Single-Cell Immune Map of Breast Carcinoma Reveals Diverse Phenotypic States Driven by the Tumor Microenvironment

Elham Azizi[1]*, Ambrose J. Carr[1,2]*, George Plitas[3,4,5,6]*, Andrew E. Cornish[1,7]*, Catherine Konopacki[3,4], Sandhya Prabhakaran[1], Juozas Nainys[2,8], Kenmin Wu[3,4,5], Vaidotas Kiseliovas[1,8], Manu Setty[1], Kristy Choi[2,9], Phuong Dao[1], Linas Mazutis[1,8], Alexander Y. Rudensky[3,4,5&], Dana Pe'er[1,10&]

[1] Program for Computational and Systems Biology, Sloan Kettering Institute, Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA
[2] Department of Biological Sciences, Columbia University, New York, NY, USA
[3] Howard Hughes Medical Institute
[4] Immunology Program, Sloan Kettering Institute, Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA
[5] Ludwig Center at Memorial Sloan Kettering Cancer Center, Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA
[6] Breast Service, Department of Surgery, Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA
[7] Columbia University College of Physicians and Surgeons, New York, NY 10032, USA
[8] Sector of Microtechnologies, Institute of Biotechnology, Vilnius University, Vilnius, Lithuania
[9] Department of Computer Science, Columbia University, New York, NY, USA
[10] Parker Institute for Cancer Immunotherapy, Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA
* These authors contributed equally
& Co-senior author and lead contact
Correspondence to peerd@mskcc.org, rudenska@mskcc.org

SUMMARY

Knowledge of the phenotypic states of immune cells in the tumor microenvironment is essential for understanding immunological mechanisms of cancer progression and immunotherapy responses, as well as the development of novel treatments. By combining single-cell RNA-seq data from over 45,000 immune cells collected from eight primary breast carcinomas, as well as matched normal breast tissue, peripheral blood, and lymph node, we created an immune map of breast cancer. We developed a preprocessing pipeline, SEQC, and a Bayesian clustering and normalization method, Biscuit, to address the computational challenges inherent to single-cell RNA-seq data, enabling integration of data across patients. This atlas revealed significant similarity between normal and tumor tissue resident immune cells. However, we observed continuous tumor-specific phenotypic expansions driven by environmental cues. Our results argue against discrete activation states in T cells and the polarization model of macrophage activation in cancer, with important implications for characterizing tumor-infiltrating immune cells.

INTRODUCTION

Recent evidence suggests that cells of the innate and adaptive immune system serve an essential accessory function in non-lymphoid normal tissues and in tumors (Arpaia et al., 2015; Dunn et al., 2004; Green et al., 2017). The prevalent view is that relatively few states of differentiation of lymphoid and myeloid cells define the local tumor environment, and that these distinct states are linked to clinical outcomes such as cancer progression and response to anti-tumor therapies.

Naïve, activated, effector, and central memory T lymphocytes, as well as chronically stimulated dysfunctional T lymphocytes, are considered to be the principal T cell differentiation states. These states are distinguished by their antigen exposure history, functional features (distinct spectra of inflammatory mediators, cytotoxic potential, self-renewal, and migratory capacity), and characteristic yet not mutually exclusive patterns of cell surface co-stimulatory and co-inhibitory receptor expression. The former molecules, such as CD28, ICOS, OX40, CD40L, and CD137, markedly enhance TCR-dependent T cell activation and effector function, whereas increasing levels of combinatorially expressed inhibitory receptors CTLA-4, PD-1, TIGIT, LAG3, TIM-3, and CD160 are characteristic of progressive T cell dysfunction and loss of self-renewal potential ("exhaustion"). The success of cancer immunotherapy based on CTLA-4 or PD-1 blockade has been attributed to prevention or reversal of intratumoral T cell exhaustion or terminal differentiation (Das et al., 2015; Pauken and Wherry, 2015). Regulatory T (Treg) cells, expressing high amounts of the transcription factor Foxp3, curtail the differentiation and activity of effector T cells and other immune cell types under physiologic conditions and are found in markedly increased numbers in solid organ tumors (Josefowicz et al., 2012; Roychoudhuri et al., 2015; Tanaka and Sakaguchi, 2017). This dedicated lineage of anti-inflammatory suppressive cells is thought to play a prominent role in cancer progression, and Treg cell targeting is considered a potentially promising strategy for tumor immunotherapy.

Likewise, two principal functional states are standardly recognized in tumor-associated macrophages: pro-inflammatory M1 macrophages, which partake in protective anti-bacterial responses and are thought to oppose tumor progression, and tissue reparative M2 macrophages, which promote tumor growth and metastasis (Mantovani and Locati, 2013; Mills et al., 2000; Murray et al., 2014). In addition, a heterogeneous group of monocytic cells and neutrophils, summarily classified as myeloid derived suppressor cells (MDSC), are capable of production of immunosuppressive mediators and have also been suggested to support tumor progression and limit tumor immunity (Kumar et al., 2016). These cell types are associated with poor clinical prognosis, and targeting them may result in therapeutic benefit (Campbell et al., 2011; Engblom et al., 2016; Gholamin et al., 2017; Pyonteck et al., 2013).

Although immunotherapy treatments targeting CTLA-4 and PD-1 have been successful in treating late-stage treatment-resistant melanoma, lung cancer, and kidney cancer (Topalian et al., 2015), meaningful clinical responses have only been observed in a subset of patients and cancer types. The observed variation in treatment efficacy has been connected to heterogeneity in the immune status and immune cell type composition of individual tumors. This is also true in breast cancer, where significant heterogeneity in immune composition is observed across subtypes as well as patients (Dushyanthen et al., 2015; Garcia-Teijido et al., 2016).

These observations raise the question of whether the aforementioned immune cell states differ in normal tissue and in the tumor and whether they represent a limited number of discrete differentiation or activation intermediates with defined gene expression characteristics. Alternatively, these cell states may represent a continuum of cell states occupying a single contiguous phenotypic spectrum, whose features, including "volume" or "size," may be affected by the tumor microenvironment.

Recent work has sought to characterize heterogeneity in cells of the innate and adaptive immune system in lung adenocarcinoma and clear cell renal cell carcinoma using CyTOF mass cytometry (Chevrier et al., 2017; Lavin et al., 2017). These studies, together with bulk RNA sequence analysis of tumor resident immune cells, have provided broad characterization of the composition and properties of main immune cell subsets (Angelova et al., 2015; Li et al., 2016; Rooney et al., 2015; Senbabaoglu et al., 2016). Further studies employing single-cell RNA-seq analysis have begun to explore finer definitions of immune cell subsets in tumors (Singer et al., 2017; Tirosh et al., 2016; Zheng et al., 2017). Yet, the scale of these recent analyses has been insufficient to address the major general questions above.

Thus, we sought to undertake a large-scale, high-dimensional analysis of cells of hematopoietic origin in human breast tumors of various types – as well as paired normal breast tissue, peripheral blood, and a lymph node – using single-cell RNA-seq combined with novel computational approaches. Our analyses revealed remarkably increased heterogeneity of intratumoral cells of both lymphoid and myeloid cell lineages, which occupy markedly expanded contiguous phenotypic space in comparison to normal breast tissue. The observed continuum of cell states likely reflects their progressive cellular activation and differentiation and argues strongly against the notion of few discrete states of differentiation or activation of individual cell types shaping the tumor microenvironment.

RESULTS

InDrop single-cell RNA-seq analysis of immune cells in breast carcinomas

To generate a deep transcriptional map of the immune cell states in human breast cancer, we constructed an atlas of the tumor immune microenvironment comprising 47,016 CD45$^+$ cells collected from 8 primary breast carcinomas from treatment naïve patients including estrogen receptor (ER$^+$) and progesterone receptor (PR$^+$) positive, human epidermal growth factor receptor 2 amplified (Her2$^+$), and triple negative (TNBC) cancers (see STAR Methods). To assess the effect of the tumor microenvironment versus normal breast tissue residence on cell phenotypes, we also analyzed CD45$^+$ cells isolated from matched normal breast tissue, peripheral blood, and lymph node obtained from fresh surgical specimens. Viable, CD45$^+$ cell populations, FACS-sorted from single-cell suspensions prepared from these samples, were subjected to single-cell barcoding and RNA sequencing (scRNAseq) using the inDrop platform (Figure 1A,B, STAR Methods) (Zilionis et al., 2017). We processed the data using our new SEQC scRNA-seq pipeline, which provides increased sensitivity and selectivity in its resulting single-cell profiles (STAR Methods).

For each patient, we first verified that major immune cell types can be accounted for based on transcriptional profiles isolated from individual CD45$^+$ cells. We began by clustering each tumor sample using PhenoGraph (Levine et al., 2015). We annotated each cluster using genome-wide correlations between the mean expression levels of each cluster and previously characterized transcriptional profiles of sorted immune cell subsets (Jeffrey et al., 2006; Novershtern et al., 2011), as well as evaluation of canonical marker expression (see STAR Methods). We were able to identify the majority of immune cell types expected to be present in human tumors, including monocytes, macrophages, dendritic cells, T cells, B cells, mast cells, and neutrophils (Figure 1C, S1C). Thus, we were able to capture a comprehensive representation of the immune ecosystem from each individual tumor.

Variation between individual tumor immune microenvironments

In agreement with the aforementioned mass cytometry analyses (Chevrier et al., 2017; Lavin et al., 2017), we found a large degree of variation in the immune cell composition of each tumor (Figure 1D). For example, the fraction of T cells varied between 21%-96% and the fraction of myeloid cells varied between 4-55%. To determine the reliability of inDrop in representative sampling of heterogeneous cell populations, we compared the proportions of cell types as measured by flow cytometry and inDrop scRNA-seq. Although a comparison of the relative representation of major immune cell types identified by scRNA-seq to those assessed by flow cytometric analysis revealed a significant bias towards monocytic lineage cell subsets relative to expected input ratios, we observed high correlation between cell type frequencies across all patient samples ($r^2 > 0.8$, Figure S1D). The observed bias, likely due to the larger cytoplasmic volume and higher RNA yield of monocytic/myeloid cells vs. T cells, was systemic and did not adversely affect our analyses.

This genome-wide view allowed us to assess system-level differences between immune cell consortia in individual patients in, for example, metabolic signatures, including hypoxia (Figure 1E). It is interesting to note that while all tumors expressed a similar average degree of a hypoxia signature, patients differed considerably in expression at the level of individual genes included in the signature. Similar variation was observed in fatty acid metabolism, glycolysis, and phosphorylation (Figure S1E-G).

Integration of data across multiple tumors

To enable an unbiased systematic comparison across patients, we merged the data from all tumors to create a map of tumor-infiltrating immune cells. However, we observed that cells from the same patient were often more similar than cells of the same lineage from another patient (Figure 2A). This was likely due both to batch effects and standard normalization procedures that conflate biological signal and technical differences, both within and between the samples. We also observed an increase in the number of molecules captured from activated immune cells, likely due to an increase in total RNA abundance upon activation (Blackinton and Keene, 2016; Cheadle et al., 2005; Marrack et al., 2000; Singer et al., 2017). In addition, our analyses showed a gradient of activation of CD8 T cells in tumors, and the most pronounced T cell activation in a TNBC tumor (BC3), in agreement with previous reports (Figure 2B) (Dushyanthen et al., 2015; Garcia-Teijido et al., 2016). To remove confounding technical effects while retaining convolved biological variation, we clustered the combined data from all samples using Biscuit, a hierarchical Bayesian model that infers clusters while simultaneously normalizing and imputing dropouts, allowing us to fit and correct for cell- and batch-intrinsic variation (Prabhakaran et al., 2016) (see STAR Methods). Cell type-specific normalization is especially crucial in cases involving vast subtype diversity, such as immune cells ranging from large macrophages to much smaller lymphocytes (Lun et al., 2016; Vallejos et al., 2017). Using Biscuit, we normalized together cells that are assigned to the same type, and extracted interpretable parameters associated with each cluster for characterization of cell subsets (STAR Methods).

After applying Biscuit to the data from all tumors (Figure 2C), we found 67 clusters covering various T cell, macrophage, monocyte, B cell, and NK cell clusters. We first asked whether individual cells tended to be most similar to cells from their own samples or if the resulting cell profiles were well mixed using an entropy measure (STAR Methods). For each cell, this measure considers the neighborhood of its most similar cells and evaluates the entropy of the sample distribution in each such neighborhood. Low entropy indicates that most neighbors come from the same sample, whereas high entropy indicates that the neighbors (most similar cells) are well distributed across the different samples. Indeed, while cells were most similar within individual samples before normalization, this was corrected after Biscuit normalization with significantly improved mixing of cells across patients when compared against standard normalization methods (Figure 2A, D) (U=1.7721e+09, p=0, Figure 2D). Using this approach, we successfully retained information on immune cell activation while stabilizing differences in library size, and uncovered a rich and robust structure in imputed data, suggesting diversity in immune cell subtypes (Figure 2A, C).

