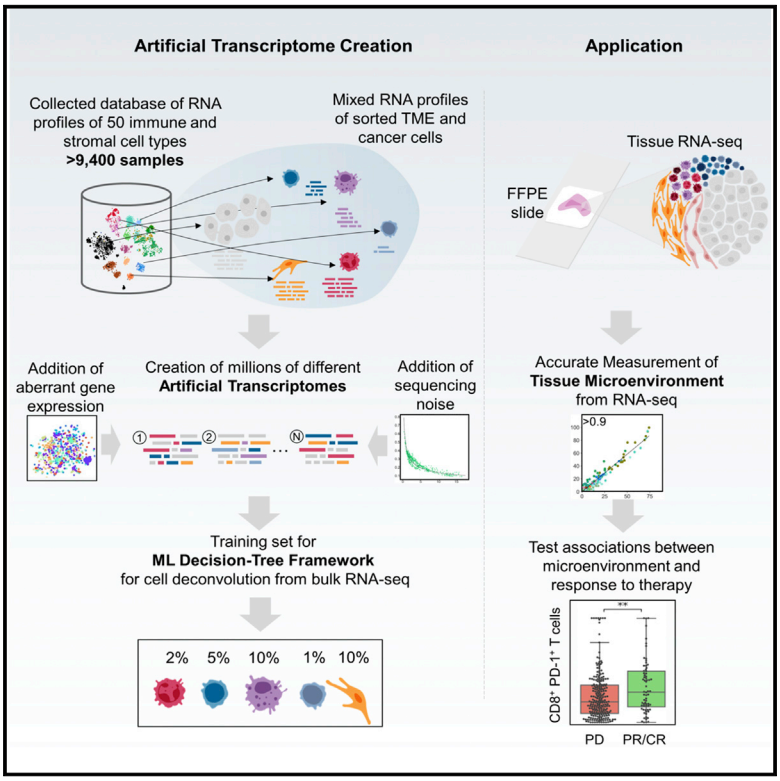
Cancer Cell

Article

Precise reconstruction of the TME using bulk RNA-seq and a machine learning algorithm trained on artificial transcriptomes

Graphical abstract



Highlights

- Development of the decision tree machine learning deconvolution algorithm Kassandra
- Creation of a broad collection of >9,400 tissue and blood sorted cell RNA profiles
- Accurate prediction of 51 unique cell subpopulations in tissue and blood
- TME reconstruction correlated PD-1-positive $CD8^+$ T cells with immunotherapy response

Authors

Aleksandr Zaitsev, Maksim Chelushkin, Daniiar Dyikanov, ..., Ravshan Ataullakhanov, Nathan Fowler, Alexander Bagaev

Correspondence

nfowler@mdanderson.org (N.F.), alexander.bagaev@bostongene.com (A.B.)

In brief

Zaitsev et al. show the accuracy and stability of TME and blood reconstruction in the recognition of 51 cell subpopulations with the decision tree machine learning deconvolution algorithm Kassandra. TME reconstruction of tumor tissues identified a correlation between PD-1-positive $CD8^+$ T cells and immunotherapy response in multiple cancers.

Zaitsev et al., 2022, Cancer Cell *40*, 879–894
August 8, 2022
https://doi.org/10.1016/j.ccell.2022.07.006

CellPress

Cancer Cell

CellPress

Article

Precise reconstruction of the TME using bulk RNA-seq and a machine learning algorithm trained on artificial transcriptomes

Aleksandr Zaitsev,[1] Maksim Chelushkin,[1] Daniiar Dyikanov,[1] Ilya Cheremushkin,[1] Boris Shpak,[1] Krystle Nomie,[1] Vladimir Zyrin,[1] Ekaterina Nuzhdina,[1] Yaroslav Lozinsky,[1] Anastasia Zotova,[1] Sandrine Degryse,[1] Nikita Kotlov,[1] Artur Baisangurov,[1] Vladimir Shatsky,[1] Daria Afenteva,[1] Alexander Kuznetsov,[1] Susan Raju Paul,[2] Diane L. Davies,[3] Patrick M. Reeves,[2] Michael Lanuti,[3] Michael F. Goldberg,[1] Cagdas Tazearslan,[1] Madison Chasse,[1] Iris Wang,[1] Mary Abdou,[1] Sharon M. Aslanian,[1] Samuel Andrewes,[1] James J. Hsieh,[4] Akshaya Ramachandran,[4] Yang Lyu,[4] Ilia Galkin,[1] Viktor Svekolkin,[1] Leandro Cerchietti,[5] Mark C. Poznansky,[2] Ravshan Ataullakhanov,[1] Nathan Fowler,[1,6,7,*] and Alexander Bagaev[1,*]

[1]BostonGene, Corp., 95 Sawyer Road, Waltham, MA 02453, USA
[2]The Vaccine and Immunotherapy Center, Massachusetts General Hospital, Boston, MA, USA
[3]Division of Thoracic Surgery, Massachusetts General Hospital, Boston, MA, USA
[4]Molecular Oncology, Division of Oncology, Department of Medicine, Washington University, St. Louis, MO, USA
[5]Division of Hematology and Medical Oncology, Weill Cornell Medicine, New York, NY, USA
[6]Department of Lymphoma and Myeloma, MD Anderson Cancer Center, 1515 Holcombe Blvd, Unit 429, Houston, TX 77030, USA
[7]Lead contact
*Correspondence: nfowler@mdanderson.org (N.F.), alexander.bagaev@bostongene.com (A.B.)
https://doi.org/10.1016/j.ccell.2022.07.006

SUMMARY

Cellular deconvolution algorithms virtually reconstruct tissue composition by analyzing the gene expression of complex tissues. We present the decision tree machine learning algorithm, Kassandra, trained on a broad collection of >9,400 tissue and blood sorted cell RNA profiles incorporated into millions of artificial transcriptomes to accurately reconstruct the tumor microenvironment (TME). Bioinformatics correction for technical and biological variability, aberrant cancer cell expression inclusion, and accurate quantification and normalization of transcript expression increased Kassandra stability and robustness. Performance was validated on 4,000 H&E slides and 1,000 tissues by comparison with cytometric, immunohistochemical, or single-cell RNA-seq measurements. Kassandra accurately deconvolved TME elements, showing the role of these populations in tumor pathogenesis and other biological processes. Digital TME reconstruction revealed that the presence of PD-1-positive $CD8^+$ T cells strongly correlated with immunotherapy response and increased the predictive potential of established biomarkers, indicating that Kassandra could potentially be utilized in future clinical applications.

INTRODUCTION

The tumor microenvironment (TME) plays an important role in disease progression and response to therapy. The TME regulates tumor survival, maintenance, growth (Hirata and Sahai, 2017), and immune surveillance (Galon et al., 2006). Elucidating the TME cellular composition and function of its varied cell populations is useful for optimizing more potent therapeutic modalities (Wei et al., 2018).

Although bulk RNA sequencing (RNA-seq) reveals the presence and quantity of all genes within a tumor and the TME at any given time, total RNA expression alone cannot identify the cellular origin of individual RNA molecules without cellular deconvolution. Primarily based on cell-type-specific gene expression profiles and linear regression algorithms to minimize the error of observed and expected expression, multiple deconvolution methods have previously been used to assess cell types (Aran et al., 2017; Becht et al., 2016; Finotello et al., 2019; Hao et al., 2019; Jew et al., 2020; Monaco et al., 2019; Nadel et al., 2021; Newman et al., 2015, 2019; Racle et al., 2017; Wang et al., 2019). Deep learning-based deconvolution methods have also been recently developed; however, these approaches often require retraining on data from single-cell RNA-seq (scRNA-seq) of the same tissue type (Torroja and Sanchez-Cabo, 2019) or paired flow cytometry data (Menden et al., 2020), limiting clinical utility.

In addition, these platforms are limited by their ability to accurately identify hierarchical subpopulations within complex cell mixtures with high precision and specificity (Liu et al., 2019a). The TME often comprises only a small fraction of the tumor and bulk RNA-seq reads; however, the precise identification of small TME cellular subsets, such as natural killer (NK) cells, is essential because they significantly impact therapeutic response

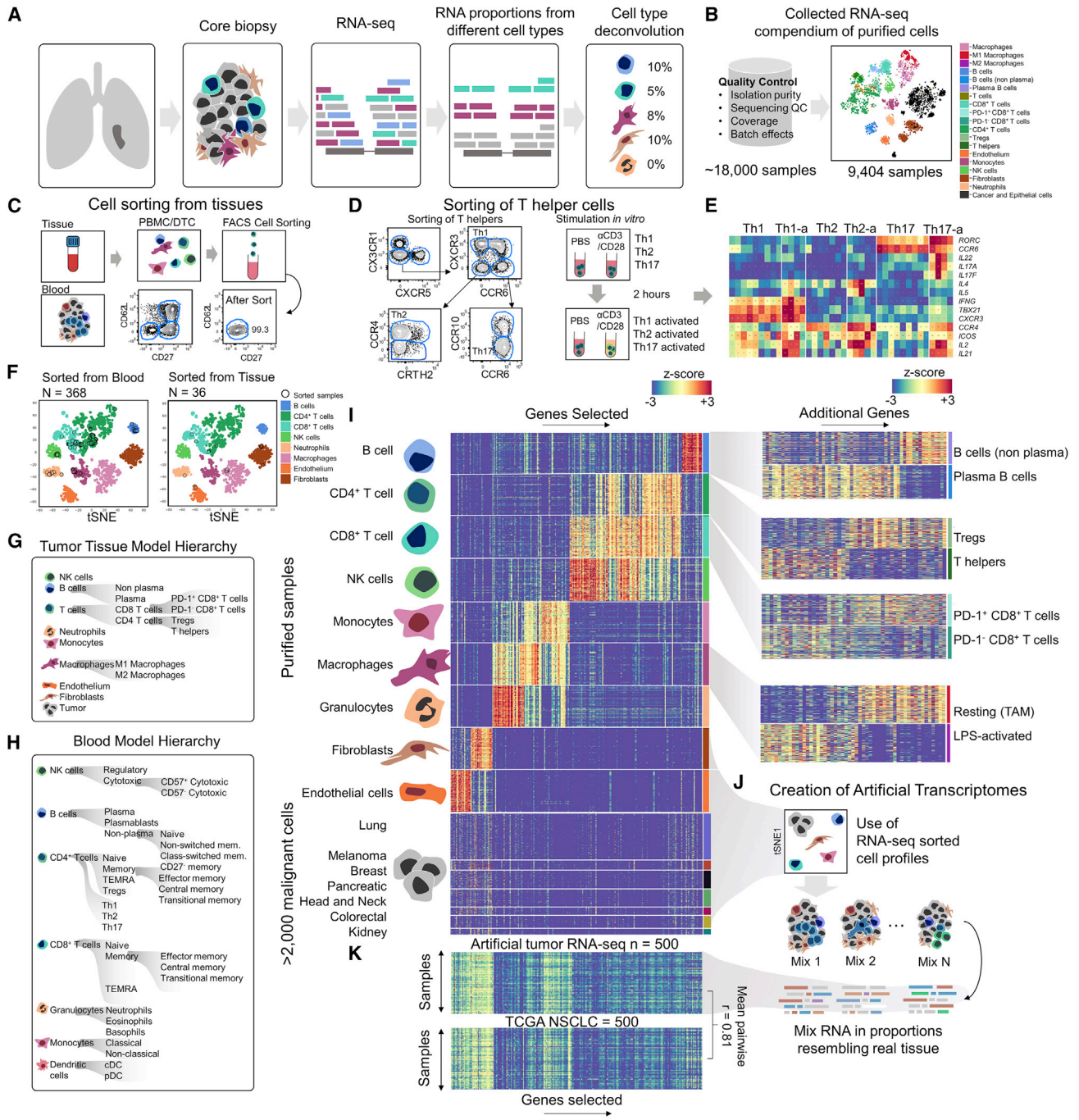


Figure 1. Defining RNA profiles of various sorted cell populations to artificially reconstruct tissues

(A) Schematic representing gene expression-dependent cell deconvolution.

(B) t-SNE of major cell types from RNA-seq samples after quality control (QC; n = 9,404).

(C) Graphic depicting blood and tissue cell sorting procedure.

(D) Schematic of sorting and stimulation of Th1, Th2, and Th17 cells.

(E) Heatmap of gene expression patterns according to different Th cell activation statuses.

(F) t-SNE plots showing overlap between the expression profiles of sorted cells from blood (n = 368) and tissue (n = 36) with the sorted cell compendium.

(G and H) Schematic of cell model trees developed for Kassandra-based deconvolution for tumor tissue (G) and blood (H).

(I) Heatmap of gene signatures associated with TME cell types for all RNA-seq samples in the database of sorted cell types. One row represents one RNA-seq sample (left). Heatmap of additional cell-type-specific gene signatures for cancer-relevant cell subpopulations (right).

(legend continued on next page)

and clinical outcome across diverse diseases. Addressing technical noise is essential during cellular deconvolution to accurately identify cell subsets from bulk RNA-seq (Ding et al., 2015; Rabadan et al., 2018).

Here, we describe the decision tree machine learning (ML) algorithm Kassandra developed for the deconvolution of cell proportions in tissue and blood on different hierarchical levels created based on the curation of a large homogeneously annotated resource of purified cell RNA profiles. The diverse and tissue-specific gene expression profiles of malignant cells were considered during the training process, increasing the stability of TME and blood reconstruction for comprehensive analysis of their impact on cancer biology and therapeutics.

RESULTS

Construction of a sorted cell RNA-seq compendium for artificial transcriptome creation

Current deconvolution algorithms utilizing RNA expression enable tissue cell composition to be determined based on the proportion of RNA sequences belonging to unique cell populations (Figure 1A). To address the complexities of cognate transcriptomic programs of cell subtypes (Figures S1A–S1D), we developed a decision tree algorithm, Kassandra, designed to accurately calculate the proportion of different cell subsets by determining the RNA fraction per cell type from RNA-seq within noncancerous and cancerous tissues. A collection of more than 18,000 bulk RNA-seq, covering numerous immune and stromal sorted cell populations and cancer cell lines, was curated using the GEO and ArrayExpress databases (Barrett et al., 2012). The raw RNA-seq datasets were combined, homogeneously annotated, and bioinformatically recalculated for comparable measurements of transcript expression within each cell type to reduce batch effects. After quality control, well-defined cell clusters were revealed, populating the Kassandra sorted cell compendium with purified RNA-seq samples (n = 9,404) of diverse immune and stromal cell populations, including malignant cells from 24 cancer types (n = 2,166) (Figures 1B, S2A, and S2B).

To deconvolve rare cell types, such as naive and memory T cell subsets, we FACS-sorted and sequenced 386 samples representing 41 subpopulations (Figures 1C and S3A–S3F). Clinically relevant T helper (Th) cells were divided into Th1, Th2, and Th17 functional phenotypes (Figures 1D and 1E), but sorting the functional phenotypes in their active states directly from tissue is complex. To circumvent this issue, we sorted Th subtypes from the blood of heathy donors and stimulated them *in vitro*, resulting in the RNA profiles of cells with active production of their designated cytokines, interleukin-2 (IL-2), IL-4, or IL-17 (Figures 1D and 1E), and enabling the use of "steady-state" and "active" samples of Th cells to train against intrinsic variability. Overall, the RNA profiles of cell types sorted from blood and cancer tissues were remarkably similar with the curated RNA profiles (Figure 1F), showing concordance of gene expression of these cell types across multiple tissues and datasets.

Final collected datasets and samples were annotated into 18 TME cell types (Figure 1G) and 41 populations present in blood (Figure 1H), a total of 51 unique populations, with RNA profiles for B cells, T cells, macrophages, NK cells, endothelial cells, and fibroblasts (Tables S1 and S2). Blood-derived cell types were divided into 16 $CD8^+$ and $CD4^+$ T cell, 7 B cell, 4 NK cell, 4 monocyte and dendritic cell, and 3 granulocyte subpopulations (Figure 1H). Cell-type-specific genes (Figures 1I and S1C; Tables S3 and S4) were selected by literature analysis, expression fold-change analysis between cells applied and correlation with our RNA-seq database, and feature importance analysis (Lundberg et al., 2018, 2020) based on ML models (STAR Methods). The selected genes were filtered to be predominantly expressed in non-malignant cells (Figure 1I).

With this comprehensive RNA collection of purified cell populations, we aimed to reconstruct tissue-like and blood-like bulk RNA profiles and imitate tissue heterogeneity by artificially mixing RNA from different purified cell subsets to create millions of artificial tumor transcriptomes for the training of Kassandra (Figure 1J). We hypothesized that expression values for multicellular tissue can be obtained by summing the gene expression of its individual cells (Zaitsev et al., 2019) to create artificial tumor RNA profiles by randomly combining RNA derived from sorted cells and tumor cell lines in proportions likely to be observed within actual tissue (STAR Methods; Figure 1J). We initially confirmed that we could develop artificial tumors after the creation of the sorted cell compendium and found that they were remarkably similar to the RNA profiles of true The Cancer Genome Atlas (TCGA) cancer types (n = 514; mean pairwise correlation value = 0.81; Figures 1K and S4A–S4C). The ability to create artificial transcriptomes enabled the development of a large-scale artificial tissue database containing the functional and phenotypic states of different cell subsets at specified proportions for Kassandra training.

Establishing TPM normalization for algorithm development

Kassandra used transcripts per million (TPMs) (Li et al., 2010; Wagner et al., 2012) as an expression unit. Often, variability in total expression belonging to short RNA transcripts strongly skews the TPM value distribution of genes of interest (Figure 2A). Noncoding RNA (e.g., microRNA and miscellaneous-RNA as previously described in the TCGA RNA pipeline) (George et al., 2017) (STAR Methods) and short transcripts of T cell receptor (TCR)- and B cell receptor (BCR)-coding genes, annotated in the transcriptome as corresponding to the V, D, or J regions, were excluded from TPM normalization (STAR Methods). Histone-coding and mitochondrial genes were excluded because of the uneven enrichment in different RNA extraction methods (e.g., poly(A) versus total RNA) (Newton et al., 2020). Unverified transcripts having low transcript support levels and transcripts with partially unknown coding sequences were also precluded

(J) Schematic representation of artificial transcriptomes from RNA-seq profiles of sorted cell populations.
(K) Heatmap comparing RNA-seq gene expression from 514 TCGA non-small cell lung carcinomas (NSCLC) and 514 artificially developed lung cancer transcriptomes.
See also Figures S1–S4 and Tables S1–S3 and S4.

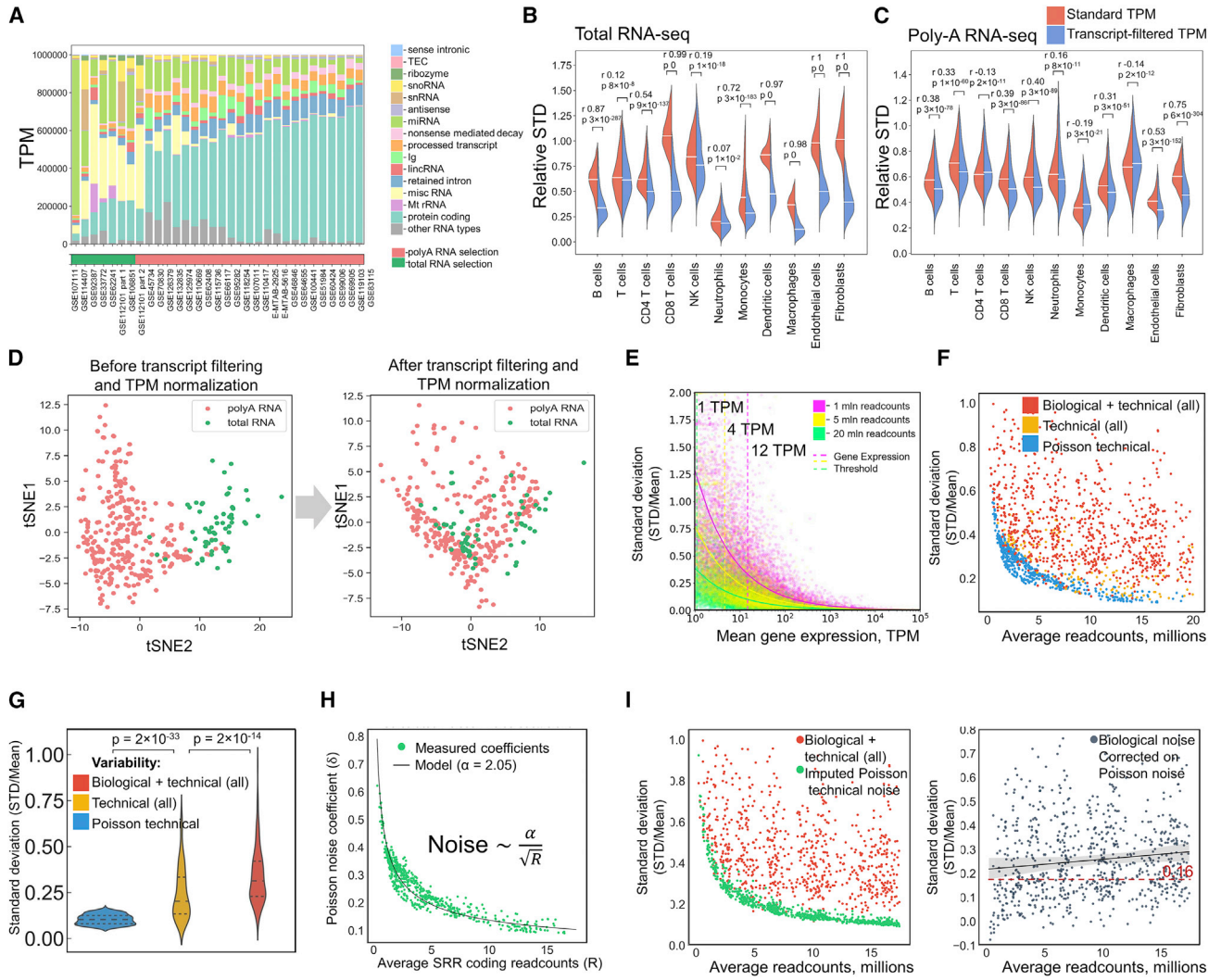


Figure 2. Establishment of expression normalization and analysis of technical noise

(A) TPM proportions of transcript types (GENCODE annotation) averaged across samples of different purified B cell datasets sequenced in different laboratories without renormalization.

(B and C) Violin plots of relative SDs in the expression of 3,515 housekeeping genes for different cell types before (red) and after (blue) transcript filtration and TPM renormalization for total or poly(A) RNA-seq. White horizontal lines depict the median values. p value was assessed by the two-tailed Wilcoxon test (p); r corresponds to rank-biserial correlation coefficient.

(D) Principal-component analysis (PCA) of sorted B cell RNA expression from either total (green) or poly(A) RNA-seq (red) before (left) and after (right) proposed transcript filtration and TPM renormalization.

