# Estimation of immune cell content in tumour tissue using single-cell RNA-seq data

**Max Schelker[1,2], Sonia Feau[1], Jinyan Du[1], Nav Ranu[1], Edda Klipp[2], Gavin MacBeath[1], Birgit Schoeberl[1], Andreas Raue[1,*]**

[1] Merrimack Pharmaceuticals, Inc., Cambridge, MA 02139, USA
[2] Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany

* araue@merrimack.com

## Abstract

**As interactions between the immune system and tumour cells are governed by a complex network of cell-cell interactions, knowing the specific immune cell composition of a solid tumour may be essential to predict a patient's response to immunotherapy. Here, we analyse in depth how to derive the cellular composition of a solid tumour from bulk gene expression data by mathematical deconvolution, using indication- and cell type-specific reference gene expression profiles (RGEPs) from tumour-derived single-cell RNA sequencing data. We demonstrate that tumour-derived RGEPs are essential for the successful deconvolution and that RGEPs from peripheral blood are insufficient. We distinguish nine major cell types as well as three T cell subtypes. As the ratios of CD4+, CD8+ and regulatory T cells have been shown to predict overall survival, we extended our analysis to include the estimation of prognostic ratios that may enable the application in a clinical setting. Using the tumour derived RGEPs, we can estimate, for the first time, the content of cancer associated fibroblasts, endothelial cells and the malignant cells in a patient sample by a deconvolution approach. In addition, improved tumour cell gene expression profiles can be obtained by this method by computationally removing contamination from non-malignant cells. Given the difficulty around sample preparation and storage to obtain high quality single-cell RNA-seq data in the clinical context, the presented method represents a computational solution to derive the cellular composition of a tissue sample.**

Enhancing a patient's immune response to cancer using immune checkpoint inhibitors is arguably the most exciting advance in the treatment of cancer in the past decade[1,2]. Unfortunately, only a subset of patients (typically ~20%) show long lasting responses post checkpoint blockade[3]. Combining prospective patient selection based on predictive response biomarkers (=precision medicine) and immunotherapy has the potential to further transform patient care. To date, it has been shown that location and abundance of immune cells are prognostic for predicting patient outcome on standard therapy[4,5]. In addition, for checkpoint inhibitors like anti-PD1, anti-PDL1 and anti-CTLA4 agents, the presence of relevant T cell populations correlates with treatment efficacy[6].

Thus, it is likely that the key to predicting response to immunotherapy lies in the patient-specific immune cell composition at the site of the tumour lesion.

In theory, it is possible to infer the immune, tumour, and stroma cell content of a solid tumour from its bulk gene expression profile if reference gene expression profiles (RGEPs) can be established for each tumour-associated cell type. Mathematically, this class of inverse problems is known as *deconvolution*[7]. To date, deconvolution of bulk gene expression has been described and validated for haematological malignancies[8,9], where RGEPs can be established from peripheral blood mononuclear cells (PBMCs). This approach has been applied theoretically to solid tumours[10], but until recently it has been impossible to validate this extrapolation experimentally. It has been difficult to obtain RGEPs for cell types that are not available in the peripheral blood, such as endothelial cells and cancer-associated fibroblasts and it remains unclear to which extent the gene expression profile of an immune cell changes upon tumour infiltration. With the advent of the single-cell RNA sequencing (scRNA-seq) technology, however, it is now possible to determine gene expression profiles for tumour-infiltrating immune cells, tumour-associated non-malignant cells, and individual tumour cells from the same solid tumour biopsy.

We collected and investigated RNA-seq gene expression profiles of more than 11,000 single cells from three distinct primary human tissue sources: To characterize cells associated with the tumour microenvironment we accessed data from 19 melanoma patients[11], to characterize the baseline immune cell gene expression we accessed data from PBMCs originating from four healthy subjects[12] and last, we generated immune and tumour cell gene expression profiles from four ovarian cancer ascites samples in-house. In the following, we show that gene expression profiles from tumour-associated immune cells and from PBMCs differ substantially. Therefore, reference profiles obtained from PBMCs are insufficient to deconvolve the bulk profile of a melanoma tumour sample. We find that indication-specific immune cell RNA-seq profiles from different patients are sufficiently similar to each other to define a *consensus* profile for each cell type, and that these consensus profiles enable accurate deconvolution of bulk tumour profiles. Our results show that the generation of specific RGEPs is both necessary and sufficient to enable reliable estimation of tumour composition from bulk gene expression data. Our approach resolves tumour-associated cell types with an unprecedented precision that considers subtle differences in the gene expression state of these cells. We can reliably identify nine different cell types including immune cells, cancer-associated fibroblasts, endothelial cells, ovarian carcinoma cells and melanoma cells. In addition, RGEPs for immune cells can be used to estimate the unknown gene expression profiles of tumour cells from bulk gene expression data patient specifically.

# Results

## Microenvironment dependent modulation of gene expression profiles of tumour-associated cells

First, to investigate the extent to which gene expression profiles change as immune cells move from peripheral blood to the tumour microenvironment, we compared immune cell scRNA-seq profiles across three human data-sets: 1) data-set of 4000 single cells derived from peripheral blood of four healthy subjects[13]; 2) data-set of 4645 tumour-derived single cells from 19 melanoma patient samples[11] and an unpublished data-set of 3114 single cells from four ovarian cancer ascites samples. Single-cell RNA-seq data requires careful data processing and normalization particularly when comparing data originating from different sources and sequencing technologies. To characterize the single cells and to illustrate genome wide similarities and differences in their gene expression profiles, we applied the dimensionality reduction technique t-distributed stochastic neighbour embedding (t-SNE)[14]. This is an unsupervised machine learning algorithm that places each single cell into a two-dimensional plane. Cells with gene expression profiles that are similar are placed close to each other and farther apart if they are more different. Fig. 1a shows that clusters associated with specific cell types and from different data sources emerge spontaneously. Using the aggregated single-cell data-set, we developed a classification approach that can identify cell types irrespective of the data source. We can identify and classify nine major cell types: T cells, B cells, macrophages/monocytes, natural killer (NK) cells, dendritic cells (DCs), cancer-associated fibroblasts (CAFs), endothelial cells (ECs), ovarian carcinoma cells and melanoma cells. All remaining cells that fail to pass the classification threshold for any specific cell type are assigned as “unknown”. Interestingly, the “unknown” cells are mostly located in the T cell clusters, suggesting that some T cells are more difficult to classify than cells from other cell types. However, the percentage of “unknown” cells per sample is generally very low (<0.03 %). Further, we could classify T cells into three subtypes: CD4+, CD8+ and regulatory T cells (Treg). The ratios of CD4+ or CD8+ T cells and immune suppressive Tregs were suggested as markers for immunologically active vs. inactive tumours[6]. Although our methods can easily be extended to include additional cells and further subdivisions, we limited ourselves to the nine major cell types to benchmark our classification algorithm. As previously reported[11] and shown in Extended Data Fig. 3, malignant tumour cells and associated fibroblasts cluster by patient and non-malignant cells cluster by cell type. Tumour biopsies should contain immune cells from tumour blood vessels and from recently extravasated immune cells. Therefore, a partial overlap between PBMC and tumour-associated immune cells is expected. We analysed pair-wise similarities between the averaged gene expression profiles of each identified cluster. This analysis is more quantitative and robust to noise as the single cell comparisons. The results shown in Fig. 1b indicate that most clusters, while distinct, are most closely associated with clusters from the same cell types. This is an important quality control step that confirms that potential batch effects are successfully alleviated by the data processing and normalization strategy (see Methods Section). Tregs seem to be most distinct across the three different data-sets potentially indicating different context-dependent subsets[15]. However, the microenvironment has a clear and quantifiable impact on gene expression. In the following, we

will address the question if gene expression profiles based on PBMCs are good approximations for what is observed in the tumour microenvironment and how the PBMC-derived gene expression profiles impact the quality of deconvolution of bulk expression data.

First, we observed that the frequency of each cell type appears to be distinct for each sample as depicted in Fig. 1c. The cellular composition of the PBMC samples from different donors is more similar to each other compared to the cellular composition across the ascites or melanoma samples. We validated the predicted cellular composition based on our scRNA-seq based classification with previously reported results for all melanoma samples [11]. Also, we compared the predicted cellular composition for all ascites samples experimentally with Fluorescence Activated Cell Sorting (FACS). As depicted in Extended Data Fig. 2 our classification is in line with previously published results and our measurements.