Breast tumor immune cell atlas reveals substantial diversity of cell states

To construct a global atlas of immune cells, enabling characterization of the impact of environment on immune cell states, we merged data from 47,016 cells across all tissues and patients revealing a diverse set of 83 clusters, each identifying a cell type or state (Figure 2E, F; S2A, B). This unexpectedly large number of clusters prompted us to test their robustness using cross-validation on subsets of the data (STAR Methods), finding assignments of cells to clusters were robust for most clusters (Figure S2C). Most clusters were shared across multiple patients, indicating similar immune states across patients, with only 10 being patient-specific (Figure 2G). We used entropy as a more stringent metric for patient mixing within clusters and found that the clusters span a range of different mixing levels (Figure S2D).

We assigned each cluster to its associated cell type by comparing cluster mean expression to bulk RNA-seq (Figure 2E, F) and found 38 T cell clusters, 27 myeloid lineage clusters, 9 B cell clusters, and 9 NK cell clusters (Table S2). By examining the expression of canonical markers in immune cell clusters, we were able to confirm and build upon predictions made by the preceding analysis (Figure 2H). Of the T cell clusters, we identified 15 CD8+ T cell clusters and 21 CD4+ T cell clusters, which were together split into 9 naive, 7 central memory, 15 effector memory, and 5 Treg clusters. We were additionally able to divide the myeloid lineage clusters into 3 macrophage, 3 mast cell, 4 neutrophil, 3 dendritic cell, 1 plasmacytoid dendritic cell, and 13 monocytic clusters. Finally, we identified 9 B cell clusters, 3 CD56^{-} NK cell clusters, and 6 CD56^{+} NK cell clusters, 2 of which are likely NKT cells (Figure 2I, J).

Since our characterization identified multiple clusters with the same cell type “label” based on surface markers and prior characterization of the corresponding peripheral blood cell phenotypes, e.g. 15 effector memory T-cell clusters (Figure 2F), we wanted to confirm that all these clusters were indeed distinct. Biscuit identifies clusters based on differences in both mean expression and covariance (i.e. co-expression) patterns for groups of genes in the same cells. The distributions defined by the Biscuit parameters identified differentially expressed genes between clusters, including canonical immune genes, and defined multiple subpopulations within each major cell type (Table S3). Final cluster annotations based on the above analysis are appended to Table S2. Moreover, we observed a prominent effect of covariance in defining the T cell clusters by comparing similarity of pairs of clusters with and without the effect of mean expression (Figure S2E, F); large differences between most clusters remained even after mean gene expression was equalized. Thus, our approach robustly identified cell states that were distinct from one another and shared across multiple tumor microenvironments. As T cell and myeloid cells represent the most abundant and diverse, and arguably most biologically significant, immune cell subsets in the tumor microenvironment, we focused our subsequent in-depth analyses on these two major cell types.

Tissue environment impacts the diversity of immune phenotypic states

A key goal of this study was to quantify the extent to which variation in immune cell phenotypes is driven by their tissue of residence, i.e. cancerous vs. normal breast tissue, using peripheral blood or the lymph node cells as references. To gain a qualitative understanding of phenotypic overlap between tissues, we carried out tSNE co-embedding (van der Maaten and Hinton, 2008) of the merged dataset annotated by clusters. This analysis showed that T cells in blood and lymph node had dramatically dissimilar phenotypes compared to cancerous or normal breast tissue resident T cells, which exhibited considerable similarity (Figure 3A, B). We observed that gene expression of T cells dramatically differed between blood and tissue resident cells, resulting in a large cluster specific to blood-derived cells being phenotypically distinct from T cells in normal and tumor tissue (shown in blue). Both T cells and myeloid lineage cells exhibited considerable overlap between tumor and normal tissue samples, with increased phenotypic heterogeneity creating an expansion of these populations observed in the tumor (Figure 3B). Figure 3C summarizes distributions of cell types across tissues.

Quantifying differences between tissues, we confirmed that naive T cells were strongly enriched in three blood-specific clusters (χ^2=361.4, df=1, p=3e-80), while B cells were more prevalent in the lymph node than in other tissues (χ^2=1737.1, df=1, p=0.0). A subset of T cell clusters was present in both tumor and normal tissue, but the cytotoxic T cell clusters were more abundant in tumor (χ^2=93.7, df=1, p=3e-25); Treg cells were also highly enriched in the tumor as expected (χ^2=336.0, df=1, p=5e-91). Moreover, some myeloid clusters were shared between normal and tumor tissue, whereas clusters of more activated monocytes and tumor-associated macrophages (TAMs) were specific to tumor (χ^2=2420.6, df=1, p=0.0). Overall, we observed much more pronounced overlap in phenotypes between tissue-resident immune cells than between any other pair of tissues (z=2.68, p=1.4e-4), highlighting that tissue of residence is a significant determinant of phenotypes of human cells of hematopoietic origin, and that states or biomarkers identified from blood immune cells may not necessarily extend to tissue embedded immune populations.

Tumor microenvironment drives an expansion of immune cell phenotypes

We observed a large number of normal breast tissue resident cell states, manifested by 13 myeloid and 19 T cell clusters that were not observed in circulation or in the secondary lymphoid tissue, i.e. lymph node. Furthermore, our data showed that the set of clusters found in normal breast tissue cells represented a subset of those observed in the tumors; 14 myeloid and 17 T cell clusters were only found in the tumor, doubling the number of observed clusters of these cell types relative to normal tissue, and there were no clusters specific to normal tissue.

This increased diversity of cell states was driven by a significant increase in the variance of gene expression in tumor compared to normal tissue (Figure 3D). We performed Gene Set Enrichment Analysis (GSEA) (Subramanian et al., 2005) on the genes with the largest increase in variance and found enrichment in targets of signaling pathways activated in the tumor environment, including Type I (IFNα) and II interferons (IFNγ), TNFα, TGFβ, IL6/JAK/STAT signaling, and the aforementioned hypoxia (Figure 3E; S3A, B; Table S5).

To understand whether the increase in variance of gene expressions in tumor tissue is due to activation of additional phenotypes that are independent from those found in normal tissue, we sought to define a metric for the “phenotypic volume” occupied by cells. Specifically, this metric uses the covariance in gene expression to measure the volume spanned by independent phenotypes, which grows with the number of active phenotypes and the degree of independence between them (see STAR methods). Moreover, our metric controls for the difference in the total number of immune cells observed in tumor versus normal tissue. Using this metric, we compared the phenotypic volume occupied by each cell type between normal and tumor tissue. Assessment of the change in volume showed a significant increase in the phenotypic volume of all major cell types, including T cells (U-test p = 0), myeloid cells (p = 0), and NK cells (p = 0) in the tumor compared to normal mammary gland tissue (Figure 3F). Precisely, the fold change in volume was 7.39e4 in T cells, 1.18e14 in myeloid cells and 6.08e4 in NK cells, indicating a massive increase in phenotypic volume in tumor compared to normal tissue. These data suggest that increased heterogeneity of cell states and marked phenotypic expansions found within the tumor in comparison to the normal tissue were likely due to more diverse local microenvironments within the tumor, which differ in the extent of inflammation, hypoxia, expression of ligands for activating and inhibitory receptors, and nutrient supply (Finger and Giaccia, 2010; Jimenez-Sanchez et al., 2017).

Intratumoral T cells reside on continuous components of variation

To further explore the most significant sources of the observed phenotypic variation, we carried out unbiased analyses by separately decomposing the gene expression of the T cell and myeloid cell lineages using diffusion maps (Coifman et al., 2005; Haghverdi et al., 2015; Haghverdi et al., 2016; Moignard et al., 2015; Setty et al., 2016). While some components distinguished discrete clusters, the majority of components defined gradual trends of variation across T cell clusters (Figure 4A, S4A). The top 3 informative components correlate, respectively, with signatures for activation, terminal differentiation, and hypoxia (Figure 4B-D; Figure S4B-F; STAR Methods).

The most informative component of variation, labeled activation trajectory, is highly correlated with gene signatures of T cell activation and progressive differentiation (p=0.0), along with IFNγ signaling (p=0.0) (see STAR Method). The mean expression of the activation signature steadily increases along the component (Figure 4B), with a concomitant gradual increase in expression of activation-related genes (Figure 4C). Intratumoral T cell populations are enriched at the positive end of the component relative to T cells found in healthy tissue (t-test p=0.0, Figure 4A, D). Specifically, tumor-resident effector memory T cells and Treg cells compose the most activated end, while mostly naïve T cells in peripheral blood congregate at the least activated terminus, consistent with their quiescent state (t-test p=0.0, Figure 4D). However, while the mean expression levels of clusters vary gradually along the component, there is also a wide range of activation states within each cluster (Figure 4D). Examining the individual genes most correlated with the component reveals a diverse set of genes whose expression is well documented to increase upon T cell activation and progressive differentiation. These include genes encoding cytolytic effector molecules granzymes A and K (GZMA and GZMK), pro-inflammatory cytokines (IL-32), cytokine receptor subunits (IL2RB), chemokines (CCL4, CCL5), and their receptors (CXCR4, CCR5) (Figure 4C).

The next most informative component of variation was labeled terminal differentiation (Figure 4E); the genes most correlated with it include co-stimulatory molecules (CD2, GITR, OX40, and 4-1BB) as well as co-inhibitory receptors (CTLA-4 and TIGIT) (Figure S4B). This set also included Foxp3, IL2RA, and Entpd1 (CD39), genes whose high expression is characteristic of Treg cells (Josefowicz et al., 2012). The same primarily Treg clusters reside at the very terminal end of both the activation and terminal differentiation components, and there is a moderate degree of overlap in the genes most correlated with the two (Figure 4A, C; S4B). However, there are also important exceptions—including the markers of exhaustion listed above—and crucially, the two trajectories traverse different paths through the rest of the clusters (Figure 4F). Indeed, some clusters—notably T cells from the lymph node (e.g. cluster 16)—express higher levels of activation than terminal differentiation (t-test p=0.0; Figure S4B-D), consistent with the idea that T cell exhaustion and terminal differentiation largely occurs in non-lymphoid tissues and not in the draining lymph node.

Interestingly, visualizing the T cell activation and terminal differentiation components together revealed remarkable continuity, in essence representing a single continuous trajectory of T cells towards a terminal state (Figure 4A, S4D). Thus, T cells reside along a broad continuum of activation, suggesting that their conventional classification into relatively few discrete activation or differentiation subtypes may grossly oversimplify the phenotypic complexity of T cell populations resident in tissues.

Response to diverse environmental stimuli and covariance of gene expression define intratumoral T cell states

Noting that only a few of the clusters were well delineated by the strongest components of variation, we sought to understand the variation driving the observed clustering. We examined the expression of gene signatures for response to environmental stimuli in each T cell cluster and found that while most clusters were arranged in a continuous fashion along the activation component, each cluster appeared unique when looking across multiple components and signatures in a combinatorial fashion. Our data show that CD4 effector and central memory clusters (Figure 5A) exhibit variable levels of expression of genes contributing to signatures for Type I and II interferon response (F-test, p=1e-54 and 0.008 respectively), hypoxia (F-test, p=4e-64), and anergy (F-test p=4e-69). Moreover, different CD8 effector and central memory clusters (Figure 5B) had different expression levels of activation (F-test p=2e-114), proinflammatory (F-test p=1e-39), and cytolytic effector pathway-related genes (F-test p=6e-32). These examples suggest that in a heterogeneous tumor microenvironment, differing in degree of inflammation, hypoxia, and nutrient availability, subpopulations of T cells either sense different environmental stimuli or respond differently to these stimuli. While many of these responses (e.g. activation or hypoxia) create phenotypic continuums, their different combinations can result in more discrete behaviors.

In contrast to effector T cells, Treg clusters displayed less variation in expression across these gene signatures: the majority of these clusters featured comparable patterns for anti-inflammatory activity, exhaustion, hypoxia, and metabolism gene sets (Figure 5C).

To identify features distinguishing the Treg clusters, we examined the Biscuit parameters differentiating them. We found that beyond mean expression levels, covariance parameters varied significantly between clusters, driving the observed differences. Specifically, two marker genes could exhibit similar mean expression in two different clusters (e.g. highly expressed in both), while the clusters show opposite signs in covariance between these genes. This occurs due to the genes typically being co-expressed in the same cells in one cluster (i.e. positive covariance), while being expressed in the other cluster in a mutually exclusive manner (i.e. negative covariance) (Fig. 5D). It is noteworthy that clusters were inferred based on the expression of over 14,000 genes; hence, negative covariance between two specific genes does not necessarily imply the existence of sub-clusters.

