U.S. HHS regulations 45 CFR 46.101(b)(4).

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 9 May 2019 Accepted: 15 July 2019
Published online: 22 August 2019

References

1. Carmona SJ, Teichmann SA, Ferreira L, Macaulay IC, Stubbington MJ, Cvejic A, Gfeller D. Single-cell transcriptome analysis of fish immune cells provides insight into the evolution of vertebrate immune cell types. Genome Res. 2017;27(3):451–461. https://doi.org/10.1101/gr.207704.116.
2. Bendall SC, Simonds EF, Qiu P, El-ad DA, Krutzik PO, Finck R, Bruggner RV, Melamed R, Trejo A, Ornatsky OI, et al. Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. Science. 2011;332(6030):687–96.

3. Shay T, Kang J. Immunological genome project and systems immunology. Trends Immunol. 2013;34(12):602–9.
4. Kinter AL, Hennessey M, Bell A, Kern S, Lin Y, Daucher M, Planta M, McGlaughlin M, Jackson R, Ziegler SF, et al. Cd25+ cd4+ regulatory t cells from the peripheral blood of asymptomatic hiv-infected individuals regulate cd4+ and cd8+ hiv-specific t cell immune responses in vitro and are associated with favorable clinical markers of disease status. J Exp Med. 2004;200(3):331–43.
5. Vegh P, Haniffa M. The impact of single-cell rna sequencing on understanding the functional organization of the immune system. Brief Funct Genomics. 2018;17(4):265–272. https://doi.org/10.1093/bfgp/ely003.
6. Kaiser JL, Bland CL, Klinke DJ. Identifying causal networks linking cancer processes and anti-tumor immunity using bayesian network inference and metagene constructs. Biotechnol Prog. 2016;32(2):470–9.
7. Finotello F, Mayer C, Plattner C, Laschober G, Rieder D, Hackl H, Krogsdam A, Posch W, Wilflingseder D, Sopper S, et al. quantiseq: quantifying immune contexture of human tumors. bioRxiv. 2017223180.
8. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, Hoang CD, Diehn M, Alizadeh AA. Robust enumeration of cell subsets from tissue expression profiles. Nat Methods. 2015;12(5):453.
9. Li B, Severson E, Pignon J-C, Zhao H, Li T, Novak J, Jiang P, Shen H, Aster JC, Rodig S, et al. Comprehensive analyses of tumor immunity: implications for cancer immunotherapy. Genome Biol. 2016;17(1):174.
10. Racle J, de Jonge K, Baumgaertner P, Speiser DE, Gfeller D. Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data. Elife. 2017;6:26476.
11. Kidd BA, Peters LA, Schadt EE, Dudley JT. Unifying immunology with informatics and multiscale biology. Nat Immunol. 2014;15(2):118.
12. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci. 2005;102(43):15545–50.
13. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP. Molecular signatures database (msigdb) 3.0. Bioinformatics. 2011;27(12):1739–40.
14. Zheng C-H, Chong Y-W, Wang H-Q. Gene selection using independent variable group analysis for tumor classification. Neural Comput Appl. 2011;20(2):161–70.
15. Wu M-Y, Dai D-Q, Shi Y, Yan H, Zhang X-F. Biomarker identification and cancer classification based on microarray data using laplace naive bayes model with mean shrinkage. IEEE/ACM Trans Comput Biol Bioinforma. 2012;9(6):1649–62.
16. Cui Y, Zheng C-H, Yang J, Sha W. Sparse maximum margin discriminant analysis for feature extraction and gene selection on gene expression data. Comput Biol Med. 2013;43(7):933–41.
17. Algamal ZY, Lee MH. Regularized logistic regression with adjusted adaptive elastic net for gene selection in high dimensional cancer classification. Comput Biol Med. 2015;67:136–45.