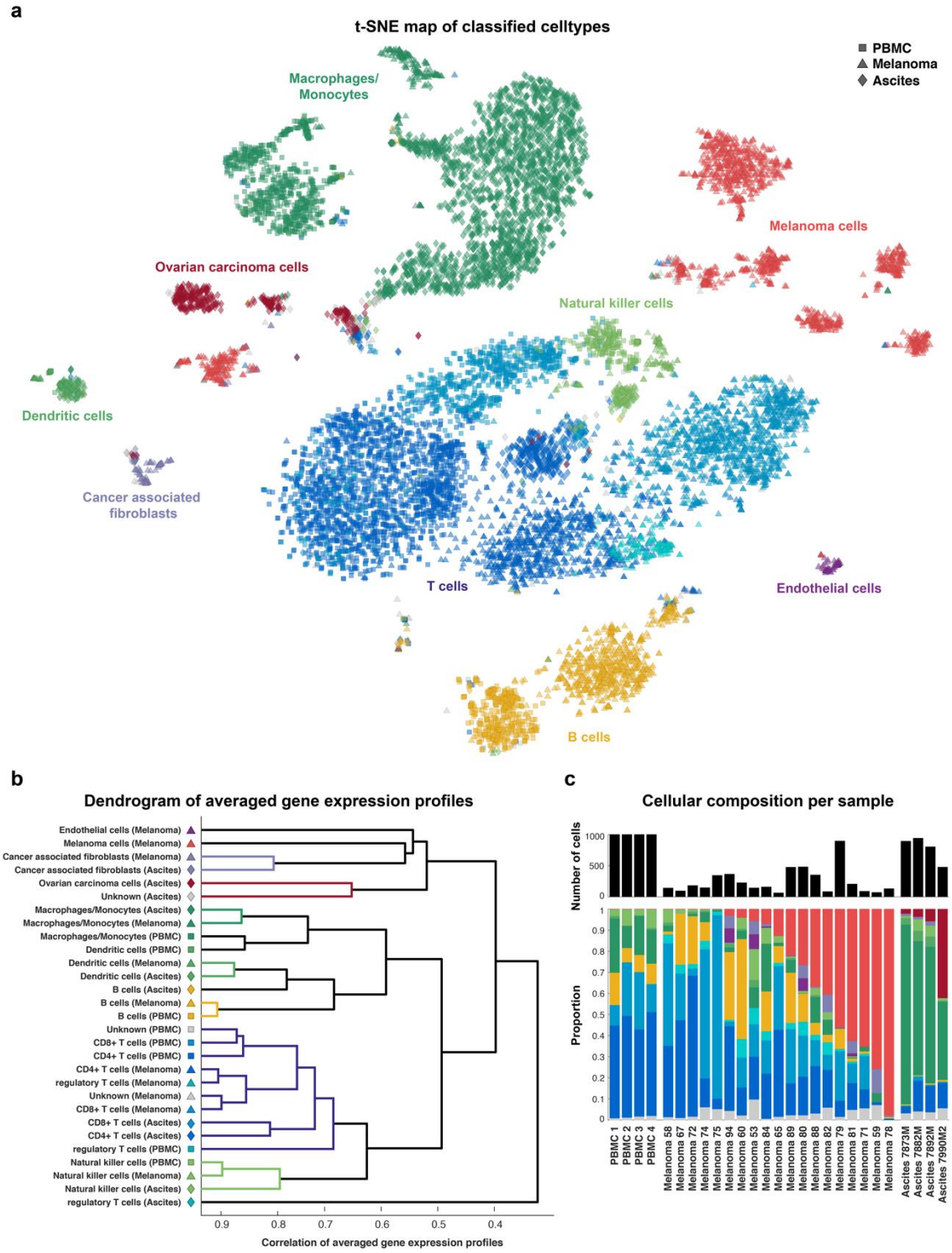



**Figure 1: Comparison of gene expression profiles of single cells from different data sources.**
**(a)** Single cells were arranged in two dimensions based on similarity of their gene expression profiles by the dimensionality reduction technique t-SNE. The clusters that emerge spontaneously can be associated with cell types (colours) and data source (symbol types: squares for PBMC-, triangles for the melanoma-, and diamonds for ascites data-sets). **(b)** Pair-wise correlation of averaged gene expression profiles of clusters encoding cell type and origin as identified in a) visualised as dendrogram. **(c)** Number of cells and cellular composition per sample.

## Using single-cell data as a benchmark for deconvolution accuracy

The microenvironment-specific gene expression profiles of immune cells as well as the true composition of a given sample can be obtained by scRNA-seq and can serve as ground truth to benchmark deconvolution approaches. We studied how the deconvolution results of bulk gene expression data, for instance of a melanoma sample, are affected by microenvironment-specific changes and by patient-to-patient variation. As benchmark for the deconvolution, we constructed "bulk" gene expression data by aggregating all single-cell gene expression data for each of the 27 samples as well as different sets of REGPs by different strategies for averaging over tissue sources and patients. We compare the inferred, *a priori* known cellular composition of a given sample using five different RGEPs (see Extended Data Fig. 4 for illustration): The first, RGEP1 is derived from the PBMC data-set only. Therefore, estimates for tumour-associated cell types will not be available in this case. The second, RGEP2, is derived for each cell type across the three data-sets (PBMC, melanoma and ascites). The third, RGEP3 is data-set/indication- and cell type-specific. As additional benchmarks, we set up two control scenarios (CNTR1 and CNTR2) that are extensions of RGEP3 and include patient-specific information. These scenarios are, of course, not applicable in the real world, but serve to evaluate the relative importance of patient-specific information. CNTR1 uses patient-specific profiles for the malignant cells only and consensus profiles for each non-malignant cell type. CNTR2 uses patient-specific profiles for all cell types. In principle, CNTR2 serves as the upper limit on what is technically possible using deconvolution approaches.

## Origin and quality of RGEP determine deconvolution results

To compare the five possible RGEPs and their impact on deconvolution accuracy, we estimated the cellular composition from the 27 constructed bulk expression datasets using the CIBERSORT deconvolution method[8]. This method is designed to be more robust against noise, unknown mixture content and closely related cell types. CIBERSORT has been shown to outperform other methods based on *in vitro* cell mixture benchmarks. All deconvolutions were performed using a collection of genes, which comprises 1076 signature genes that were found to maximally differentiate various cell types[8,11]. For each cell type, the estimated proportion was compared to the true proportion in the 27 constructed samples (Fig. 2a). The Pearson correlation coefficients between estimated and true cellular composition were used as a measure of prediction accuracy (Fig. 2b). The different T cell subsets are considered later in Fig. 3. Overall, estimations based on RGEP1 were less accurate ($\rho$=0.82) than for RGEP2 and RGEP3 or for CNTR1 and CNTR2 ($\rho \geq 0.98$). For RGEP1, due to the unavailable reference profiles for tumour-associated cell types the true proportion of unknown cells is larger than for the other RGEPs and the estimation quality is mediocre ($\rho$=0.65). For RGEP2 and RGEP3 as well as for CNTR1 and CNTR2, the true proportion of unknown cells is negligibly small. Correlation is not a good measure of accuracy in case the true proportion of cells is small. For RGEP1 the estimation performs well for T cells ($\rho$=0.88, not distinguished into subtypes here), B cells ($\rho$=0.99) and macrophages/monocytes ($\rho$=0.99). However, the accuracy improves further for all other settings ($\rho \geq 0.99$). For RGEP1 the estimation for DCs ($\rho$=-0.04) is poor and mediocre for NK cells ($\rho$=0.78). The estimation for DCs improves considerably for RGEP2 ($\rho$=0.82) and

RGEP3 (ρ=0.95). The estimation for DCs still improves slightly for CNTR1 (ρ=0.97) but reaches its maximum only for CNTR2 (ρ=1.00), indicating that gene expression of DCs is heavily dependent on the source of isolation which is in agreement with the evidence that distinct subsets of DCs are highly specialized in the generation of immunity[16]. The estimation for NK cells improves slightly for RGEP2 (ρ=0.82) and reaches close to optimal in RGEP3 (ρ=0.95) compared to CNTR1 (ρ=0.96) and CNTR2 (ρ=1.00). For RGEP2 to CNTR2, estimates for the tumour-associated cell types (CAFs, ECs and the malignant cells) become available and are estimated accurately (ρ≥0.95). Interestingly, the estimation for the malignant cells does not improve much upon inclusion of patient-specific information, suggesting that deconvolution using consensus profiles is feasible. This is possible because the tumour cells are in general very different from the non-malignant cells which make their deconvolution easier (see Fig. 1b). For CNTR2, ECs and CAFs have an increased accuracy (ρ=1.00) compared to the other settings (ρ~0.95), indicating that gene expression of those cell types is influenced by patient-specific microenvironment.