As an example, our analysis showed that the CTLA-4 gene, which encodes a prototypical inhibitory checkpoint receptor that is highly expressed in Tregs and activated T cells, exhibited rich covariance patterns with other mechanistically related genes (Figure 5E-G; S5A, B). CTLA-4 co-varied strongly with TIGIT and co-stimulatory receptor GITR in Treg clusters 46, 56, and 87; with CD27 in clusters 46 and 80; and with co-stimulatory receptor ICOS only in cluster 80 (Figure 5F,G). We observed considerable differences in covariance patterns between numerous pairs of other checkpoint genes across Treg clusters (Figure 5G). Additionally, covariance between other key immune genes in Treg clusters exhibited modular structures, with groups of genes co-expressed together, suggesting co-regulation and potential involvement in similar functional modalities (Figure 5H). Since varied proportions of Treg clusters were observed in individual patient samples, the differences in gene co-expression were present across patients as well as clusters within a given patient (Figure 5I). We observed that the majority of patients did not have all 5 subtypes of Treg cells, and in fact most were dominated by only one subtype (cluster). It must be noted that we also observed similar differences in co-variation patterns across activated T cell clusters, even if these did not play as essential a role in their delineation (Figure S5C). Thus, co-variation of genes has a role in defining T cell clusters, in particular Treg clusters (Figure 5G,H, SF2).

Activation and differentiation define components of variation of intratumoral myeloid cells

Although myeloid lineage cells are commonly thought to be highly diverse and able to markedly influence the state of the tumor microenvironment and, thereby, impact clinical outcomes, the heterogeneity of intratumoral monocytes and macrophages remains insufficiently characterized (Campbell et al., 2011; De Henau et al., 2016; Engblom et al., 2016; Eppert et al., 2011; Gholamin et al., 2017; Pyonteck et al., 2013). A broad survey of the major monocytic subsets suggests the existence of both gradual and abrupt phenotypic shifts (Figure 6A). As with the T cells above, we employed diffusion maps to assess heterogeneity in and across these monocytic populations, excluding neutrophils and mast cells, which formed much more distinct clusters (Figure 6B). This analysis revealed four major branches that displayed clearer segregation of cell states, and moderately less continuity, than the analogous T cell maps (Figure S6A).

The first branch almost entirely comprises intratumoral macrophages from three clusters (23, 25, and 28) (Figure 6B-F). Among the top genes correlated with the branch are macrophage activation-associated genes APOE, CD68, TREM2, and CHIT1 (Figure S6B); the branch thus likely reflects progression towards a distinct state resulting from the differentiation and activation of either recruited or tissue-resident macrophages in the tumor microenvironment. Additionally, expression of genes typically implicated in a polarization model of tissue-reparative and immunosuppressive M2 macrophage activation, including scavenger receptor MARCO, extracellular matrix component FN1, pro-angiogenic receptor NRP2, SPP1 (osteopontin), and inhibitory molecule B7-H3 (CD276), increased along this branch (Figure S6B). Concomitantly, pro-inflammatory and immunostimulatory genes, including chemokine CCL3 (MIP-1a), typically associated with M1 macrophages, likewise increased along the branch. Quite strikingly, we found that M1 and M2 gene signatures in fact positively correlated in the myeloid populations (Figure 6G). These findings support the idea that macrophage activation is markedly impacted by the tumor microenvironment in a manner that does not comport with the polarization model, either as discrete states or along a spectrum of alternative polarization trajectories.

The second and third branches together captured a more gradual trajectory from blood monocytes (mainly cluster 42, 97.5% present in blood) to intratumoral monocytes (clusters 67, 91, 68 and 94) (Figure 6B-D; S6B-D). The “blood terminus” of the trajectory correlated with expression of co-stimulatory gene ITGAL, but also with several tumor growth-promoting genes, i.e. fibroblast and epidermal growth factors, as well as IL-4 (Figure S6B). The latter has been proposed to support the M2 type of macrophage activation (Mantovani and Locati, 2013; Mills et al., 2000; Murray et al., 2014). The other end of the trajectory, populated by intratumoral monocytes, was characterized by high expression of activation and antigen presentation-related genes encoding CD74 and HLA-DRA, but also an IFN-inducible gene encoding ISG15, which has been described to be secreted by TAMs and enhance stem-like phenotypes in pancreatic tumor cells (Figure S6B) (Sainz et al., 2014).

The fourth branch correlated with canonical plasmacytoid dendritic cell (pDC) markers such as LILRA4, CLEC4C (CD303), and IL3RA. The most discrete of the myeloid components, this branch separated the lone pDC cluster (41) from the other myeloid-monocytic cell clusters (Figure 6B, E, G, H; S6C). This subset was also the only monocytic cluster common between the tumor and the lymph node; it featured high levels of granzyme B (GZMB) (Figure S6B), which has been proposed to be a means, by which pDCs may suppress T cell proliferation in cancer (Jahrsdorfer et al., 2010; Swiecki and Colonna, 2015).

Covariance patterns help distinguish TAM subpopulations

While the TAM clusters projected to a distinct region in the diffusion component, separating them from other monocytic cells, they appeared very similar to one another (Figure 6F; S6A). This similarity was supported at the genomic scale by shared patterns of DEGs (Table S3) and short pairwise distances (Figure S2B). However, similarly to the intratumoral Treg cells, co-variation patterns defined distinctions between intratumoral myeloid cell subsets. Specifically, co-variation of canonical genes for M1 or M2 macrophages distinguished the TAM clusters. All three of the TAM populations, particularly clusters 23 and 28, were among the monocytic lineage clusters that exhibited the most similarity to the canonical M2 signature (Figure 6G). However, both of these clusters also expressed high levels of the M1 signature genes, and significant expression of the two signatures was often coincident (Figure 6G, H).

We observed pronounced inter-cluster differences in co-expression patterns in TAM clusters. One example among many was co-expression of two M2-type markers, MARCO and B7-H3. In an unexpected manner, while TAM clusters 23, 25, and 28 all expressed high levels of both genes, they co-varied positively in clusters 23 and 25, but negatively in 28 (p = 0, p = 5e-06, p = 0, respectively; Figure 7A-C; S7A,B). The differing covariance patterns were not an artifact of modeling as they were also significant in raw un-normalized data (Figure S7A, STAR Methods).

The degree of co-expression of genes associated with M1 and M2 signatures also varied widely within clusters in a manner not fitting the functional M1/M2 annotation. For example, expression of CD64 in cluster 23 exhibited varying degrees of positive co-variance with FN1, MMP14, MSR1, cathepsins, MARCO, and VEGFB, but co-varied slightly negatively with chemokine CCL18 (Figure 7C). Taken together, these findings demonstrate that co-variation patterns define TAM clusters, and further highlight the lack of mutual exclusivity between the proposed prototypical M1 and M2 states.

DISCUSSION

Despite major clinical advances in cancer immunotherapy, our ability to understand its mechanisms of action or predict its efficacy is confounded by the complex, heterogeneous, and poorly understood composition of immune cells within tumors. Since cancer is generally a disease that affects older, post-reproductive individuals, with the exception of inherited genetic predispositions it is unlikely that specialized mechanisms of adaptive or innate immunity evolved to facilitate tumor immune surveillance. It seems reasonable to suggest that immune mechanisms affecting tumor progression must also operate in non-cancerous tissues to maintain organismal integrity and tissue function in the face of infection, stress, inflammation, and injury. A corollary to this notion is that features of immune cells in tumors must, by and large, resemble features of cells in non-cancerous tissues. A recent population-level RNA-seq analysis of Treg cells and effector CD4 T cells in breast cancer and normal breast tissue identified a high level of phenotypic similarity between tissue and tumor-resident T cells, thus providing experimental support for this idea (Plitas et al., 2016). A similar RNA-seq study focusing primarily on Treg cell analysis in colorectal and lung cancer suggested that cancer-resident Treg cells differ considerably from those found in the normal tissue (De Simone et al., 2016). Despite seeming differences in conclusions, distinguishing features of intratumoral Treg cells as compared to normal tissue-resident ones detected in these two reports were associated with their heightened activation and thus can not distinguish between differences in immune states themselves or differences in immune state proportions. Thus, the averaging of gene expression features in bulk cell population analyses and the lack of the assessment of a broad spectrum of immune cell subsets do not allow for a definitive investigation of specific effects of the tumor environment on immune cells.

To address this question, we undertook an unbiased comparative single-cell RNA-seq analysis of all tumor versus normal tissue-resident immune cell subsets and constructed an immune atlas in breast carcinomas, combining immune cells isolated from normal and cancerous breast tissue, as well as peripheral blood and the lymph node. Our analysis was empowered by a suite of novel computational tools for single-cell RNA-seq data, including a data processing pipeline more sensitive in its ability to detect immune molecules, a powerful clustering and normalization algorithm, and new metrics for volume of phenotypic space. In particular, our computational method Biscuit infers subpopulations of cells according to similarity in gene expression as well as gene co-expression patterns, and iteratively normalizes cells based on their assignment to subpopulations. This method allowed us to overcome strong batch effects typical of clinical samples that would have otherwise dominated the signal and obscured the identification of shared cell states across tumors. Moreover, Biscuit's parametric model enabled us to characterize covariance patterns and more accurately capture complex relationships between pairs of genes of interest.

The atlas revealed vast diversity in the repertoire of immune cells representative of both the adaptive and innate immune systems. Our examination of hematopoietic nucleated cells from treatment-naïve human breast cancer and normal breast tissue across different patients revealed that the biggest change to the immune cells was linked to the tissue environment, resulting in cell states that are substantially different than those present in the blood and lymph node. Interestingly, immune cell subpopulations in normal tissue were observed to be a subset of those found in tumor tissue, an observation that could not have been found with bulk gene expression measurements. Furthermore, the diversity of cell states significantly expanded between normal tissue and tumor, as quantified with our metric of "phenotypic volume" occupied by immune cell states. We observed tremendous expansion of the immune phenotypic space occupied by all major cell types in breast tumors as compared to normal breast tissue. It seems reasonable to speculate that the majority, if not all, immune cell states found in cancer can be found in corresponding non-cancerous tissues in response to different stresses such as infection, wound healing, or inflammation.

The observation of an expanding T cell "phenotypic space" in the tumor argues against the view of activated T cells rapidly traversing through sparse transitional cell states towards a few predominant, discrete, and stable states, including Treg, effector, memory, and exhausted T cells. Three major components contributed to this phenotypic expansion in tumor tissue that helped explain the heterogeneity of T cells, including T cell activation, terminal differentiation, and hypoxic response. The strongest of these components was a predominant trajectory of progressive T cell activation and differentiation across 38 T cell clusters, including Treg and terminally differentiated T cell clusters, found at the extreme activation terminus. One obvious explanation for the

”continuity” of intratumoral T cell activation is the presence of increasingly diverse environments defined by a multitude of gradients including growth, pro-inflammatory, and tissue repair factors, as well as oxygen, nutrient, and metabolite gradients which exist to a lesser extent in healthy breast tissue (Buck et al., 2017). Indeed, we found groups of genes within the corresponding signaling pathways, most prominently immune activation (IFN/IL6/JAK/STAT) and hypoxia, to be differentially expressed across T cell clusters.

A non-mutually exclusive possibility is that the wide range of TCR signal strengths afforded by a diverse repertoire of T cell receptors (TCR) accounts for the continuous spectrum of T cell activation, obscuring the transitional states. The latter may also be accounted for by asynchrony in polyclonal T cell activation or heterogeneity in the types of antigen-presenting cells, their activation status, and their anatomical distribution. Unlike polyclonal T cell populations, activation of a monoclonal T cell population with a “fixed” specificity for tumor “self” or neo-antigen may yield sparse discontinuous “phenotypic” spaces reflecting discrete functional T cell states. In support of the latter possibility, recent bulk gene expression and chromatin accessibility analyses showed that cognate tumor neo-antigen recognition by TCR transgenic T cells results in an orderly progression of activated T cells through a reversible dysfunctional intermediate state towards an irreversible dysfunctional terminal state (Philip et al., 2017). Additionally, diverse TCR specificities can contribute to the spatial distribution of T cells and, therefore, facilitate their exposure to the distinct environments (“mini-niches”) discussed above.

While T cells of various cell types exhibit continuous levels of activation, our inferred subsets further show variable levels of responses to environmental stimuli, and the combinations of these environmental exposures jointly define the identity of discrete CD4+/CD8+ T cell subsets. We also identified 5 Treg subsets that showed similar responses to environmental pressures and shared differentially expressed genes, but exhibited drastic differences in gene covariance patterns. Particularly noteworthy was co-expression of checkpoint receptor genes in some Treg subpopulations as compared to mutually exclusive expression of the same genes in other Treg clusters. In this regard, co-variant expression of CTLA-4, TIGIT, and co-stimulatory receptor GITR and other co-receptors in multiple Treg cell clusters suggests that these Treg cell populations may occupy different functional niches. Cells co-expressing CTLA-4 and TIGIT have been demonstrated to selectively inhibit pro-inflammatory Th1 and Th17 responses but not Th2 responses, promoting tissue remodeling (Joller et al., 2014). The observed co-expression of functional cell surface and signaling molecules by intratumoral Treg cells may enable targeted modulation of Treg cell activity in the tumor microenvironment using combinatorial therapeutic approaches (Mantovani and Locati, 2013). We also observed considerably different proportions of Treg clusters across patients, suggesting that multi-dimensional profiling might be necessary to personalize such therapies. These findings have implications in the way that tumor immune responses are described and interrogated. It must be noted that the discrete cell states that are commonly utilized to describe immune responses are largely defined from highly polarizing conditions such as acute and chronic infections.