(E) Gene SD in dependence of gene expression at total coverage of 1 (pink), 5 (yellow), and 10 (green) million read counts.

(F) Dot plot of samples with sequential additions of the shown noise levels: technical within one replicate and technical across multiple replicates and biological.

(G) Violin plot of the distribution of the same SDs of gene expression calculated within samples possessing different types of noise with a two-tailed Mann-Whitney test for significance. Dashed and dotted lines represent the median and the interquartile range (IQR), respectively.

(H) Measured Poisson noise coefficients for technical replicates of RNA-seq experiments with different total read count coverage.

(I) Dot plots showing dependence of gene expression mean SD on the total coverage of RNA-seq read counts for one cell type within one dataset (left) with biological and technical noise and (right) after subtraction of the imputed Poisson noise.

STD, standard deviation. See also Figure S5.

from normalization because of increased noise in the gene expression calculations. Finally, 48 additional transcripts were removed according to other annotation tags reporting a lack of evidence or quality (STAR Methods). Our advanced TPM normalization decreased variation of housekeeping genes (Eisenberg and Levanon, 2013) (Figures 2B and 2C). Transcript filtering and TPM renormalization made gene expression more comparable across different datasets (e.g., B cell datasets; Figure S5A). Similarly, principal-component analysis (PCA) of B cell RNA-seq derived from poly(A) RNA and total RNA enrichment indicated a separation of expression before, but not after, renormalization (Figure 2D). TPM normalization reduced expression batch effects across different datasets belonging to the curated sorted cell RNA profile compendium (Figures 2B–2D).

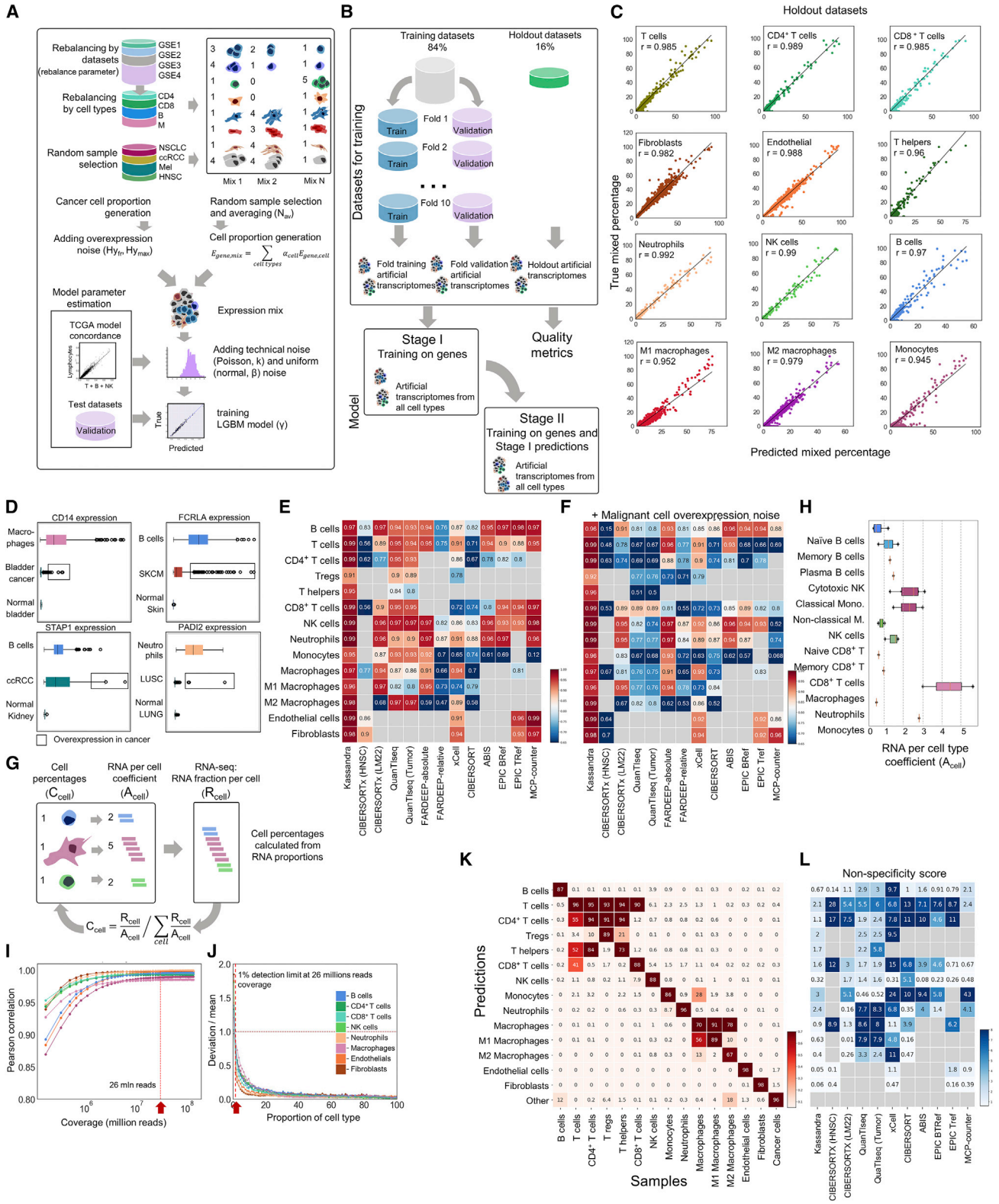


Figure 3. Training and performance of Kassandra deconvolution on artificial transcriptomes

(A) Workflow depicting the development of artificial transcriptomes of different cellular populations, training of tree-based ML procedure (LightGBM) to predict proportions of admixed cells, and estimation of model parameters using TCGA validation datasets and tissue samples.

(legend continued on next page)

Variability of sequencing technology influences cell deconvolution

Gene expression variability depends on the total read counts of RNA-seq and gene expression level or the number of read counts aligned on the transcript (Figure 2E). The 40% standard deviation (SD) of expression was observed at one TPM at 20 million reads coverage or at 125% SD at 1 million reads coverage (Figure 2E). Higher coverage reduces gene expression technical variability, including clinically relevant genes like PD-1/PD-L1 expressed at two to four TPMs. Inherent biological variability attributed to dynamic cell states contributes to transcript variance among samples. Variations were assessed by examining the data in the database with multiple replicates and experiments (Figure 2F). The noise increased from 10% to 26% from technical (library prep sequenced twice) to biological replicates (multiple experiments for a dataset) for certain cell types (Figure 2G).

To circumvent errors in analysis caused by technical variability (non-Poisson and Poisson noise), we formulated a TPM-based mathematical noise model (STAR Methods). The resulting dependence of Poisson technical noise (δ_{P_i}) on coverage and gene expression for gene i is expressed as

$$\text{Poisson technical noise } \delta_{P_i} = \alpha\sqrt{\frac{1}{l_i \underline{T}_i R}}, \qquad \text{(Equation 1)}$$

where l_i is an effective gene length, $\underline{T_i}$ represents mean expression in TPM units in technical replicates, R represents sample total read count coverage, and α is the proportional coefficient. This formula demonstrates lower coverage results in higher variability. The proposed formula correctly explained gene expression variability from expression levels and coverage measured within technical replicates of purified cell populations (Figure 2H). By plotting replicate coefficient of variation (CV) values dependent on read counts, we calculated the noise coefficient (α) (Figure 2H). The technical noise for each sample and gene was inferred according to Equation 1. Technical Poisson noise was subtracted from all technical noise (Figure S5B) and biological noise (Figure 2I), obtaining a non-Poisson additive to noise of approximately 16% (Figures 2I and S5B; STAR Methods).

The creation of artificial tumor transcriptomes adjusted for technical noise and aberrant gene expression to train Kassandra

The amount of available TME RNA-seq datasets with known cell composition are very limited; therefore, artificial RNA-seq of tissues and blood was developed *in silico* and utilized to train the Kassandra algorithm to robustly recognize diverse cell populations. Combinations of all available sorted cell populations and cancer cell lines were generated to develop cancer-specific artificial transcriptomes, imitating biological variability. In total, 18 million and 8 million transcriptomes were generated to train Kassandra-Tumor (Figure 1G) and Kassandra-Blood models (Figure 1H), respectively. For artificial transcriptome creation, different cell types (RNA-seq from sorted microenvironment populations and cancer cell lines/sorted cancer cells) were randomly selected in proportions closely resembling real tissue (Figure 3A; STAR Methods), and artificial technical noise was added to each artificial transcriptome as we demonstrated that technical sequencing variability influences gene expression measurements (Figures 2E–2G). In addition, all datasets from the sorted cell compendium were divided prior to creation of the artificial transcriptomes into training plus validation (84%) and holdout (16%) datasets to confirm the algorithm remained unaffected by batch effects across different datasets (Figure 3B).

After the addition of technical noise to the artificial tumor transcriptomes, in the first training stage, the cell-type models were trained on TPM-calibrated expression values to return RNA fraction per cell type from RNA-seq. In the second stage, the training data consisted of gene expression combined with the predicted RNA percentages per cell type obtained in the first stage. This stepwise training method enabled the model to adapt using information from other cell types and subtypes for their corresponding models and allowed the utilization of all datasets hierarchically for artificial transcriptomes (Figure 3B; STAR Methods). Knowledge regarding the percentages of "other" cell types allowed the second-stage model to adjust subpopulation percentages. Removal of the second stage resulted in decreased correlation between predicted and true percentages and increased mean absolute error (MAE) for holdout validation

(B) All datasets from our database were divided into training and validation datasets (84% of all datasets) and holdout datasets (validation only, 16% of all datasets) for final evaluation of algorithm performance. In stage I training, the model is trained on gene expression data. In stage II, the model receives predicted cell percentages as a result from stage I training.
(C) Performance of the algorithm measured as a Pearson correlation on holdout artificial transcriptomes. The significance of Pearson correlation (r) to be nonzero was assessed by the use of the exact distribution of r (two-tailed test) (p). For all comparisons, p values were $<10^{-300}$.
(D) *CD14*, *FCRLA*, *STAP1*, and *PADI2* expression in normal tissue (GTEx), immune cell types (from collected database), and cancerous tissue (TCGA). The box indicates overexpression outliers in cancerous tissue, and the error bars represent SD.
(E and F) Heatmap representing the Pearson correlation of different cell types between predicted and true artificial mix values without (E) or with (F) random overexpression noise added. Gray boxes indicate that the algorithm does not provide a predicted value for this cell type.
(G) Schematic showing the variability in the number of RNA molecules for different cell types.
(H) Boxplots depicting the RNA per cell-type coefficients for the listed cell types. In the boxplots, the right whisker indicates the maximum value or 75th percentile +1.5 IQR; the left whisker indicates the minimum value or 25th percentile −1.5 IQR. The central line indicates the coefficient median value of RNA per cell type.
(I) The dependence of the Pearson correlation of predicted cell percentages by Kassandra with true cell percentages dependent on the total read count coverage measured using holdout mixes. The red line indicates the RNA-seq coverage of data for which the algorithm was trained and optimized.
(J) Dependence of deconvolution error on overall cell percentage in RNA-seq. Red arrow indicates cell percentage (about 1%) where error reaches 100% (e.g., 1 + −1%).
(K) Graph of the prediction accuracy of Kassandra of pure samples (cell percentage).
(L) Non-specificity score for the listed 11 deconvolution algorithms. Values are percentages of nonspecific (false-positive) predictions relative to specific (true positive) predictions for different cell types.
See also Figures S6–S11 and Table S5.

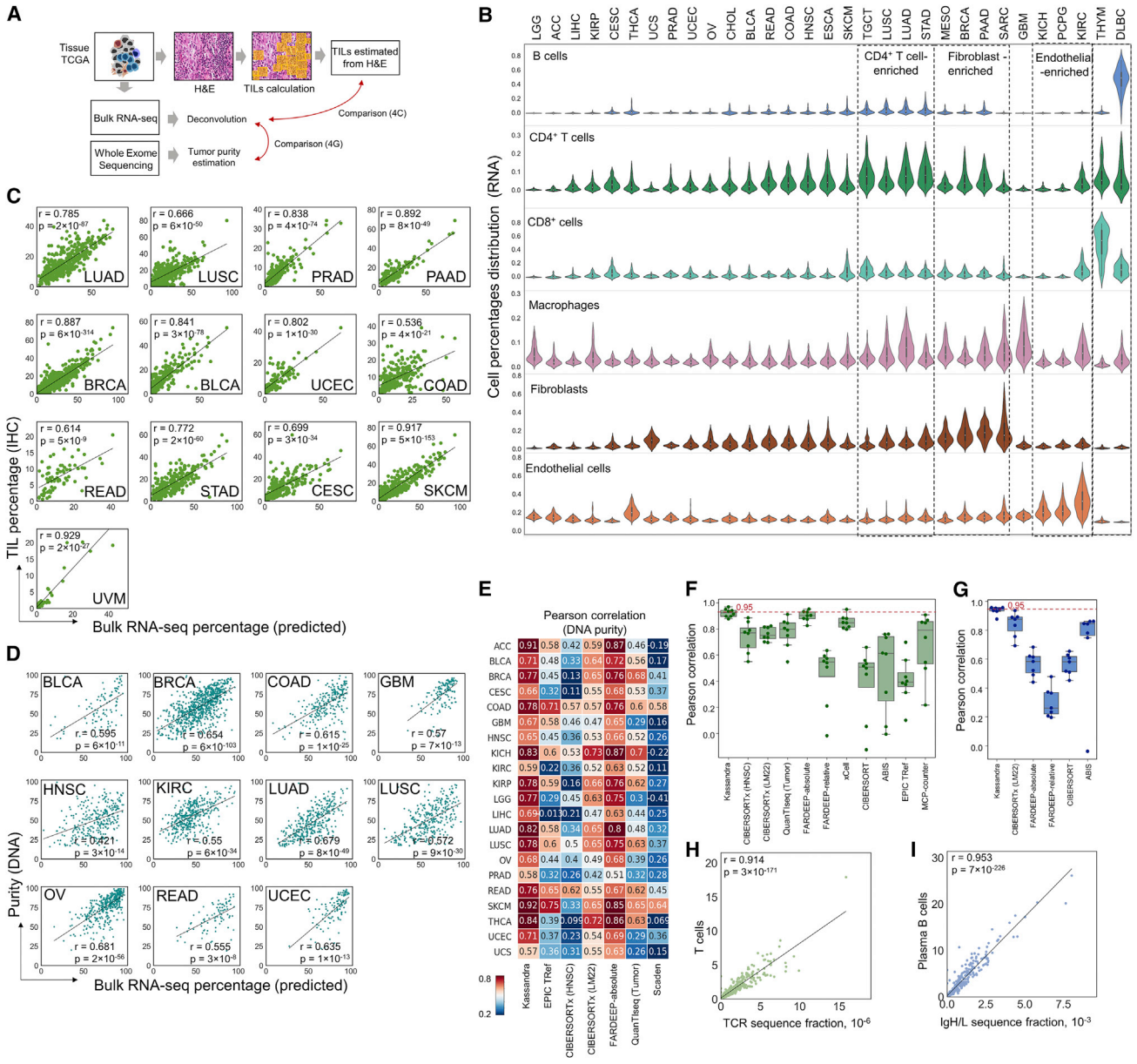


Figure 4. Large-scale TME deconvolution from RNA-seq of healthy and tumor tissues

(A) Algorithm validation scheme based on TCGA data using lymphocyte percentages recognized from H&E staining and tumor purity estimated from whole-exome sequencing (WES).

(B) Violin plots with internal boxplots indicate the differences between cell-type percentages deconvolved by Kassandra on TCGA RNA-seq data (n = 10,489). $CD4^+$ T cell, fibroblast, and endothelial-enriched groups of cancer types are indicated with dashed lines. Each data point corresponds to a predicted cell population fraction for a single sample. The boxes in the boxplots represent the IQR, where the center lines depict the median. The upper whiskers indicate the maximum values or 75th percentile +1.5 IQR; the lower whiskers indicate the minimum values or 25th percentile −1.5 IQR.

(C) Pearson correlation between percentages of lymphocytes predicted by Kassandra on TCGA RNA-seq data and calculated by machine analysis of histological TCGA slides.

(D) Correlation of predicted percentages of malignant cells from RNA-seq by Kassandra with tumor purity estimated from WES for 11 TCGA cancer types.

(E) Pearson correlation values between tumor purity (CPE) and predicted percentages of malignant cells based on TCGA RNA-seq data by different deconvolution algorithms.

(F and G) Boxplots showing Pearson correlation values for predicted T cell RNA percentage by different deconvolution algorithms with T cell receptor (CDR3 region of TCR) reads and for predicted plasma B cell RNA percentage by different deconvolution algorithms with B cell receptor (CDR3 region of IgH) reads in different TCGA cancer types. Each data point corresponds to a different cancer type. The boxes in the boxplots represent the interquartile range (IQR), where the center lines depict the median. The upper whisker indicates the maximum value or 75th percentile +1.5 IQR; the lower whisker indicates the minimum value or 25th percentile −1.5 IQR.

(legend continued on next page)

datasets (Figures S6A–S6D). The predicted RNA percentages strongly correlated with the actual admixed RNA percentage in the artificial transcriptomes for multiple cell types in holdout datasets ($r > 0.95$; Figure 3C).

However, after this training process, we applied Kassandra for the reconstruction of tumor tissues and found that aberrant gene expression, a hallmark of cancer cells, may have interfered with TME deconvolution as illustrated by the unexpected expression of various non-malignant markers in different cancers (Figure 3D). To train the algorithm to ignore tumor aberrant overexpression, we added random expression noise to the artificial transcriptomes (admixed cancer cell lines with TME cell populations), imitating patient-specific gene overexpression present in tumors. Using this approach, Kassandra was stable in predicting cell types within the mixtures with aberrant expression noise ($r = 0.92–0.99$), whereas other methods produced weaker correlations ($r < 0.70$; Figures 3E, 3F, and S7A–S7D). Removing noise and aberrant gene expression from Kassandra decreased the correlation coefficients and increased the MAE (Figure S6).

We randomly brute forced combinations of LightGBM parameters, selecting metrics resulting in the highest correlation and lowest MAE to develop Kassandra (Figures S8A–S8D). We compared the performance of the final version of Kassandra with other ML algorithms: a support vector regression (SVR) (e.g., CIBERSORTx; Newman et al., 2019) and a deep learning neural network (NN)-based model (e.g., Scaden; Menden et al., 2020) (Figures S9A–S9D). The NN could potentially lead to superior performance in comparison with tree-based methods, but NN requires significant training data and parameter optimization efforts, which may cause overtraining. The SVR was trained on the RNA profiles of the sorted cell Kassandra compendium, and the Scaden NN was trained on the same artificial mixes as Kassandra and a scRNA-seq lung cancer dataset because it was primarily developed using scRNA-seq (Figure S9A; STAR Methods). Both the SVR and Scaden showed lower Pearson and concordance correlation coefficients between predicted and true cell percentages (e.g., average r: Kassandra 0.83 versus SVR 0.69 versus Scaden 0.71) and higher MAEs (e.g., Kassandra 4.1 versus SVR 5.5 versus Scaden 6.8) when tested on validation and holdout datasets compared with Kassandra (Figures S9B and S9C). Kassandra also outperformed the Scaden NN trained on scRNA-seq as shown in a cytometry time of flight validation experiment (e.g., average r: Kassandra 0.85 versus Scaden 0.77 and average ccc: Kassandra 0.64 versus Scaden 0.37; Figure S9D).

Calculation of the limit of detection of Kassandra

Kassandra accurately predicts the RNA levels in a sample as demonstrated; however, Kassandra was developed for the enumeration of the correct percentage of a cell type, which relies on the RNA concentration per cell and, in turn, depends on cell size (Monaco et al., 2019). Tissue-specific or varied coefficients can be used to convert RNA levels to cell numbers (Figures 3G). For some immune subpopulations, we experimentally measured the RNA per cell-type coefficients relative to T cells (Figures 3H and S10A; STAR Methods). For stromal cells, we hypothesized the TME qualitatively represents the same cellular phenotypes across cancer types, so over 10,000 TCGA pan-cancer samples were used to amend previously defined coefficients (Monaco et al., 2019; Racle et al., 2017) and obtain values by fitting. For fitting, random RNA-per-cell coefficient sets were generated, and for each tumor and cell type, the RNA-per-cell coefficient set with the best correlation with the "other" cell fraction and tumor purity obtained from TCGA DNA analysis was chosen (Figures S10B and S10C; Table S5). For blood subpopulations where coefficients were not measured, the coefficients were calculated by fitting to 45 independent FACS experiments (analogously to ABIS; Monaco et al., 2019) relative to naive Th cells (Figure S10D; Table S5; STAR Methods).

By creating holdout artificial transcriptomes with low numbers of total read counts, we tested the Kassandra limit of detection (LOD) (Figure 3I). The accuracy decreased with coverage under 26 million reads (Figure 3I). The average LOD for cell types was approximately 0.5%–1% (Figure 3J), depending on the cell type. Fibroblasts were detected with variability up to 0.3%, whereas NK cells, which share the expression of multiple genes with $CD8^+$ T cells (Figure S1A), were reliably detected beginning at only 1%. False-positive and nonspecific detection by Kassandra was found to be the lowest among all tested approaches, including CIBERSORT (Newman et al., 2019), EPIC (Racle et al., 2017), MCP-counter (Becht et al., 2016), and quanTIseq (Finotello et al., 2019) (Figures 3K, 3L, and S11). To assess specificity, we used the RNA profiles of one specific cell type, and the background signal of other cell types was treated as a nonspecific signal (Cossarizza et al., 2019) (Figures 3K and 3L).

Kassandra reconstruction of tumor and healthy tissues

After Kassandra development and training, the cellular composition of diverse TCGA tumors (Figures 4A and S12) and healthy tissues (Figures S13A and S13B) was analyzed with Kassandra-Tumor (Figure 4B). The highest proportion of fibroblasts was observed in pancreatic adenocarcinoma, known to contain vast amounts of cancer-associated fibroblasts (CAFs) (Norton et al., 2020). Macrophages, crucial factors in prognosis and treatment, were predicted to be 10% or more in the TME of glioblastoma multiforme and lung adenocarcinoma (Rakaee et al., 2019). When applying Kassandra to Genotype-Tissue Expression (GTEx) datasets (GTEx Consortium, 2013), mononuclear cells were primarily found in blood samples (Figures S13A and S13B). These results show that deconvolution by Kassandra identified and enumerated cell types from a diverse array of tissues in the expected ranges.

To calculate the number of tumor-infiltrating lymphocytes (TILs) in tissue samples, we applied Kassandra to TCGA data with matching H&E slides, enabling validation via histological review (Figure 4A) (Saltz et al., 2018). Kassandra deconvolution highly correlated with H&E-assessed cell types in 13 different cancer types for a total of 4,035 samples ($r > 0.7$ for 10/13 analyzed cancers; Figures 4C, S14, and S15A–S15C). CIBERSORTx,

(H) Pearson correlation of predicted T cell RNA percentage by Kassandra with T cell receptor (CDR3 region of TCR) reads by MiXCR (Bolotin et al., 2015) in LUSC TCGA data.