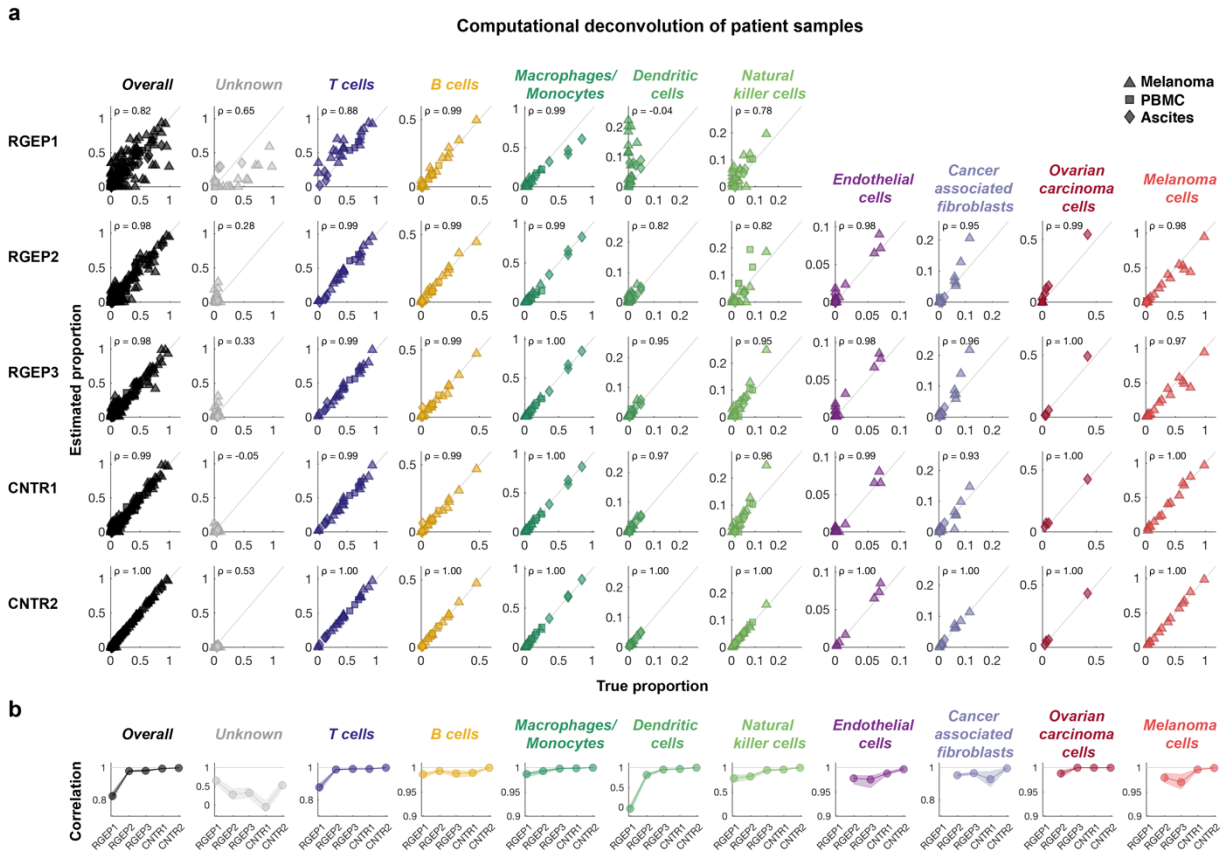



**Figure 2: Estimation accuracy of cellular composition is dependent on the origin and quality of RGEPs.**

**(a)** Scatter plot of true and estimated cell proportions for all 27 patient samples. Each dot represents one patient sample. Values close to the diagonal correspond to high deconvolution accuracy. Columns depict cell types; rows describe the five different configurations (REGP1-3 and CNTR1-2). *ρ* denotes the Pearson's correlation coefficient. In configuration REGP1, estimates for tumour-associated cell types are not available. **(b)** Pearson's correlation coefficient between estimated and true cell fraction for all five

configurations. Dots denote the median of the correlation coefficient; the shading represents the uncertainty based on bootstrapping (upper and lower quartile). (Please note the different scaling of the figure axes.)

Given the importance of T-cell ratios for treatment outcome[6], we further analysed the estimation accuracy for T cell subsets as well as for therapeutically relevant T cell ratios (Fig. 3). Surprisingly, for CD8+ T cells, the estimation results are accurate (ρ~0.95) for all RGEPs. For CD4+ and regulatory T cells, the estimation results using RGEP1 are only mediocre (ρ=0.63 and ρ=0.43) but improve significantly for RGEP2 (ρ=0.87 and ρ=0.94). This is also reflected in the ratios of Treg/CD4+, CD8+/Treg and CD4+/CD8+ T cells that reach accurate estimations for RGEP2 (ρ=0.94, ρ=0.96 and ρ=0.93). The estimation for all T cell subsets and ratios does not significantly improve for CNTR1 but does improve for CNTR2 (ρ=1.00), indicating that gene expression of T cells is influenced by the patient-specific microenvironment. In summary, deconvolution using consensus gene expression profiles based on indication-specific gene expression profiles (RGEP3) were sufficient to obtain reliable estimates of the cellular composition of the samples without requiring patient-specific data on the individual cell types. Deconvolution using gene expression profiles based on data from peripheral blood (RGEP1) or based on averages across all three data-sets/indications (RGEP2) was considerably less accurate.

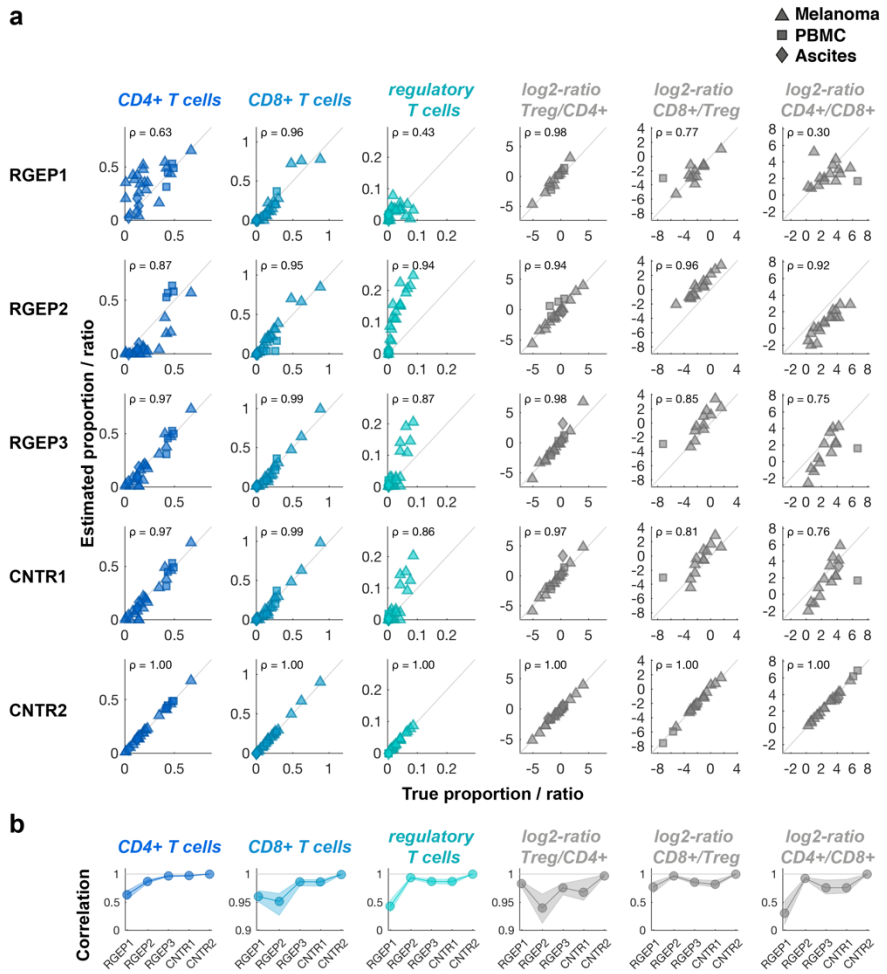


**Figure 3: Estimation accuracy of cellular composition for T cell subsets and ratios depends on the origin and quality of RGEPs.**
Description as for Fig. 2, ratios are indicated on a log2-scale.

To determine the impact of using alternative gene sets for the deconvolutions, we repeated the analyses using the best performing RGEP3 and four additional gene sets as well as three alternative deconvolution algorithms. Interestingly, the impact of different gene sets and deconvolution algorithms was relatively small compared to the impact of the origin and quality of the RGEPs (see Extended Data Fig. 5). CIBERSORT in conjunction with the *Merged* gene set provided the best overall results.

## Estimation of patient-specific tumour cell gene expression profiles

Although using RGEP3 that is indication- but not patient-specific enables the accurate estimation of cellular composition of any given patient biopsy from bulk gene expression data, the gene expression profile of the malignant cells varies the most from patient to patient. Differences in gene expression in tumour cells are expected to play a key role in predicting response to traditional therapies, including both targeted and chemotherapies. As such, it is also of interest to estimate the patient-specific tumour cell profile following deconvolution. If consensus profiles exist for every non-malignant cell type and indication, the patient-specific tumour cell profile can be obtained by simply subtracting the profile of each non-malignant cell type from the bulk profile, weighted by its inferred proportion. In practice, however, the bulk profile will always be "contaminated" by cells for which consensus profiles do not exist ("unknown" cells). For example, neutrophils are not represented in scRNA-seq data, as they are difficult to isolate, highly labile *ex vivo* and therefore are difficult to preserve with current single cell isolation methods[17]. Using scRNA-seq data, we calculated the estimated tumour cell expression profiles for each patient sample and compared them to the true tumour cell profile (Fig. 4a). As some genes, such as housekeeping genes, correlate between all cells irrespective of cell type, a certain baseline correlation is expected. We estimated this baseline correlation by correlating the gene expression profiles of the non-malignant cells with the true tumour cell gene expression profiles. We observe a baseline correlation of 0.7-0.8 for all samples, irrespective of the estimated proportion of tumour cells in the samples. As expected, the estimation accuracy of the tumour cell expression improved with increasing tumour cell content (Fig. 4b). Notably, when the estimated proportion of tumour cells in the samples exceeded 20%, the estimated tumour cell gene expression profiles exhibited a correlation of $\rho>0.9$ with the true profile. The predicted tumour cell gene expression profiles in samples with more than 20% but less than 70% tumour cells correlate significantly better with the true tumour cell gene expression profiles compared to the uncorrected overall gene expression profiles. If a sample contains more than 70% tumour cells the gene expression profile of the whole sample is dominated by the tumour cells already and does not require any subtraction. For samples with less than 20% tumour cells, the subtraction does not improve the estimation because the signal of the tumour cell gene expression is low. In summary, for samples with a tumour cell content between 20-70% deconvolution results in significantly improved gene expression profiles.