Our analyses appear to offer a more nuanced view of tumor and normal tissue-resident myeloid lineage cells, in comparison to T cells, in terms of continuity vs. separation of cell states. Unlike T cells, which primarily displayed continuous activation transitions, we observed sharper state delineations in myeloid populations. This difference between T cells and myeloid cells was likely due to a markedly less appreciated developmentally established myeloid cell heterogeneity, whose understanding has started to emerge only recently (Perdiguero and Geissmann, 2016). Indeed, the phenotypic expansion in myeloid cells was associated with activation of macrophages and monocytes and emergence of pDC subsets distinct from mDCs (myeloid DCs, also known as conventional DCs or cDCs). However, our analyses also showed common features to those in T cells, including gene expression covariance identifying cell clusters, and an expansion of immune phenotypic space in breast tumor as compared to normal breast tissue.

Similarly to T cells, we have not observed discrete states of myeloid cell activation/differentiation such as M1 or M2 macrophages or myeloid derived suppressor cells. In contrast, we found both M1 and M2 associated genes frequently expressed in the same cells, positively correlated with one another and following the same activation trajectory. Furthermore, we found that covariance patterns between gene markers associated with the M1 and M2 model show rich diversity, and help distinguish the three TAM clusters. These results challenge the prevailing model of macrophage activation, wherein M1 and M2 activation states either exist as mutually exclusive discrete states or macrophages reside along a spectrum between the two states with negatively correlated expression of M1 and M2-associated genes. Our findings solidify and reinforce previous reports from the bulk analysis of tumor-associated macrophages in mouse models of oncogene-driven breast cancer and analysis of myeloid cells in lung and kidney cancer using mass cytometry (Chevrier et al., 2017; Franklin et al., 2014; Lavin et al., 2017). Notably, we observed more patient-specific variation in myeloid lineage cells than in T cells, with the frequency of the former ranging from just over 10% to over 50% in individual patients. Individual clusters similarly exhibited ranges of patient specificity. The large patient effect in myeloid cells suggests that attempts at generalized targeting or reprogramming of suppressive myeloid cell populations are not likely to yield uniform responses and personalization at the patient level may be needed.

In conclusion, these findings show that studying average gene expression across groups of cells fails to characterize heterogeneity in co-expression of genes, and by extension their potential suitability as therapeutic co-targets. Single-cell RNA-sequencing analyzed using Biscuit, as shown here, allows for inference of accurate and meaningful covariance parameters; indeed, the algorithm takes into consideration these covariance values when defining clusters. This makes it possible to query in a precise manner how numerous functionally and therapeutically important immune markers are co-expressed at the level that matters: that of individual cells. Our characterization of the immune cell subsets inhabiting primary solid tumor and the corresponding normal tissue, and their heterogeneity within a given patient and between different patients, revealed expansions of a continuous "phenotypic space" as a principal feature of the two main cellular targets of cancer immunotherapy - T cells and myeloid cells. These observations, along with the resulting extensive immune single-cell RNA-seq datasets and the comprehensive analytical platform, will facilitate better understanding of potential mechanisms behind immune cell contributions to promoting and opposing tumor progression.

Acknowledgements

We thank Barbara Engelhardt, David Blei, Frederic Geissmann, and Roshan Sharma for valuable conversations related to this manuscript. We thank the Flow Cytometry Core, Tissue Procurement Service, and Integrated Genomics Operation at MSKCC for assistance. We thank the specimen donors at MSKCC. This study was supported by NIH grants R37 AI034206 (A.Y.R), NIH DP1- HD084071 (D.P), NIH R01CA164729 (D.P). Cancer Center Support Grant P30 CA008748, the Hilton-Ludwig Cancer Prevention Initiative funded by the Conrad N. Hilton Foundation, Ludwig Cancer Research, and Gerry Center for Metastasis and Tumor Ecosystems. G.P. is the recipient of a Breast Cancer Alliance Young Investigator Grant and the Society of MSK Research Grant. C.K. was a recipient of a Hutham S. Olayan Graduate Fellowship. A.J.C is a recipient of an HHMI international graduate fellowship. A.Y.R. is an investigator with the Howard Hughes Medical Institute.

References

Angelova, M., Charoentong, P., Hackl, H., Fischer, M.L., Snajder, R., Krogsdam, A.M., Waldner, M.J., Bindea, G., Mlecnik, B., Galon, J.*, et al.* (2015). Characterization of the immunophenotypes and antigenomes of colorectal cancers reveals distinct tumor escape mechanisms and novel targets for immunotherapy. Genome biology *16*, 64.

Arpaia, N., Green, J.A., Moltedo, B., Arvey, A., Hemmers, S., Yuan, S., Treuting, P.M., and Rudensky, A.Y. (2015). A Distinct Function of Regulatory T Cells in Tissue Protection. Cell *162*, 1078-1089.

Blackinton, J.G., and Keene, J.D. (2016). Functional coordination and HuR-mediated regulation of mRNA stability during T cell activation. Nucleic Acids Res *44*, 426-436.

Buck, M.D., Sowell, R.T., Kaech, S.M., and Pearce, E.L. (2017). Metabolic Instruction of Immunity. Cell *169*, 570-586.

Campbell, M.J., Tonlaar, N.Y., Garwood, E.R., Huo, D., Moore, D.H., Khramtsov, A.I., Au, A., Baehner, F., Chen, Y., Malaka, D.O.*, et al.* (2011). Proliferating macrophages associated with high grade, hormone receptor negative breast cancer and poor clinical outcome. Breast Cancer Res Treat *128*, 703-711.

Cheadle, C., Fan, J., Cho-Chung, Y.S., Werner, T., Ray, J., Do, L., Gorospe, M., and Becker, K.G. (2005). Control of gene expression during T cell activation: alternate regulation of mRNA transcription and mRNA stability. BMC Genomics *6*, 75.

Chevrier, S., Levine, J.H., Zanotelli, V.R.T., Silina, K., Schulz, D., Bacac, M., Ries, C.H., Ailles, L., Jewett, M.A.S., Moch, H.*, et al.* (2017). An Immune Atlas of Clear Cell Renal Cell Carcinoma. Cell *169*, 736-749 e718.

Coifman, R.R., Lafon, S., Lee, A.B., Maggioni, M., Nadler, B., Warner, F., and Zucker, S.W. (2005). Geometric diffusions as a tool for harmonic analysis and structure definition of data: diffusion maps. Proc Natl Acad Sci U S A *102*, 7426-7431.

Das, R., Verma, R., Sznol, M., Boddupalli, C.S., Gettinger, S.N., Kluger, H., Callahan, M., Wolchok, J.D., Halaban, R., Dhodapkar, M.V.*, et al.* (2015). Combination therapy with anti-CTLA-4 and anti-PD-1 leads to distinct immunologic changes in vivo. Journal of immunology *194*, 950-959.

De Henau, O., Rausch, M., Winkler, D., Campesato, L.F., Liu, C., Cymerman, D.H., Budhu, S., Ghosh, A., Pink, M., Tchaicha, J.*, et al.* (2016). Overcoming resistance to checkpoint blockade therapy by targeting PI3Kgamma in myeloid cells. Nature *539*, 443-447.

De Simone, M., Arrigoni, A., Rossetti, G., Gruarin, P., Ranzani, V., Politano, C., Bonnal, R.J., Provasi, E., Sarnicola, M.L., Panzeri, I.*, et al.* (2016). Transcriptional Landscape of Human Tissue Lymphocytes Unveils Uniqueness of Tumor-Infiltrating T Regulatory Cells. Immunity *45*, 1135-1147.

Dunn, G.P., Old, L.J., and Schreiber, R.D. (2004). The immunobiology of cancer immunosurveillance and immunoediting. Immunity *21*, 137-148.

Dushyanthen, S., Beavis, P.A., Savas, P., Teo, Z.L., Zhou, C., Mansour, M., Darcy, P.K., and Loi, S. (2015). Relevance of tumor-infiltrating lymphocytes in breast cancer. BMC Med *13*, 202.

Engblom, C., Pfirschke, C., and Pittet, M.J. (2016). The role of myeloid cells in cancer therapies. Nature reviews Cancer *16*, 447-462.

Eppert, K., Takenaka, K., Lechman, E.R., Waldron, L., Nilsson, B., van Galen, P., Metzeler, K.H., Poeppl, A., Ling, V., Beyene, J.*, et al.* (2011). Stem cell gene expression programs influence clinical outcome in human leukemia. Nature medicine *17*, 1086-1093.

Finger, E.C., and Giaccia, A.J. (2010). Hypoxia, inflammation, and the tumor microenvironment in metastatic disease. Cancer metastasis reviews *29*, 285-293.

Franklin, R.A., Liao, W., Sarkar, A., Kim, M.V., Bivona, M.R., Liu, K., Pamer, E.G., and Li, M.O. (2014). The cellular and molecular origin of tumor-associated macrophages. Science *344*, 921-925.

Garcia-Teijido, P., Cabal, M.L., Fernandez, I.P., and Perez, Y.F. (2016). Tumor-Infiltrating Lymphocytes in Triple Negative Breast Cancer: The Future of Immune Targeting. Clin Med Insights Oncol *10*, 31-39.

Gholamin, S., Mitra, S.S., Feroze, A.H., Liu, J., Kahn, S.A., Zhang, M., Esparza, R., Richard, C., Ramaswamy, V., Remke, M.*, et al.* (2017). Disrupting the CD47-SIRPalpha anti-phagocytic axis by a humanized anti-CD47 antibody is an efficacious treatment for malignant pediatric brain tumors. Science translational medicine *9*.

Green, J.A., Arpaia, N., Schizas, M., Dobrin, A., and Rudensky, A.Y. (2017). A nonimmune function of T cells in promoting lung tumor progression. J Exp Med.

Haghverdi, L., Buettner, F., and Theis, F.J. (2015). Diffusion maps for high-dimensional single-cell analysis of differentiation data. Bioinformatics *31*, 2989-2998.

Haghverdi, L., Buttner, M., Wolf, F.A., Buettner, F., and Theis, F.J. (2016). Diffusion pseudotime robustly reconstructs lineage branching. Nature methods *13*, 845-848.

Jahrsdorfer, B., Vollmer, A., Blackwell, S.E., Maier, J., Sontheimer, K., Beyer, T., Mandel, B., Lunov, O., Tron, K., Nienhaus, G.U.*, et al.* (2010). Granzyme B produced by human plasmacytoid dendritic cells suppresses T-cell expansion. Blood *115*, 1156-1165.

Jeffrey, K.L., Brummer, T., Rolph, M.S., Liu, S.M., Callejas, N.A., Grumont, R.J., Gillieron, C., Mackay, F., Grey, S., Camps, M.*, et al.* (2006). Positive regulation of immune cell function and inflammatory responses by phosphatase PAC-1. Nat Immunol *7*, 274-283.

Jimenez-Sanchez, A., Memon, D., Pourpe, S., Veeraraghavan, H., Li, Y., Vargas, H.A., Gill, M.B., Park, K.J., Zivanovic, O., Konner, J.*, et al.* (2017). Heterogeneous Tumor-Immune Microenvironments among Differentially Growing Metastases in an Ovarian Cancer Patient. Cell *170*, 927-938 e920.

Joller, N., Lozano, E., Burkett, P.R., Patel, B., Xiao, S., Zhu, C., Xia, J., Tan, T.G., Sefik, E., Yajnik, V.*, et al.* (2014). Treg cells expressing the coinhibitory molecule TIGIT selectively inhibit proinflammatory Th1 and Th17 cell responses. Immunity *40*, 569-581.

Josefowicz, S.Z., Lu, L.F., and Rudensky, A.Y. (2012). Regulatory T cells: mechanisms of differentiation and function. Annu Rev Immunol *30*, 531-564.

Kumar, V., Patel, S., Tcyganov, E., and Gabrilovich, D.I. (2016). The Nature of Myeloid-Derived Suppressor Cells in the Tumor Microenvironment. Trends in immunology *37*, 208-220.

Lavin, Y., Kobayashi, S., Leader, A., Amir, E.D., Elefant, N., Bigenwald, C., Remark, R., Sweeney, R., Becker, C.D., Levine, J.H.*, et al.* (2017). Innate Immune Landscape in Early Lung Adenocarcinoma by Paired Single-Cell Analyses. Cell *169*, 750-765 e717.

Levine, J.H., Simonds, E.F., Bendall, S.C., Davis, K.L., Amir el, 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., Severson, E., Pignon, J.C., Zhao, H., Li, T., Novak, J., Jiang, P., Shen, H., Aster, J.C., Rodig, S.*, et al.* (2016). Comprehensive analyses of tumor immunity: implications for cancer immunotherapy. Genome biology *17*, 174.