18. Liang Y, Liu C, Luan X-Z, Leung K-S, Chan T-M, Xu Z-B, Zhang H. Sparse logistic regression with a l 1/2 penalty for gene selection in cancer classification. BMC Bioinformatics. 2013;14(1):198.
19. Bielza C, Robles V, Larrañaga P. Regularized logistic regression without a penalty term: An application to cancer classification with microarray data. Expert Syst Appl. 2011;38(5):5110–8.
20. Cawley GC, Talbot NL. Gene selection in cancer classification using sparse logistic regression with bayesian regularization. Bioinformatics. 2006;22(19):2348–55.
21. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw. 2010;33(1):1.
22. Zou H, Hastie T. Regularization and variable selection via the elastic net. J Royal Stat Soc: Ser B (Stat Methodol). 2005;67(2):301–20.
23. Tirosh I, Izar B, Prakadan SM, Wadsworth MH, Treacy D, Trombetta JJ, Rotem A, Rodman C, Lian C, Murphy G, et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell rna-seq. Science. 2016;352(6282):189–96.
24. Chung W, Eum HH, Lee H-O, Lee K-M, Lee H-B, Kim K-T, Ryu HS, Kim S, Lee JE, Park YH, et al. Single-cell rna-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. Nat Commun. 2017;8:15081.
25. Caligiuri MA. Human natural killer cells. Blood. 2008;112(3):461–9.
26. Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV, Clark NR, Ma'ayan A. Enrichr: interactive and collaborative html5 gene list enrichment analysis tool. BMC Bioinformatics. 2013;14(1):128.
27. Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, Koplev S, Jenkins SL, Jagodnik KM, Lachmann A, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res. 2016;44(W1):90–7.
28. Sharma P, Hu-Lieskovan S, Wargo JA, Ribas A. Primary, adaptive, and acquired resistance to cancer immunotherapy. Cell. 2017;168(4):707–23.
29. Jerby-Arnon L, Shah P, Cuoco MS, Rodman C, Su M-J, Melms JC, Leeson R, Kanodia A, Mei S, Lin J-R, et al. A cancer cell program promotes t cell exclusion and resistance to checkpoint blockade. Cell. 2018;175(4):984–97.
30. Charrad M, Ghazzali N, Boiteau V, Niknafs A, Charrad MM. Package 'nbclust'. J Stat Softw. 2014;61:1–36.
31. Newman AM, Steen CB, Liu CL, Gentles AJ, Chaudhuri AA, Scherer F, Khodadoust MS, Esfahani MS, Luca BA, Steiner D, Diehn M, Alizadeh AA. Determining cell type abundance and expression from bulk tissues with digital cytometry. Nat Biotechnol. 2019;37(7):773–782. https://doi.org/10.1038/s41587-019-0114-2.
32. Finotello F, Trajanoski Z. Quantifying tumor-infiltrating immune cells from transcriptomics data. Cancer Immunol Immunother. 2018;67(7):1031–40.
33. Xu C, Su Z. Identification of cell types from single-cell transcriptomes using a novel clustering method. Bioinformatics. 2015;31(12):1974–80.
34. Grün D, Lyubimova A, Kester L, Wiebrands K, Basak O, Sasaki N, Clevers H, van Oudenaarden A. Single-cell messenger rna sequencing reveals rare intestinal cell types. Nature. 2015;525(7568):251.
35. Hu Y, Hase T, Li HP, Prabhakar S, Kitano H, Ng SK, Ghosh S, Wee LJK. A machine learning approach for the identification of key markers involved in brain development from single-cell transcriptomic data. BMC Genomics. 2016;17(13):1025.
36. Yao F, Zhang C, Du W, Liu C, Xu Y. Identification of gene-expression signatures and protein markers for breast cancer grading and staging. PloS ONE. 2015;10(9):0138213.
37. Linsley PS, Speake C, Whalen E, Chaussabel D. Copy number loss of the interferon gene cluster in melanomas is linked to reduced t cell infiltrate and poor patient prognosis. PloS ONE. 2014;9(10):109760.