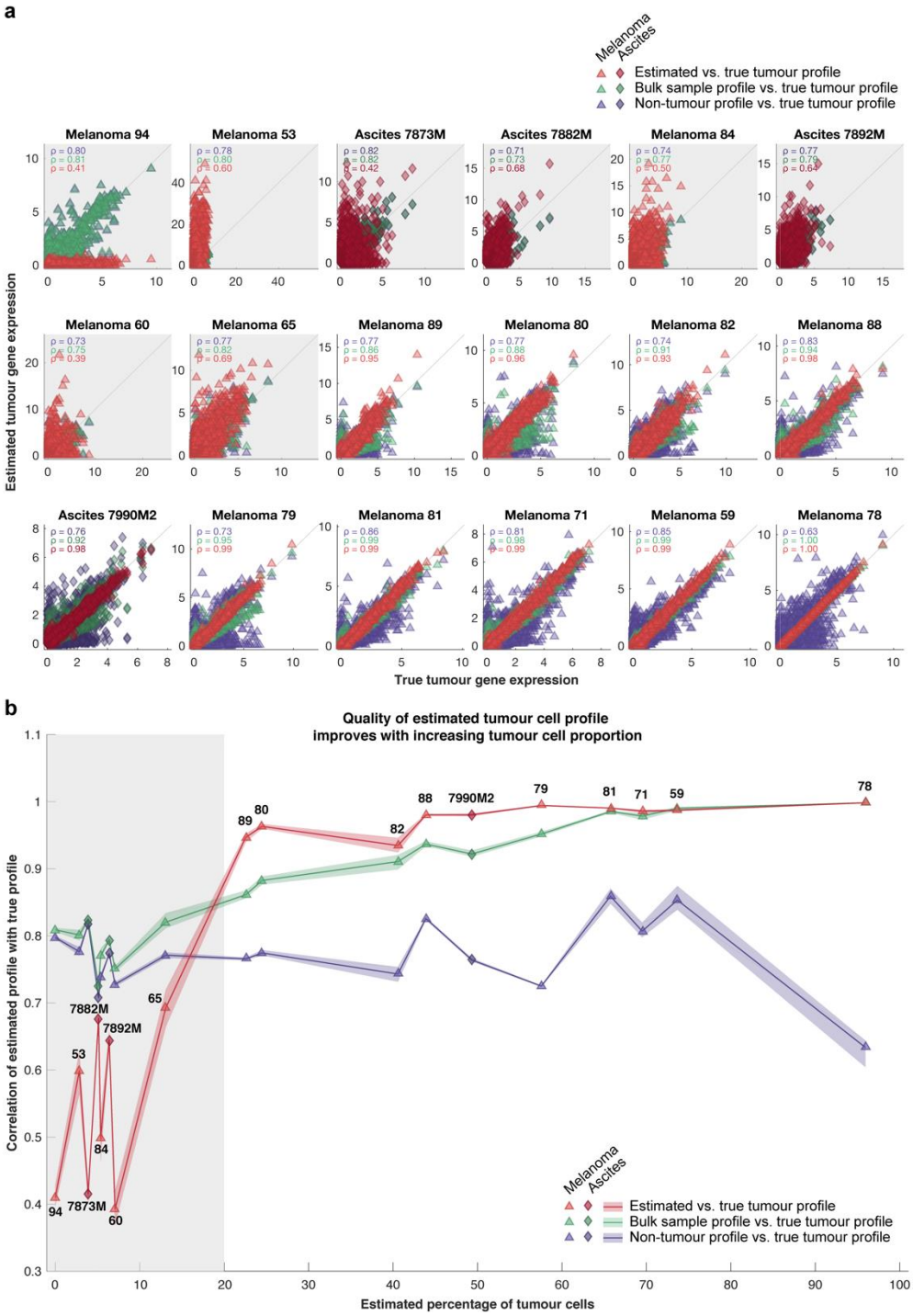



**Figure 4: Estimation accuracy of patient-specific tumour cell gene expression profiles.**
**(a)** Scatter plot of estimated vs. true transcriptome wide gene expression (17,933 genes) of the tumour cells for individual patient samples. Patient samples without any tumour cells have been excluded from this analysis. $\rho$ denotes the Pearson's correlation. Correlation plots with grey background indicate patient samples with less than 20% tumour cell content. Colours according to legend in panel b). **(b)** Correlation values from panel a) plotted against the estimated proportion of tumour cells for each patient sample. Shading represents uncertainty based on bootstrapping. Symbols and numbering denote individual patient samples.

## Discussion

Cellular heterogeneity is present in any biological sample. Single cell RNA-seq allows us to understand how cellular heterogeneity contributes to function or patient outcome. However, it is still much easier to obtain bulk gene expression data. The work presented here shows how deconvolution approaches can be applied to bulk gene expression data to infer cellular composition and to provide a tool to link cellular heterogeneity to biological function or drug response from bulk gene expression data. We show that with indication- and cell type-specific reference gene expression profiles deconvolution methods like CIBERSORT can accurately estimate the cellular composition of a given biopsy sample and in addition give us more accurate information about the tumour cell gene expression profiles by eliminating contamination from non-malignant cells. This is most relevant if the tumour cell content ranges between 20-70%.

Benchmarking different gene expression reference profiles and different deconvolution algorithms, we showed that the estimation accuracy is ultimately limited by the origin and quality of the reference gene expression profiles. Reference gene expression profiles derived from PBMCs are insufficient to enable accurate deconvolution of tumour bulk gene expression data[18]. By combining well-established deconvolution algorithms with state-of-the-art single-cell RNA-seq data of tumour biopsies, we showed that indication-specific consensus profiles of immune, stromal, and tumour cells, obtained directly from the tumour microenvironment, can be used to obtain accurate estimates of the cellular composition of a given sample. Overall, we found that the origin and quality of the reference profiles play a more dominant role than the deconvolution algorithm or gene set that is used, although gene sets designed to address as many cellular subsets as possible are clearly needed for accurate estimations of cellular heterogeneity.

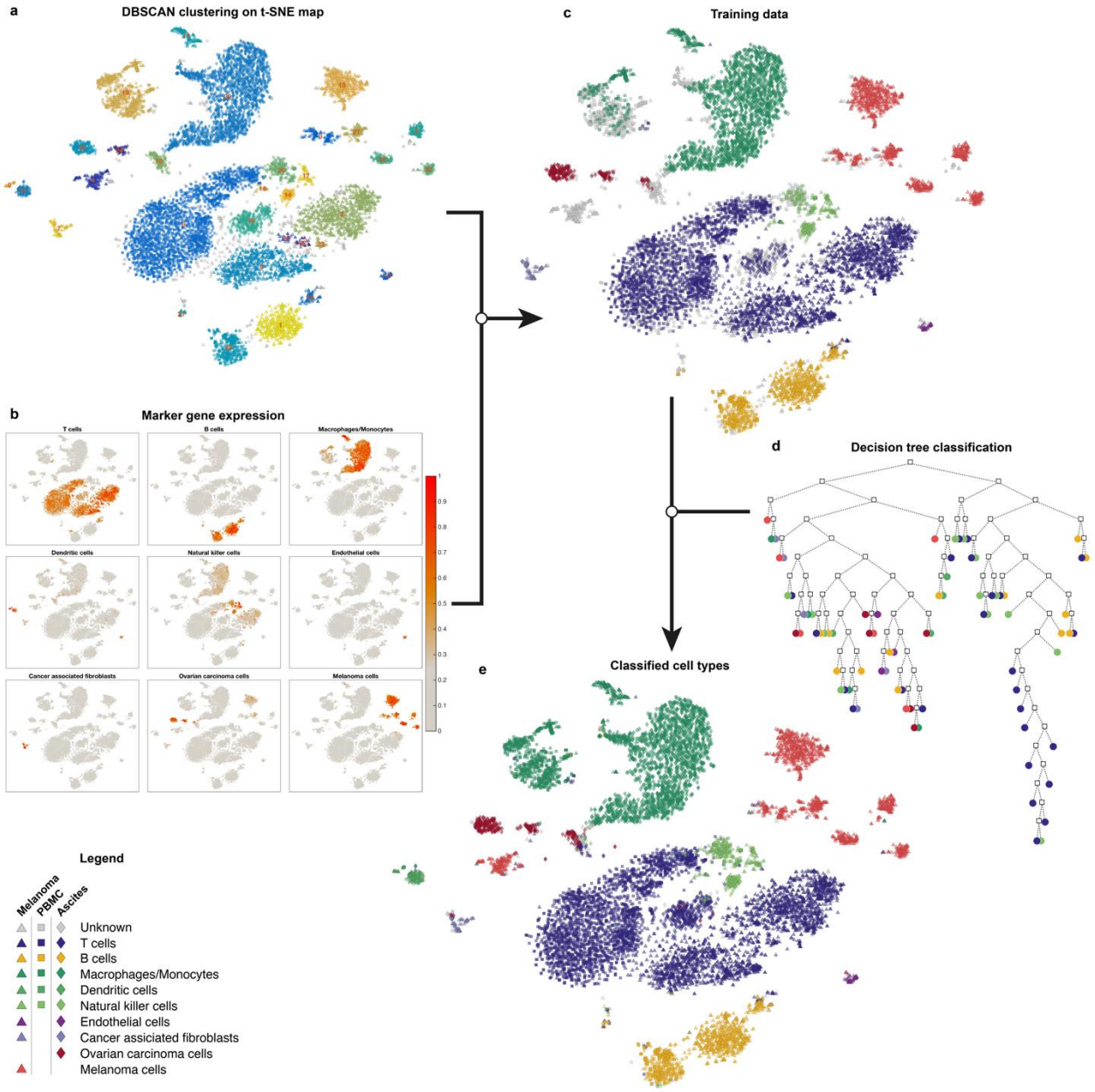
With the availability of public scRNA-seq data from PBMCs and melanoma samples as well as the ability to generate scRNA-seq data ourselves, we found that the gene expression profiles of tumour-associated immune cells differ considerably from those of blood-derived immune cells. Despite this systematic modulation, we found that patient-to-patient differences do not confound the deconvolution of bulk expression data and that consensus reference profiles can be established for each cell type, including tumour cells, for each specific microenvironment and indication.