Lun, A.T., Bach, K., and Marioni, J.C. (2016). Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. Genome biology *17*, 75.

Mantovani, A., and Locati, M. (2013). Tumor-associated macrophages as a paradigm of macrophage plasticity, diversity, and polarization: lessons and open questions. Arteriosclerosis, thrombosis, and vascular biology *33*, 1478-1483.

Marrack, P., Mitchell, T., Hildeman, D., Kedl, R., Teague, T.K., Bender, J., Rees, W., Schaefer, B.C., and Kappler, J. (2000). Genomic-scale analysis of gene expression in resting and activated T cells. Curr Opin Immunol *12*, 206-209.

Mills, C.D., Kincaid, K., Alt, J.M., Heilman, M.J., and Hill, A.M. (2000). M-1/M-2 macrophages and the Th1/Th2 paradigm. Journal of immunology *164*, 6166-6173.

Moignard, V., Woodhouse, S., Haghverdi, L., Lilly, A.J., Tanaka, Y., Wilkinson, A.C., Buettner, F., Macaulay, I.C., Jawaid, W., Diamanti, E.*, et al.* (2015). Decoding the regulatory network of early blood development from single-cell gene expression measurements. Nat Biotechnol *33*, 269-276.

Murray, P.J., Allen, J.E., Biswas, S.K., Fisher, E.A., Gilroy, D.W., Goerdt, S., Gordon, S., Hamilton, J.A., Ivashkiv, L.B., Lawrence, T.*, et al.* (2014). Macrophage activation and polarization: nomenclature and experimental guidelines. Immunity *41*, 14-20.

Novershtern, N., Subramanian, A., Lawton, L.N., Mak, R.H., Haining, W.N., McConkey, M.E., Habib, N., Yosef, N., Chang, C.Y., Shay, T.*, et al.* (2011). Densely interconnected transcriptional circuits control cell states in human hematopoiesis. Cell *144*, 296-309.

Pauken, K.E., and Wherry, E.J. (2015). Overcoming T cell exhaustion in infection and cancer. Trends Immunol *36*, 265-276.

Perdiguero, E.G., and Geissmann, F. (2016). The development and maintenance of resident macrophages. Nat Immunol *17*, 2-8.

Philip, M., Fairchild, L., Sun, L., Horste, E.L., Camara, S., Shakiba, M., Scott, A.C., Viale, A., Lauer, P., Merghoub, T.*, et al.* (2017). Chromatin states define tumour-specific T cell dysfunction and reprogramming. Nature *545*, 452-456.

Plitas, G., Konopacki, C., Wu, K., Bos, P.D., Morrow, M., Putintseva, E.V., Chudakov, D.M., and Rudensky, A.Y. (2016). Regulatory T Cells Exhibit Distinct Features in Human Breast Cancer. Immunity *45*, 1122-1134.

Prabhakaran, S., Azizi, E., Carr, A., and Pe'er, D. (2016). Dirichlet Process Mixture Model for Correcting Technical Variation in Single-Cell Gene Expression Data. In Proceedings of The 33rd International Conference on Machine Learning, B. Maria Florina, and Q.W. Kilian, eds. (Proceedings of Machine Learning Research: PMLR), pp. 1070--1079.

Pyonteck, S.M., Akkari, L., Schuhmacher, A.J., Bowman, R.L., Sevenich, L., Quail, D.F., Olson, O.C., Quick, M.L., Huse, J.T., Teijeiro, V.*, et al.* (2013). CSF-1R inhibition alters macrophage polarization and blocks glioma progression. Nature medicine *19*, 1264-1272.

Rooney, M.S., Shukla, S.A., Wu, C.J., Getz, G., and Hacohen, N. (2015). Molecular and genetic properties of tumors associated with local immune cytolytic activity. Cell *160*, 48-61.

Roychoudhuri, R., Eil, R.L., and Restifo, N.P. (2015). The interplay of effector and regulatory T cells in cancer. Current opinion in immunology *33*, 101-111.

Sainz, B., Jr., Martin, B., Tatari, M., Heeschen, C., and Guerra, S. (2014). ISG15 is a critical microenvironmental factor for pancreatic cancer stem cells. Cancer research *74*, 7309-7320.

Senbabaoglu, Y., Gejman, R.S., Winer, A.G., Liu, M., Van Allen, E.M., de Velasco, G., Miao, D., Ostrovnaya, I., Drill, E., Luna, A.*, et al.* (2016). Tumor immune microenvironment characterization in clear cell renal cell carcinoma identifies prognostic and immunotherapeutically relevant messenger RNA signatures. Genome biology *17*, 231.

Setty, M., Tadmor, M.D., Reich-Zeliger, S., Angel, O., Salame, T.M., Kathail, P., Choi, K., Bendall, S., Friedman, N., and Pe'er, D. (2016). Wishbone identifies bifurcating developmental trajectories from single-cell data. Nature biotechnology *34*, 637-645.

Singer, M., Wang, C., Cong, L., Marjanovic, N.D., Kowalczyk, M.S., Zhang, H., Nyman, J., Sakuishi, K., Kurtulus, S., Gennert, D.*, et al.* (2017). A Distinct Gene Module for Dysfunction Uncoupled from Activation in Tumor-Infiltrating T Cells. Cell *171*, 1221-1223.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S.*, et al.* (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences of the United States of America *102*, 15545-15550.

Swiecki, M., and Colonna, M. (2015). The multifaceted biology of plasmacytoid dendritic cells. Nature reviews Immunology *15*, 471-485.

Tanaka, A., and Sakaguchi, S. (2017). Regulatory T cells in cancer immunotherapy. Cell research *27*, 109-118.

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.

Topalian, S.L., Drake, C.G., and Pardoll, D.M. (2015). Immune checkpoint blockade: a common denominator approach to cancer therapy. Cancer cell *27*, 450-461.

Vallejos, C.A., Risso, D., Scialdone, A., Dudoit, S., and Marioni, J.C. (2017). Normalizing single-cell RNA sequencing data: challenges and opportunities. Nature methods *14*, 565-571.

van der Maaten, L., and Hinton, G. (2008). Visualizing Data using t-SNE. J Mach Learn Res *9*, 2579-2605.

Zheng, G.X., 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. Nature communications *8*, 14049.

Zilionis, R., Nainys, J., Veres, A., Savova, V., Zemmour, D., Klein, A.M., and Mazutis, L. (2017). Single-cell barcoding and sequencing using droplet microfluidics. Nat Protoc *12*, 44-73.

Figure Legends

Figure 1: Single-Cell RNA Seq Experimental Design and Initial Data Exploration

(A) Flow chart displaying experimental design and analysis strategy.

(B) Summary of samples obtained and patient metadata; more details in S1A.

(C) t-SNE projection of complete immune systems from two example breast cancer tumors. scRNA-seq data for each tumor is processed with pipeline in Figure S1B and library size-normalized; each dot represents a single-cell colored by PhenoGraph clustering, and clusters are labeled by inferred cell types. Additional tumors are presented in in Figure S1D.

(D) Pie charts of cell type fractions for each patient's tumor-infiltrating immune cells, colored by cell type.

(E) Left: Boxplots showing expression of Hypoxia signature (defined as the mean normalized expression of genes in the hypoxia signature listed in Table S4) across immune cells from each patient. Right: Heatmap displaying z-scored mean expression of genes in hypoxia signature. Top: Barplot showing total expression of each gene indicated in the heatmap, across all patients. See Figure S1E-G for additional signatures.

Figure 2: Unbiased Characterization of the Immune System Across Breast Cancer Patients

(A) t-SNE projection of tumor-infiltrating immune cells from 4 breast cancer patients after library-size normalization (left panel) and Biscuit normalization and imputation (right panel). Cells are colored by tumor (patient). Less mixing of tumors indicates either batch effects or patient-specific cell states.

(B) Left: Boxplots showing expression of CD8 T cell activation signature (defined as the normalized mean expression of genes in the activation signature listed in Table S4) across immune cells from each patient. Right: Heatmap displaying z-scored mean expression of genes in activation signature. Top: Barplot showing total expression of each gene indicated in the heatmap across all patients. Expression of T cell activation signature shows variability across patients and increased expression in patients BC6 and BC3.

(C) t-SNE map of breast tumor-infiltrating immune cells from all 8 patients after Biscuit normalization and imputation showing rich structure and diverse cell types. Cells colored by Biscuit clusters and labeled with inferred cell types.

(D) Histogram depicting entropy of the patient distribution as a measure of sample mixing. Entropy is computed per cell, based on the distribution of patients in (30-NN) local cell neighborhoods after library-size normalization (left panel) as compared to Biscuit (right panel).

(E) t-SNE projection of complete atlas of immune cells, post-Biscuit normalization, from all patients and all tissues including tumor, blood, lymph, and contralateral normal tissue,labeled by inferred cell type (left panel) and normalized expression of 8 immune cell markers (right panel). Figure S2 presents further details on inferred clusters with complete annotations in Table S2.

(F) Pearson correlations of cluster expression centroids to bulk RNA-seq data from purified immune populations (from Jeffrey 2006 and Novershtern 2011). Scale bar displays r-values.

(G) Histogram of frequency of patients contributing to each cluster showing that 19 clusters (out of 95) are present in all 8 patients and 10 clusters are patient-specific.

(H) Expression of canonical and cell type markers across clusters, z-score normalized across clusters. T-exhausted denotes the mean expression of terminal differentiation signature listed in Table S4.

(I,J) Differentially expressed genes in B cell (I) and NK cell (J) clusters, standardized by z-scores within cell type. As an example, the expression of CD19 is standardized across all B cell clusters to highlight clusters with higher or lower expression of the marker compared to the average B cell cluster, but is highly expressed in nearly all B cell clusters (refer to Table S3 for all DEGs in these and other clusters).

Figure 3: Impact of Environment on Breast Immune Cells

(A) Breast immune cell atlas inferred from combining all patient samples and tissues, presented after Biscuit and projected with t-SNE. Each dot represents a cell and is colored by cluster label; major cell types are marked according to Figure 2F,H and Table S2,3.

(B) Subsets of immune atlas t-SNE projection in (A) showing cells from each tissue presented separately on the same coordinates as 3A to highlight the differences between tissues compartments.

(C) Proportions of cell types across tissue types in pie charts.

(D) Distribution of variance of normalized expression computed for each gene across all immune cells (all patients) from tumor tissue compared to that in normal breast tissue.

(E) Hallmark GSEA enrichment results on genes with highest difference in variance in tumor T cells vs normal tissue T cells. See Figure S3 for enrichment in monocytic and NK cells. Most significant results are shown; full lists of enrichments are presented in Table S5.

(F) Phenotypic volume in log-scale (defined as determinant of gene expression covariance matrix, detailed in STAR methods) of T cells, monocytic cells, and NK cells, comparing tumor immune cells and normal breast immune cells after correcting for differences in number of cells. Massive expansion of volume spanned by independent phenotypes active in tumor compared to normal tissue is shown for all three major cell types.

Figure 4: Detailed Characterization of T Cells

(A) Visualization of all cells from T Cell clusters using first, second, and third informative diffusion components (two uninformative components denoting isolated NKT and blood-specific clusters were removed from further analysis). Each dot represents a cell colored by cluster, and by tissue type in insert. The main trajectories are indicated with arrows and annotated using the signature most correlated with each component. See Figure S4D for additional components.

(B) Traceplot of CD8 T cell activation signature (defined as mean expression across genes in signature listed in Table S4) for all T cells along first informative diffusion component. Cells are sorted based on their projection along the diffusion component (x-axis), and the blue line indicates moving average over normalized and imputed expression, using a sliding window of length equal to 5% of total number of T cells; shaded area displays standard error (y-axis).

(C) Heatmap showing expression of immune-related genes with the largest positive correlations with activation component, averaged per cluster and z-score standardized across clusters; columns (clusters) are ordered by mean projection along the component. See Figure S4 for heatmaps for additional components.

(D) Violin plot showing the projection of T-cells along activation component aggregated by total density (left), tissue type (middle), and cluster (right). See Figure S4 for violin plots for additional components. Number of dots inside each violin are proportional to number of cells.

(E) Trace-plots (as in B) of (left) terminal differentiation signature along second informative component and (right) hypoxia signature along third informative component, labeled respectively as terminal differentiation and hypoxia components. List of genes associated with signatures are presented in Table S4.

(F) Heatmap of cells projected on each diffusion component (rows) averaged by cluster (columns).

Figure 5. Covariance Patterns Help Define Distinct T Cell Clusters

(A, B, C) Heatmaps showing normalized and imputed mean expression levels for a curated set of transcriptomic signatures (rows) important to T Cells (listed in Table S4) for (A) CD4 memory clusters, (B) CD8 memory clusters, and (C) T Regulatory clusters. Only signatures with high expression in at least one T cell cluster are shown. Signature expression values are z-scored relative to all T cell clusters but only shown for clusters of the same cell type for ease of visualization.

(D) Cartoon illustration of two clusters of cells showing similar mean expression for two example marker genes but opposite covariance between the same two genes.