(I) Pearson correlation of predicted plasma B cell RNA percentage by Kassandra with B cell receptor (CDR3 region of IgH) reads by MiXCR in TCGA-LUSC data. See also Figures S7 and S12–S16.

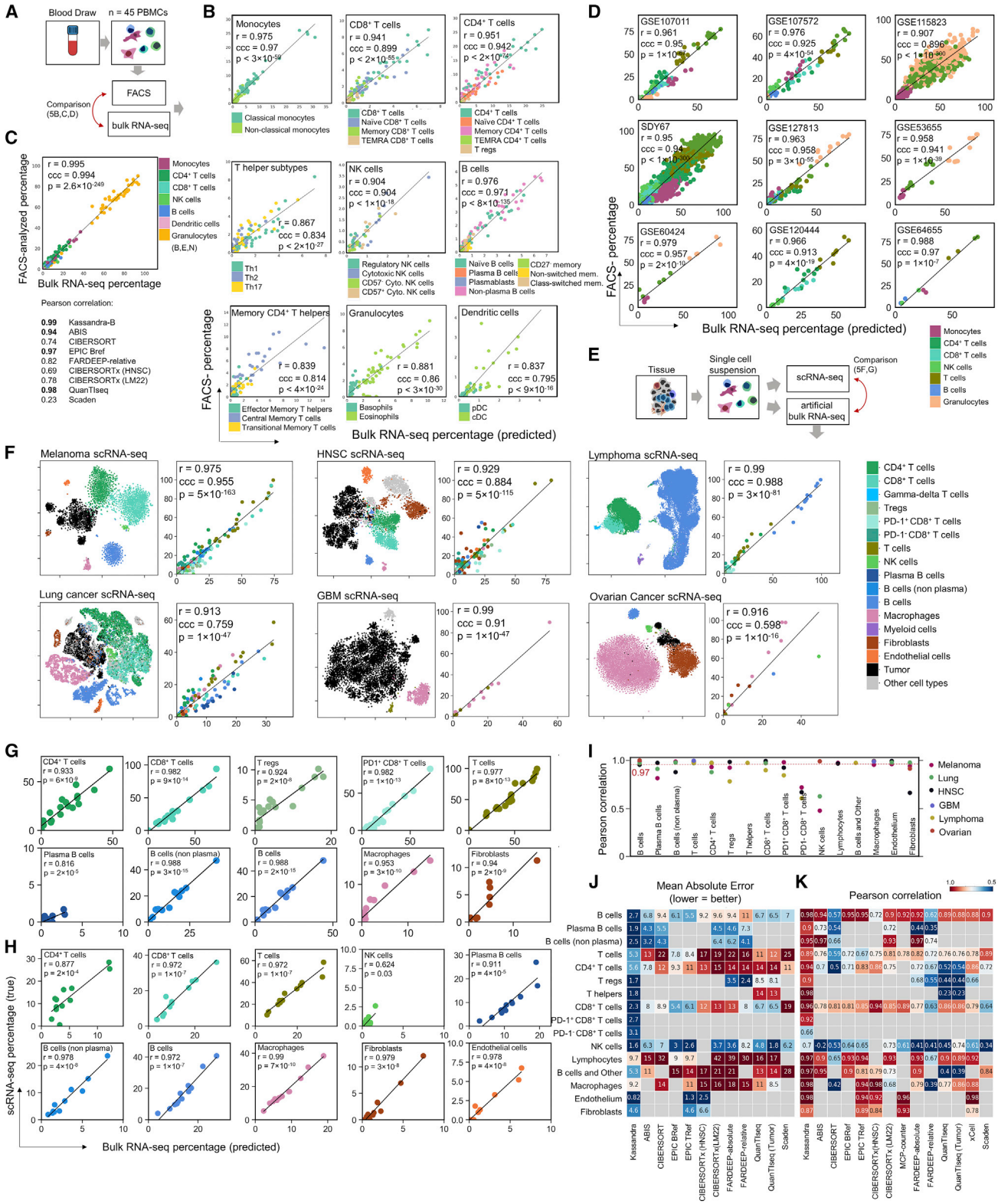


Figure 5. Validation of cellular composition deconvolution and TME reconstruction by Kassandra

(A) Schematic representation of a validation experiment comparing bulk RNA-seq and FACS for the same PBMC samples extracted from whole blood.

(B) Pearson correlation and concordance correlation coefficient (ccc) of true RNA percentages for cell-type identification among FACS analysis of PBMCs and Kassandra predictions from the bulk RNA-seq results.

(legend continued on next page)

QuanTIseq, and FARDEEP-absolute showed lower Pearson and concordance correlation values across the cancer types (Figures S15A–S15C). The Kassandra-predicted tumor purity better correlated with the best practice DNA-calculated purity values than other deconvolution methods (Aran et al., 2015) (Figures 4D, 4E, S7C, and S16).

The proportion of expressed TCR and IgH/L (BCR) sequences correlate with the presence of T or plasma B cells (Reuben et al., 2020) actively producing immunoglobulins (Sharonov et al., 2020). We realigned sequences using MIXCR to characterize CDR3 transcripts, which are associated with different T and plasma B cell clones. Kassandra, but not other deconvolution tools, showed a strong correlation of predicted T cell percentages with the number of identified TCRs within the sample (Figures 4F and 4H) and plasma B cell percentages with IgH/L transcripts (Figures 4G and 4I).

Reconstruction of blood cellular composition by Kassandra

To validate the ability of Kassandra-Blood to predict 38 cell types derived from blood (Figure 1H), we performed extensive FACS analysis (Table S6; Figures 5A and S17A) and sequencing of 45 peripheral blood (PB) mononuclear cell (PBMC) or PB lymphocyte fractions of whole blood from different donors (Figure 5A). For the majority of cell subpopulations, Pearson correlation coefficients were greater than 0.9 (Figures 5B and S17B), and for major cell types, overall Pearson and concordance correlation coefficients reached 0.995 (Figure 5C). Other algorithms showed lower correlation (Figures S18 and S19). Kassandra also accurately deconvolved 11 novel cell types (e.g., $CD27^{-}$ memory B cells, non-switched memory B cells, $CD57^{+/-}$ cytotoxic NK cells, TEMRA, and transitional CD4/$CD8^{+}$ T cells; $r > 0.84$; Figures 5B, S17B, and S19). Even at very low total cellular percentages, deconvolved Th and granulocytic subsets had correlations of 0.87 and 0.88, respectively.

Pseudobulk RNA-seq datasets built from nine independent PBMC scRNA-seq datasets (Avila Cobos et al., 2020) were deconvolved using Kassandra (Figure S20A; STAR Methods). Cells from scRNA-seq were manually phenotyped (Figure S20B; STAR Methods), and a significant correlation value ($r = 0.97$, $p = 8 \times 10^{-52}$) was obtained when aligning the true scRNA-seq percentages with the Kassandra-predicted cell percentages from pseudobulk (Figures S20C and S20D). Next, eight independent PBMC datasets containing more than 867 samples were analyzed comparing FACS-based percentages with Kassandra cell prediction (Figure 5D). Notably, Kassandra-based deconvolution and FACS strongly correlated ($r = 0.907$–0.988; Figure 5D). Correlations of 0.97 with FACS were also measured for bone marrow validation cohorts (GEO: GSE120444; $p = 4 \times 10^{-19}$) and TIL/CAF mixes (GEO: GSE121127; $p = 2 \times 10^{-12}$; Figure S20E). Normal lymph node cellular percentages from CyTOF also strongly correlated with RNA-seq reconstruction by Kassandra ($r = 0.95$; Figures S21A–S21C).

Validation of Kassandra TME reconstruction across different tumor types

To confirm the ability of Kassandra to deconvolve the cell populations from varied tumor types, we compared Kassandra cell percentages predictions with scRNA-seq data derived from six tumor types (Figures 5E, 5F, and S22–S25). Cells from scRNA-seq were annotated manually (STAR Methods), and RNA from all cells of each patient was mixed to imitate tumor bulk RNA-seq (Figure 5E; STAR Methods). Notably, in melanoma and lung cancer, Kassandra accurately predicted $CD4^{+}$ T cells ($r = 0.93$ and 0.88), T regulatory cells (Tregs; $r = 0.92$), plasma ($r = 0.91$ and 0.82), and non-plasma B cells ($r = 0.99$ and 0.98; Figures 5G, 5H, S24B, and S24C), even though these cell types express overlapping gene signatures. The median correlation of each cell type reconstruction across the six scRNA-seq datasets reached 0.97 (Table S7; Figure 5I). Kassandra correctly predicted the most cell types with the lowest MAEs (Figure 5J) and strongest correlations (Figures 5K, S7D, and S23) compared with other deconvolution tools.

To further demonstrate the clinical utility of Kassandra in reconstructing various tumor types, we tested its ability to accurately reconstruct the TME using primary early-stage non-small cell lung carcinoma (NSCLC) and clear cell renal cell carcinoma (ccRCC) tumors of varying grades collected in a clinical setting. NSCLC tumor samples were processed for bulk RNA-seq and CyTOF analysis. We performed RNA-seq of the same cellular suspensions prepared for CyTOF using more than 40 cellular markers (Figures 6A, 6B, S26A, and S26B). Kassandra strongly correlated

(C) Comparison of predicted cell percentages by Kassandra from bulk RNA-seq with actual cell percentages obtained by flow cytometry measurements. Mean Pearson correlation coefficients (r) and concordance correlation coefficients (ccc) are shown. The significance of Pearson correlation (r) to be nonzero was assessed by the use of the exact distribution of r (two-tailed test) (p).

(D) Comparison of predicted cell percentages by Kassandra from bulk RNA-seq with actual cell percentages obtained by flow cytometry measurements. The Pearson correlation (r) significance was assessed using a two-tailed test. Mean Pearson correlation coefficients (r) and concordance correlation coefficients (ccc) are shown. The significance of Pearson correlation (r) to be nonzero was assessed by the use of the exact distribution of r (two-tailed test) (p).

(E) Schematic representation of the workflow of a validation experiment using scRNA-seq samples from PBMCs and solid tumors. scRNA-seq data were artificially mixed to create a bulk RNA-seq dataset.

(F) t-SNE plots of cell phenotyping and overall correlation of true scRNA percentage values derived from scRNA-seq data with deconvolution predictions by Kassandra from pseudobulk RNA-seq data. Mean Pearson correlation coefficients (r) and concordance correlation coefficients (ccc) are shown. The significance of Pearson correlation (r) to be nonzero was assessed by the use of the exact distribution of r (two-tailed test) (p).

(G and H) Scatterplot of true RNA percentages per cell type from scRNA-seq data with Kassandra deconvolution predictions from artificial bulk RNA-seq data. Correlations are shown for different cell subpopulations in melanoma (G) and lung cancer (H) from (F). The Pearson correlation (r) significance was assessed using a two-tailed test.

(I) Pearson correlation of true cell-type RNA percentage values derived from scRNA-seq data (F) with deconvolution predictions by Kassandra from pseudobulk RNA-seq data. The median correlation is 0.97.

(J and K) Mean absolute error (MAE) scores (J) and mean Pearson correlation (K) values between predicted values from artificial bulk RNA-seq data with true values derived from all analyzed scRNA-seq datasets from (F) for different deconvolution algorithms.

See also Figures S6, S7, S9, S17–S25, and Tables S6 and S7.

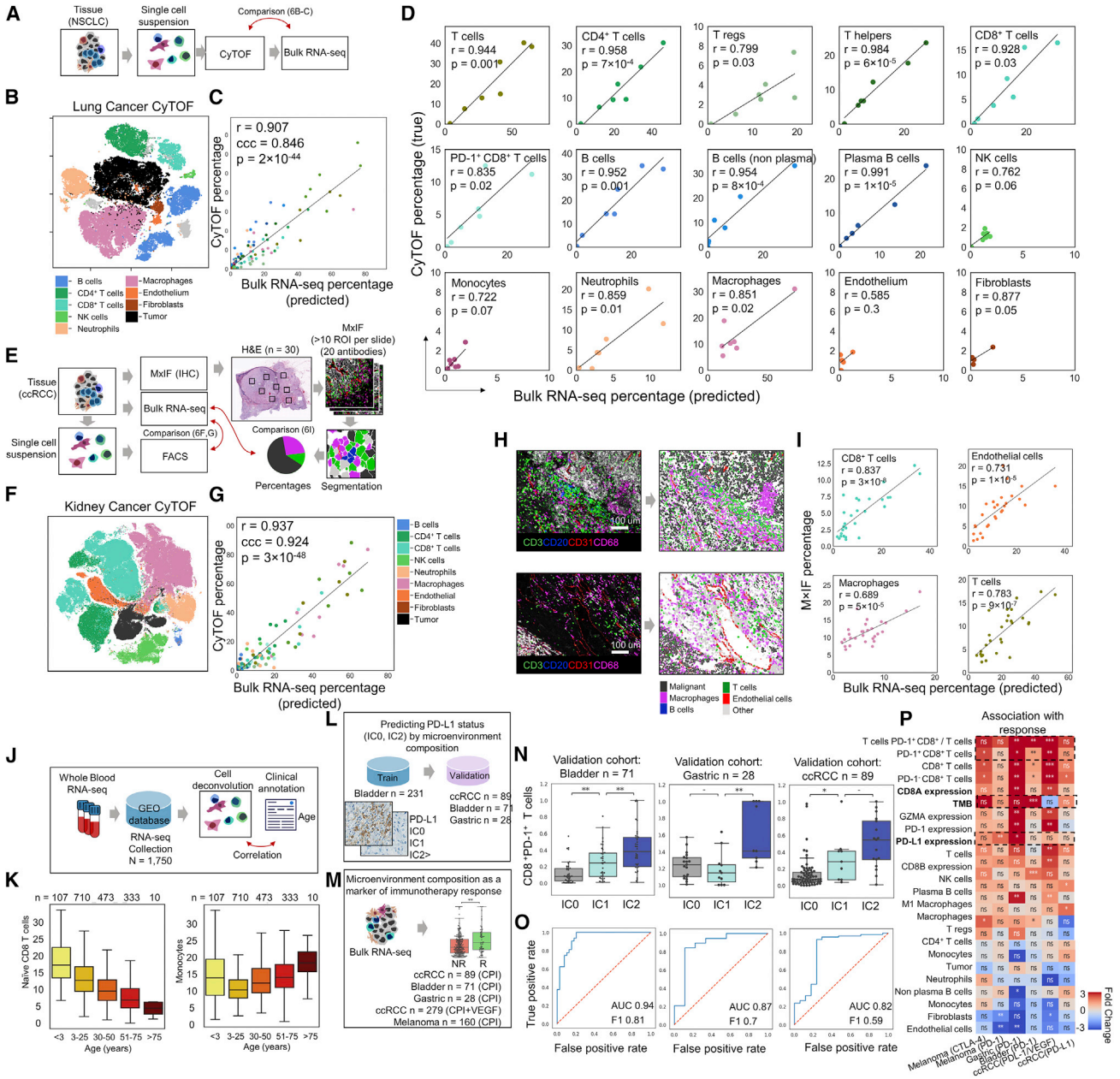


Figure 6. Validation of microenvironment deconvolution and applicability of Kassandra for immunotherapy response prediction

(A) Schematic representation of a validation experiment using bulk RNA-seq and CyTOF from the same lung adenocarcinoma biopsies.

(B) t-SNE plot of cell phenotyping from CyTOF of lung adenocarcinoma samples.

(C) Correlation of cell percentages measured by CyTOF (n = 7) with cell percentages predicted by Kassandra from bulk RNA-seq of the same NSCLC biopsies. Pearson correlations values were calculated for all cell subtypes combined.

(D) Correlation of cell-type percentages measured by CyTOF with cell percentages predicted by Kassandra from bulk RNA-seq.

(E) Schematic representation of a validation experiment using bulk RNA-seq, CyTOF, and MxIF of the same ccRCC biopsies.

(F) t-SNE plot of cell phenotyping by CyTOF analysis of the ccRCC samples.

(G) Correlation of cell percentages measured by CyTOF (n = 8) with cell percentages predicted by Kassandra from bulk RNA-seq.

(H) Multiplex immunofluorescence (MxIF) images (20 markers) of ccRCC samples with cell segmentation and typing.

(I) Correlation of immune and endothelial cell percentages measured by MxIF (n = 28) with Kassandra predictions from bulk-RNA seq of the same tissue region. The Pearson correlation (r) significance was assessed using a two-tailed test.

(J) Schematic representation of associations of immune cell composition determined via deconvolution of whole blood bulk RNA-seq (n = 1,750) with aging.

(K) Boxplots showing the Kassandra-predicted percentages of naive CD8 T cells (left) and monocytes (right) in blood samples collected from patients at different ages. The numbers of patients in each age group are indicated above the corresponding boxplots. Each data point corresponds to a predicted cell population fraction for a single sample. The boxes in the boxplots represent the IQR, where the center lines depict the median. The upper whiskers indicate the maximum values or 75th percentile +1.5 IQR; the lower whiskers indicate the minimum values or 25th percentile −1.5 IQR.

(legend continued on next page)

with CyTOF detection of T and B cell, neutrophil, macrophage, Treg, NK cell, endothelial cell, and fibroblast populations ($r = 0.907$, $p = 2 \times 10^{-44}$) compared with other approaches (Figures 6C, 6D, and S26D). Kassandra predicted the presence of low-abundance cell types in this dataset, such as NK cells, monocytes, endothelial cells, and fibroblasts; however, with the exception of fibroblasts, the correlation with CyTOF was lower compared with abundant cell populations ($r = 0.6–0.8$) (Figure 6D). The ccRCC patient samples were collected and processed for comparative bulk RNA-seq and CyTOF (Figure 6E). t-SNE analysis showed only immune cells (ICs) were efficiently recovered from the tumor samples (Figure 6F), and Kassandra accurately predicted the immune populations in the ccRCC samples with the strongest correlation ($r = 0.937$, $p = 3 \times 10^{-48}$; Figures 6G and S26E).

Moreover, multiplex immunofluorescence (MxIF) provided spatial analysis of 28 ccRCC samples with more than 10 tissue regions of interest (ROIs) (Figure 6E). Intratumoral ROIs were selected for comparison, and cell segmentation was performed on MxIF images (Jackson et al., 2020; Pachynski et al., 2021) to reconstruct single-cell proteomic data (Figure 6H). Mean fluorescent intensity within a cell segment was used for cell typing, and the proportions of macrophages, $CD8^+$ T cells, NK cells, and B cells were significantly similar to the Kassandra-predicted percentages from RNA-seq ($r = 0.7–0.8$, $p < 0.0001$; Figures 6I and S26C). The relative amount of blood vessels within a tissue was calculated as an area of CD31 marker because of the intricate shape of endothelial cells causing segmentation difficulty. Nevertheless, the endothelial cell area correlated with predicted endothelial cell percentages from RNA-seq ($r = 0.731$, $p = 0.00001$; Figure 6I).

As mentioned, to make Kassandra stable against cancer-specific noise and expression, we added both cancer cell lines and sorted malignant cells to the artificial transcriptomes (STAR Methods). The t-SNE plot shows the partial overlap of their expression profiles (Figure S27A,B). Kassandra was not intended to predict exact tumor purity and outputs the "other" fraction with all uncharacterized cells (including cancer cells). To address the stability of Kassandra, previously unseen cancer cell lines (e.g., COLO829, MCF7, and K562) were admixed with PBMCs at different ratios ranging from 100:0 to 12:88 (cell line: PBMC). Kassandra reconstructed the percentages of all non-malignant IC types from PBMCs in correct proportions in all mixes and ratios despite low mRNA content from the PBMCs (Figure S27). Moreover, the percentages of "other" previously unseen cell types were calculated with the overall Pearson correlation coefficient of 0.94 ($p < 0.001$).

The application of Kassandra to measure blood immune composition of archival samples

The ability to perform scRNA-seq or flow cytometry on archived blood samples is often limited because of multiple factors (Zheng et al., 2017). To demonstrate that Kassandra can provide value to samples unsuitable for other analyses, we collected whole blood and PBMC RNA-seq profiles of 1,750 healthy donors from multiple studies for which multicolor flow cytometry was not performed (Figure 6J). Kassandra-Blood digitally profiled the immune composition of the blood samples to reveal additional clinical associations from this metacohort analysis related to aging. For example, the naive $CD8^+$ T cell population significantly decreased with age, while the monocyte percentage increased with age as previously observed (Britanova et al., 2014, 2016) (Figure 6K), supporting the application of Kassandra on archival samples for hypothesis validation and to infer novel findings in tissue samples where conventional cytometry methods were unsuitable.

The Kassandra-reconstructed TME predicts PD-L1 immunohistochemistry (IHC) status and correlates with response to immunotherapy

Upregulation of PD-L1 expression within the tumor is heavily dependent on the specific immune suppressive context of the TME. In particular, macrophages express PD-L1 to suppress T cells via the PD-1/PD-L1 axis (Liu et al., 2020; Wei et al., 2019). We correlated Kassandra-Tumor deconvolution of the TME with PD-L1 IHC to predict PD-L1 status, an important therapy-associated biomarker (Morsch et al., 2020), or therapy response, potentially providing a complementary method to predict immunotherapy response with RNA-seq (Figures 6L and 6M). In addition, this analysis would assess whether microenvironment composition can predict PD-L1 status. For the bladder cancer discovery cohort (Mariathasan et al., 2018) and the bladder cancer (Mariathasan et al., 2018), gastric cancer (Kim et al., 2018), and ccRCC validation cohorts, all tested independently, deconvolved $PD\text{-}1^+$ $CD8^+$ T cells significantly correlated with PD-L1 IC IHC levels represented as the percentage of the positive tumor area: $IC2^+$: $\geq$5% PD-L1; IC1: $\geq$1% but <5% PD-L1; and IC0: <1% PD-L1 ($p < 0.001$; Figures 6N, S28A, and S28B). A regression model was trained on RNA-seq from the bladder cancer cohort (Mariathasan et al., 2018) to predict PD-L1 IHC status (Figure 6L). The percentage of $PD\text{-}1^+$ $CD8^+$ T cells was assessed in additional tumor types and correlated with PD-L1 IC status in bladder cancer, gastric cancer, and ccRCC (Figure 6O). The generated TME regression model independently predicted PD-L1 status with 0.92, 0.87, and 0.82 area under the curve (AUC) performances scores in the validation cohorts (Figures 6O and S28C–S28E).