We restricted our analyses to nine major cell types and three T cell subsets. Additional subdivisions can be added by defining these cell types in the scRNA-seq data-set and by choosing an appropriate gene set to enable these subdivisions. In practice, it is best to limit the number of subdivisions as much as possible, as more uncertainty is introduced when attempting to distinguish cell types with similar profiles.

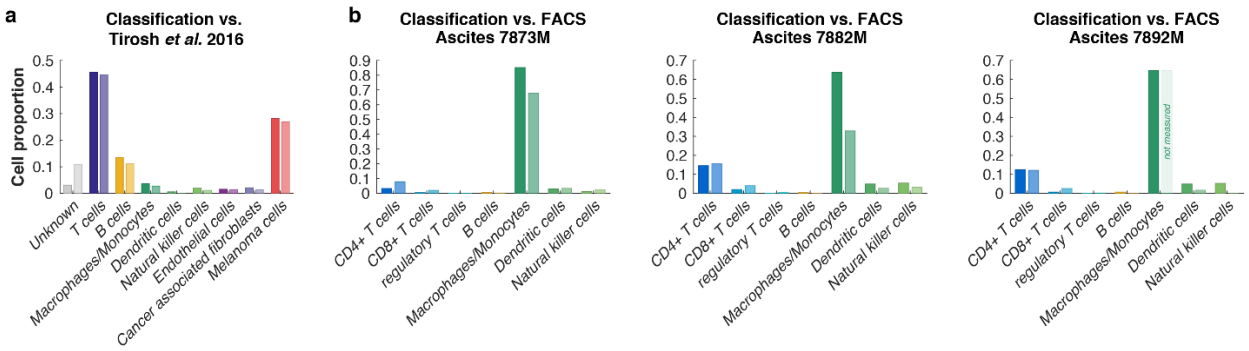
Given that we obtained the best results using indication- and cell-specific reference gene expression profiles, it is likely that consensus reference profiles for immune and stromal cells will need to be established from scRNA-seq data for every solid tumour indication. Why, then, are these deconvolution approaches necessary? At this time, scRNA-seq experiments are difficult to perform in a routine clinical setting. Tumour samples need to be acquired and analysed within hours. As

opposed to PBMCs, these samples can neither be fixed nor frozen. The reference scRNA-seq data-sets can be obtained, but only in very controlled settings. With the appropriate data-sets in place, however, deconvolution approaches enable routine clinical samples to be analysed both for cell content and patient-specific tumour cell gene expression profiles. Bulk RNA-seq data can easily be obtained from either flash-frozen or formalin-fixed, paraffin-embedded tissue samples, including both surgically resected material and core needle biopsies. The deconvolution approach presented here enables the estimation of immune cell content and improved tumour gene expression profiles in a clinical trial setting, which is necessary to link cellular content with treatment response. Therefore, we anticipate the discovery of novel predictive response biomarkers for both conventional and immune-directed therapy by taking the cellular composition into account.

# Extended data figures

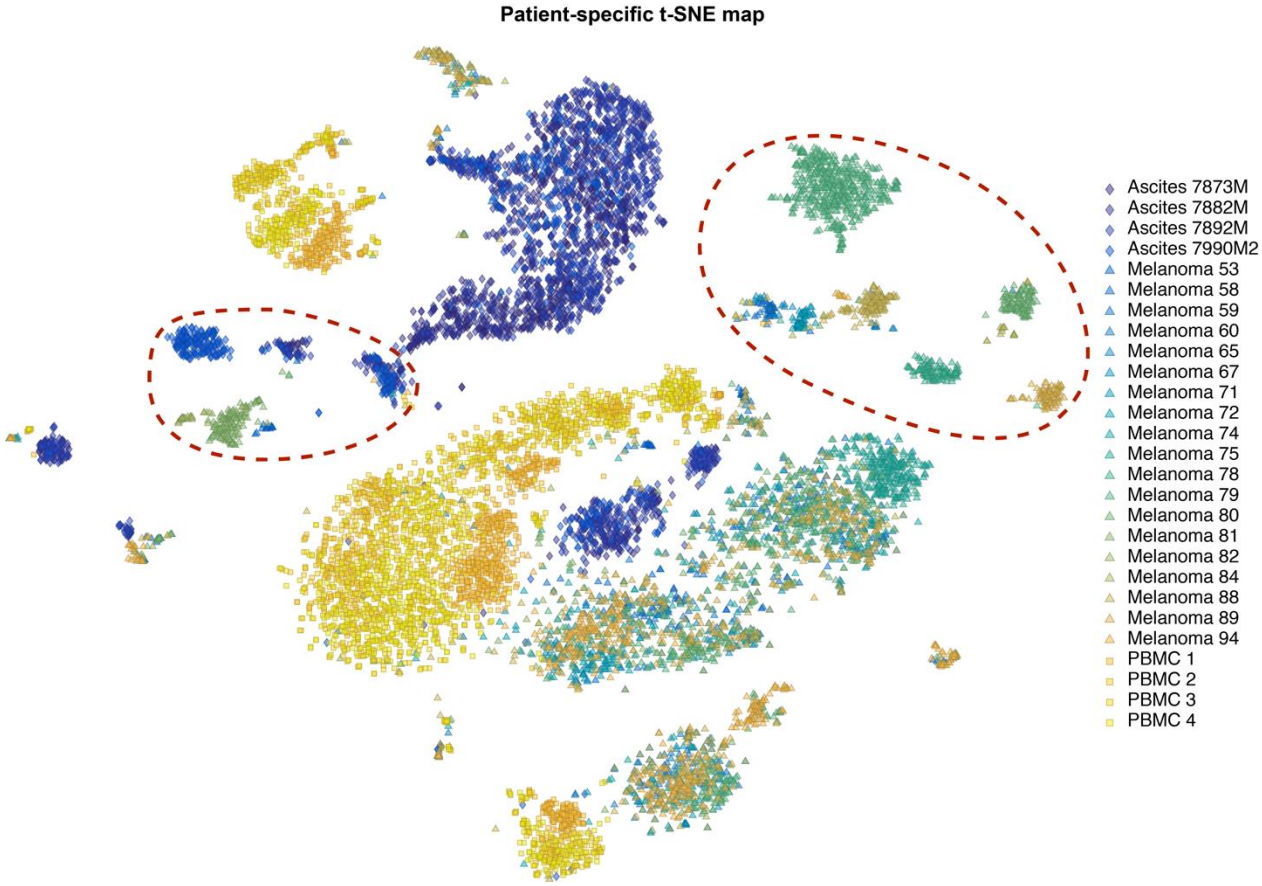



**Extended Data Figure 1: Classification of cell types from scRNA-seq expression profiles using decision trees.**
**(a)** DBSCAN clustering is performed on the t-SNE map to identify distinct cell clusters with high similarity. **(b)** The expression of 45 marker genes is evaluated based on three logical gates (AND, OR, NOT) and shown on top of the t-SNE mapping. **(c)** Predominant cell types within each cluster are identified based on the marker gene expression and used as a training set for unsupervised classification. **(d)** A decision tree classifier is trained and utilized to predict the cell types of all individual cells. **(e)** The resulting map indicates all nine major cell types by colour and by indication (melanoma, ascites) or location (PBMCs) by symbols.