38. Hoek KL, Samir P, Howard LM, Niu X, Prasad N, Galassie A, Liu Q, Allos TM, Floyd KA, Guo Y, et al. A cell-based systems biology assessment of human blood to monitor immune responses after influenza vaccination. PloS ONE. 2015;10(2):0118528.
39. Beyer M, Mallmann MR, Xue J, Staratschek-Jox A, Vorholt D, Krebs W, Sommer D, Sander J, Mertens C, Nino-Castro A, et al. High-resolution transcriptome of human macrophages. PloS ONE. 2012;7(9):45466.
40. Şenbabaoğlu Y, Gejman RS, Winer AG, Liu M, Van Allen EM, de Velasco G, Miao D, Ostrovnaya I, Drill E, Luna A, et al. Tumor immune microenvironment characterization in clear cell renal cell carcinoma identifies prognostic and immunotherapeutically relevant messenger rna signatures. Genome Biol. 2016;17(1):231.
41. Corces MR, Buenrostro JD, Wu B, Greenside PG, Chan SM, Koenig JL, Snyder MP, Pritchard JK, Kundaje A, Greenleaf WJ, et al. Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. Nat Genet. 2016;48(10):1193.
42. Kumar NA, Cheong K, Powell DR, da Fonseca Pereira C, Anderson J, Evans VA, Lewin SR, Cameron PU. The role of antigen presenting cells in the induction of hiv-1 latency in resting cd4+ t-cells. Retrovirology. 2015;12(1):76.
43. Zhang H, Xue C, Shah R, Bermingham K, Hinkle CC, Li W, Rodrigues A, Tabita-Martinez J, Millar JS, Cuchel M, et al. Functional analysis and transcriptomic profiling of ipsc-derived macrophages and their application in modeling mendelian disease. Circ Res. 2015;117(1):17–28. https://doi.org/10.1161/CIRCRESAHA.117.305860.
44. Kanduri K, Tripathi S, Larjo A, Mannerström H, Ullah U, Lund R, Hawkins RD, Ren B, Lähdesmäki H, Lahesmaa R. Identification of global regulators of t-helper cell lineage specification. Genome Med. 2015;7(1):122.
45. Spurlock III CF, Tossberg JT, Guo Y, Collier SP, Crooke III PS, Aune TM. Expression and functions of long noncoding rnas during human t helper cell differentiation. Nat Commun. 2015;6:6932.
46. Schmidt A, Marabita F, Kiani NA, Gross CC, Johansson HJ, Éliás S, Rautio S, Eriksson M, Fernandes SJ, Silberberg G, et al. Time-resolved transcriptome and proteome landscape of human regulatory t cell (treg) differentiation reveals novel regulators of foxp3. BMC Biol. 2018;16(1):47.

47. Edgar R, Domrachev M, Lash AE. Gene expression omnibus: Ncbi gene expression and hybridization array data repository. Nucleic Acids Res. 2002;30(1):207–10.
48. Oshlack A, Wakefield MJ. Transcript length bias in rna-seq data confounds systems biology. Biol Dir. 2009;4(1):14.
49. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of rna-seq data. Genome Biol. 2010;11(3):25.
50. Risso D, Schwartz K, Sherlock G, Dudoit S. Gc-content normalization for rna-seq data. BMC Bioinformatics. 2011;12(1):480.
51. Schwalie PC, Ordóñez-Morán P, Huelsken J, Deplancke B. Cross-tissue identification of somatic stem and progenitor cells using a single-cell rna-sequencing derived gene signature. Stem Cells. 2017;35(12): 2390–402.
52. García-Laencina PJ, Sancho-Gómez J-L, Figueiras-Vidal AR. Pattern classification with missing data: a review. Neural Comput Appl. 2010;19(2):263–82.
53. Efron B, Tibshirani R. An Introduction to the Bootstrap. London: Chapman and Hall; 1993.
54. Chernick MR. Bootstrap Methods: A Practitioner's Guide. New York: Wiley; 1999, pp. 150–1.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.