Beyond the correlation between PD-L1 IHC and $PD\text{-}1^+$ $CD8^+$ T cell percentages, the relationship between TME deconvolution and response to immunotherapy was examined in bladder cancer (anti-PD-L1) (Mariathasan et al., 2018), gastric cancer (anti-PD-1) (Kim et al., 2018), ccRCC (anti-PD-L1, anti-PD-L1+BEVA)

(L) Schematic representation of comparison of deconvolution of bulk RNA-seq with PD-L1 IHC status.
(M) The association of immune cell composition and immunotherapy response across different cohorts.
(N) Boxplots showing the level of $PD\text{-}1^+$ $CD8^+$ T cells in bladder cancer, gastric cancer, and ccRCC cohorts stratified by PD-L1 immune cell (IC) immunostaining levels represented as the percentage of positive tumor area. The boxes in the boxplots represent the IQR, where the center lines depict the median. The upper whisker indicates the maximum value or 75th percentile +1.5 IQR; the lower whisker indicates the minimum value or 25th percentile −1.5 IQR.
(O) Receiver operating characteristics of the Kassandra-predicted $PD\text{-}1^+$ $CD8^+$ T cells with PD-L1 IHC status for the three validation cohorts from (N).
(P) Heatmap of the significance of the denoted cell types to immunotherapy response across the listed cohorts.
See also Figures S9 and S26–S29.

(Pal et al., 2020), melanoma (anti-PD-1) (Gide et al., 2019; Hugo et al., 2016; Liu et al., 2019b), and anti-CTLA-4 (Nathanson et al., 2017; Van Allen et al., 2015) cohorts (Figures S29A–S29D, J). Notably, Cox proportional hazard models showed that the ratios of PD-1$^+$ CD8$^+$ T cells to all T cells were significantly associated with immunotherapy responders in all cohorts independently of TMB and PD-L1 expression values (Figures 6P and S29E–S29I). ML-based response prediction models that combined Kassandra-reconstructed TME percentages and TMB and PD-L1 expression demonstrated greater predictive power in comparison with single metrics when training with cross-validation was applied, resulting in improved performance characteristics on the unseen melanoma cohort (receiver operating characteristics [ROC] AUC 0.64 [TMB] to 0.75 [Kassandra + TMB + PD-L1; Figure S29J), indicating deconvolution could be applied to increase the accuracy of immunotherapy response predictive models.

DISCUSSION

We developed the decision tree ML-based algorithm Kassandra to reconstitute the cellular composition of both tumor biopsies and blood using bulk RNA-seq. An extensive database containing RNA-seq data of more than 9,400 sorted samples of various cell populations was compiled and employed to create artificial transcriptomes to train and develop Kassandra. This comprehensive manually harmonized compendium of sorted cell types provides an extensive resource of RNA expression data. Kassandra training on this comprehensive database enabled the accurate prediction of cell percentages of new RNA-seq samples without pre-education on scRNA-seq or other experimental data. In addition, unlike NN models (Menden et al., 2020), the decision tree model provides predictive behavior of the final predictions and allows the analysis of important genes via feature importance estimation. The use of NN might lead to superior performance in comparison with tree-based methods; however, in our experience, NN requires substantial training data and optimization efforts. In initial tests, the NN model exhibited more overtraining behavior, highlighting the value of a more stable and predictable decision tree model.

Kassandra deciphered cellular proportions even when challenged with mixtures containing phenotypically similar cell types using a biology- and bioinformatics-driven approach to select genes uniquely expressed in certain cell types. In contrast with earlier described deconvolution methods, Kassandra does not use a gene set enrichment analysis (GSEA)-like approach (xCell (Aran et al., 2017), MCP-counter (Becht et al., 2016)), least square linear optimization (Finotello et al., 2019; Hao et al., 2019; Monaco et al., 2019; Racle et al., 2017), or linear matrix-based methods (Newman et al., 2015, 2019; Wang et al., 2019). Rather, Kassandra divides samples into subpopulations based on significant differentiation in input variables, leading to superior accuracy ($r < 0.97$, $p < 10^{-10}$) and stable performance in various malignant tissues. We also used a two-step training architecture that allows the use of maximum information from different levels of the cellular hierarchy. Finally, we established a coefficient that converts the RNA levels to absolute cell numbers because different cell types contain different amounts of RNA, and this coefficient allows quantitative analysis of the cellular components of any given tissue, enabling the comparison of different samples. Kassandra is a tissue agnostic algorithm that uses a tissue agnostic transformation coefficient that accurately predicts cell percentages in different tissues.

The addition of functional and technical noise to the artificial transcriptomes used to train Kassandra resulted in the successful management of the noise observed in real samples. Indeed, excluding noise from the final Kassandra algorithm, aberrant cancer cell expression, and the second-stage model resulted in decreased performance. Moreover, Kassandra can be further optimized and improved by simple retraining on an extended dataset in the future, which would increase the overall accuracy of predictions and account for additional cell states.

This tree-based comprehensive approach allowed Kassandra to identify 18 TME subpopulations in the examined tissues and 41 populations derived from PBMCs, resulting in a total of 51 unique cell types that can be identified from bulk RNA-seq data, including Th subsets, which are notoriously difficult to deconvolve. Accurate recognition of this diverse cellular repertoire remains unparalleled compared with other platforms, such as CIBERSORT, ABIS, and EPIC, which were designed to deconvolve IC populations, and MCP-counter and quanTIseq, which can also recognize fibroblasts and endothelial cells. Kassandra can precisely deconvolve many cell types, including CD4$^+$ and CD8$^+$ T subpopulations (Tregs, PD-1$^+$ cells), relevant for predicting immunotherapy response. Using Kassandra, we found a positive correlation of PD-1$^+$ CD8$^+$ T cells with PD-L1 IHC levels, a clinical biomarker of immunotherapy response, and combining Kassandra with TMB and PD-L1 expression status improved immunotherapy response prediction. The universal approach used for the development of Kassandra can be applied to any cell type based on its unique RNA profile; therefore, future application of this workflow can precisely predict diverse cell types, such as epithelial and glandular cells, glial cells, neurons, and other cell types integral to a wide array of pathologies. This computational tool could lead to an improved and more comprehensive understanding of biology in archival samples that have only RNA-seq data available, ultimately supporting clinical applications for diverse diseases in the future.

STAR★METHODS

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

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Cell lines
  - Blood samples
  - Tissue samples
- METHOD DETAILS
  - Datasets
  - RNA-seq processing and normalization
  - Gene selection
  - Generation of artificial transcriptomes

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.ccell.2022.07.006.

ACKNOWLEDGMENTS

This work was supported by BostonGene, Corp. M.C.P. and J.J.H. receive research funding from BostonGene, Corp. The results shown here are in whole or part based on data generated by the TCGA Research Network (http://cancergenome.nih.gov/) and The Cancer Imaging Archive (TCIA) (Clark et al., 2013), GEO (Barrett et al., 2012), ArrayExpress (Athar et al., 2019), and GTEx (GTEx Consortium, 2013). We thank Grigorii Nos for his help with the scRNA-seq lymphoma dataset processing.

AUTHOR CONTRIBUTIONS

A. Bagaev, N.F., and R.A. conceived and jointly supervised the study. A. Zaitsev, M. Chelushkin, I.C., B.S., and A. Bagaev conceived the deconvolution model. A. Zaitsev, M. Chelushkin, D.D., A. Bagaev, E. Nuzhdina, I.C., and A. Zotova implemented the deconvolution algorithm and performed analyses. A. Zaitsev, M. Chelushkin, D.D., V.Z., V.S., A. Zotova, B.S., and D. Afenteva participated in the database generation. A. Zaitsev, M. Chelushkin, D.D., A. Baisangurov, A. Zotova, and B.S. performed major database analytics. M. Chelushkin and Y.L. performed scRNA-seq analysis. V.Z. performed the CyTOF analysis. N.K. performed analysis of PD-L1 associations with TME. S.R.P., D.L.D., P.M.R., M.L., and M.C.P. obtained the NSCLC biopsies and conducted the CyTOF and bulk RNA-seq experiments. J.J.H., A.R., and Y.L. obtained the ccRCC biopsies and conducted the bulk RNA-seq, CyTOF, and MxIF experiments. L.C. provided the lymph nodes for deconvolution validation. I.G. and V.S. conducted the MxIF analysis. M.F.G., C.T., M. Chasse, I.W., M. Abdou, S.A., and S.M.A. performed the RNA-seq, cell sorting, and flow cytometry experiments. A. Bagaev, K. Nomie, A. Zaitsev, I.C., D.D., M. Chelushkin, S.D., B.S., and A.K. contributed to figure and table generation for the manuscript. A. Bagaev, A. Zaitsev, D.D., M. Chelushkin, S.D., K. Nomie, R.A., and N.F. wrote and revised the manuscript and prepared the figures.

DECLARATION OF INTERESTS

N.F. is the Chief Medical Officer of BostonGene, Corp. and a professor at the University of Texas MD Anderson Cancer Center. A. Zaitsev, M. Chelushkin, V.Z., B.S., D.D., E. Nuzhdina, A. Bagaev, and R.A. are inventors on patent applications related to Kassandra. All other authors declare no competing interests.

Received: December 1, 2021
Revised: May 10, 2022
Accepted: July 12, 2022
Published: August 8, 2022

REFERENCES

Altman, M.C., Gill, M.A., Whalen, E., Babineau, D.C., Shao, B., Liu, A.H., Jepson, B., Gruchalla, R.S., O'Connor, G.T., Pongracic, J.A., et al. (2019). Transcriptome networks identify mechanisms of viral and nonviral asthma exacerbations in children. Nat. Immunol. *20*, 637–651.

Aran, D., Sirota, M., and Butte, A.J. (2015). Systematic pan-cancer analysis of tumour purity. Nat. Commun. *6*, 8971.

Aran, D., Hu, Z., and Butte, A.J. (2017). xCell: digitally portraying the tissue cellular heterogeneity landscape. Genome Biol. *18*, 220.

Athar, A., Füllgrabe, A., George, N., Iqbal, H., Huerta, L., Ali, A., Snow, C., Fonseca, N.A., Petryszak, R., Papatheodorou, I., et al. (2019). ArrayExpress update - from bulk to single-cell expression data. Nucleic Acids Res. *47*, D711–D715.

Avila Cobos, F., Alquicira-Hernandez, J., Powell, J.E., Mestdagh, P., and De Preter, K. (2020). Benchmarking of cell type deconvolution pipelines for transcriptomics data. Nat. Commun. *11*, 5650.

Barrett, T., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M., Holko, M., et al. (2012). NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Res. *41*, D991–D995.

Becht, E., Giraldo, N.A., Lacroix, L., Buttard, B., Elarouci, N., Petitprez, F., Selves, J., Laurent-Puig, P., Sautès-Fridman, C., Fridman, W.H., et al. (2016). Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. Genome Biol. *17*, 218.

Bolotin, D.A., Poslavsky, S., Mitrophanov, I., Shugay, M., Mamedov, I.Z., Putintseva, E.V., and Chudakov, D.M. (2015). MiXCR: software for comprehensive adaptive immunity profiling. Nat. Methods *12*, 380–381.

Bray, N.L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. Nat. Biotechnol. *34*, 525–527.

Britanova, O.V., Putintseva, E.V., Shugay, M., Merzlyak, E.M., Turchaninova, M.A., Staroverov, D.B., Bolotin, D.A., Lukyanov, S., Bogdanova, E.A., Mamedov, I.Z., et al. (2014). Age-related decrease in TCR repertoire diversity measured with deep and normalized sequence profiling. J. Immunol. *192*, 2689–2698.

Britanova, O.V., Shugay, M., Merzlyak, E.M., Staroverov, D.B., Putintseva, E.V., Turchaninova, M.A., Mamedov, I.Z., Pogorelyy, M.V., Bolotin, D.A., Izraelson, M., et al. (2016). Dynamics of individual T cell repertoires: from cord blood to centenarians. J. Immunol. *196*, 5005–5013.

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat. Biotechnol. *36*, 411–420.

Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., Moore, S., Phillips, S., Maffitt, D., Pringle, M., et al. (2013). The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. J. Digit. Imaging *26*, 1045–1057.

Cossarizza, A., Chang, H.-D., Radbruch, A., Acs, A., Adam, D., Adam-Klages, S., Agace, W.W., Aghaeepour, N., Akdis, M., Allez, M., et al. (2019). Guidelines for the use of flow cytometry and cell sorting in immunological studies (second edition). Eur. J. Immunol. *49*, 1457–1973.

Ding, B., Zheng, L., Zhu, Y., Li, N., Jia, H., Ai, R., Wildberg, A., and Wang, W. (2015). Normalization and noise reduction for single cell RNA-seq experiments. Bioinformatics *31*, 2225–2227.

Eisenberg, E., and Levanon, E.Y. (2013). Human housekeeping genes, revisited. Trends Genet. *29*, 569–574.

Finotello, F., Mayer, C., Plattner, C., Laschober, G., Rieder, D., Hackl, H., Krogsdam, A., Loncova, Z., Posch, W., Wilflingseder, D., et al. (2019). Molecular and pharmacological modulators of the tumor immune contexture revealed by deconvolution of RNA-seq data. Genome Med. *11*, 34.

Frankish, A., Diekhans, M., Ferreira, A.-M., Johnson, R., Jungreis, I., Loveland, J., Mudge, J.M., Sisu, C., Wright, J., Armstrong, J., et al. (2019). GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res. *47*, D766–D773.

Galon, J., Costes, A., Sanchez-Cabo, F., Kirilovsky, A., Mlecnik, B., Lagorce-Pagès, C., Tosolini, M., Camus, M., Berger, A., Wind, P., et al. (2006). Type, density, and location of immune cells within human colorectal tumors predict clinical outcome. Science *313*, 1960–1964.

George, B., Ashokachandran, V., Paul, A.M., and Girijadevi, R. (2017). Transcriptome sequencing for precise and accurate measurement of transcripts and accessibility of TCGA for cancer datasets and analysis. In Applications of RNA-Seq and Omics Strategies - From Microorganisms to Human Health, F.A. Marchi, P.D.R. Cirillo, and E.C. Mateo, eds. (InTech).

Gide, T.N., Quek, C., Menzies, A.M., Tasker, A.T., Shang, P., Holst, J., Madore, J., Lim, S.Y., Velickovic, R., Wongchenko, M., et al. (2019). Distinct immune cell populations define response to anti-PD-1 monotherapy and anti-PD-1/anti-CTLA-4 combined therapy. Cancer Cell *35*, 238–255.e6.

Griffiths, J.A., Richard, A.C., Bach, K., Lun, A.T.L., and Marioni, J.C. (2018). Detection and removal of barcode swapping in single-cell RNA-seq data. Nat. Commun. *9*, 2667.

GTEx Consortium (2013). The genotype-tissue expression (GTEx) project. Nat. Genet. *45*, 580–585.

Hagberg, A., Swart, P., and S Chult, D. (2008). Exploring network structure, dynamics, and function using networkx. In Proceedings of the 7th Python in Science Conference (Los Alamos, NM (United States): Los Alamos National Lab. (LANL)).

Hao, Y., Yan, M., Heath, B.R., Lei, Y.L., and Xie, Y. (2019). Fast and robust deconvolution of tumor infiltrating lymphocyte from expression profiles using least trimmed squares. PLoS Comput. Biol. *15*, e1006976.

Hirata, E., and Sahai, E. (2017). Tumor microenvironment and differential responses to therapy. Cold Spring Harb. Perspect. Med. *7*, a026781.

Hoek, K.L., Samir, P., Howard, L.M., Niu, X., Prasad, N., Galassie, A., Liu, Q., Allos, T.M., Floyd, K.A., Guo, Y., et al. (2015). A cell-based systems biology assessment of human blood to monitor immune responses after influenza vaccination. PLoS One *10*, e0118528.

Hugo, W., Zaretsky, J.M., Sun, L., Song, C., Moreno, B.H., Hu-Lieskovan, S., Berent-Maoz, B., Pang, J., Chmielowski, B., Cherry, G., et al. (2016). Genomic and transcriptomic features of response to anti-PD-1 therapy in metastatic melanoma. Cell *165*, 35–44.

Hunter, J.D. (2007). Matplotlib: A 2D Graphics Environment. Comput. Sci. Eng. *9*, 90–95. https://doi.org/10.1109/MCSE.2007.55.

Izar, B., Tirosh, I., Stover, E.H., Wakiro, I., Cuoco, M.S., Alter, I., Rodman, C., Leeson, R., Su, M.-J., Shah, P., et al. (2020). A single-cell landscape of high-grade serous ovarian cancer. Nat. Med. *26*, 1271–1279.

Jackson, H.W., Fischer, J.R., Zanotelli, V.R.T., Ali, H.R., Mechera, R., Soysal, S.D., Moch, H., Muenst, S., Varga, Z., Weber, W.P., et al. (2020). The single-cell pathology landscape of breast cancer. Nature *578*, 615–620.

Jew, B., Alvarez, M., Rahmani, E., Miao, Z., Ko, A., Garske, K.M., Sul, J.H., Pietiläinen, K.H., Pajukanta, P., and Halperin, E. (2020). Accurate estimation of cell composition in bulk expression through robust integration of single-cell information. Nat. Commun. *11*, 1971.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In Advances in Neural Information Processing Systems (Curran Associates, Inc.).

Kim, S.T., Cristescu, R., Bass, A.J., Kim, K.-M., Odegaard, J.I., Kim, K., Liu, X.Q., Sher, X., Jung, H., Lee, M., et al. (2018). Comprehensive molecular characterization of clinical responses to PD-1 inhibition in metastatic gastric cancer. Nat. Med. *24*, 1449–1458.

Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., Baglaenko, Y., Brenner, M., Loh, P.-R., and Raychaudhuri, S. (2019). Fast, sensitive and accurate integration of single-cell data with Harmony. Nat. Methods *16*, 1289–1296.

Lambrechts, D., Wauters, E., Boeckx, B., Aibar, S., Nittner, D., Burton, O., Bassez, A., Decaluwé, H., Pircher, A., Van den Eynde, K., et al. (2018). Phenotype molding of stromal cells in the lung tumor microenvironment. Nat. Med. *24*, 1277–1289.

Levine, J.H., Simonds, E.F., Bendall, S.C., Davis, K.L., Amir, E.-A.D., Tadmor, M.D., Litvin, O., Fienberg, H.G., Jager, A., Zunder, E.R., et al. (2015). Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. Cell *162*, 184–197.

Li, B., Ruotti, V., Stewart, R.M., Thomson, J.A., and Dewey, C.N. (2010). RNA-Seq gene expression estimation with read mapping uncertainty. Bioinformatics *26*, 493–500.

Linsley, P.S., Speake, C., Whalen, E., and Chaussabel, D. (2014). Copy number loss of the interferon gene cluster in melanomas is linked to reduced T cell infiltrate and poor patient prognosis. PLoS One *9*, e109760.

Liu, C.C., Steen, C.B., and Newman, A.M. (2019a). Computational approaches for characterizing the tumor immune microenvironment. Immunology *158*, 70–84.

Liu, D., Schilling, B., Liu, D., Sucker, A., Livingstone, E., Jerby-Arnon, L., Zimmer, L., Gutzmer, R., Satzger, I., Loquai, C., et al. (2019b). Integrative molecular and clinical modeling of clinical outcomes to PD1 blockade in patients with metastatic melanoma. Nat. Med. *25*, 1916–1927.

Liu, Y., Zugazagoitia, J., Ahmed, F.S., Henick, B.S., Gettinger, S.N., Herbst, R.S., Schalper, K.A., and Rimm, D.L. (2020). Immune cell PD-L1 colocalizes with macrophages and is associated with outcome in PD-1 pathway blockade therapy. Clin. Cancer Res. *26*, 970–977.

Lun, A.T.L., Riesenfeld, S., Andrews, T., Dao, T.P., and Gomes, T.; participants in the 1st Human Cell Atlas Jamboree (2019). EmptyDrops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. Genome Biol. *20*, 63.

Lundberg, S.M., Nair, B., Vavilala, M.S., Horibe, M., Eisses, M.J., Adams, T., Liston, D.E., Low, D.K.-W., Newman, S.-F., Kim, J., and Lee, S.I. (2018). Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. Nat. Biomed. Eng. *2*, 749–760.

Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., and Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. Nat. Mach. Intell. *2*, 56–67.

Mariathasan, S., Turley, S.J., Nickles, D., Castiglioni, A., Yuen, K., Wang, Y., Kadel, E.E., III, Koeppen, H., Astarita, J.L., Cubas, R., et al. (2018). TGFβ attenuates tumour response to PD-L1 blockade by contributing to exclusion of T cells. Nature *554*, 544–548.

McKinney, W. (2011). pandas: a Foundational Python Library for Data Analysis and Statistics, *9*.

Melsted, P., Ntranos, V., and Pachter, L. (2019). The barcode, UMI, set format and BUStools. Bioinformatics *35*, 4472–4473.

Menden, K., Marouf, M., Oller, S., Dalmia, A., Magruder, D.S., Kloiber, K., Heutink, P., and Bonn, S. (2020). Deep learning–based cell composition analysis from tissue expression profiles. Sci. Adv. *6*, eaba2619.

Monaco, G., Lee, B., Xu, W., Mustafah, S., Hwang, Y.Y., Carré, C., Burdin, N., Visan, L., Ceccarelli, M., Poidinger, M., et al. (2019). RNA-seq signatures normalized by mRNA abundance allow absolute deconvolution of human immune cell types. Cell Rep. *26*, 1627–1640.e7.

Morsch, R., Rose, M., Maurer, A., Cassataro, M.A., Braunschweig, T., Knüchel, R., Vögeli, T.A., Ecke, T., Eckstein, M., Weyerer, V., et al. (2020). Therapeutic implications of PD-L1 expression in bladder cancer with squamous differentiation. BMC Cancer *20*, 230.

Nadel, B.B., Lopez, D., Montoya, D.J., Ma, F., Waddel, H., Khan, M.M., Mangul, S., and Pellegrini, M. (2021). The Gene Expression Deconvolution Interactive Tool (GEDIT): accurate cell type quantification from gene expression data. GigaScience *10*, giab002.

Nathanson, T., Ahuja, A., Rubinsteyn, A., Aksoy, B.A., Hellmann, M.D., Miao, D., Van Allen, E., Merghoub, T., Wolchok, J.D., Snyder, A., et al. (2017). Somatic mutations and neoepitope homology in melanomas treated with CTLA-4 blockade. Cancer Immunol. Res. *5*, 84–91.

Neftel, C., Laffy, J., Filbin, M.G., Hara, T., Shore, M.E., Rahme, G.J., Richman, A.R., Silverbush, D., Shaw, M.L., Hebert, C.M., et al. (2019). An integrative

model of cellular states, plasticity, and genetics for glioblastoma. Cell *178*, 835–849.e21.

Newman, A.M., Liu, C.L., Green, M.R., Gentles, A.J., Feng, W., Xu, Y., Hoang, C.D., Diehn, M., and Alizadeh, A.A. (2015). Robust enumeration of cell subsets from tissue expression profiles. Nat. Methods *12*, 453–457.