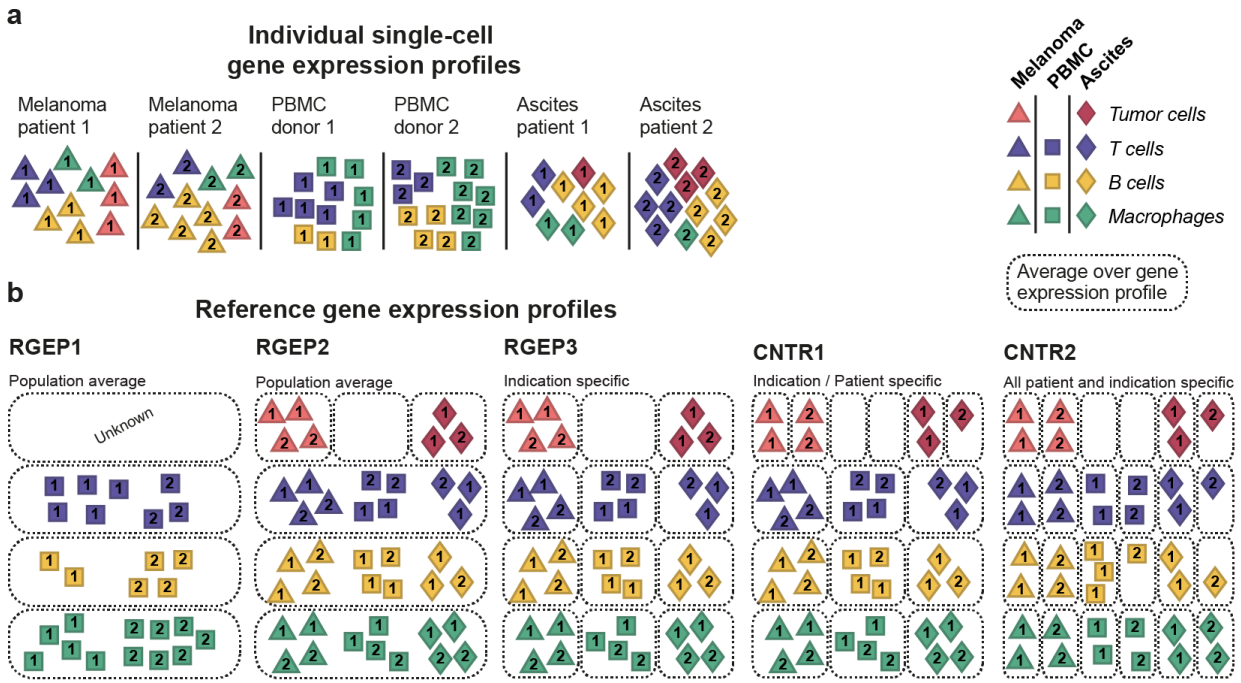

**Extended Data Figure 2: Benchmarking the cell type classification to literature and experimental FACS analysis.**

**(a)** The result of our cell type classification (left bars, dark colours) compared to the cell types provided across all melanoma samples in the data-set by Tirosh *et al.*[19] (right bars, light colours). **(b)** Cell type classification (left bars, dark colours) compared to FACS data (right bars, light colours) for three ovarian ascites patient samples. For sample 7892M, macrophages/monocytes could not be detected by FACS.

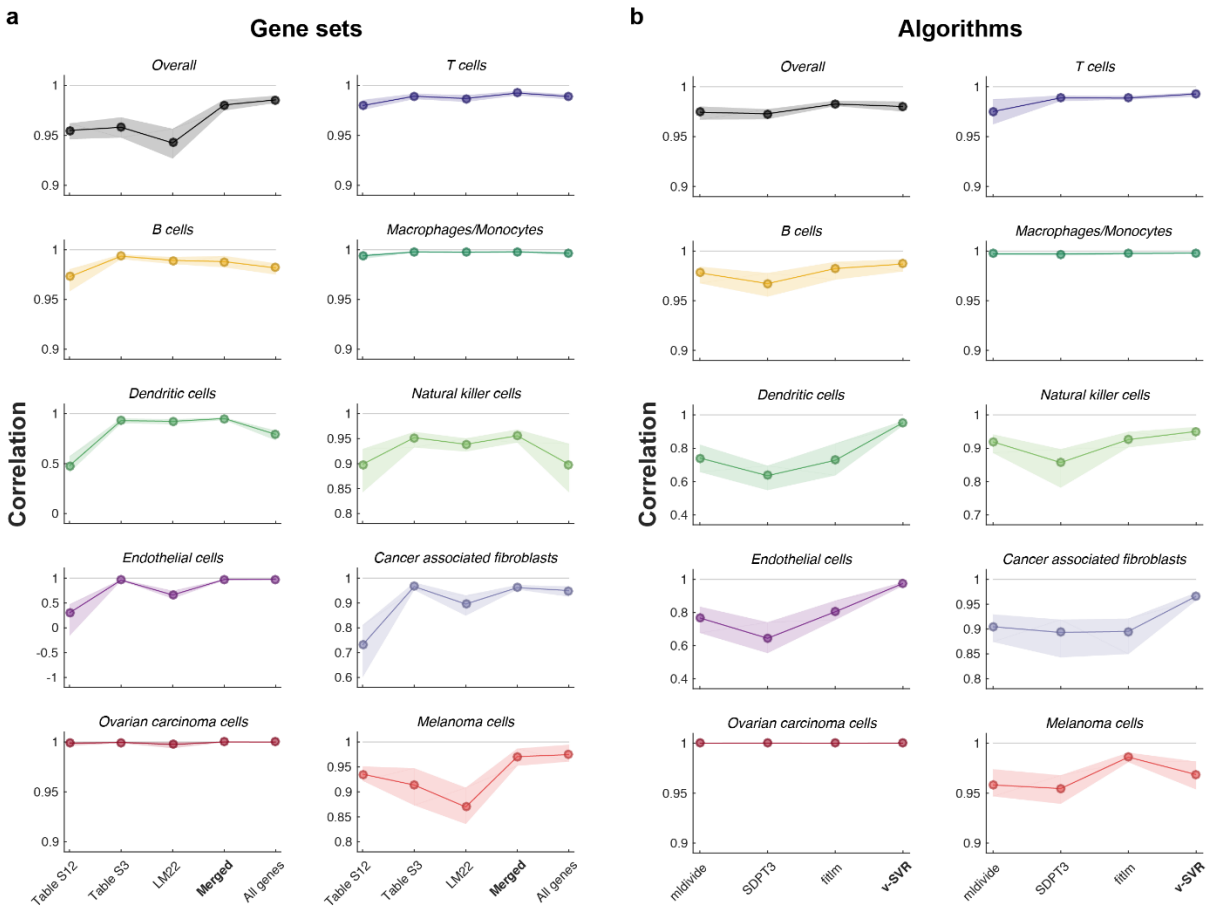



**Extended Data Figure 3: t-SNE map with patient-specific colour-coding.**

Single cells (symbols) were arranged in two dimensions based on similarity of their gene expression profiles by the dimensionality reduction technique t-SNE. Colours indicate the patient sample and symbols show the source location (triangles for melanoma, squares for PBMCs, and diamonds for ascites). Red dashed ellipses indicate clusters of malignant tumour cells.




**Extended Data Figure 4: Construction of five RGEPs for benchmarking the estimation accuracy.** **(a)** For each source location (melanoma, ascites, PBMC) individual single cell gene expression profiles are collected for multiple patients. Colours indicate the cell type, numbers indicate the patient sample and symbols show the source location (triangles for melanoma, squares for PBMCs, and diamonds for ascites). **(b)** Construction of REGPs from three single cell data-sets: RGEP1 bases on the population average of PBMC data; RGEP2 takes all three source locations into account; RGEP3 is indication- and location-specific; CNTR1 is patient-specific for tumour cells and indication/location-specific for non-malignant cells; CNTR2 is fully patient-specific.




**Extended Data Figure 5: Estimation accuracy is dependent on different signature gene sets and deconvolution algorithms.**

Dots denote the mean of the correlation coefficient; the shading represents the uncertainty based on bootstrapping. **(a)** Comparison of five different signature gene sets used for generating the reference gene expression profiles. Table S12: a published set of 244 marker genes differentially expressed in regulatory T cell subpopulations based on the scRNA-seq melanoma data[11]; Table S3: a published set of 385 marker genes based on the scRNA-seq melanoma data[11]; LM22: gene set which comprises 547 signature genes that were found to maximally differentiate various cell types[8]; Merged: a list of 1076 unique genes combined from the LM22, Table S12 and Table S3 gene lists as well as the 45 marker genes used for classification; All genes: a list of 17,933 genes that have a non-zero expression in at least one sample. **(b)** Comparison of four different algorithms used in the deconvolution approach. mldivide: exact solution using an algorithm for matrix inversion in MATLAB (The MathWorks, Inc.); SDPT3: a semidefinite-quadratic-linear programming algorithm from the CVX package[20]; fitlm: fitting a linear model (y = a*x+b) to the data based on least-squares in MATLAB; $\nu$-SVR: a support vector regression algorithm used in the CIBERSORT method.