(E) Scatter plot showing mean expression of GITR vs. CTLA-4 for each T cell cluster (represented by a dot). Treg clusters, labeled in red, have high mean expression levels of both genes.

(F) Distribution of covariance between GITR and CTLA-4 across all T cell clusters (purple), with values for Treg clusters labeled in red. Note that Treg cluster covariance values are present as both positive (46, 56, 87) and negative (80) outliers, exhibiting differences in covariance despite sharing high mean expression levels. See Figure S5A for similar computation on the raw, un-normalized, and un-imputed data, verifying the result.

(G) Network visualization illustrating strength of covariance between pairs of checkpoint receptor genes in Treg clusters. Edge width denotes absolute magnitude of covariance and color denotes sign of covariance (red positive and blue negative); the example case of CTLA-4 and GITR is highlighted in yellow. Distinct patterns in gene covariance, in addition to mean expression, drive the definition of clusters in the Biscuit model. See figure S5C for similar networks for T cell effector and effector memory populations.

(H) Heatmaps showing covariance between immune genes in T reg clusters 56 (left panel), and 87 (right panel). Note different modules of covarying genes.

(I) Pie charts showing proportion of the five Treg clusters in each patient, indicating that differences in covariance patterns between clusters also translate to patients.

Figure 6: Detailed Characterization of Myeloid Cells

(A) t-SNE map projecting only myeloid cells across all tissues and patients. Cells are colored by Biscuit cluster and cell types are circled and labeled based on bulk RNA-seq correlation-based annotations.

(B through E) Projection of cells in myeloid clusters on macrophage activation, pDC, and monocyte activation (first, second, and fourth) diffusion components. Cells are colored by (B) cluster, (C) tissue type, (D) cell type (as explained in STAR Methods), and (E) expression of example lineage demarcating genes. The main trajectories are indicated with arrows and labeled in (B).

(F) Violin plots showing the density of cells along macrophage activation component and organized by overall density (left panel), tissue type (middle panel), and cluster (right panel).

(G) Scatter plot of normalized mean expression of M1 and M2 signatures per cell (dark blue); cells assigned to 3 TAM clusters have been highlighted by cluster (light blue, pink, yellow); each dot represents a cell and cells are plotted in randomized order.

(H) Heatmap showing imputed mean expression levels in myeloid clusters for a curated set of transcriptomic signatures important to myeloid cells (listed in Table S4), z-score normalized per signature.

See also Figure S6 for additional violin plots and heatmaps representing the other components.

Figure 7: Covariance Patterns Help Define Distinct Macrophage Clusters

(A) Scatterplot of mean expression of MARCO and CD276 in each myeloid cluster; each dot represents a cluster. Average expression levels for the three TAM clusters (23, 25, and 28) are marked in red, indicating high expression of both markers in macrophage clusters.

(B) Distribution of covariance between MARCO and CD276 across all myeloid clusters. TAM clusters(23, 25, and 28) are marked in red and present substantial outliers. See Figure S7A for similar computation on the raw, un-normalized, and un-imputed data, verifying the result.

(C) Heatmaps showing covariance patterns of M1 and M2 macrophage polarization marker genes (including many current or potential drug targets) in 3 TAM clusters (23, 25, and 28).

Supplementary Figure Legends

Figure S1: Additional details on samples and individual immune systems, related to Figure 1.

(A) All clinical and related metadata for all 8 patients.

(B) Bioinformatics pipeline for processing single-cell RNA-seq data; inputs boxed on left, outputs labeled in red. Each rectangle represents a processing step. Braces on bottom display file formats as data moves through the pipeline. Quality control and metrics are provided in Table S1.

(C) t-SNE projection of complete immune systems from six breast cancer tumors. scRNA-seq data for each tumor is processed with pipeline in Figure S1B and library size-normalized; each dot represents a single-cell colored by PhenoGraph clustering, and clusters are labeled by inferred cell types. Two additional tumors are presented in Figure 1C.

(D) Regression of flow cytometry cell type percentages in each patient against RNA-seq cell type percentages for B cells (blue), monocytic cells (orange), and T cells (green).

(E-G) Expression of metabolic signatures: fatty acid metabolism (E), glycolysis (F), and phosphorylation (G), summarized as boxplots (left) showing expression of each respective signature (defined as the mean normalized expression of genes in each signature listed in Table S4) across immune cells from each patient; and heatmap (right) displaying z-scored mean expression of genes in each signature; (top) barplot showing total expression of each gene indicated in the heatmap across all patients. See Figure 1E for one additional signature.

Figure S2: Details of inferred parameters for clustering and normalization using Biscuit, related to Figure 2.

(A) Posterior probability of assignment of cells to clusters in the Biscuit model in the full immune cell atlas of combined tissues and patients presented in Figure 2E; note broad distributions in assignment of naive T cells (bottom) as compared to other cell types.

(B) Bhattacharyya pairwise distances between clusters of Figure 2F, H (blue: small distance to yellow: large distance).

(C) Robustness analysis of clusters performed with 10-fold cross-validation; boxplots summarize the probability of a pair of cells being assigned to the same final cluster across all 10 subsets.

(D) Boxplots showing entropy of distribution of patients in each cluster, computed with bootstrapping to correct for cluster size. Note that cluster labels are given by size (cluster 1 has the most number of cells and cluster 95 has the fewest) and ordering clusters by mean entropy in this plot indicates that entropy does not correlate with size.

(E) Violin plot of pairwise Bhattacharyya distances between distribution of expression of each gene between all pairs of clusters in the same or different cell types considering mean and covariance of expression, averaged across all genes.

(F) Same as (E), but after removing the effect of cluster mean in computing similarity, thus considering only covariance.

(G) Distribution of Biscuit alpha parameters per cell vs log of library size, with cells colored by clusters; Biscuit alpha parameters correct for differences in library size across and within clusters.
(H) Distribution of inferred cell-specific parameters alpha and (I) beta in Biscuit across cells from each patient. These differences were corrected in normalizing with alpha and beta parameters.

Figure S3. Analysis of differences in gene expression variance between immune cells in breast tumor compared to normal breast tissue, related to Figure 3.
Hallmark GSEA enrichment results on genes with highest difference in variance in tumor vs normal tissue in (A) NK and (B) monocytic cells. See Figure 3E for enrichment in T cells; complete lists of enrichments are presented in Table S5.

Figure S4. Additional details on T cell diffusion components, related to Figure 4.
(A) Hartigan's dip test on density of cells projected on diffusion components, showing statistically significant continuity (lack of "dips") in cells along T cell activation component (component 3, third panel from left), whereas other components exhibit more defined states (multimodality).
(B) Violin plot of cells projected on terminal differentiation diffusion component: terminal differentiation components organized by tissue type (left panels) and cluster (center panel). Also, heatmap showing expression of immune-related genes with the largest positive correlations with component, averaged per cluster and z-score standardized across clusters; columns (clusters) are ordered by mean projection along the component.
(C) Scatter plot showing mean expression levels of T Cell activation and terminal differentiation signatures for all T Cell clusters. Red triangles denote T Reg clusters showing high expression in both signatures. The dotted line denotes y = x.
(D) Visualization of all T Cell clusters using activation (component 3), terminal differentiation (component 4), and tissue specificity (component 6) diffusion components. Cells are colored by clusters, and by tissue type in insert. The main trajectories are indicated with arrows and annotated with labels.
(E,F) Same as (B) for hypoxia (E) and tissue specificity (F) components.
See Figure 4 for additional component.

Figure S5. Details of covariance patterns in T cell clusters, related to Figure 5.
(A) Displaying null distributions and observed covariances between CTLA-4 and GITR in raw, un-normalized data using hypothesis testing, subsampling, and permutation (see STAR methods); shows that the differences in covariance shown in Figure 5F,G are also present in un-normalized and un-imputed data, and hence are not an artifact of computation.
(B) Bivariate plots of expression levels of GITR and CTLA-4 in Treg clusters based on inferred mean and covariance parameters from Biscuit. Dark blue color indicates the highest density of cells and light yellow the lowest density of cells.
(C) Network graphs showing covariance between checkpoint receptors in activated T cell clusters. Edge width denotes absolute magnitude (strength) of covariance and color denotes sign of covariance (red positive and blue negative). Note diversity across clusters. Similar graphs for T reg clusters are shown in 5G.

Figure S6: Additional details on diffusion component analysis of myeloid cells, related to Figure 6.
(A) Hartigan's dip test on density of cells projected on diffusion components indicating no diffusion components across myeloid cells show statistically significant continuity, implying myeloid cells reside in defined (multimodal) states along major components explaining variation.

(B) Heatmap showing expression of immune-related markers with the largest positive correlation with TAM activation, pDCs, and monocyte activation components.
(C) Violin plot showing the density of cells projected along pDC component and organized by tissue type and cluster.
(D) Violin plot showing the density of cells projected along monocyte activation component and organized by tissue type and cluster.

Figure S7. Details of covariance patterns in myeloid clusters
(A) Displaying null distributions and observed covariances between MACRO and CD276 in raw, unnormalized data using hypothesis testing, subsampling, and permutation (see STAR methods), showing that the differences in covariance in normalized data as shown in Figure 7B are also present in un-normalized and un-imputed data, and hence is not an artifact of computation.
(B) Bivariate plots of expression levels of MARCO and CD276 in Treg clusters based on inferred mean and covariance parameters from Biscuit. Dark blue color indicates the highest density of cells and light yellow the lowest density of cells.

Methods Figure Legends

Figure M1. Data Driven Pipeline Construction
(A) Visualization of cell barcode GC content (percentage, x-axis) versus cell barcode read coverage (y) displaying higher coverage at 50% GC content. Yellow to purple color represents density of cell barcodes.
(B) Schematic of capture primer displaying amplification machinery, cell barcodes, UMIs, and poly-T capture site.
(C) Cartoon showing common sources of barcode errors including (i) breakage and (ii) substitution errors.
(D) Comparison of complete GENCODE annotation against a reduced annotation containing only GENCODE-annotated lincRNA and protein coding RNA. Displaying drop-out events occurring on x-axis as well as masking events on y-axis.
(E) Example cell filtering plot showing the empirical cumulative density of molecules (y-axis) per cell barcode (x-axis). Note that a small number of cell barcodes contain most of the molecules in the experiment. Dashed black lines represent cut-off points after which cell barcodes are considered to consist of contamination. Red barcodes are excluded.
(F) Coverage plot comparing the total molecules in each cell (x axis) against the average coverage in each cell (y axis). Densities of cells with aberrantly low coverage such as those with lower than 5 reads / molecule are considered likely errors and are discarded.
(G) Mitochondrial (MT) RNA fraction plot displaying the total number of molecules in each cell vs the fraction of those molecules that come from mitochondrial sources. Cells in red consist of more than 20% MT-RNA and are considered to be likely dying cells. These cells (red) are discarded.
(H) Complexity plot displaying the number of detected molecules (x axis) vs genes (y axis). The relationship is fit with linear regression and cells (red) whose residual gene detection is greater than 3 standard deviations are removed.
(I) Heatmap of pairwise pseudo-bulk sample-sample correlations (r2) across all samples and replicates in the experiment.

Figure M2. (A) Stochastic data generative process for Biscuit illustrated with a toy example. **Top panel**: Left: shows 3 multivariate Gaussian densities with no technical variation. Middle: An ideal cell (yj) is simulated as a random draw from any of these 3 Gaussians. Right: The covariance matrix across 10 such randomly-drawn cells showing 3 block covariances across the diagonal corresponding to three clusters. **Bottom panel**: Left: shows 3 multivariate Gaussian densities with means and covariances scaled using (j, j) to handle cell-specific variations. Middle: A cell (lj) is simulated as a random draw from any of these 3 scaled Gaussians. Right: The covariance matrix across 10 such randomly-drawn cells showing loss of signal in the 3 block diagonal covariances. We assume the model for lj captures real single-cell measurements and the goal is to normalize data by converting it to follow the model for yj.

(B). Finite state automata for Biscuit. The shaded circle denotes lj, which is observed gene expression for cell j, white circles show latent variables of interest, rectangles depict the number of replications at the bottom right corner, diamonds are hyper-parameters, and double diamonds are hyper-priors obtained empirically. Inference equations are obtained by inverting the date generative process.

(C) Left panel: Input count matrix to Biscuit. Middle panel: Inference algorithm with Gibbs iterations are depicted where cell-specific (j, j) and cluster-specific (k, k) parameters are iteratively inferred leading to cell assignments to clusters. Right panel: Output from Biscuit, which is the normalized and imputed count matrix.

Supplementary Tables

Table S1. Aggregated quality control and metric information for each sequencing sample. Columns A-C contain metadata on the patient, tissue of origin, and sample number. Related to Figure S1B.

Table S2. Annotations of clusters inferred in full breast immune atlas (across all patients and tissues) and their proportions across tissues and patients.

Table S3. List of differentially expressed genes in clusters listed in Table S2 (sheet 1); the subset of differentially expressed immune-related genes (sheet 2).