Newman, A.M., Steen, C.B., Liu, C.L., Gentles, A.J., Chaudhuri, A.A., Scherer, F., Khodadoust, M.S., Esfahani, M.S., Luca, B.A., Steiner, D., et al. (2019). Determining cell type abundance and expression from bulk tissues with digital cytometry. Nat. Biotechnol. *37*, 773–782.

Newton, Y., Sedgewick, A.J., Cisneros, L., Golovato, J., Johnson, M., Szeto, C.W., Rabizadeh, S., Sanborn, J.Z., Benz, S.C., and Vaske, C. (2020). Large scale, robust, and accurate whole transcriptome profiling from clinical formalin-fixed paraffin-embedded samples. Sci. Rep. *10*, 17597.

Norton, J., Foster, D., Chinta, M., Titan, A., and Longaker, M. (2020). Pancreatic cancer associated fibroblasts (CAF): under-explored target for pancreatic cancer treatment. Cancers *12*, E1347.

Pachynski, R.K., Kim, E.H., Miheecheva, N., Kotlov, N., Ramachandran, A., Postovalova, E., Galkin, I., Svekolkin, V., Lyu, Y., Zou, Q., et al. (2021). Single-cell spatial proteomic revelations on the multiparametric MRI heterogeneity of clinically significant prostate cancer. Clin. Cancer Res. *27*, 3478–3490.

Pal, S.K., McDermott, D.F., Atkins, M.B., Escudier, B., Rini, B.I., Motzer, R.J., Fong, L., Joseph, R.W., Oudard, S., Ravaud, A., et al. (2020). Patient-reported outcomes in a phase 2 study comparing atezolizumab alone or with bevacizumab vs sunitinib in previously untreated metastatic renal cell carcinoma. BJU Int. *126*, 73–82.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine Learning in Python (Mach. Learn). PYTHON 6.

Puram, S.V., Tirosh, I., Parikh, A.S., Patel, A.P., Yizhak, K., Gillespie, S., Rodman, C., Luo, C.L., Mroz, E.A., Emerick, K.S., et al. (2017). Single-cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer. Cell *171*, 1611–1624.e24.

Rabadan, R., Bhanot, G., Marsilio, S., Chiorazzi, N., Pasqualucci, L., and Khiabanian, H. (2018). On statistical modeling of sequencing noise in high depth data to assess tumor evolution. J. Stat. Phys. *172*, 143–155.

Racle, J., de Jonge, K., Baumgaertner, P., Speiser, D.E., and Gfeller, D. (2017). Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data. Elife *6*, e26476.

Rakaee, M., Busund, L.-T.R., Jamaly, S., Paulsen, E.-E., Richardsen, E., Andersen, S., Al-Saad, S., Bremnes, R.M., Donnem, T., and Kilvaer, T.K. (2019). Prognostic value of macrophage phenotypes in resectable non-small cell lung cancer assessed by Multiplex immunohistochemistry. Neoplasia *21*, 282–293.

Reuben, A., Zhang, J., Chiou, S.-H., Gittelman, R.M., Li, J., Lee, W.-C., Fujimoto, J., Behrens, C., Liu, X., Wang, F., et al. (2020). Comprehensive T cell repertoire characterization of non-small cell lung cancer. Nat. Commun. *11*, 603.

Roider, T., Seufert, J., Uvarovskii, A., Frauhammer, F., Bordas, M., Abedpour, N., Stolarczyk, M., Mallm, J.-P., Herbst, S.A., Bruch, P.-M., et al. (2020). Dissecting intratumour heterogeneity of nodal B-cell lymphomas at the transcriptional, genetic and drug-response levels. Nat. Cell Biol. *22*, 896–906.

Saltz, J., Gupta, R., Hou, L., Kurc, T., Singh, P., Nguyen, V., Samaras, D., Shroyer, K.R., Zhao, T., Batiste, R., et al. (2018). Spatial organization and molecular correlation of tumor-infiltrating lymphocytes using deep learning on pathology images. Cell Rep. *23*, 181–193.e7.

Sharonov, G.V., Serebrovskaya, E.O., Yuzhakova, D.V., Britanova, O.V., and Chudakov, D.M. (2020). B cells, plasma cells and antibody repertoires in the tumour microenvironment. Nat. Rev. Immunol. *20*, 294–307.

Shin, H., Shannon, C.P., Fishbane, N., Ruan, J., Zhou, M., Balshaw, R., Wilson-McManus, J.E., Ng, R.T., McManus, B.M., and Tebbutt, S.J.; PROOF Centre of Excellence Team (2014). Variation in RNA-Seq transcriptome profiles of peripheral whole blood from healthy individuals with and without globin depletion. PLoS One *9*, e91041.

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W.M., 3rd, Hao, Y., Stoeckius, M., Smibert, P., and Satija, R. (2019). Comprehensive integration of single-cell data. Cell *177*, 1888–1902.e21.

Tirosh, I., Izar, B., Prakadan, S.M., Wadsworth, M.H., 2nd, Treacy, D., Trombetta, J.J., Rotem, A., Rodman, C., Lian, C., Murphy, G., et al. (2016). Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science *352*, 189–196.

Torroja, C., and Sanchez-Cabo, F. (2019). Corrigendum: digitaldlsorter: deep-learning on scRNA-seq to deconvolute gene expression data. Front. Genet. *10*, 1373.

Van Allen, E.M., Miao, D., Schilling, B., Shukla, S.A., Blank, C., Zimmer, L., Sucker, A., Hillen, U., Foppen, M.H.G., Goldinger, S.M., et al. (2015). Genomic correlates of response to CTLA-4 blockade in metastatic melanoma. Science *350*, 207–211.

Van Gassen, S., Callebaut, B., Van Helden, M.J., Lambrecht, B.N., Demeester, P., Dhaene, T., and Saeys, Y. (2015). FlowSOM: using self-organizing maps for visualization and interpretation of cytometry data. Cytometry A. *87*, 636–645.

Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat. Methods *17*, 261–272. https://doi.org/10.1038/s41592-019-0686-2.

Vivian, J., Rao, A.A., Nothaft, F.A., Ketchum, C., Armstrong, J., Novak, A., Pfeil, J., Narkizian, J., Deran, A.D., Musselman-Brown, A., et al. (2017). Toil enables reproducible, open source, big biomedical data analyses. Nat. Biotechnol. *35*, 314–316.

Wagner, G.P., Kin, K., and Lynch, V.J. (2012). Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. Theory Biosci. *131*, 281–285.

Wang, X., Park, J., Susztak, K., Zhang, N.R., and Li, M. (2019). Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. Nat. Commun. *10*, 380.

Wei, S.C., Duffy, C.R., and Allison, J.P. (2018). Fundamental mechanisms of immune checkpoint blockade therapy. Cancer Discov. *8*, 1069–1086.

Wei, Y., Zhao, Q., Gao, Z., Lao, X.-M., Lin, W.-M., Chen, D.-P., Mu, M., Huang, C.-X., Liu, Z.-Y., Li, B., et al. (2019). The local immune landscape determines tumor PD-L1 heterogeneity and sensitivity to therapy. J. Clin. Invest. *129*, 3347–3360.

Zaitsev, K., Bambouskova, M., Swain, A., and Artyomov, M.N. (2019). Complete deconvolution of cellular mixtures based on linearity of transcriptional signatures. Nat. Commun. *10*, 2209.

Zheng, G.X.Y., Terry, J.M., Belgrader, P., Ryvkin, P., Bent, Z.W., Wilson, R., Ziraldo, S.B., Wheeler, T.D., McDermott, G.P., Zhu, J., et al. (2017). Massively parallel digital transcriptional profiling of single cells. Nat. Commun. *8*, 14049.

Zimmermann, M.T., Oberg, A.L., Grill, D.E., Ovsyannikova, I.G., Haralambieva, I.H., Kennedy, R.B., and Poland, G.A. (2016). System-wide associations between DNA-methylation, gene expression, and humoral immune response to influenza vaccination. PLoS One *11*, e0152034.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Mouse Anti-Human CD13-BB700, clone: L138, cat#:746057	BD Biosciences	RRID: AB_2743440
Mouse Anti-Human CCR3-BV421, clone: 5E8, cat#:310714	Biolegend	RRID: AB_2561886
Mouse Anti-Human CD123-BV605, clone: 6H6, cat#:306026	Biolegend	RRID: AB_2563826
Mouse Anti-Human HLA-DR-BV650, clone: L243, cat#:307650	Biolegend	RRID: AB_2563828
Mouse Anti-Human CD3-BV711, clone: UCHT1, cat#:563725	BD Biosciences	RRID: AB_2744392
Mouse Anti-Human CD45-BV786, clone: HI30, cat#:304048	Biolegend	RRID: AB_2563129
Mouse Anti-Human CD66b-PE, clone: 6/40c, cat#:392904	Biolegend	RRID: AB_2750202
Mouse Anti-Human CD56-PE-Dazzle 594, clone: 5.1H11, cat#:362544	Biolegend	RRID: AB_2565922
Mouse Anti-Human CD11c-PE-Cy7, clone: 3.9, cat#:301607	Biolegend	RRID: AB_389350
Mouse Anti-Human CD19-PE-Cy5, clone: HIB19, cat#:302210	Biolegend	RRID: AB_314240
Mouse Anti-Human FcεR1-Alexa Flour 488, clone: AER-37, cat#:334640	Biolegend	RRID: AB_2721290
Mouse Anti-Human CD10-BB700, clone: MEM-78, cat#:746101	BD Biosciences	RRID: AB_2743472
Mouse Anti-Human CD125-BV421, clone: A14, cat#:743927	BD Biosciences	RRID: AB_2741855
Mouse Anti-Human CD16-BV650, clone: 3G8, cat#:302042	Biolegend	RRID: AB_2563801
Mouse Anti-Human CD64-BV711, clone: 10.1, cat#:305042	Biolegend	RRID: AB_2800778
Mouse Anti-Human CCR3-PE-Dazzle 594, clone: 5E8, cat#:310728	Biolegend	RRID: AB_2687007
Mouse Anti-Human CD117-PE-Cy7, clone: 104D2, cat#:313211	Biolegend	RRID: AB_893228
Mouse Anti-Human CD3-PE-Cy5, clone: HIT3a, cat#:300310	Biolegend	RRID: AB_314046
Mouse Anti-Human CD56-PE-Cy5, clone: 5.1H11, cat#:362516	Biolegend	RRID: AB_2564089
Mouse Anti-Human CD14-PE-Cy5, clone: M5E2, cat#:301864	Biolegend	RRID: AB_2860767
Mouse Anti-Human CD14-Alexa Flour 488, clone: M5E2, cat#:301811	Biolegend	RRID: AB_493159
Mouse Anti-Human CD9-BB700, clone: M-L13, cat#:745827	BD Biosciences	RRID: AB_2743276
Mouse Anti-Human CD16-BV421, clone: 3G8, cat#:562874	BD Biosciences	RRID: AB_2716865
Mouse Anti-Human CD3-BV510, clone: OKT3, cat#:317332	Biolegend	RRID: AB_2561943
Mouse Anti-Human CCR3-BV510, clone: 5E8, cat#:310722	Biolegend	RRID: AB_2571977
Mouse Anti-Human CD19-BV510, clone: HIB19, cat#:302242	Biolegend	RRID: AB_2561668
Mouse Anti-Human CD7-BV510, clone: M-T701, cat#:563650	BD Biosciences	RRID: AB_2713913
Mouse Anti-Human FcεR1-BV605, clone: AER-37, cat#:334628	Biolegend	RRID: AB_2566506
Mouse Anti-Human CD33-BV711, clone: WM53, cat#:303424	Biolegend	RRID: AB_2565775
Mouse Anti-Human CD84-PE, clone: CD84.1.21, cat#:326008	Biolegend	RRID: AB_2229003
Mouse Anti-Human CD15-PE-dazzle 594, clone: W6D3, cat#:323038	Biolegend	RRID: AB_2564103
Mouse Anti-Human CD169-PE-Cy7, clone: 7-239, cat#:346014	Biolegend	RRID: AB_2750264
Mouse Anti-Human CD206-PE-Cy5, clone: 15-2, cat#:321108	Biolegend	RRID: AB_571919
Mouse Anti-Human CD1c-BV421, clone: L161, cat#:331525	Biolegend	RRID: AB_10933249
Mouse Anti-Human CD15-BV510, clone: W6D3, cat#:563141	BD Biosciences	RRID: AB_2738025
Mouse Anti-Human CD19-BV510, clone: SJ25C1, cat#:363020	Biolegend	RRID: AB_2564229
Mouse Anti-Human CCR3-BV510, clone: 5E8, cat#:310721	Biolegend	RRID: AB_2571976
Mouse Anti-Human CD16-BV711, clone: 3G8, cat#:302044	Biolegend	RRID: AB_2563802
Mouse Anti-Human CLEC9A-PE, clone: 8F9, cat#:353804	Biolegend	RRID: AB_10965546
Mouse Anti-Human CD141-PE-Dazzle 594, clone: M80, cat#:344120	Biolegend	RRID: AB_2687144

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Mouse Anti-Human CD19-BB515, clone: HIB19, cat#:564456	BD Biosciences	RRID: AB_2744309
Mouse Anti-Human IgD-BB700, clone: IA6-2, cat#:566538	BD Biosciences	RRID: AB_2744486
Mouse Anti-Human CD138-BV421, clone: MI15, cat#:356515	Biolegend	RRID: AB_2562659
Mouse Anti-Human CD13-BV510, clone: WM15, cat#:740162	BD Biosciences	RRID: AB_2739915
Mouse Anti-Human IgG-BV605, clone: G18-145, cat#:563246	BD Biosciences	RRID: AB_2738092
Mouse Anti-Human CD39-BV650, clone: TU66, cat#:563681	BD Biosciences	RRID: AB_2738370
Mouse Anti-Human CD24-BV711, clone: ML5, cat#:311136	Biolegend	RRID: AB_2566579
Mouse Anti-Human CD10-BV786, clone: HI10a, cat#:564960	BD Biosciences	RRID: AB_2739025
Goat Anti-Human IgA-PE, polyclonal, cat#: 2050-09	Southern Biotech	RRID: AB_2795707
Mouse Anti-Human IgM-PE-Dazzle 594, clone: MHM-88, cat#:314529	Biolegend	RRID: AB_2566482
Mouse Anti-Human CD27-PE-Cy7, clone:M-T271, cat#:356412	Biolegend	RRID: AB_2562258
Mouse Anti-Human CD38-PE-Cy5, clone:HIT2, cat#:303508	Biolegend	RRID: AB_314360
Mouse Anti-Human CD45-BB515, clone:H130, cat#:564585	BD Biosciences	RRID: AB_2732068
Mouse Anti-Human NKp44-BB700, clone:p44-8	BD Biosciences	cat#:624381
Mouse Anti-Human CD123-BV510, clone:6H6, cat#:306022	Biolegend	RRID: AB_2562068
Mouse Anti-Human NKG2A-BV605, clone:131411, cat#:747921	BD Biosciences	RRID: AB_2872382
Mouse Anti-Human CD158-BV650, clone:HP-MA4	BD Biosciences	cat#:752506
Mouse Anti-Human NKG2C-BV711, clone:134591, cat#:748164	BD Biosciences	RRID: AB_2872625
Mouse Anti-Human CD57-BV786, clone:QA17A04, cat#:393329	Biolegend	RRID: AB_2860967
Mouse Anti-Human CD161-PE, clone:HP-3G10, cat#:339904	Biolegend	RRID: AB_1501083
Mouse Anti-Human NKG2D-PE-Cy7, clone:1D11, cat#:320812	Biolegend	RRID: AB_2234394
Mouse Anti-Human CD107a-PE-Cy5, clone:eBioH4A3, cat#:15-1079-42	ThermoFisher Scientific	RRID: AB_10547280
Mouse Anti-Human CD27-BB515, clone:M-T271, cat#:564643	BD Biosciences	RRID: AB_2744354
Mouse Anti-Human CD8-BB700, clone:RPA-T8, cat#:566452	BD Biosciences	RRID: AB_2744459
Mouse Anti-Human γδTCR-BV421, clone:11F2, cat#:744870	BD Biosciences	RRID: AB_2742548
Mouse Anti-Human CD3-BV605, clone:OKT3, cat#:317322	Biolegend	RRID: AB_2561911
Mouse Anti-Human iNKT-BV650, clone:6B11, cat#:744000	BD Biosciences	RRID: AB_2741919
Mouse Anti-Human TCR Vδ2-BV711, clone:B6, cat#:331412	Biolegend	RRID: AB_2565421
Mouse Anti-Human TCR Vα7.2-PE-Cy7, clone:3C10, cat#:351712	Biolegend	RRID: AB_2561994
Mouse Anti-Human CD45RA-PE-Cy5, clone:HI100, cat#:304110	Biolegend	RRID: AB_314414
Mouse Anti-Human CXCR3-BV421, clone:G025H7, cat#:353716	BD Biosciences	RRID: AB_2561448
Mouse Anti-Human CD4-BV510, clone:L200, cat#:563094	BD Biosciences	RRID: AB_2738001
Mouse Anti-Human CD62L-BV650, clone:DREG-56, cat#:304832	Biolegend	RRID: AB_2563821
Mouse Anti-Human CD95-BV711, clone:DX2, cat#:305644	Biolegend	RRID: AB_2632623
Mouse Anti-Human CX3CR1-PE, clone:2A9-1, cat#:341604	Biolegend	RRID: AB_1595456
Mouse Anti-Human PD-1-PE-Dazzle 594, clone:EH12.2H7, cat#:329940	Biolegend	RRID: AB_2563659
Mouse Anti-Human CXCR5-PE-Cy7, clone:J252D4, cat#:356924	Biolegend	RRID: AB_2562355
Mouse Anti-Human ICOS-BB515, clone:DX29, cat#:564549	BD Biosciences	RRID: AB_2738840
Mouse Anti-Human Tim3-BV421, clone:F38-2E2, cat#:345008	Biolegend	RRID: AB_11218598
Mouse Anti-Human γδTCR-BV510, clone:11F2, cat#:745026	BD Biosciences	RRID: AB_2742655
Mouse Anti-Human CD27-BV711, clone:M-T271, cat#:356430	Biolegend	RRID: AB_2650751
Mouse Anti-Human Lag3-BV786, clone:11C3C65, cat#:369322	Biolegend	RRID: AB_2716127
Mouse Anti-Human TIGIT-PE, clone:A15153G, cat#:372704	Biolegend	RRID: AB_2632730
Mouse Anti-Human CD39-PE-Cy7, clone:A1, cat#:328212	Biolegend	RRID: AB_2099950
Mouse Anti-Human CD4-BB515, clone:RPA-T4, cat#:564419	BD Biosciences	RRID: AB_2744419
Mouse Anti-Human CCR6-BB700, clone:11A9, cat#:746139	BD Biosciences	RRID: AB_2743501

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Mouse Anti-Human CXCR3-BV421, clone:1C6/CXCR3, cat#:562558	BD Biosciences	RRID: AB_2737653
Mouse Anti-Human CD8-BV510, clone:RPA-T8, cat#:301048	Biolegend	RRID: AB_2561942
Mouse Anti-Human CD45RA-BV786, clone:HI100, cat#:304140	Biolegend	RRID: AB_2563816
Mouse Anti-Human IL-7RA-PE, clone:A019D5, cat#:351340	Biolegend	RRID: AB_2564136
Mouse Anti-Human CCR4-PE-Dazzle 594, clone:L291H4, cat#:359420	Biolegend	RRID: AB_2564095
Mouse Anti-Human CD25-PE-Cy5, clone:BC96, cat#:302608	Biolegend	RRID: AB_314278
Mouse Anti-Human CTLA-4-BB515, clone:BNI3, cat#:566918	BD Biosciences	RRID: AB_2869947
Mouse Anti-Human CD4-BB700, clone:L200, cat#:566479	BD Biosciences	RRID: AB_2739738
Mouse Anti-Human CD25-BV421, clone:BC96, cat#:302630	Biolegend	RRID: AB_11126749
Mouse Anti-Human CD8-BV510, clone:RPA-T8, cat#:563256	BD Biosciences	RRID: AB_2738101
Mouse Anti-Human CD3-BV605, clone:UCHT1, cat#:300460	Biolegend	RRID: AB_2564380
Armenian Hamster Anti-Human ICOS-BV650, clone:C398.4A, cat#:313550	Biolegend	RRID: AB_2749929
Mouse Anti-Human Lag3-BV786, clone:11C3C65, cat#:369322	Biolegend	RRID: AB_2716127
Mouse Anti-Human CD127-PE-CF594, clone:HIL-7R-H21, cat#:562397	BD Biosciences	RRID: AB_11154212
Mouse Anti-Human CD4-APC-H7, clone:RPA-T4, cat#:560158	BD Biosciences	RRID: AB_1645478
Mouse Anti-Human CRTH2-R718, clone:BM16,	BD Biosciences	cat#:751948
Mouse Anti-Human CCR10-APC, clone:IB10, cat#:564771	BD Biosciences	RRID: AB_2738943
Mouse Anti-Human Tim3-Alexa Flour 647, clone:7D3, cat#:565559	BD Biosciences	RRID: AB_2744367
Mouse Anti-Human CD39-APC-FIRE750, clone:A1, cat#:328230	Biolegend	RRID: AB_2650839
Mouse Anti-Human CD27-BV786, clone:O323, cat#:302832	Biolegend	RRID: AB_2562674
Mouse Anti-Human CD69-APC-R700, clone:FN50, cat#:565154	BD Biosciences	RRID: AB_2744449
Mouse Anti-Human NKp80-PE, clone:5D12, cat#:566329	BD Biosciences	RRID: AB_2739689
Mouse Anti-Human CD14-PE-CF594, clone:MϕP9, cat#:562335	BD Biosciences	RRID: AB_11153663
Mouse Anti-Human CD169-APC, clone:7-239, cat#:346008	Biolegend	RRID: AB_11147948
Mouse Anti-Human CD13-BV711, clone:WM15, cat#:301722	Biolegend	RRID: AB_2687015
Mouse Anti-Human CD10-BV786, clone:HI10A, cat#:564960	BD Biosciences	RRID: AB_2739025
Mouse Anti-Human HLA-DR-APC-FIRE750, clone:LN2, cat#:327024	Biolegend	RRID: AB_2810492
Mouse Anti-Human CD36-PE-Cy7, clone:5-271, cat#:336222	Biolegend	RRID: AB_2716142
Mouse Anti-Human CD11c-Alexa Flour 700, clone:B-ly6, cat#:561352	BD Biosciences	RRID: AB_10612006
Mouse Anti-Human CD3-Alexa Flour 488, clone:UCHT1, cat#:300415	Biolegend	RRID: AB_389310
Mouse Anti-Human CCR4-BV605, clone:L291H4, cat#:359418	Biolegend	RRID: AB_2562483
Unlabeled Normal Mouse IgG, cat#: 0107-01	Southern Biotech	RRID: AB_2732898
Human TrueStain FcX, cat#: 422302	Biolegend	RRID: AB_2818986
Insulin solution human CAS No.11061-68-0 Sigma	Sigma	CAS No.11061-68-0
Biological samples		
Whole blood from healthy donors, collected in K2-EDTA vacutainers (Purple top)	Research Blood Components	Item#: 016-018
ccRCC biopsies	James Hsieh	
NSCLC biopsies	Mark Poznansky	
Throat biopsies	Leandro Cerchietti	
Chemicals, peptides, and recombinant proteins		
Brilliant stain buffer, cat#:566385	BD Biosciences	RRID: AB_2869761
Monocyte Blocker	Biolegend	cat#:426103