# Methods

## Data sources

Ovarian cancer ascites of four patients were obtained from University of Massachusetts – Worcester. Samples were shipped on ice on the same day and processed upon receiving. Each sample was filtered through a 70μm filter and the cells were centrifuged down at ~300xG in a swing bucket centrifuge. Cells were resuspended in PBS-5% FBS and counted. 1e7 viable cells were frozen down in 90% FBS + 10% DMSO and stored at -80C until use. Cryopreserved cells were thawed at 37C water bath. Cells were spun down and resuspended in PBS-0.1% BSA and stored on ice. Viable cell number was measured and cells were diluted to 1.6 to 2e5 cells per ml in PBS-0.1% BSA. About 3,000 cells were encapsulated using the InDrops procedure[21,22] at the Single Cell Core at Harvard Medical School. The libraries from about 1,000 cells per sample were sequenced with the Illumina Nextseq 500 method at the Molecular Biology Core Facility at Dana-Farber Cancer Institute. Single-cell melanoma data were obtained from Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/gds) under accession number GSE72056 in a pre-processed format. Single-cell RNA-seq data of PBMCs from patient blood samples were downloaded from the 10x Genomics website (https://support.10xgenomics.com/single-cell/datasets; "4k/8k PBMCs from a Healthy Donor", "Frozen PBMCs (Donor A/B/C)"[13]) and 1,000 random cells were selected randomly for each donor to ensure similar size as for the melanoma and ascites data-sets. The *LM22* gene set was taken from the supplementary information of Newman *et al.* 2015[8]. *Table S3* and *Table S12* were obtained from the supplementary information of Tirosh *et al.* 2016[11].

## Data processing

Gene expression values were used on the transcripts per million (TPM) scale as provided by current quantification methods [23–25]. The expression values were transformed to

$$y = \log_2(\mathrm{TPM} + 1).$$

To ensure cross-sample comparability, all single-cell melanoma, PBMC and ascites samples were normalized to the average expression of 3559 housekeeping genes[26] by

$$\tilde{y}_i = y_i \cdot \frac{\overline{\mathrm{HK}}}{\mathrm{HK}_i},$$

where $y_i$ represents the gene expression profile of the $i$-th sample, $\mathrm{HK}_i$ denotes the average gene expression over all housekeeping genes of the $i$-th sample and $\overline{\mathrm{HK}}$ is the average expression over all housekeeping genes and samples. Other normalization methods like upper quartile or median normalization could not be applied to scRNA-seq data as the single-cell measurements contain too many genes with zero expression leading to a zero-upper quartile and median for several samples. Gene symbols of the single-cell melanoma data were corrected to account for automatic conversion into dates by Microsoft Excel[27].

## Flow cytometry analysis

Ascites were stained with either anti-human CD45 (H30, APC), CD3 (OKT3, Alexa-Fluor 700), CD4 (SK3, APC-Cy7), CD8 (RPA-T8, BV510), CD25 (BC96, Percp-Cy5.5), CD56 (5.1H11, BV570), CD127 (A019D5, BV421) CD16 (3G8, BV650) Abs or with anti-human CD45 (H30, APC), CD1c (L161, BV421), HLA-DR (L243, APC-Cy7), CD14 (M5E2, Alexa-Fluor 700), CD15 (W6D3, BV605) CD16 (3G8, BV650) Abs in PBS 2% FBS for 20 min at 4°C. The antibodies were purchased from BD Pharmingen or Biolegend. The samples were acquired on an LSRFortessa flow cytometer (Becton Dickinson) and analysed using FlowJo software.

## Classification of cell types based in single-cell

For classification of scRNA-seq data, a multi-step approach was developed. In contrast to the classification approach presented by Tirosh *et al.*[11], malignant and non-malignant cells were not treated separately, and only the scRNA-seq gene expression data were used for generating a training set. A workflow chart of our classification approach is depicted in Extended Data Fig. 1. After normalizing all data as described above, t-SNE mapping was performed on the Merged gene set in order to identify clusters of similar cells. Subsequently, the DBSCAN algorithm[28] was used to identify clusters based on the t-SNE map as shown in Extended Data Fig 1a. The parameters of DBSCAN were set manually to `MinPts=25` and `Eps=1.5` with the aim that each larger cell group on the map is assigned to a separate cluster. For each cell, the expression of a total of 45 marker genes (see Supplementary Table 1) was normalized to [0, 1] and evaluated based on three categories of genes: (1) AND genes that are all required, (2) OR genes where only the expression of one of them is necessary, and (3) NOT genes where the expression is a negative selection criterion. Evaluating all three categories for each cell type led to a score describing the likelihood of each cell to belong to a certain cell type. The resulting score is depicted as a heat map on top of the t-SNE map in Extended Data Fig. 1b. In each DBSCAN cluster, a total cell type score was calculated and only cell types with a predominant total score (>75% of the maximal score) were assigned preventing the misclassification of closely related cell types (e.g. Natural killer cells and T cells). This initial cell type assignment led to a sparse training set as depicted in Extended Data Fig. 1c. A decision tree classifier was trained based on these training data (also based on the Merged gene set). Using the trained classifier, the identity of all cells was predicted and validated based on 5-fold cross-validation (Extended Data Fig. 1d) showing a high accuracy (98.06%) for the classification of the major cell types. Cells with a posterior probability lower than 0.99 were marked as “unknown”. The resulting classification is shown in Extended Data Fig. 1e. Further, three sub-types of T cells (CD4+, CD8+ and regulatory T cells) were classified based on the T cell population defined in the first round of classification. The same procedure as explained above was repeated on the T cell subtypes (as shown in Supplementary Figure 1). This was necessary because the similarity of sub-types is much higher than for distinct cell types. Only the parameters for DBSCAN needed to be adjusted (`MinPts=25` and `Eps=1.75`) to account for the smaller sample size and the different distances on the t-SNE map of the T cells. The cross-validation of the

classification resulted in an accuracy of 93.88% for the T cell sub types. The resulting t-SNE map indicating all cell types and T cell subtypes is depicted in Fig. 1a.

## Construction of "bulk" gene expression data from single cell data

The "bulk" gene expression data that was used for testing the deconvolution approach was generated from single-cell RNA sequencing data by aggregating reads from all cell barcodes for each patient sample. As single-cell and conventional bulk sequencing differ in their quantification biases, we cannot assume that single-cell based reference gene expression profiles are applicable for deconvolution of conventional bulk sequencing data. Therefore, to apply the deconvolution based on single-cell reference gene expression profiles, conventional sequencing must be adapted to closely mimic the quantification process in single-cell sequencing, however, without the cell barcoding that would be problematic in a clinical trial setting.

## Deconvolution algorithms

Computational approaches to decipher the relative immune cell content in the tumour environment from microarray or RNA-seq gene expression data have been proposed[8,9,29,30] and have been validated on blood samples[8,9] or *in-vitro* cell mixtures[30]. Detailed reviews on this topic are available[7,31]. A method called CIBERSORT was proposed and its performance was compared to previously existing methods[8]. Using blood samples of a total of 41 patients, the authors could show that CIBERSORT outperforms the other methods when comparing the deconvolution results to flow cytometry data. The authors of the study also released a set of 547 genes (called LM22) which was used for their deconvolution approach. The CIBERSORT method (and most of the other methods mentioned above) assumes that the gene expression profile of an unknown bulk sample can be explained by the weighted sum of the cell type specific profiles of which it is composed. The weights vector leading to the linear combination can be obtained by solving a linear equation system computationally. As biological data can be obscured by technical and biological variability, methods for deconvolution need to be robust against noise. The contamination of the sample with unknown cell types can be a further source of noise. A method called $\nu$-Support vector regression ($\nu$-SVR) combines feature-selection with a linear loss function and $L_2$-regularization[32] and is therefore robust against noise[31]. The performance of deconvolution approaches has been widely demonstrated on *in vitro* mixtures and setting where the cellular gene expression profiles were directly measured such as for peripheral blood mono-nuclear cell (PBMC) content in blood. Although there have been attempts to use RNA-seq data for deconvolution of cell mixtures[30,33], so far, the accuracy of the approach has not been evaluated in a realistic setting and in a systematic manner. The potential of having absolute expression values from RNA-seq data rather than relative data from microarray has not been fully exploited.

## Signature gene sets

The basis for an accurate deconvolution is the choice of the *signature gene set*. The gene expression levels of these genes need to be informative enough to distinguish between cell types contained in

the mixture/bulk sample. For our comparison, we chose five different signature gene sets. The *LM22* gene set[8] consists of 547 genes of which 496 are contained in the scRNA-seq PBMC, melanoma and ascites data-sets. The *Table S12* gene set[11] contains 244 genes that are preferentially expressed in regulatory T cells of which 239 are present in all three data-sets. The *Table S3* gene set[11], a list of 391 genes that have been identified as differentially expressed among the cell types in scRNA-seq data. Therefore, 374 genes are contained across all three scRNA-seq data-sets. A *Merged* gene set, generated by merging all genes from the LM22, the Table S3 and the Table S12 gene sets and adding the 45 marker genes used for classification training. It consists of 1076 unique genes, with 1015 genes in common with the scRNA-seq data. An *All genes* gene set, consisting of 17,936 genes that are contained in the all three scRNA-seq data-sets with 17,933 non-zero genes for at least one single-cell profile.