Table S4. List of gene signatures (sources listed in STAR Methods)

Table S5. Full hallmark GSEA enrichment results for gene variance in T-cells (Sheet 1), NK cells (Sheet 2), and Monocytic Cells (Sheet 3). These enrichments expand upon partial lists shown in Figure 3 and S3.

Table S6. List of capture reads, molecules, cells, and genes per sample replicate

STAR Methods

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological Samples		
Breast carcinoma	Fresh operative tissue samples	MSKCC
Normal breast tissue	Fresh operative tissue samples	MSKCC
Lymph node	Fresh operative tissue samples	MSKCC
Peripheral blood	Patient venipuncture	MSKCC
Antibodies		
anti-CD45 ab	Biolegend	RRID:AB_2566372
DAPI Stain	Calbiochem	SCR_014366
anti-Foxp3 ab	Thermo Fisher	RRID:AB_1834364
anti-CD3 ab	Biolegend	RRID:AB_1575008
anti-CD4 ab	Biolegend	RRID:AB_571945
anti-CD16 ab	Biolegend	RRID:AB_314207
anti-CD56 ab	BD Biosciences	RRID:AB_396853
anti-CD8 ab	Biolegend	RRID:AB_528885
anti-CD19 ab	Biolegend	RRID:AB_2562015
anti-CD11b ab	Biolegend	RRID:AB_2563395
Critical Commercial Assays		
inDrop platform	Custom build droplet microfluidics platform	Zilionis et al., 2017; Klein et al., 2015
HiSeq 2500	Illumina	MSKCC
NEBNext mRNA Second Strand Synthesis Module	NEB	cat no. E6111S

HiScribe T7 High Yield RNA Synthesis Kit	NEB	cat no. E2040S
Kapa 2× HiFi HotStart PCR mix	Kapa Biosystems	cat no. KK2601
Software and Algorithms		
SEQC	This paper	https://github.com/ambrosejcarr/seqc.git
Biscuit	This paper and Prabhakaran, S., Azizi, E., Carr, A., & Pe'er, D, 2016	https://github.com/sandhya212/BISCUIT_SingleCell_IMM_ICML_2016
t-SNE	van der Maaten and Hinton 2008	https://lvdmaaten.github.io/software/
PhenoGraph	Levine et al 2015	https://github.com/jacoblevine/PhenoGraph
Diffusion map	Coifman et al 2010	http://www.math.jhu.edu/~mauro/#tab_DiffusionGeom
FACSDiva	BD Biosciences	RRID:SCR_001456
FlowJo	BD Biosciences	RRID:SCR_008520 https://www.flowjo.com/
FastQC	Andrews 2010	https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
STAR 2.5.3a	Dobin et al 2013	https://github.com/alexdobin/STAR

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources should be directed to and will be fulfilled by Lead Contacts Dana Pe'er (peerd@mskcc.org) and Sasha Rudensky (rudenska@mskcc.org).

METHOD DETAILS

Sample Collection

Tissues were collected from women undergoing surgery for primary breast cancer. Normal tissue was obtained from prophylactic mastectomies of the same cancer patients, and peripheral blood mononuclear cells (PBMCs) were obtained from patients prior to their surgical procedures. All samples were obtained after informed consent and approval from the Institutional Review Board (IRB) at Memorial Sloan Kettering Cancer Center.

Lymphocytes from tumor and normal tissues were isolated by mincing the freshly obtained surgical specimens into 1 mm cubic pieces and subsequent enzymatic digestion using Liberase TL (Sigma) for 20 min at 37°C. The digested tissues were then passed through a 100M filter and washed twice with PBS prior to surface staining. Lymphocytes were stained at 1*x*10^6 cells per ml for 20 min with anti-CD45 and DAPI for live-dead discrimination following Fc receptor blockade (BioLegend). Viable lymphocytes (CD45+DAPI-) were sorted on a FACSARIA sorter (BD Biosciences). Post-sort purity was routinely > 95% for the sorted populations.

Each sample was then divided into 10,000 cell aliquots and at least 2 technical replicates were processed through the complete experimental protocol (see below), with the exception that technical replicates were often processed on the same sequencing lane.

Library Preparation

We employed inDrop (Klein et al., 2015; Zilionis et al., 2017), a droplet-based single cell RNA sequencing technology. inDrop was selected over alternative technologies because it makes use of closely packed deformable hydrogel beads, ensuring that 75-90% of the input cells are paired with a unique barcode. This efficiency allowed deep sampling of many immune cells from individual patients, even in cases where immune infiltration was relatively low. As a result, we were able to compare triple-negative breast cancer (TNBC) samples with Her2+ and ER+ samples, which in some cases had as few as 50,000 tumor infiltrating immune cells.

Isolated, FACS-sorted CD45+ cells were suspended in ice-cold 1X PBS supplemented with 16% (v/v) Optiprep and 0.05% (w/v) BSA and encapsulated into 1.5 nL droplets together with custom-made DNA barcoding hydrogel beads and RT/lysis reagents. The microfluidics chip was operated at a throughput of ~30,000 cells per hour, and over 75% of cells entering microfluidics chips were co-encapsulated with one DNA barcoding hydrogel bead. The frequency of cell doublets (droplets having two cells) was low (~0.59%) due to highly diluted cell suspensions used for encapsulation, corresponding to approximately 1 cell for every 12th droplet. In general, single-cell RNA-Seq library preparation was carried out following the protocol reported recently (Zilionis et al., 2017), with some modifications as described below.

Construction of new barcode sets

GC content has a known impact on PCR efficiency (Mamedov et al., 2008): high or low fractions of G and C nucleotides reduce sequence amplification efficiency. We analyzed data produced with earlier version of DNA barcodes (Klein et al., 2015; Zilionis et al., 2017) and observed that barcodes with balanced GC content achieved higher molecule number (Figure M1A). We reasoned that balancing GC content across our barcodes would decrease variance across our libraries, thus increasing the average number of mRNA observed per cell. Further, we observed that the original barcode sequences had a minimum Hamming distance of 2. This is adequate to identify but not to correct single-base errors. We redesigned a library so that all barcodes had balanced GC content, with Hamming distance of >= 3, such that all single base errors are correctable, and with an average Hamming distance between pairs of barcodes of 13.3. This was done by performing a constrained optimization over barcodes of various lengths obtained from Edittag (Faircloth and Glenn, 2012). As a result, the vast majority of barcode errors are correctable and, as our results showed, the single-cell RNA-Seq libraries generated with new DNA barcoding hydrogel beads produced an overall increase in molecules/million sequencing reads of 5.3%.

The custom-made hydrogel beads carrying new DNA barcode sets were synthesized using the Agilent Bravo Automated Liquid Handling Platform following the previously described protocol (Zilionis et al., 2017). Before loading the DNA barcoding beads into the chip, they were washed twice in 1X SuperScript-III RT buffer and lysis reagent (1% (v/v) Igepal-CA630). In contrast to the approach in Zilionis et al., the Illumina PE Read 1 sequence was placed on the RT primer, thus the full-length primer sequence was as follows:

/5Acryd/PC/CGATGACG**TAATACGACTCACTATAG**GGATACCACCATGGCTCTTTCCCTACACGACGCTCTTCCGATCT[12345678901]GAGTGATTGCTTGTGACGCCTT[12345678]NNNNNNNNTTTTTTTTTTTTTTTTTTTTV,

where 5Acryd is an acrydite moiety, PC is a photo-cleavable spacer, the letters in bold indicate T7 RNA promoter sequence, and underlined letters indicate the site for Illumina PE Read 1 Sequencing primer. The numbers indicate cell barcodes, which were specifically designed for this experiment to have 50% GC content and Hamming distance of >= 3 between each pair of barcode. Fluorescent in situ hybridization (FISH) analysis confirmed that hydrogel beads carried ~10^8 covalently-attached and photo-releasable barcoding DNA primers.

Increasing the throughput

To increase the cell isolation throughput we used a cell barcoding chip (v2) (Droplet Genomics) and flow rates for cell suspension at 250 µl/hr, for RT/lysis mix at 250 µl/hr, and for barcoded hydrogel beads at 75 µl/hr. The flow rate for droplet stabilization oil was 550 µl/hr. Such flow parameters generated approximately 40,000 droplets an hour. After loading all components (cells, beads and RT/lysis reagents) into droplets, the final composition of a reaction under which cDNA synthesis was carried out was 155 mM KCl, 50 mM NaCl, 11 mM $MgCl_2$, 135 mM Tris-HCl [pH 8.0], 0.5 mM KH_2PO_4, 0.85 mM Na_2HPO_4, 0.35 % (v/v) Igepal-CA630, 0.02 % (v/v) BSA, 4.4% (v/v) Optiprep, 2.4 mM DTT, 0.5 mM dNTPs, 1.3 U/ml RNAsIN Plus, and 11.4 U/ml SuperScript-III RT enzyme. After emulsion collection on ice the tube was exposed to 350 nm UV-light to photo-release DNA barcoding primers attached to the hydrogel beads. The RT reaction was initiated by transferring the tube to 65ºC for 1 min followed by a 1-hour incubation at 50ºC and 15 min at 75ºC. Post-RT droplets were chemically broken to release barcoded cDNA, which was then purified and amplified as described previously (Zilionis et al., 2017)). At the final step, libraries were amplified using trimmed PE Read 1 primer (PE1):

5’-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGA

and indexing PE Read 2 primer (PE2):

5'-CAAGCAGAAGACGGCATACGAGAT[index]GTGACTGGAGTTCAGACGTGTGCTCTTCCG ATCT,

where [index] encoded one of the following sequences: CGTGAT, ACATCG, GCCTAA, TGGTCA, CACTGT or ATTGGC).

Multiplexing of PCR libraries allowed for the pooling of different samples onto one lane of Illumina HiSeq2500 flow cell when desired.

Sequencing and fastq quality control

Data were sequenced on Illumina HiSeq 2500 instruments using paired-end sequencing (PE1 54 bp and PE2 66 bp). Each replicate was sequenced on one half of a HiSeq lane, at an initial depth of approximately 100 million reads. scRNA-seq produces lower-complexity libraries than bulk sequencing techniques, which can infrequently lead to reduced base quality. Because each patient sample is precious, and because it is difficult to compare libraries with different average molecule counts, we verified the quality of each sequencing library with FastQC (Andrews, 2010), a software package that estimates the number of un-callable and low quality bases. Libraries that displayed significant (>25%) low quality bases were re-sequenced to maximize inter-sample comparability. See Supplementary Table S1 for sample sequencing depths.

QUANTIFICATION AND STATISTICAL ANALYSIS

Data preprocessing: SEQC

Overview: At the time of data collection adequate analysis methods available to construct count matrices from sequencing files for this data type were lacking. We therefore designed SEquence Quality Control (SEQC), a package that takes Illumina fastq or bcl files and generates a count matrix that is carefully filtered for errors and biases. The process is outlined in figure S1B. We developed SEQC into a general purpose method to build a count matrix from single cell sequencing reads, able to process data from inDrop, drop-seq, 10X, and Mars-Seq2 technologies.

Briefly, SEQC begins by extracting the cell barcode and UMI from the forward read and storing these data in the header of the reverse read. This produces a single fastq file containing alignable sequence and all relevant metadata. The merged file is carefully filtered for cell barcode substitution errors, broken barcodes, and low-complexity polymers to eliminate errors early in the pipeline, saving analysis cost.

Filtered reads are aligned against the genome with STAR (Dobin et al., 2013), a high performance community-standard aligner. After alignment, minimal representations of sequencing reads are translated into an hdf5 read store object, where cell barcodes are represented in reduced 3-bit coding. Reads are annotated with a reduced set of exon and gene ids representing gene features—only the ones that are possible to detect with poly-A capture based droplet RNA sequencing—and SEQC attempts to resolve reads with multiple equal-scoring alignments.

In cases where genomic and transcriptomic alignments are present, the transcriptomic alignments are retained. Unique alignments from the previous step are corrected for errors using an enhancement of the method designed in Jaitin et al. (Jaitin et al., 2014), with an additional probability model to constrain the false positive rate. The error-reduced, uniquely-aligned data are grouped by cell, molecule, and gene annotation, and compressed into count matrices containing (1) reads and (2) molecules. This matrix is thresholded in a similar manner to what has been previously described (Macosko et al., 2015; Zheng et al., 2017).

Finally, SEQC outputs a series of QC metrics in an HTML archive that can be used to evaluate the quality of the library and the success of the run. SEQC is fully modular, and as such has been adapted to process drop-seq, 10x, and mars-seq data by switching de-multiplexing modules. In addition, it can be configured either to run on a local high-performance cluster, or can automatically initiate runs on Amazon Web Services compute platforms, for those without access to local compute servers. The SEQC code is free and open-source, and can be found at https://github.com/ambrosejcarr/seqc.git, licensed under the MIT license.