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
ImmunoCult™ Human CD3/CD28 T Cell Activator	Stem Cell Technologies	cat#:10971
SYTOX™ Red Dead Cell Stain	ThermoFisher Scientific	cat#:S34859
ViaStain™ AOPI Staining Solution	Nexcelom	cat#:CS2-0106
Ghost Dye Violet 510 Viability Dye	Tonbo Biosciences	cat#:13-0870-T100
CytoFIX/CytoPERM	BD Biosciences	cat#:554722
Ficoll-Paque PLUS	Cytiva	cat#:17144003
CryoStor-CS10	BioLife Solutions	cat#:210102
Newborn Calf Serum	MiliporeSIGMA	cat#:N4762
Fetal Bovine Serum	Corning	cat#:35-011-CV
Ethylenediaminetetraacetic acid (EDTA) 0.5M	ThermoFisher Scientific	cat#:15575-038
Penicillin-Streptomycin-Glutamine (100X)	ThermoFisher Scientific	cat#:10378016
RPMI Medium 1640 (1X)	ThermoFisher Scientific	cat#:11875-093
IMDM, no phenol red	ThermoFisher Scientific	cat#:21056023
CellGenix® T Cell Medium GMP-Prototype	CellGenix	cat#:24814-0500
Phosphate Buffered Saline (1X)	ThermoFisher Scientific	cat#:20021-027
Critical commercial assays		
EasySep™ Human $CD4^+$ T Cell Enrichment Kit,	Stem Cell Technologies	cat#:19052
EasySep™ Human Basophil Isolation Kit	Stem Cell Technologies	cat#:17969
EasySep™ Human Pan-DC Pre-Enrichment Kit	Stem Cell Technologies	cat#:19251
RosetteSep™ Human Granulocyte Depletion Cocktail	Stem Cell Technologies	cat#:15664
RosetteSep™ Human $CD8^+$ T Cell Enrichment Cocktail	Stem Cell Technologies	cat#:15063
RosetteSep™ Human NK Cell Enrichment Cocktail	Stem Cell Technologies	cat#:15065
RosetteSep™ Human $CD4^+$ T Cell Enrichment Cocktail	Stem Cell Technologies	cat#:15065
Universal Mycoplasma Detection Kit	ATCC	cat#: 30-1012K
Deposited data		
Sorted cell compendium of RNA profiles	This paper	https://science.bostongene.com/kassandra/
Sequenced raw RNA-seq data of cellular subpopulations	This paper	EGA: EGAS00001006272
Full list of datasets used for training Kassandra-Tumor		Table S8
The list of datasets used to train Kassandra-Blood		Table S8
Full list of holdout datasets		Table S8
The list of whole blood RNA-Seq datasets		Table S8
Anti-PDL-1-treated bladder cancer	Mariathasan et al., 2018	EGA: EGAS00001002556
Anti-PD-L1-treated gastric cancer	Kim et al., 2018	ERP107734
ccRCC immotion150	Pal et al., 2020	EGA: EGAC00001000946
Anti-CTLA-4-treated melanoma	Van Allen et al., 2015	dbGAP: phs000452
Anti-CTLA-4-treated melanoma	Nathanson et al., 2017	SRA: SRP067586
Anti-PD-1^+ anti-CTLA-4 or anti-PD-1-treated melanoma	Gide et al., 2019	ENA: ERP105482
Anti-PD-1-treated metastatic melanoma	Liu et al., 2019b	dbGAP: phs001036
Anti-PD-1-treated melanoma	Hugo et al., 2016	GEO: GSE78220, GEO: GSE96619
Melanoma scRNA-seq validation dataset	Tirosh et al., 2016	GEO: GSE72056
Head and neck carcinoma scRNA-seq validation dataset	Puram et al., 2017	GEO: GSE103322
Glioblastoma scRNA-seq validation dataset	Neftel et al., 2019	
PBMC scRNA-seq validation datasets		10X Genomics
PBMC scRNA-seq validation dataset	Zheng et al., 2017	
Lung cancer scRNA-seq validation datasets	Lambrechts et al., 2018	ArrayExpress: E-MTAB-6149, ArrayExpress: E-MTAB-6653
Ovarian cancer scRNA-seq validation datasets	Izar et al., 2020	GEO: GSE146026
B-cell lymphoma scRNA-seq validation dataset	Roider et al., 2020	

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Flow cytometry validation datasets	Monaco et al., 2019;	
Flow cytometry validation datasets	Finotello et al., 2019	
Flow cytometry validation datasets	Newman et al., 2019	
Flow cytometry validation datasets	Linsley et al., 2014	
Flow cytometry validation datasets	Hoek et al., 2015	
Flow cytometry validation datasets	Zimmermann et al., 2016	
Experimental models: Cell lines		
K562 Chronic Myelogenous Leukemia, Lymphoblast, cat#: CCL-243	ATCC	CVCL_0004
COLO829, skin fibroblast, cat: CRL-1974	ATCC	CVCL_1137
MCF7, epidermal adenocarcinoma, cat#:HTB-22	ATCC	CVCL_003
Software and algorithms		
Kassandra code	This paper	https://github.com/BostonGene/Kassandra
scikit-learn	Pedregosa et al., 2011	https://github.com/scikit-learn/scikit-learn
Pandas	McKinney, 2011	https://github.com/pandas-dev/pandas
scipy	Virtanen et al., 2020	https://github.com/scipy/scipy
numpy	-	https://github.com/numpy/numpy
matplotlib	Hunter, 2007	
pyyaml	-	https://github.com/yaml/pyyaml
matplotlib	-	https://github.com/matplotlib/matplotlib
networkx	Hagberg et al., 2008	https://github.com/networkx
lightgbm	Ke et al., 2017	https://github.com/microsoft/LightGBM
ABIS	Monaco et al., 2019	https://github.com/giannimonaco/ABIS
EPIC	Racle et al., 2017	https://github.com/GfellerLab/EPIC/releases/tag/v1.1
CIBERSORT	Newman et al., 2015	https://cibersort.stanford.edu/
CIBERSORTx	Newman et al., 2019	https://cibersortx.stanford.edu/
QuanTIseq	Finotello et al., 2019	https://icbi.i-med.ac.at/software/quantiseq/doc/index.html#quanTIseq
xCell	Aran et al., 2017	https://github.com/dviraran/xCell
FARDEEP	Hao et al., 2019	https://github.com/YuningHao/FARDEEP.git
MCP-counter	Becht et al., 2016	https://github.com/ebecht/MCPcounter
Scaden	Menden et al., 2020	https://github.com/KevinMenden/scaden
Other		
BD FACSAria™ III Cell Sorter	BD Biosciences	Part #: 648282C2
BD FACSCelesta™ Cell Analyzer	BD Biosciences	Part #: 660345
NovaSeq 6000	Illumina	Ref.#: 20012850

RESOURCE AVAILABILITY

Lead contact

Further information and requests should be directed to and will be fulfilled by the lead contact, Nathan Fowler (nfowler@mdanderson.org).

Materials availability

This study did not generate new unique reagents.

Data and code availability

The analyzed data and the training data are available at our open website at https://science.bostongene.com/kassandra/ and https://github.com/BostonGene/Kassandra. Additionally, sequenced raw RNA-seq data of cellular subpopulations is deposited EGA: EGAS00001006272, a link to EGA dataset will also be available at https://science.bostongene.com/kassandra/). ccRCC, lung, and lymphoma datasets analyzed in this report will be made available upon reasonable request. A user-friendly web-based Kassandra tool at https://science.bostongene.com/kassandra/ was developed. With this tool, the user can upload available RNA-seq data and employ Kassandra-based deconvolution to their own data as well as supplied RNA-seq data as a test file. We have also deposited all code at https://github.com/BostonGene/Kassandra.

Accessions for the datasets used in this study are listed in Table S8.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Cell lines

Cell culture

MCF-7, K562, and Colo829 cell lines were all purchased from ATCC. K562 and Colo829 were cultured in RPMI-1640 with Glutamax-I, 10 mM HEPES, 100 U penicillin and 0.1 mg streptomycin, and 10% v/v heat inactivated fetal bovine serum (FBS, Corning, Corning, NY, USA). MCF-7 cells were cultured in EMEM with L-glutamine, 10 μg/mL insulin (BioXtra, Regina, SK, Canada), 100 U penicillin and 0.1 mg streptomycin, and 10% v/v FBS. All cells at 37°C in 5% CO_2 K562 were grown in suspension in upright T-75 flasks (Corning) and passaged 1:8 every 3 days before harvesting. MCF-7 and Colo829 were grown in 100 cm tissue culture-grade petri-plates until 80% confluent, detached with trypsin-EDTA, and sub-cultured at a ratio of 1:4. For these experiments, early passage cells were passaged an additional 3–4 times to generate roughly 10 million cells for RNA extraction. Cell lines were authenticated in house through whole-exome sequencing (WES) of DNA and aligned with the reference genomes of the parent cell lines (NCBI GEO Accession). Mycoplasma contamination was routinely tested using the PCR-based Universal Mycoplasma Detection kit (ATCC, Manassas, VA, USA).

Ex vivo *stimulation of T-helper cell subsets*

Th1, Th2, and Th17 cells were sorted from healthy donor PBMC using a combination of chemokine receptors. Live $CD45RA^{-}$ $CXCR5^{-}$ $CX3CR1^{-}$ $CD3^{+}$ $CD4^{+}$ memory T cells were sorted based on the expression of chemokine receptors CXCR3, CCR6, and CCR4. Sorted populations of Th1 ($CXCR3^{+}$ $CCR6^{-}$ $CCR4^{-}$), Th2 ($CCR4^{+}$ $CCR6^{-}$ $CXCR3^{-}$), and Th17 ($CCR6^{+}$ $CCR4^{+/-}$ $CXCR3^{-}$) cells were stimulated with ImmunoCult™ (Stem Cell Technologies, Vancouver, BC, Canada) Human CD3/CD28 T Cell Activator reagent (25 uL/mL), or left unstimulated in 0.2 mL of CellGenix media (Freiburg, Germany) for two hours in a 96-well deep well plate at 37°C with 5% CO_2. After incubation, cells were centrifuged for three minutes, 450 x G. Supernatant was removed by gentle aspiration. Cell pellets were resuspended in 0.2 mL Maxwell Homogenization Buffer (Promega, Madison, WI, USA) for storage at −80°C until RNA extraction using the Maxwell simply cells RNA kit followed by Illumina Stranded mRNA library preparation and RNA-seq as described in the RNA-seq section of the STAR Methods.

Blood samples

The peripheral blood of healthy donors was obtained from Research Blood Components (Watertown, MA, USA). For cell sorting from blood and tissue, PBMCs were prepared from peripheral blood, labeled with monoclonal antibodies to identify populations of interest, and sorted using a BD FACSAria III through a 100 μm nozzle. For flow cytometry, peripheral blood from 45 healthy donors was collected in K2-EDTA vacutainers and processed within 24 hours of collection as described in the Flow Cytometry section of the STAR Methods.

Tissue samples

NSCLC biopsies (n = 7) of early stage lung tumors were collected by resection (Mark Poznansky, VIC, Mass General Hospital, Boston, MA). Single-cell suspensions were prepared, and the same sample was subdivided for RNA-seq and CyTOF (n = 40 markers) analysis. Three RNA-seq/CyTOF normal tonsils samples were obtained from Dr. Leandro Cerchietti, WCMC. ccRCC samples for RNA-seq/MxIF (n = 28) and RNA-seq/CyTOF (n = 8) were also collected (Dr. Hsieh, Washington University in St. Louis). All tumor samples were collected under IRB-approved protocols at each institution.

METHOD DETAILS

Datasets

In addition to novel datasets, open source databases including ArrayExpress (Athar et al., 2019), GEO (Barrett et al., 2012), The Cancer Genome Atlas (TCGA) and The Genotype-Tissue Expression (GTEx) were used (Vivian et al., 2017; Aran et al., 2015; Saltz et al., 2018).

Sorted cell RNA-seq were collected from ArrayExpress and GEO databases (Table S8). All collected datasets included RNA-seq (read length higher than 31 bp) without polyA depletion and without the use of targeted panels. Several quality checks were performed. Samples with a total number of coding counts (of sequenced fragments) of less than 4 million were excluded. Samples contaminated with microorganisms such as mycoplasma and bacteria were excluded. Datasets containing ‘‘monocytes’’

differentiated into macrophages for 7 days were labeled as "macrophages" and were retained in the database. Other datasets derived from pluripotent stem cells were excluded. Expression QC analysis was performed to exclude abnormal or unreliable datasets. Widely used cell-specific genes (e.g., CD4, CD3, and CD45) were analyzed, and datasets were excluded when the expression of these genes did not agree with the dataset cell type label. Datasets of cell subtypes were re-labeled based on the lack of cell-specific gene expression. For example, when a dataset labeled as Treg lacked the expression of FOXP3 and IL2RA but expressed $CD4^+$ T cell-specific genes, the dataset was re-labeled as $CD4^+$ T cells.

A total of 18,193 samples of sorted cells derived from microenvironment or blood and cancer cell samples, both line and sorted were collected. Based on the quality checks described above, 9,041 samples were selected. Additionally, we used our dataset comprising RNA-seq of sorted cells from both tissue and blood (n = 348). Also, 15 samples of plasma and non-plasma B cells were obtained from Dr. Leandro Cerchietti, WCMC (denoted as BGD000001 dataset in Table S8). In total, the Kassandra database contains 9,404 samples annotated into 18 TME cell types (apart from 6 used in training indirectly) and 38 populations present in blood (Tables S1, S2, and S8).

RNA-seq processing and normalization

Bulk RNA-seq processing

Bulk RNA-seq fastq files were processed by Kallisto version 0.42.4 (Kallisto for Linux) (Bray et al., 2016). The Kallisto index file was downloaded from the Xena project to be consistent with TCGA and GTEx expression data we used (Vivian et al., 2017). This index file was built based on GENCODE transcriptome annotation version 23 (Frankish et al., 2019) and the human reference genome GRCh38 with genes from the PAR locus removed (chrY:10,000-2,781,479 and chrY:56,887,902-57,217,415) (Vivian et al., 2017). In contrast to paired-end fastq files, single-end fastq files were processed by Kallisto with additional options -l 200 -s 15 in line with Xena. The processing resulted in TPM transcript expression. All cell type datasets obtained from GEO or ArrayExpress were recalculated in the same way.

Fastq files were subjected to quality control measures *via* our pipeline employing FastQC (v0.11.5 or later), FastQ Screen (v0.11.1 or later) and MultiQC (v1.4 or later) tools. The reference genomes utilized for the creation of BWA aligner indices (for FastQ Screen) included *Homo sapiens* (GRCh38), *Mus musculus*, *Danio rerio*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Saccharomyces cerevisiae*, *Arabidopsis thaliana*, *Mycoplasma arginini*, Escherichia virus phiX174, microbiome (downloaded from NIH Human Microbiome Project website), adapters (provided with FastQC v0.11.5), and UniVec (NCBI).

scRNA-seq processing

Fastq files from scRNA-seq lung cancer datasets (E-MTAB-6149 and E-MTAB-6653 (Lambrechts et al., 2018)) obtained by the 10x Genomics experimental protocol, were processed by the Kallisto scRNA-seq pipeline ("bus" mode of Kallisto for raw reads pseudoalignment with -x flag accounting for the assay chemistry, additionally BUStools utils) (Melsted et al., 2019). Nine scRNA-seq PBMC datasets were downloaded from the 10x Genomics website as raw count matrices (see "validation" section). Empty droplets were filtered by means of barcodeRanks function of DropletUtils package (Griffiths et al., 2018; Lun et al., 2019). Raw count matrices for the scRNA-seq dataset of B-cell lymphomas (BCL) (Roider et al., 2020) were acquired from heiDATA under accession code VRJUNV. For further analysis, the Seurat package (Butler et al., 2018; Stuart et al., 2019) was used. UMI counts were normalized using SCTransform for lung/PBMC datasets and by the total expression within a cell for BCL. Batch correction for the BCL datasets was performed using Harmony package (Korsunsky et al., 2019) after genes from TCR and IG chains loci were excluded. Principal component analysis (PCA) was performed for dimensional reduction (RunPCA function) and the first 20 and 90 components were selected for further processing for lung/PBMC and BCL datasets, respectively. The Shared Nearest Neighbor graph was constructed (FindNeighbors function), and the Louvain community detection algorithm was applied to identify cell clusters (FindClusters function). t-SNE plots of lung/PBMC processed data (additionally, with Seurat batch correction) and a UMAP plot for BCL were used for visualization. Appropriate labels were assigned to each cluster based on the expression of marker genes.

Tables with log2(TPM/10+1) values for four scRNA-seq datasets were acquired from GEO: melanoma (GSE72056) (Tirosh et al., 2016), head and neck carcinoma (GSE103322) (Puram et al., 2017), glioblastoma (GSE131928) (Neftel et al., 2019) and ovarian cancer (GSE146026, 10x Genomics) (Izar et al., 2020). t-SNE plots of log2(x+1) transformed TPM expression values of cell-type-specific genes were utilized for visualization. Cell type annotations provided by the authors of the original studies were used for further analysis and elaboration. To define the T-cell, B-cell and fibroblast subpopulations that were not provided by the original studies, cells derived from different samples of melanoma (Tirosh et al., 2016), HNSC (Puram et al., 2017; Tirosh et al., 2016), and lung (Lambrechts et al., 2018) datasets were combined based on their annotation (for each dataset independently) and clustered by PhenoGraph (Levine et al., 2015), obtaining additional clusters that were manually assigned to the specific cellular population. For the refinement process of cell typing of the melanoma, HNSC, ovarian cancer and lung datasets, a set of genes for reclustering and nearest neighbors numbers as a PhenoGraph parameter were selected (Table S8). The resulting cell typing with key cell markers is shown in Figures S22, S24, and S25.

Normalization

Transcript groups presented in Table S8 were excluded from the TPM dataframe resulting from bulk RNA-seq processing (as explained in the main text). Non-coding RNA (e.g., micro-RNA and misc-RNA as previously described in the TCGA RNA pipeline (George et al., 2017)) (Figure S30A) and short transcripts of TCR- and BCR-coding genes, annotated in the transcriptome as corresponding to the V, D or J regions, were excluded from TPM normalization. Histone-coding and mitochondrial genes were omitted due to the uneven enrichment in different RNA extraction methods (e.g., PolyA vs Total RNA) (Newton et al., 2020). Unverified transcripts

having low transcript support level, and transcripts with partially unknown coding sequences were also precluded from normalization. Finally, 48 additional transcripts were removed according to other annotation tags reporting a lack of evidence or quality.

The sum of expression of all retained transcripts was normalized to 1,000,000, which resulted in adjusted TPM values. Finally, each gene was assigned a TPM expression value by summing the TPM values of its transcripts according to the GENCODE transcriptome annotation (Frankish et al., 2019) (Figure S30B). The full lists of retained transcripts and genes and an overview of different transcript and gene categories before and after filtering are provided in Table S8. For bulk RNA-seq created from scRNA-seq data (for all of which transcript expression or/and raw data are unavailable) and bulk RNA datasets without available raw data (SDY67 (Zimmermann et al., 2016), GSE127813 (Newman et al., 2019)), gene expression was (re)normalized to TPM within the intersection of the set of all genes from adjusted TPM dataframes described above and a geneset of a corresponding dataset.

Gene selection

Immune, stromal and endothelial cell type-specific genes were pre-selected by literature analysis, expression fold change analysis with statistical testing and correlation analysis using collected RNA-seq samples of sorted cell types. Expression fold change analysis with statistical testing was performed in three stages according to the level of phenotypic hierarchy of cell types. After transcript filtering and TPM renormalization, the first stage of analysis was performed on the following 8 sets of sorted cell samples: $CD4^+$ T cells, $CD8^+$ T cells and NK cells combined, B cells, Neutrophils, Macrophages and Monocytes combined, Endothelium, Fibroblasts, and malignant cells (cancer cell lines). The Kruskal-Wallis test (nonparametric ANOVA analogue) was performed on all genes for which at least one median within cell-type groups was greater than 1 TPM, and genes unsatisfied a p-value threshold of 0.05 (adjusted by Bonferroni correction) were removed. The Conover-Iman test (nonparametric pairwise test for multiple comparisons) was performed on the remaining genes of the 8 sorted cell sample sets (with p-value threshold of 0.05 adjusted by Bonferroni correction for multiple comparisons within a current geneset). Finally, fold change (FC) analysis of median expressions of these sets was conducted. Genes with log2(FC+1) less than 2 were excluded.

The second and the third stages of fold change and statistical analysis for feature selection were the same as the first one using another sample grouping. The second stage was performed on the following groups: T cells, NK cells, B cells, neutrophils, macrophages, monocytes, endothelial cells, fibroblasts, and malignant cells (cancer cell lines). The third stage was performed on the following groups: Tregs, T helpers, PD-1^+ $CD8^+$ T cells, PD-1^- $CD8^+$ T cells, NK cells, B cells (non-plasma), plasma B cells, neutrophils, macrophages, monocytes, endothelial cells, fibroblasts, and malignant cells (cancer cell lines). Finally, gene sets obtained during each stage were unified.

Independently, correlation analysis of artificial transcriptomes comprising non-malignant cell types was performed. Genes whose expression in mixes correlated with RNA percentage of a given cell type better than Pearson correlation coefficient 0.6, were selected as candidates to be used as specific genes of this cell type. Additionally, we conducted feature importance analysis (the SHAP approach (Lundberg et al., 2020) from our tree-based machine learning models (i.e., allowing the algorithm to determine the most significant genes for cell identification). Results of all computational analyses were combined and revisited based on literature analysis. The final sets of preselected genes are listed in Tables S3 and S4.

Generation of artificial transcriptomes

Using the collection of 9,414 samples (Tables S1–S3, S4, S5, S6 and S8), we created a variety of artificial transcriptomes. Here, we assume that the gene expression profile of a tissue is the result of a linear combination of individual cell expression profiles within that tissue (Zaitsev et al., 2019). The pipeline of creating artificial transcriptomes from purified RNA-seq samples of cell types is shown in Figure 1J. The samples of different cell types are randomly selected from the common pool, averaged within the cell type, and summed into a final expression file in proportions that resemble real tissues. The overall process is described in detail below.