## Settings for algorithm comparison

For deconvolution of the bulk patient profiles, the data was filtered to the Merged gene set and one of three deconvolution algorithms was applied. For $\nu$-Support vector regression ($\nu$-SVR) we used the implementation of libSVM[34] for MATLAB (version R2016a, The MathWorks Inc., Natick, MA, USA). The parameters were set to "`-s 4 -t 0 -n 0.50 -h 0 -c 1 -q`". The `mldivide` function from MATLAB uses the pseudo-inverse of the matrix $\underline{B}$ for solving for `w = pinv(B)*m`. This is equivalent to using `w = mldivide(B, m)`. The `fitlm` function from MATLAB fits a linear model to the data based on a least-squares fit. The main difference to the `mldivide` function is that for `fitlm` an intercept is taken into account. For the CVX package for MATLAB the problem was defined as:

```
cvx_begin quiet
    cvx_solver sdpt3;
    variable w(size(B, 2)) nonnegative;
    minimize( norm((B*w - m), 2) + lambda*norm(f, 2))
    subject to
        w <= 1;
        sum(w) <= 1;
cvx_end
```

with `lambda=1` and solved using the `SDPT3` algorithm[35] for semidefinite-quadratic-linear programming problems.

## Processing of estimation results

The results for the proportions of known cell types $\vec{w}$ as obtained by one of the above mentioned algorithms are processed by replacing negative numbers by zeros[8]. The proportion of unknown other cell types $\widetilde{w}$, i.e. cell types for which no reference profile was available, is calculated by taking the difference between one and the sum of all $m$ known cell proportions:

$$\widetilde{w} = 1 - \sum_{i=1}^{m} w_i.$$

## Estimation quality assessment

To assess the quality of our deconvolution results, we compared the true cellular fractions, as calculated from the number of single-cell measurements for each cell type and patient, with the estimation result by calculating Pearson's correlation coefficient $\rho$ for all patients. We quantified the uncertainty of our quality measure by performing bootstrap re-sampling (100 replications) of our deconvolution results and calculated the median and lower and upper quartiles.

## Dimensionality reduction

To obtain a low-dimensional representation of high-dimensional data, dimensionality reduction methods can be applied. T-distributed stochastic neighbour embedding (t-SNE) enables the reduction from many to two dimensions while keeping local neighbourhoods[14]. For removing noise and improving the performance, a principle component analysis (PCA) can be used to reduce the initial dimensionality before running t-SNE. We used the Barnes-Hut implementation of t-SNE[36] with the default settings to analyse our data. The result is a map that reflects the similarities between the high-dimensional input data as depicted in Fig. 1 and Extended Data Figs. 3 and 4.

## Tumour gene expression profile estimation

To calculate the gene expression profile of an average tumour cell for each individual patient, we need to subtract the explained portion of gene expression from the patient's bulk sample gene expression profile and rescale the expression with the estimated tumour proportion, i.e.

$$\vec{t}_i = \frac{\vec{m}_i - \underline{B}_{\text{non-tumor}}\,\vec{w}_{i,\text{non-tumour}}}{w_{i,\text{tumour}}}$$

## References


1. Drake, C. G., Lipson, E. J. & Brahmer, J. R. Breathing new life into immunotherapy: review of melanoma, lung and kidney cancer. *Nat. Rev. Clin. Oncol.* **11,** 24–37 (2014).

2. Intlekofer, A. M. & Thompson, C. B. At the bench: preclinical rationale for CTLA-4 and PD-1 blockade as cancer immunotherapy. *J. Leukoc. Biol.* **94,** 25–39 (2013).

3. Schumacher, T. N., Kesmir, C. & van Buuren, M. M. Biomarkers in cancer immunotherapy. *Cancer Cell* **27,** 12–14 (2015).

4. Galon, J. *et al.* Type, density, and location of immune cells within human colorectal tumors predict clinical outcome. *Science (80-. ).* **313,** 1960–1964 (2006).

5. Ino, Y. *et al.* Immune cell infiltration as an indicator of the immune microenvironment of pancreatic cancer. *Br. J. Cancer* **108,** 914–923 (2013).

6. Shang, B., Liu, Y., Jiang, S. & Liu, Y. Prognostic value of tumor-infiltrating FoxP3+ regulatory T cells in cancers: a systematic review and meta-analysis. *Sci. Rep.* **5,** 15179 (2015).

7. Mohammadi, S., Zuckerman, N., Goldsmith, A. & Grama, A. A Critical Survey of Deconvolution Methods for Separating cell-types in Complex Tissues. *Prepr. http//arxiv.org/abs/1510.04583* (2015).

8. Newman, A. M. *et al.* Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* **12,** 453–457 (2015).

9. Qiao, W. *et al.* PERT: A Method for Expression Deconvolution of Human Blood Samples from Varied Microenvironmental and Developmental Conditions. *PLoS Comput. Biol.* **8,** (2012).

10. Gentles, A. J. *et al.* The prognostic landscape of genes and infiltrating immune cells across human cancers. *Nat. Med.* **21,** 938–945 (2015).

11. Tirosh, I. *et al.* Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science (80-. ).* **352,** 189–196 (2016).

12. Zheng, G. X. Y. *et al. Massively parallel digital transcriptional profiling of single cells*. *bioRxiv* (2016). doi:10.1101/065912

13. Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8,** 14049 (2017).

14. Van Der Maaten, L. J. P. & Hinton, G. E. Visualizing high-dimensional data using t-sne. *J. Mach. Learn. Res.* **9,** 2579–2605 (2008).

15. Ulges, A., Schmitt, E., Becker, C. & Bopp, T. in *Advances in Immunology, Volume 132* (ed. Alt, F.) 1–46 (2016). doi:10.1016/bs.ai.2016.08.002

16. Collin, M., Mcgovern, N. & Haniffa, M. Human dendritic cell subsets. *Immunology* **140,** 22–30 (2013).

17. Thomas, H. B., Moots, R. J., Edwards, S. W. & Wright, H. L. Whose gene is it anyway? the effect of preparation purity on neutrophil transcriptome studies. *PLoS One* **10,** 1–15 (2015).

18. Aran, D. *et al.* Digitally deconvolving the tumor microenvironment. *Genome Biol.* **17,** 175 (2016).

19. Tirosh, I. *et al.* Single-cell RNA-seq supports a developmental hierarchy in human oligodendroglioma. *Nature* **539,** 309–313 (2016).

20. Grant, M. & Boyd, S. CVX: Matlab software for disciplined convex programming. *Available at http://cvxr.com/cvx/* (2008).

21. Klein, A. M. *et al.* Droplet barcoding for single cell transcriptomics applied to embryonic stem cells HHS Public Access. *Cell. May* **21,** 1187–1201 (2015).

22. Mazutis, L. *et al.* Single-cell analysis and sorting using droplet-based microfluidics. *Nat. Protoc.* **8,** 870–891 (2013).

23. Li, B. & Dewey, C. N. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12,** 323 (2011).

24. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides accurate, fast, and bias-aware transcript expression estimates using dual-phase inference. *Prepr. http//biorxiv.org/content/early/2016/08/30/021592* (2015). doi:10.1101/021592

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

26. Eisenberg, E. & Levanon, E. Y. Human housekeeping genes, revisited. *Trends in Genetics* **29,** 569–574 (2013).

27. Ziemann, M. *et al.* Gene name errors are widespread in the scientific literature. *Genome Biol.* **17,** 177 (2016).

28. Tran, T. N., Drab, K. & Daszykowski, M. Revised DBSCAN algorithm to cluster data with dense adjacent clusters. *Chemom. Intell. Lab. Syst.* **120,** 92–96 (2013).

29. Abbas, A. R., Wolslegel, K., Seshasayee, D., Modrusan, Z. & Clark, H. F. Deconvolution of blood microarray data identifies cellular activation patterns in systemic lupus erythematosus. *PLoS One* **4,** (2009).

30. Gong, T. & Szustakowski, J. D. DeconRNASeq: A statistical framework for deconvolution of heterogeneous tissue samples based on mRNA-Seq data. *Bioinformatics* **29,** 1083–1085 (2013).

31. Newman, A. M. & Alizadeh, A. A. High-throughput genomic profiling of tumor-infiltrating leukocytes. *Current Opinion in Immunology* **41,** 77–84 (2016).

32. Schölkopf, B., Smola, A. J., Williamson, R. C. & Bartlett, P. L. New Support Vector Algorithms. *Neural Comput.* **12,** 1207–1245 (2000).

33. Mehnert, J. M. *et al.* Immune activation and response to pembrolizumab in POLE-mutant endometrial cancer. *J. Clin. Invest.* **126,** 2334–2340 (2016).

34. Chang, C.-C. & Lin, C.-J. LIBSVM: a library for support vector machines. *ACM Trans. Intell. Syst. Technol.* **2,** 1–27 (2011).

35. Toh, K. C., Todd, M. J. & Tütüncü, R. H. SDPT3 — A Matlab software package for semidefinite programming, Version 1.3. *Optim. Methods Softw.* **11,** 545–581 (1999).

36. Van Der Maaten, L. Accelerating t-SNE using tree-based algorithms. *J. Mach. Learn. Res.* **15,** 3221–3245 (2014).