Rebalancing number of samples by datasets and cell subpopulations

We rebalanced the number of samples per datasets and cell types to create optimal mixes/artificial transcriptomes. The number of sorted cell samples in a single dataset ranged from one to several hundreds. Datasets with too many samples can lead to overtraining of the models to the specific experiment if samples are randomly selected from the common pool. To balance the impact of different datasets, a number of samples within the dataset were resampled (Figure S31A). For each cell type, samples were resampled according to the formula below. Where random $N_{dataset_{new}}$ samples from dataset are repeatedly taken from $N_{dataset_{old}}$.

$$N_{dataset_{new}} = N_{\max}\left(\frac{N_{dataset_{old}}}{N_{\max}}\right)^{1-r} \quad \text{(Equation 2)}$$

where N_{max} is the number of samples in the largest dataset for one particular cell type, $N_{dataset_{old}}$ is the original number of samples in the dataset, and the rebalance parameter r is in the range [0, 1], in which 0 indicates no change in the number of samples, and 1 indicates that each dataset will contain the same number of samples. The rebalancing parameter r was set to 0.43 (Table S8).

Within a cell type (e.g., B cells), the number of sorted samples that belong to subpopulations (e.g., plasma/non-plasma B cells) varies. Uneven selection of subpopulation samples could also lead to overtraining of the models if they are randomly selected from the common pool. Therefore, the samples of each subpopulation $P_{subtype} \cdot msize/min_P + 1$ samples are resampled with

replacement. $P_{subtype}$ is a number reflecting the proportion of a given subtype, *msize* is the maximum between the number of samples for each subpopulation, and min_P is the minimum number $P_{subtype}$ between all subpopulations. The resampling with replacement was performed recursively for all nested subpopulations.

Averaging of samples

When generating artificial tissues, we averaged expression values from randomly selected samples of the same cell type prior to the mix. This minimizes batch effects, reduces noise while increasing the read coverage, and allows for the development of a greater diversity of artificial samples. However, averaging too many samples leads to a decrease in the biological variability within a given cell type, which will affect the learning outcome. Therefore, the number of samples for averaging (*Nav*) was used as a parameter, which, together with other parameters, was selected during hyperparameter optimization described below. In the final implementation of the models, $Nav = 9$ for each cell type was used to create artificial tissues (Table S8).

Generation of tissue cell proportions

To create a large number of artificial transcriptomes, selected cell types were mixed in various ratios. The random proportion of cells for each cell type and mix was generated using the following formula:

$$f_{cell} = \frac{R_{cell}K_{cell}}{\sum_{cell} R_{cell}K_{cell}} \quad \text{(Equation 3)}$$

where f_{cell} is the generated mRNA fraction for a particular cell type, R_{cell} is the random number uniformly distributed from 0 to 1, and K_{cell} is the coefficient from Table S8 for this cell type (coefficients for the most likely ratios of cell types in the tissue). For a specific set of mixtures, only those cell types that were not nested within each other were used (Figures S31B and S31C). While the proportions of cells in artificial transcriptomes were primarily generated to resemble the cellular composition of the real tumor tissue (Equation 4), mixes with cells ranging from 0 to 100% were developed to train the model for outliers. These outlier transcriptomes were generated by selecting random proportions from the Dirichlet distribution:

$$F = Dir(H) \quad H_{cell} = \frac{1}{N_{cells}} \quad \text{(Equation 4)}$$

where F is the vector of cell mRNA fractions with the length equal to cell type number N_{cell}, and H is the vector with concentration parameters with the same length. Each concentration parameter H_{cell} was equal to the inverse number of cell types. The proportion of mixes generated using the Dirichlet distribution was controlled by the parameter *Dp*, which was set to 0.335 (Table S8). Finally, the artificial transcriptomes generated based on random and Dirichlet distribution were combined.

Cancer cell selection and hyperexpression noise

To each artificial TME mix one tumor sample was added. This sample was randomly selected out of 2,166 sorted cancer cells and cell lines (Tables S1 and S8). The proportion of cancer cells was generated from a normal distribution $N(Tl, Tw^2)$ with parameters of variance (tumor width, *Tw*) equal to 1 and mean (tumor level, *Tl*) equal to 0.5 (Table S8). If the generated value was greater than one or less than zero, it was set at zero.

In tumor sample datasets, overexpression or amplification of genes is observed. To make robust models that consider the variability observed in real tumor biopsies, the aberrant expression of genes was imitated in our models. Hyperexpression noise was added to each cancer cell expression in an artificial tumor tissue mix. For the small number of genes controlled by hyperexpression fraction (*Hf*), noise values were added to the gene expression profiles (Figure S31D). With noise value selected as a random value from a uniform distribution from zero to maximum hyperexpression level (*Mhl*). Hyperexpression fraction in the final model was set to be equal to 0.03, and the maximum hyperexpression level was set to 3,428 TPM.

Generation of artificial tissue TME expression profiles

To create the artificial tissues, the expression vectors of each cell type were summed with coefficients reflecting the fractions of mRNA of the cells (the sum of the fraction is equal to one).

$$T_i^{mix_{before}} = \sum_{cell\ types} f_{cell} T_i^{cell} \qquad \sum_{cell\ types} f_{cell} = 1 \quad \text{(Equation 5)}$$

where $T_i^{mix_{before}}$ and T_i^{cell} is expression of gene *i* in the mix and cell in TPM units. Finally, simulated noise is added to get the resulting mix expression values $T_i^{mix_{after}}$:

$$T_i^{mix_{after}} = T_i^{mix_{before}} + Noise\left(T_i^{mix_{before}}\right) \quad \text{(Equation 6)}$$

For creation of artificial replicas of TCGA data samples, we deconvolved TCGA data with our trained model. Then, using our cell compendium data we performed the same procedure as for other artificial tissues with known RNA proportions of each cell type for each sample. Artificial TCGA samples were similar to the real ones (Figure S4).

For the PBMC models, the artificial transcriptomes were constructed exactly as the solid tumor models but models were trained on artificial transcriptomes with tumors and the cell type ratios being drawn from the uniform distribution.

Noise models

Expression of a single gene is presented as a sum of true expression μ_{T_i} plus sequencing error, which is a sum of Poisson technical noise P_i^j, normally distributed noise derived from sequencing library preparation, N_{prep_i}, and the most variable biological noise, resulting from different functional states of the specimens, N_{bio_i}:

$$T_i^j = \mu T_i + P_i^j + N_{prep_i} + N_{bio_i} \quad \text{(Equation 7)}$$

Quantitative relative standard deviation (*SD*/*mean*) of noise δ_i for gene *i* was calculated by the formula:

$$\delta_i = \sqrt{\delta_{P_i}^2 + \delta_{N_i}^2} \quad \text{(Equation 8)}$$

where δ_{P_i} is the relative standard deviation of Poisson technical noise and δ_{N_i} is the relative standard deviation of the normally distributed noise.

Assuming that technical replicates (*j*) of the same sample were sequenced with the same total coverage in readcounts ($R_j = R$). By definition, the expression value in TPM units is calculated by the formula:

$$T_i^j = \frac{C_i^j}{l_i k^j} \cdot 10^6,\ k^j = \sum_m \frac{C_m^j}{l_m} \quad \text{(Equation 9)}$$

where C_i^j is the expression in counts of gene *i* in the sequencing replicate *j*, and l_i is the effective length of gene *i*. Because the number of total read counts (*R*) are the same, then $k^{j_1} \approx k^{j_2} = \langle k^j \rangle_j = K$ for every j_1, j_2 (where the operation $\langle\rangle_i$ is averaging across index *i*).

Using (8):

$$\overline{T}_i = \left\langle \frac{C_i^j}{l_i k^j} \right\rangle_j \cdot 10^6 = \frac{\overline{C}_i}{K l_i} \cdot 10^6 \Rightarrow \overline{C}_i = \frac{\overline{T}_i l_i K}{10^6} \quad \text{(Equation 10)}$$

where $\underline{C_i}$ is the the average expression value in counts of gene *i*, similar to $\underline{T_i}$. From the Poisson distribution:

$$\mu c_i = \sigma_{c_i}^2 = \overline{\left(C_i^j - \mu c_i\right)^2} = \left(\frac{K l_i}{10^6}\right)^2 \overline{\left(T_i^j - \mu T_i\right)^2} = \left(\frac{K l_i}{10^6}\right)^2 \sigma_{T_i}^2 \quad \text{(Equation 11)}$$

Let

$$\beta^2 = \frac{10^6}{K} \quad \text{(Equation 12)}$$

then from Equations (10) and (12) follows:

$$\overline{T}_i = \sigma_{T_i}^2 \frac{K l_i}{10^6} = \sigma_{T_i}^2 \frac{l_i}{\beta^2} \Rightarrow \beta = \sqrt{\frac{l_i}{\overline{T}_i}} \sigma T_i \quad \text{(Equation 13)}$$

Thus, for gene *i* in TPM units, the standard deviation (*SD*) and noise (*SD*/*mean*) are expressed using the following formulas:

$$\sigma T_i = \beta \sqrt{\frac{\overline{T}_i}{l_i}} \quad \text{(Equation 14)}$$

$$\delta_{P_i} = \beta \sqrt{\frac{1}{l_i \overline{T}_i}} \quad \text{(Equation 15)}$$

K is proportional to the total number of sample coverage in readcounts (*R*) assuming:

$$K = \frac{10^6}{\alpha^2} R \quad \text{(Equation 16)}$$

Following from Equations (9), (12), and (15):

$$\beta \equiv \beta(R) = \frac{\alpha}{\sqrt{R}} \Rightarrow \delta_{P_i} = \alpha \sqrt{\frac{1}{l_i \overline{T}_i R}} \quad \text{(Equation 17)}$$

For noise modeling (Figures 2E–2I), data analysis only included samples with $5 \cdot 10^6 < R < 17.5 \cdot 10^6$ read counts and genes with mean expression value >5 TPM. The obtained model was used to simulate technical noise in artificial transcriptomes to mimic sequencing at $R = 30 \cdot 10^6$ read counts. Noise was added as two separate summands (technical and biological) according to the formula:

$$T_i^{mix_{after}} = T_i^{mix_{before}} + \beta\sqrt{\frac{T_i^{mix_{before}}}{l_i}}\xi_P + \gamma T_i^{mix_{before}}\xi_N \quad \text{(Equation 18)}$$

where ξ_P, $\xi_N \sim N(0, 1)$ and γ is the coefficient of uniform level of non-poisson noise, which we assumed has a standard normal distribution. Expression values lower than 0 TPM were rounded to 0 TPM.

To estimate coefficient α from (Equation 17) we collected 648 sets of technical replicates of samples with nearly the same read-count (±10% from the average value). For each set, β was approximated for each gene using Equations (14) and averaged. Coefficient $\alpha = 2.05$ was estimated according to (Equation 17). Also, from (Equation 17) coefficient $\beta(30 \cdot 10^6) = 0.37$. The γ coefficient was estimated to be equal to 0.168618 (Table S8). After technical correction, the measured variation of replicates lost dependence on the coverage. Subsequently, this noise was used to add technical variation to the artificial mixtures, resulting in better mimicking of real tissues, ensuring stability when encountering real-world sequencing variability.

Model training

Each model was trained to predict the percent RNA fraction of each cell type represented in the mix using LightGBM version 2.3.1 (https://github.com/microsoft/LightGBM). LightGBM models for each cell type were trained in two stages, with each stage creating a separate model. For each cell type model, mixes were generated from samples of the cell types according to their hierarchy (Figures 1G and 1H) and the model training illustrated in Figures S31B and S31C. The input training data for the first stage was a set of 150,000 artificial mixes for each cell type, using gene expression as training features. In this first stage, median expression values were calculated as an additional feature. In the second stage, predictions for each cell type calculated using first stage models were used as additional features (Figure S32A). For each cell type and at each stage, 10 independent models were trained using different random subsets of datasets (Figure S32B). Predictions from final 10 trained second stage models were averaged to obtain the final RNA percentage of a cell type (Figure S32B). As a result, a total of 18 million artificial RNA-seq mixes were generated for the training of 420 LightGBM models. Ultimately, the Kassandra algorithm was implemented using Python 3 using the following libraries: pandas, scikit-learn, SciPy, NumPy, matplotlib, seaborn, LightGBM. In addition, we implemented a measure of prediction uncertainty (SD of predictions) for each cell type by revoking the code on an ensemble of 10 independent models and calculating the SD across the 10 models as visualized for scRNA-seq and CyTOF experiments (Figure S33). Kassandra-Blood models were trained separately by the same procedure described above, with the exception that only one model per cell type was trained, resulting in a total of 8 million artificial RNA-seq mixes generated for the training of 40 LightGBM models.

Parameter optimization

The parameters utilized for mixture generation (*Nav*, γ, *Dp*, *r*, *Hf*, *Mhl*) could not be selected on artificial mixtures. Therefore, these parameters were selected using an indirect method on real tissues. Schema of parameter optimization is shown in Figure S34A. We trained LightGBM models for each cell type separately. Parameters *Nav*, γ, *Dp*, *r*, *Hf*, *Mhl* (Table S8) have been varied, and the predicted fractions for a major cell type were compared with predictions of the sum of the it's subpopulation (Figure S34B). Each parameter was selected randomly with a uniform distribution from the range specified in Table S8. For each set of parameters, 6 groups of 50,000 mixes were generated for model training. First-stage (LightGBM) trained models were applied to TCGA and GTEx samples to obtain predictions of cell type mRNA percentages. Then, the following groups of models (concordance groups) were compared with the MAE metric: Immune cells vs Lymphocytes + Myeloid cells, Myeloid cells vs Macrophages + Monocytes + Neutrophils, Lymphocytes vs NK cells + T cells + B cells, T cells vs $CD4^+$ T cells + $CD8^+$ T cells, $CD4^+$ T cells vs Tregs + T helpers, $CD8^+$ T cells vs PD-1^+ $CD8^+$ T cells + PD-1^- $CD8^+$ T cells, B cells vs Plasma B cells + Non plasma B cells. A total of 2,097 sets of parameters spanning approximately 629 million artificial mixes were tested. A set of parameters with the best MAE (Figure S34C) was selected for each concordance group (Figures S35A and S35B), and the selected parameter sets were averaged, with the final parameter values shown in Table S8.

LightGBM hyperparameter optimization

All training datasets were randomly divided into training and validation samples; subsequently, mixes were generated from these datasets (Figure S36A). Specific parameters were optimized for model training (Table S8). The evaluation metric was the mean absolute error (MAE), and the evaluation score of the vector was the averaged score between folds (Figure S36B). Overall, 40,000 sets of parameter vectors were generated, with each vector consisting of the 9 most important parameters for LightGBM (Figure S36C) with a unique parameter range (Table S8). Every cell type model was trained on these data according to its unique list of genes. Models for each parameter vector were trained separately on 3 folds and evaluated on the validation set of mixes. Overall, 100 parameter vectors for each cell type was chosen as a starting population for the genetic algorithm (Figure S8C). In brief, the algorithm performed crossover (random exchange of parameters between two vectors) and mutations (random modification of one random parameter in a vector). If the parameter is max_depth, it can mutate its value only by $\Delta_p = \pm 1, 0$. Mutation value Δ_p for other parameters was calculated according to the formula:

$$\Delta_p = \pm\theta(p_{\min} + \xi p_{\max}) \quad \text{(Equation 19)}$$

where $\theta = 0.5$ is the power level of mutation, p_{min} and p_{max} are the boundary values of the parameter, $\xi \sim N(0, 1)$, the sign is chosen randomly. The generated vectors were then evaluated to form a new generation of 100 vectors for each cell type for the next iteration across all vectors used in the genetic algorithm, which was repeated if the algorithm was not manually stopped (Figure S8B). Using

these generated vectors, we predicted the percent of each cell type in the mix. This procedure was repeated using new mixtures with the same cell ratio concatenated with predicted features, and the vectors were sorted by their evaluation score. Ultimately, the best parameter vector was selected for each model cell type for two steps (Table S8).

Deconvolution specificity

The specificity analysis was performed on holdout samples that were not used in the development of Kassandra (Table S8). The expression of 10 random samples for each cell type were averaged, and each deconvolution algorithm was applied to the expression values (Figures 3K and S11). For all the algorithms, the values represent the percentages of cells, with the exception of MCP-counter and xCell, where values for each cell type were normalized to the maximum and multiplied by 99. To calculate a non-specificity score, 50 random mixes were created from samples of other cells, including cancer cells. Mixes were created using the method described above to prepare artificial tumor samples, except that only fractions for cell types from the Dirichlet distribution for all cell types were used, and each deconvolution algorithm was applied to the mixes (Figure 3L). The scores were calculated using the following formula:

$$Score_{cell} = \frac{F_{cell_{nonspecific}}}{F_{cell_{specific}}} \cdot 100 \quad \text{(Equation 20)}$$

where $F_{cell_{specific}}$ is the value predicted for the cell type by the algorithm on 10 averaged random samples of this cell type described above (true positive signal). And $F_{cell_{nonspecific}}$ is the average value for this cell type on mixes where cells of this type were absent (false positive signal).

Limit of detection

Limit of Detection (LOD) was assessed on holdout samples (Table S8). To investigate the dependence of the algorithm on the number of readcounts, mixes should be generated with a determined number of readcounts. To create these mixes, fastq files were fragmented, and each file contained approximately 50,000 target readcounts (single-end reads or pairs of paired-end reads) after transcript filtering. A total of 74,340 fastq file fragments were prepared, expression for each was calculated using Kallisto and normalized into TPM as described above.

For each total coverage from 0.15 to 125 mln reads, sets of 150,000 artificial tumors were created from small fragments of fastq files of sorted cells RNA-seq. A total of 7.8 million mixes were created. In each artificial tumor the fractions of all cell types, including cancer cells, were selected from a Dirichlet distribution with concentration parameters inversely proportional to the number of types. The dependence of the Pearson correlation on the number of readcounts is shown in Figure 3I, and the dependence of the prediction variation on the fraction of cell RNA in the mix is shown in Figure 3J.

Quantitative cell estimation

Various cells contain different amounts of total RNA due to their different sizes and functions (Racle et al., 2017; Monaco et al., 2019). The predicted mRNA fractions (R_{cell}) of the main cell types were normalized to 1.0, taking into account their subtypes. If their sum was less than one, their values were unchanged, and the remainder up to one was denoted as the "Other" cell type: $R_{Other} = 1 - \sum_{cell} R_{cell}$. If the sum was greater than one, then the fractions were normalized to one, and "Other" was written to zero. To calculate cell percentages from RNA fractions, we used the following formula:

$$C_{cell} = \frac{\frac{R_{cell}}{A_{cell}}}{\sum_{cell} \frac{R_{cell}}{A_{cell}}} \quad \text{(Equation 21)}$$

where C_{cell} is the cell fraction of the cell type, R_{cell} is the RNA fraction of the cell type, A_{cell} is the relative RNA per cell coefficient and $\sum_{cell} R_{cell} = 1$. For recalculations, no cell subpopulations of the included cell types were used. For neutrophils, monocytes, B cells, T cells and NK cells, coefficients were taken from previous studies (Racle et al., 2017; Monaco et al., 2019). We utilized intermediate values for those cells and integrated the values into our recalculation. For other cell types, including myeloid cell and stromal cells, 7 million sets of random coefficient values were generated for each corresponding population. For each set of values, we applied the values to RNA deconvolution of tumor tissues from the TCGA RNA-seq data. The "Other" cell type also has its own RNA per cell coefficient. In bulk tumor tissue, the "Other" cell type calculated by the model included malignant cells and benign epithelial cells not deconvolved by Kassandra. Cell types or subtypes that included or were a subset of the utilized types were recalculated according to the change in used types. We calculated the percentage of "Other" cell fraction (meaning cancer cells) and estimated its correlation with tumor purity of TCGA samples calculated by whole exome sequencing (WES) by the ABSOLUTE algorithm (Aran et al., 2015) (Figure 4D). For each cancer type, we selected a coefficient set with the best correlation value. Final coefficients were obtained by averaging values for each cell type across cancers (Table S5).

To validate the coefficients, equal cellular proportions (50:50) of T cells and a specific cell type were mixed and sequenced, and the algorithm was employed to calculate the relative RNA per cell coefficients. T cells or naive $CD4^+$ T cells were used as a reference having a coefficient equal to 1, allowing the calculation of the coefficients for a specific cell type as a ratio of its RNA percentage to the RNA percentage of admixed naive $CD4^+$ T cells. The coefficients for neutrophils, monocytes, and T cells were concordant with other algorithms; coefficients for macrophages were in good concordance with the predictions from TCGA (Figure S10A).

This method was also utilized for blood-derived naïve B cells, memory B cells, plasma B cells, cytotoxic NK cells, classical and non-classical monocytes, NK cells, naïve $CD8^+$ T cells, and memory $CD8^+$ T cells (Figure 3G). For blood subpopulations where coefficients were not measured, we calculated the coefficients by fitting RNA percentages to 45 independent FACS experiments analogously as described in ABIS (Monaco et al., 2019) (Figure S10D). For coefficients fitting linear regression with fixed intercept were used (scikit-learn package python). Each cell type coefficient was iteratively fitted with other cell type coefficients fixed during the whole iteration. In the next iteration, coefficients from the previous iteration were used to recalculate new cell proportions and update coefficients, and the iterations were continued until all the coefficients converged. This resulted in fitted and measured coefficients for all cell types concordant in the order of magnitude with other algorithms and experimentally measured values (Figure S10D).

Other deconvolution algorithms

The Kassandra algorithm was compared with 9 different published deconvolution algorithms: EPIC (Racle et al., 2017), CIBERSORT (Newman et al., 2015), CIBERSORTx (Newman et al., 2019) with available matrices LM22 and HNSC, FARDEEP in relative and absolute modes (Hao et al., 2019), quanTIseq in default and tumor modes (Finotello et al., 2019), ABIS (Monaco et al., 2019), and MCP-counter (Becht et al., 2016), xCell (Aran et al., 2017) and Scaden (Menden et al., 2020). FARDEEP version 1.0.1 returned an error in several analyses; therefore, no results could be used for comparison for certain analyses. Comparison with the MAE metric for the xCell and MCP-counter algorithms were not included because they produce scores as the result, not percentages of cells.

For the CIBERSORTx algorithm, the default LM22 and HNSC matrices were used. Processed gene expression values were uploaded as gene expression mixtures and calculated in absolute mode (without batch correction, without permutations, and with quantile normalization disabled as recommended for RNA-seq data) on the official website (https://cibersortx.stanford.edu/index.php). The ABIS deconvolution algorithm was launched as a shiny application (https://github.com/giannimonaco/ABIS) in RStudio (v1.1.463, R version 3.5.1). Expression values processed as described above (see "RNA-seq processing and normalization") were uploaded to the application and were analyzed in RNA-seq mode (not microarray mode). MCP-counter was designed to predict the abundance of certain subtypes of cells in artificial *in vitro* mixes and was trained on microarray data, not on RNA-seq, and MCP-counter does not predict the absolute values of cells, only units. xCell uses gene enrichment scores to predict the amount of certain cell types in a sample.

Scaden algorithm

We calculated cell percentages for validation datasets using the Scaden (PBMC) model *via* the web interface (https://scaden.ims.bio). Second, Scaden was trained on our scRNA-seq lung cancer dataset (E-MTAB-6149 and E-MTAB-6653), and *via* the Scaden API, a total of 500,000 mixes were simulated. Third, Scaden was trained on the same artificial mixes as Kassandra to compare LightGBM models with NN models. Expression values for the training were processed as described in their documentation. In all cases, NN was trained in 5,000 steps as suggested by the authors (https://scaden.readthedocs.io/en/latest/usage/).

Validation

Histologically defined percentages of tumor-infiltrating lymphocytes (TILs) and macrophages for TCGA H&E slides (Saltz et al., 2018) were used for the initial validation of Kassandra predictions. Relative percentages recovered by neural nets from H&E images were directly correlated with Kassandra predictions from RNA-seq of the same tumor samples. To calculate the "TIL percentage", T cells, B cells and NK cells were summed.

Three scRNA-seq datasets of Smart-Seq2 experimental design were also used for validation: melanoma (GSE72056) (Tirosh et al., 2016), head and neck carcinoma (GSE103322) (Puram et al., 2017) and glioblastoma (GSE131928, Smart-Seq2) (Neftel et al., 2019) datasets. Additionally, 13 10X Genomics experimental design datasets were also used: 8 PBMC (from healthy donors) demo datasets (10X Genomics company; https://support.10xgenomics.com/single-cell-gene-expression/datasets/10k PBMCs (v3 chemistry), 1k PBMCs (v2 chemistry), 1k PBMCs (v3 chemistry), 3k PBMCs (v1 chemistry), k PBMCs (v1 chemistry), k PBMCs (v2 chemistry), k PBMCs (v2 chemistry), 5k PBMCs (v3 chemistry), and a dataset with 68k PBMC cells from methodological 3′ scRNA-Seq study (Zheng et al., 2017), two lung cancer datasets (12 adenomatous or squamous tumor samples with 10X Genomics Chemistry v2 from E-MTAB-6149 and E-MTAB-6653) (Lambrechts et al., 2018), an ovarian cancer (GSE146026, 10X Genomics part) (Izar et al., 2020) dataset and a B-cell lymphoma dataset (Roider et al., 2020) (additionally containing 3 reactive non-malignant lymph node samples). Single-cell data were processed as described above, transformed to pseudobulk RNA-seq samples and deconvolved with Kassandra. Smart-Seq2 artificial bulks were constructed as a mean expression vector of all TPM expression vectors of cells belonging to each patient sample, which is equivalent to the sum of all cell vectors and subsequent TPM normalization. Cell percentages in each artificial bulk were calculated and set as true values for comparison with deconvolution algorithm predictions of RNA percentages (because TPM normalization within each single cell eliminates difference in amounts of RNA per cell). 10X Genomics artificial bulks were constructed as the sum of all expression vectors (umi counts) of cells belonging to each patient sample and then normalized to the sum of umi counts and multiplied by 1 million; therefore, the resulting artificial bulk expression vector was approximate to TPM. True cell type RNA proportions for each artificial bulk was calculated as the umi count sum of each cell type within this pseudobulk divided by the total number of umi counts of this artificial bulk.

In addition to scRNA-seq, Kassandra was also validated on flow cytometry data. We performed both flow cytometry and RNA-seq analysis on 45 paired PBMC samples. Kassandra's prediction was compared with cell quantities obtained by flow cytometry (Monaco et al., 2019; Finotello et al., 2019; Newman et al., 2019; Linsley et al., 2014; Hoek et al., 2015; Zimmermann et al., 2016) in 6 public datasets (total number of samples = 517) and from 2 datasets processed by automated hematology analyzer (total number

of samples = 350) (Altman et al., 2019; Shin et al., 2014). Moreover, 7 lung adenocarcinoma samples for subsequent RNA-seq and CyTOF analyses were obtained (Dr. Mark Poznansky, VIC, Mass General Hospital). In brief, biopsies of early stage lung tumors were collected by resection. Single-cell suspensions were prepared, and the same sample was subdivided for RNA-seq and CyTOF (n = 40 markers) analysis. Three RNA-seq/CyTOF normal tonsils samples were obtained from Dr. Leandro Cerchietti, WCMC. Cellular percentages obtained from CyTOF were directly compared with cellular percentages predicted by Kassandra from RNA-seq data. Multiplex immunofluorescence (MxIF) was also employed for Kassandra validation where 28 RNA-seq/MxIF and 8 RNA-seq/CyTOF clear cell renal cell cancer (ccRCC) samples were collected (Dr. Hsieh, Washington University in St. Louis). Tumor samples were divided and sent for RNA-seq, MxIF, and CyTOF analysis. Cellular percentages obtained from MxIF and CyTOF were directly compared with cellular percentages predicted by Kassandra from RNA-seq. These tumor samples were collected under IRB-approved protocols at the listed institutions.

Prediction of IHC PD-L1 expression by Kassandra-based TME reconstruction

The bladder cancer (n = 348, EGA: EGAS00001002556 (Mariathasan et al., 2018)) and gastric cancer (n = 34, ERP107734 (Kim et al., 2018)) datasets were obtained from the EGA or SRA databases. The ccRCC expression dataset with corresponding IHC values (PD-L1 IC: ranging from 0 - 40%) was provided by Dr. Hsieh (Washington University in St. Louis). Samples PB-16-054, PB-16-043, PB-16-066 belonging to the Kim et al. dataset were excluded due to poor coverage or because the samples were outliers on the PCA. Samples PB-16-006, PB-16-007, PB-16-008, PB-16-010, PB-16-011, PB-16-013, PB-16-014, PB-16-015, PB-16-016, PB-16-026, PB-16-047, PB-16-048, PB-16-049, PB-16-051, PB-16-052, PB-16-055, PB-16-056, PB-16-057 from the Kim et al. dataset were excluded due to HLA mismatch with WES from the same patients (Kim et al., 2018). All the PD-L1 positivity scores were unified to IC0 (<1% positive immune cells); IC1 (1–5% positive immune cells); and IC2+ (>5% positive immune cells). The bladder cancer dataset was randomly divided into training (n = 235) and test (n = 115) cohorts. We applied an ordinal regression model on the cell percentages reconstructed by Kassandra from RNA-seq on the training set to predict PD-L1 IHC status (IC0, IC1, IC2+). Next, we created a logistic regression model (for classification of the IC0 and IC2+ IHC expression levels) on the bladder cancer training cohort. We assessed the model performance on the bladder test, gastric and ccRCC cohorts using AUC values (Figure 6O).

PD-1^{+} CD8^{+} T cell percentage association with response to immunotherapy

The bladder cancer (n = 348, EGA: EGAS00001002556 (Mariathasan et al., 2018)), gastric cancer (n = 34, ERP107734 (Kim et al., 2018)), ccRCC immotion150 (EGAC00001000946 (Pal et al., 2020)) datasets were obtained from the EGA or SRA databases. Immotion150 was divided into the atezolizumab (ccRCC PD-L1) and atezolizumab+bevacizumab (ccRCC PD-L1/VEGF) and sunitinib (not shown) cohorts and analyzed separately. We also curated only pre-treatment samples collected less than 200 days before the start of therapy. For anti-CTLA-4-treated patients from the Van Allen et al., 2015 (n = 40) and Nathanson et al., 2017 (n = 20) cohorts, only non-acral or mucosal samples were analyzed. Gide (ERP105482) (Gide et al., 2019), Liu (Liu et al., 2019b) (phs001036), Hugo (Hugo et al., 2016) (GSE78220, GSE96619) cohort samples were combined. Obtaining a total of 6 cohorts of patients for analysis. For survival analysis, Cox hazard regression modeling was conducted with cell percentages, PD-L1 expression and Tumor Mutational Burden (TMB) as parameters.

ML-based immunotherapy response prediction

The bladder (anti-PD-L1), gastric cancer (anti-PD-1), ccRCC (anti-PD-L1, -PD-L1+BEVA) cohorts were used to evaluate the predictive power of TME percentages predicted by Kassandra as a therapy response classifier. Both cross-validation and a separate melanoma cohort (anti-PD-1) were used to evaluate classifier performance (Figure S28J). Separate LightGBM classifiers were trained with the following features: 1) Z-score of PD-L1 expression (TPM) within one cohort, 2) TMB, 3) cell percentages predicted by Kassandra: M1 macrophages, endothelial cells, NK cells, and 4) all of the above features in different combinations.

Database of blood RNA-seq samples

Whole blood RNA-seq were collected from ArrayExpress and GEO databases. All collected datasets included RNA-seq (read length higher than 31 bp) without polyA depletion and without the use of targeted panels. Several quality checks were performed. Samples with a total number of coding counts (of sequenced fragments) of less than 1 million were excluded. Samples contaminated with microorganisms such as mycoplasma and bacteria were excluded. Samples without explicit age annotations were excluded from the analysis. This resulted in 1,750 samples of whole blood RNA-seq (Table S8).

Samples were split into 5 groups by the donor's reported age- ['<3′, '3–25′, '30–50′, '51–75′, '>75']. Cell fractions were predicted by the Kassandra Blood deconvolution model. Granulocytes and their subtypes were excluded from predictions and fractions were renormalized to add up to 1. Differences in predicted fractions of monocytes and differences in predicted fractions of Transitional Memory T-helpers relative to total predicted T cells between age groups were tested for statistical significance by the two-tailed Mann-Whitney test.

CyTOF data processing

The following markers were used in the CyTOF panel for the analysis of the non-small cell lung carcinoma (NSCLC) cell suspension: CD11b, CD11c, CD127, CD137, CD14, CD152, CD154, CD16, CD184, CD19, CD197, CD20, CD206, CD223, CD25, CD3, CD31, CD33, CD38, CD39, CD4, CD44, CD45, CD45RA, CD45RO, CD56, CD66b, CD68, CD69, CD8, CD90, CD95, EPCAM, FoxP3, HLA-DR, PD-1, PD-L1, T-bet, Tim3. For the analysis of the renal cell carcinoma (RCC) tissue, the following markers were used in

the CyTOF panel: CD45, CD3, HLA-A_B_C, CD57, CD69, CD4, CD8, CD11c, CD16, CD25 (IL-2R), CD10 7(LAMP1), CD66b, CD45RA, CD163, CD86, CD27, CD197 (CCR7), CD14, FoxP3, CD127 (IL-7Ra), CD141 (Thrombomodulin), CD38, CD19, HLA-DR, CD68, GranzymeB, PD-1, PD-L1, CD56 (NCAM), CD11b (Mac-1). The samples were gated for separation of beads, reference PBMCs, and isolation of living cells (Cisplatin/DNA1) (Figure S37A). The CyTOF signal values were subjected to standard hyperbolic arcsine (archsin(x/5)) transformation before clustering (Figure S37B). To obtain the primary cell populations, we clustered and assigned a cell type to the cells of each sample independently. Clusterization was performed using FlowSOM 1.20.0 (Van Gassen et al., 2015). For clustering stability, we repeated clustering of each sample 30 times using 90% of randomly selected cells. Clusterization was performed using FlowSOM, with default parameters except for the grid size (xdim = 10, ydim = 10).

As a reference, a NSCLC (VIC26) and a ccRCC (WUR120_A1) sample was used. Clusters for VIC26 and WUR120_A1 were manually chosen as the main populations in accordance with the following rules: $CD8^+$ T cells ($CD45^+$ $CD3^+$ $CD8^+$ $CD4^-$ $CD56^-$), $CD4^+$ T cells ($CD45^+$ $CD3^+$ $CD4^+$ $CD8^-$ $CD56^-$), NK cells ($CD45^+$ $CD56^+$ $CD16^+$), neutrophils ($CD45^+$ $CD66^+$ $CD16^+$), macrophages/monocytes ($CD45^+$ HLA-DR^+ $CD11c^+$), fibroblasts ($CD45^-$ $CD90^+$ $EPCAM^-$), endothelial cells ($CD45^-$ $CD31^+$ $EPCAM^-$ for NSCLC and $CD45^-$ $CAIX^-$ $CD107^+$ for RCC), B cells ($CD45^+$ $CD19^+$ $CD20^+$), $CD20^-$ B cells/Plasma cells ($CD45^+$ $CD19^+$ $CD20^-$), Tumor ($EPCAM^+$ for NSCLC/$CAIX^+$ for RCC).

For the remaining samples, the resulting clusters for each repeat were typed by comparing the average values with pre-marked reference populations. For every cluster, the mean vector of signals was calculated. The resulting vector was correlated using Pearson correlation with known vectors for pre-marked populations. For every cluster (and all cells in it), the population was typed with a pre-marked population with maximum correlation. If the most correlated population had a correlation coefficient <0.6, the cell type for each individual cell was defined as a consensus of repeated clustering, as the most frequent type for that cell in repeats. To determine the subpopulation of $CD8^+$ T cells, $CD4^+$ T cells, B cells, and macrophages, cells of these types were separated from others and subsequent analysis was performed independently for each cell type. Also, for noise reduction, only selected markers were employed to determine subpopulations. The following markers were used to determine subpopulations of T cells, B cells and macrophages: For $CD4^+$ T cells and $CD8^+$ T cells (CD69, CD25, CD223, CD95, CD45RA, Tim3, FoxP3, CD39, T-bet, CD45RO, CD127, PD-1, CD38, CD4, CD8); for NSCLC and RCC (CD45RA, CD197, CD127, CD25, FoxP3, CD57, Granzyme B, CD27, CD69, PD-1, CD38); for B cells (CD95, CD39, CD69, CD184, CD38, CD20, HLA-DR), B cell subpopulations were determined only for NSCLC samples; for macrophages and monocytes (CD206, CD68, CD11c, CD14, PD-L1, CD16, HLA-DR, CD44, CD11b, CD127, CD38, CD33); for NSCLC and RCC (CD14, CD11c, CD11b, HLA-DR, CD16, CD163, CD68, CD4, CD38, PD-L1, CD86, CD107a, CD69).

Multiplex imaging

Staining

Five μm FFPE ccRCC tissue mounted onto Superfrost™ Ultra Plus adhesion slides were baked at 60ºC for 1 hour, deparaffinized with 2 washes of fresh-xylene and rehydrated with ethanol washes 100% (2x), 95% (2x), 70% (2x), 50% (2x), 1X PBS (1x), and 0.3% Triton X100 in 1X PBS (1x) and was subjected to a two-step antigen retrieval process. Next, the tissue sections underwent repeated cycles of staining, imaging and signal removal. The sections were stained with antibodies directly conjugated with either Cy3 or Cy5 dye at a previously optimized concentration. All antibody mixes used for seven incubation rounds were incubated at room temperature for 1 hour in a humid chamber. After incubation for all rounds, slides were washed in 1X PBS for 5 min (3x). The tissue sections were then stained with DAPI solution (1 ug/mL) for 15 min. The slides were washed with 1X PBS and the coverslip was added immediately using mounting media. Antibodies against BAP1, CAIX, CCASP3, CD11B, CD11C, CD16, CD206, CD20, CD31, CD3, CD45, CD4, CD56, CD68, CD8, DAPI, GRB, H3K36TM, HLA, KI67, NAKATPASE, PBRM1, PCK26, PD-L1, PTEN, S6 were used.

Imaging and image processing

Immunofluorescence confocal imaging of human ccRCC tissue slides was performed using a GE IN Cell Analyzer 2200 equipped with x20 objective. Images were captured with a CMOS camera under the following exposure settings. In total, we captured 145 regions of interest (ROI) from 28 different patient samples. MxIF imaging was performed with the following high-efficiency fluorochrome-specific filter sets specific for DAPI, CY3, CY5. Image processing and deconvolution were performed with NIS Elements. After this pre-processing, a large data set of images were obtained at the same resolution and 16-bitness. The images were cropped into small regions of the same size, while maintaining marker names and numbering of regions for the purpose of original image restoration at the final steps.

Cell segmentation and typing

Cell segmentation was performed using UNet semantic segmentation neural network and watershed post-processing of the identified cell masks to reduce under segmented cell counts. For the cell segmentation neural network, 2 markers were utilized in the training set as follows: (i) a region was designated as a cell only in the presence of one nucleus; (ii) a closed NAKATPASE border around the nucleus was designated as a membrane; (iii) if no NAKATPASE border was found, we defined the cell membrane as an area at a distance up to 15 pixels from the nucleus, depending on the proximity of neighbors. Mean fluorescent intensity of fluorescent markers for each cell segment was calculated, allowing analysis of MxIF data as a single cell proteomics dataset with total number of cells = 1,084,511. Cells were clustered using Phenograph, and each cluster was manually assigned with a specific cell type. Cluster of cells expressing CD3, CD4, CD45 was annotated as $CD4^+$ T cells, cluster expressing CD3, CD8, CD45 was annotated as $CD8^+$ T cells, cluster expressing CD206, CD68, CD11c was annotated as macrophages. To estimate the proportion of blood vessels, we used CD31 marker expression. We created a binary mask and measured the percentage of endothelium area that covers the tissue for each ROI. All of the processing steps utilized the OpenCV library in Python language.

Flow cytometry

Peripheral blood was collected from 45 healthy donors (Research Blood Components, Watertown, MA, USA) in K2-EDTA vacutainers and processed within 24 hours of collection. For the generation of PBMC fractions, red blood cells and granulocytes were removed using density gradient centrifugation and SepMate tubes (Stem Cell Technologies, Vancouver, Canada) by layering blood over Ficoll-Paque Plus (Cytiva, Marlborough, MA, USA). For analysis of complete peripheral blood leukocytes, RBC from undiluted whole blood was lysed using RBC lysis buffer (ThermoFisher, Waltham, MA, USA). Both sample preparations were subsequently processed similarly; after several washes in flow cytometry staining buffer (PBS + 2% newborn calf serum (v/v) + 1 mM EDTA), cells were counted, resuspended in staining buffer containing TrueStain FcX, Monocyte blocker (Biolegend, San Diego, CA, USA) and 10% (v/v) Brilliant stain buffer (BD Biosciences, San Jose, CA, USA) to block non-specific labeling and Ghost Dye Aqua (Tonbo Biosciences San Diego, CA, USA) to assess viability. Two million cells were then collected and lysed for RNA extraction using the RNAeasy mini kit (Qiagen, Hilden, Germany). The remaining cells were then labeled with different antibodies to resolve subpopulations of $CD4^+$ T cells, $CD8^+$ T cells, B cells, monocytes, NK cells, granulocytes, and dendritic cells (Table S6). The cells were then fixed in BD CytoFix/CytoPerm for stabilization prior to acquisition. Up to one million events per panel were acquired on a BD FACSCelesta flow cytometer and populations of interest were identified using manual bivariate gating on FlowJo Software V10 (BD Biosciences).

Cell lines and creation of mixtures with PBMCs

The COLO829, MCF7 and K562 cancer cell lines were purchased from American Type Culture Collection (ATCC, Manassas, VA, USA) and maintained according to the vendor's instructions. PBMCs were obtained from fresh peripheral blood of one healthy donor (Research Blood Components, Watertown, MA, USA), and isolated using Ficoll-Paque as described above. Cell lines and PBMCs were counted using a Cellometer Auto 2000 (Nexcelcom, Lawrence, MA, USA) using acridine orange and propidium iodide to enumerate viable cells. Live cancer cells were then mixed with live PBMCs in ratios of 12:88, 25:75, 50:50 and 100:0, respectively, in duplicate (5×10^5 total cells, each). RNA was then prepared from each of the cell mixes for subsequent sequencing.

To promote Kassandra stability against cancer-specific noise and expression, both cancer cell lines and sorted malignant cells were added to the mixtures (Figure S27). Kassandra was not intended to predict exact tumor purity and outputs the "other" fraction with all uncharacterized cells (including cancer cells). To address the stability of Kassandra, previously unseen cancer cell lines (e.g., COLO829, MCF7, and K562) were admixed with PBMCs at different ratios ranging from 100:0 to 12:88 (cell line: PBMC). Kassandra reconstructed the percentages of all non-malignant immune cell types from PBMCs in correct proportions in all mixes and ratios despite low PBMC mRNA content (Figure S27). The percentages of "other" previously unseen cell types were calculated with the overall Pearson correlation coefficient of 0.94 ($p < 0.001$).

Cell sorting from blood and tissue

Different subtypes of $CD4^+$ and $CD8^+$ T cells, B cells, NK cells, monocytes, granulocytes, and dendritic cells (Table S6) were sorted from the peripheral blood of healthy donors (Research Blood Components, Watertown, MA, USA). Briefly, PBMCs were prepared from peripheral blood, labeled with monoclonal antibodies to identify populations of interest and sorted using a BD FACSAria III through a 100 μm nozzle. The gating strategy is depicted in Figures 3A–3E. Post-sort purity was verified for each population and found to be >95% for each subset (Figure S3F). The collected data were analyzed with FlowJo Software V10 (BD Biosciences, Franklin Lakes, NJ, USA). Fresh tumor biopsies were obtained from Tissue for Research (New Orleans, LA, USA) and single cell suspension were made using the Human Tumor Dissociation kit and gentle MACS Octo Dissociator with heaters (Miltenyi, Auburn, CA, USA) according to the manufacturer's instructions. Live ($DAPI^{neg}$) $CD4^+$ ($CD45^+$, $CD3^+$, $CD4^+$), and $CD8^+$ ($CD45^+$, $CD3^+$, $CD8^+$) T cells, macrophages ($CD45^+$, $CD14^+$, $CD64^+$, $CD15^-$, $CD3^-$), and fibroblasts ($CD45^-$, $EpCAM^-$, $CD31^-$, $CD90^+$, $Podoplanin^+$) were sorted through the 100 μm nozzle. Post-sort purity was checked immediately after sorting and found to be >96% for cells sorted from primary tumor tissue. All cells were then transferred to RNAeasy lysis buffer and processed for sequencing.

RNA sequencing

RNA was extracted using the RNAeasy mini kit (Qiagen, Hilden, Germany). Libraries were prepared with Illumina TruSeq® Stranded mRNA Library Prep (Poly-A mRNA; stranded). Libraries were sequenced on NovaSeq 6000 as Paired-End Reads (2x150) with targeted coverage of 50 mln reads.

STATISTICAL ANALYSIS

Statistics were calculated using the *scipy.stats* module in Python 3. Correlations were Pearson unless otherwise stated. The significance of Pearson correlations (r) to be nonzero was assessed by the use of the exact distribution of r (two-tailed test). All graphs were plotted using custom implementation of *matplotlib* and *seaborn* libraries of Python 3. The diagrams are drawn on the website www.draw.io.