Fastq Demultiplexing: The first stage of SEQC takes multiple fastq files containing genomic information and barcoding metadata spread across multiple sequencing files, and merges that information into a single fastq file using a "platform" class that comprises the locations of the cell barcodes and UMIs, the type of barcode and UMI correction to be run, and the number of T-nucleotides that are expected to be read from the capture primer. The merged fastq file contains genomic, alignable sequence in the sequence field, and has read metadata prepended to the name field, separated by colons. This step can be adjusted for novel sequencing approaches by adding a new platform class, often with only 10 lines of code. This allows the complete SEQC pipeline to be rapidly tested on iterations of new technologies.

For inDrop, which has variable-length cell barcodes, the description defines an additional method to localize the constant spacer sequence, which is flanked on both sides by the cell barcode (Figure M1B). The cell barcodes and molecular identifier are then extracted relative to the position of the spacer. Finally, we count the number of T-nucleotides that follow the UMI, where the poly-T spacer is supposed to be, and store this information for downstream filtering steps. The generated fastq file has the following format:

```
@<CELLBARCODE>:<UMI>:<#T>;GENOMIC READ NAME (read 2)

<GENOMIC SEQUENCE (read 2)>

+
<GENOMIC QUALITY (read 2)>
```

Substitution Error Rate Estimation: Two pieces of information *not* retained by the demultiplexing module are the cell barcode and UMI quality scores. Some pipelines, such as 10x Genomics' Cell Ranger, posit that sequencing error is the major source of substitution mutations in 3' sequencing data. Our inDrop data does not support this view of library construction. In inDrop, each read contains a 16-19 bp cell barcode selected from a whitelist of known barcodes. By examining barcodes for single base mutations, we estimated a positional, nucleotide-specific error rate for each sample (Table S1). E.g. to calculate the probability of a conversion from adenosine to cytosine, where $A \to C$ denotes this nucleotide conversion:

$$P_{A\to C} = \frac{1}{n\cdot m} \sum_{j=1}^{m} \sum_{i=1}^{n} \{1 \; if \; x_{ij\,:\,A\to C} \; else \; 0\}$$

where x_j is a barcode, $j \in \{1, ..., m\}$ and each barcode has n bases.

The average observed per-barcode error rates are 4%, a number far in excess of the abundance reported by the Illumina sequencer, which can be reliably calculated from errors in phiX included in sequencing runs (mean error rate 0.2% ∓ 0.1%) (Manley et al., 2016); a 4% error rate is more in line with aggregate error rates of the enzymes used in the preparation of sequencing libraries.

To verify that quality scores do not predict error rates, we tested the correlation between the error state of the cell barcode (1 if the base contains an error and 0 otherwise) with Illumina quality scores. If quality were predictive of substitution errors, we would expect to observe strong negative correlations, suggesting that low quality implies high error probability. However, we observed no relationship (mean r^2=0.04, max r^2=0.06; ‘C’ errors) on either inDrop or 10x data.

In contrast, mutations to N bases produce the expected relationship, with base quality negatively correlating with → N substitutions (r^2=-0.87). However, N base errors made up less than 1 / 100,000 of the observed errors in our experiment, and we conclude (1) that base quality is not meaningfully predictive of error rates, and (2) that most sequenced error is derived from upstream library construction steps.

While a 4% barcode error rate is higher than the error rate observed by other technologies, the use of linear amplification means that the errors we observe are non-cumulative: each transcript is generated from the original captured mRNA molecule, and as a result, 99.94% of observed barcodes have one or zero errors, all of which are correctable. This is in contrast to PCR-based amplification approaches which propagate errors that occur in early cycles, requiring more complex, graph-based correction methods, and larger Hamming buffers (see https://github.com/vals/umis).

Pre-alignment filtering: This module takes as input a raw merged fastq file and outputs a merged fastq file with corrected cell barcodes and full length UMI sequences (that may still contain substitution errors).

Cell barcode correction: Cell barcode errors in inDrop are easy to detect by design: we have a whitelist of 147,456 barcodes, each with Hamming distance >= 3. Thus, any single base substitution error is resolved by creating a lookup table for all barcodes and all single substitutions. If found in the table, the barcode is corrected. If not, it is discarded. As estimated above, the probability of a cell barcode containing an error is ~2%, and thus the expected rate of barcodes accruing 2 errors in a barcode is 1 / 2500. A 2-error lookup table has a very large memory footprint and would significantly increase computational cost of processing each experiment. Alternative algorithms have greater complexity and would increase run time. Thus, we accept this low rate of loss and proceed to correct single base errors, recovering approximately 2% additional data for each sample.

UMI validation: In contrast to cell barcodes, UMIs are random, and correction cannot proceed by the same strategy, so we devote a section later in the pipeline to the detection of UMI errors after the gene and mapping position of a fragment are identified.

Another source of error in scRNA-seq experiments, including inDrop, cel-seq, mars-seq, and likely drop-seq and 10x genomics, is the fracturing and random-priming of capture primers (Figure M1C) (Jaitin et al., 2014). We often observe cell-barcode prefixes followed by randomers. When fragmentation occurs at the cell barcode level, we can remove the fragments using the whitelist approach above. To remove barcodes that break in the UMI, we determined that we would sequence 5 bases into the poly-T tail of the primer, which we expect to be all T-nucleotides. By excluding reads with more than 1 non-T nucleotide, we are able to exclude most broken UMIs.

In aggregate, the filters in this section remove an average of 36% of reads (sd = 9.3%), depleting the count matrix of spurious molecules (see Table S1 for detailed values). These values are consistent with the results of running SEQC on drop-seq or MARS-seq datasets (data not shown).

Annotation Construction

Because the genome annotation is designed to be broadly applicable across sequencing modalities, it contains many features that are theoretically undetectable by inDrop and other 3' sequencing technologies. To address this, we constructed a custom annotation by starting with the current GENCODE genome and GTF file and removing all feature annotations that are not theoretically detectable by inDrop.

Two characteristics of inDrop limit its ability to capture certain gene biotypes. First, it employs poly-A capture, and thus will not detect non-polyadenylated transcripts. Second, it uses SPRIselect beads at several stages to deplete primers from reaction media. These beads carry out size selection, preferentially depleting primers but also small RNA species such as snoRNA, miRNA, and snRNA. Thus, libraries are expected to contain only transcribed, polyadenylated RNA of length > ~200 nt. Examining gene biotypes, this meant retaining protein coding and lncRNA biotypes, and excluding others.

To determine the impact of this change of reference on our data, we aligned the same single-cell immune dataset against the full reference and the reduced reference described above. We constructed a pseudo-bulk dataset for each reference by summing the molecules across all cells, producing an expression vector that contained a single value for each gene. We hypothesized that the reduction in reference features would result in a concentration of alignments in biologically relevant genes by depleting non-specific features, and that there would be many drop-out events where genes would be detected in the complete reference, but not the subset.

This is exactly what was observed (Figure M1D). The overall r2 value between the references is 0.94, with 93% of genes holding the exact same values in both reference alignments. In addition, information is concentrated in 35% fewer features, despite losing only 8% of the total molecules. There is also a large drop-out contingent present only when aligned against the complete reference. Gene ontology enrichment against this reference revealed high-level biologically agnostic enrichments, such as “protein coding,” “translation,” and other enrichments, which suggest a random sampling of high-expression genes.

Surprisingly, there was also a contingent of genes present only in the reduced alignment. These genes were highly enriched for immunological pathways, including JAK/STAT signaling, cytokine production, cytokine receptors, and immune growth factors, and further included critical immune genes such as IL3RA, a plasmacytoid dendritic cell marker (Figure M1D). This suggests that they are likely to represent true annotations for genes in this dataset, and that reducing the annotation produces a gain in specificity. We reasoned that these genes were uncovered in the reduced annotation because there are features in the complete set, such as pseudogenes, which have high homology to transcribed genes. Including these annotations, which should not be detectable, produces illogical multi-alignment to multiple genetic locations. When such multi-alignment cannot be resolved, most pipelines (including this one) exclude those multi-aligned reads, losing valuable signal. Given these results, we believe that the 8% molecule loss is the result of correctly discarding low-complexity alignments that were spuriously assigned a low-quality transcriptomic feature.

We note that Cell Ranger, the most commonly used 10x pipeline, carries out an extreme version of this redesign: it removes any gene that is not protein coding. We believe that this is too harsh: it excludes numerous transcribed pseudogenes and lincRNA which have been previously shown to be expressed, have biological functionality, and to be detectable in scRNA-seq.

The extreme case of annotation redesign is to exclusively align to expected features. Several methods exist to accomplish this, including Kallisto (Bray et al., 2016) and Tophat2 (Kim et al., 2013). However, 3’ scRNA-seq data typically contains between 10-30% genomic contamination, as identified by intergenic alignments. When we aligned directly against the transcriptome, we found that approximately 1% of intergenic reads were mistakenly aligned to exonic locations despite having higher alignment scores to genetic regions (data not shown). As a result, we align to the genome, considering only detectable features, and prefer transcriptomic alignments in cases where there are equivalent genomic and transcriptomic alignments, but remove reads that score highest against the genome.

Alignment

The merged, single-ended fastq files are aligned to hg38 with STAR using the annotation file as described. We selected STAR because it is a fast, highly parallel, cloud-scalable aligner that benchmarks well against existing aligners (Illicic 2016). We note that STAR automatically trims bases as necessary to find alignments, and as such no pre-trimming of reads based on quality is carried out. Alignment parameters used are as follows: –outFilterType BySJout, –outFilterMultimapNmax 100, –limitOutSJcollapsed 2000000 –alignSJDBoverhangMin 8, –outFilterMismatchNoverLmax 0.04, –alignIntronMin 20, –alignIntronMax 1000000, –readFilesIn fastqrecords, –outSAMprimaryFlag AllBestScore, –outSAMtype BAM Unsorted

This module thus takes as input a fastq file and produces a bam file containing up to 20 multiple alignments per input fastq record and with all unaligned reads contained in the same file. This format is useful for archival purposes, as it can be used to reconstruct the original merged fastq file without data loss.

Multi-Alignment Correction

Alignment algorithms aim to identify the unique portion of the genome that was transcribed to generate the read that is being aligned. In some cases, this unique source cannot be identified, and in these cases multiple possible sources are reported. These are commonly termed "multi-alignments", and because 3' ends of genes have higher homology than other parts of the genome, multi-alignments are more common in 3' sequencing data than in approaches that cover the full-transcriptome, such as Smart-seq2. Despite the increased frequency, most 3' pipelines discard multi-alignments and deal exclusively with unique genes. This module is designed to resolve all multiple alignments, producing an output that contains resolved (now unique) alignments.

There are several existing approaches to resolving multi-alignments, of which transcriptomic pseudo-alignment, such as that done by Salmon (Patro et al., 2017) and Kallisto (Bray et al., 2016), and EM approaches, such as RSEM (Li and Dewey, 2011), are the most common. However, both methods have the effect of decreasing signal-to-noise ratios for inDrop sequencing data. For RSEM, low-coverage 3' sequencing data contains considerable uncertainty, which RSEM is designed to pass into the count matrix. This uncertainty is normally removed by UMI-aware count based methods, although it incurs some data loss. Kallisto and Salmon, in contrast, both pseudo-align and resolve multi-alignments, but only against the transcriptome. This causes alignments from contamination that is of genomic origin to be pseudo-aligned to transcriptomic positions, producing inflated and spurious alignments for low-homology genes.

Of the high-throughput droplet-based approaches, inDrop has a unique combination of linear amplification and UMIs, which produces high fragment coverage per UMI. Although individual alignments are often ambiguously aligned to more than one location, it is often possible to look at the set of fragments assigned to an UMI and to identify a unique gene that is compatible with all the observed fragments. Here we implement an efficient method to find the unique genes that generate each fragment set. When a fragment set cannot be attributed to a specific gene, it is discarded.

Starting with all reads attributed to a cell, we begin by grouping reads according to their UMI, producing "fragment sets" S. Typically, these fragment sets represent trivial problems, such as $s_1 = \{A, A, AB\}$, a set with two unique alignments to gene A and a third ambiguous alignment to genes A and B. In this case all three observations support the gene A model, while only one observation supports the gene B model.

In cases of UMI collisions, where two mRNA molecules were captured by different primers that happen to share the same UMI sequence, this can lead to problems wherein reads from these merged fragment sets are mistakenly discarded as multi-aligning. However, because the probability of two genes sharing significant homology is low, it is usually possible to recover these molecules by first separating fragment sets into disjoint sets. For example, if a fragment set

$s_2 = \{A,\ AB,\ B,\ B,\ C,\ CD,\ ABC,\ E,\ EF,\ EF\}$,

it is broken into two disjoint groups:

$s_2 = s_3 \cup s_4;\ s_3 \cap s_4 = \varnothing$ where: $s_3 = \{A,\ AB,\ B,\ B,\ C,\ CD,\ ABC\}$ and $s_4 = \{E,\ EF,\ EF\}$.

This is biologically reasonable, as molecule collisions are the only way to reasonably obtain a group of molecules that covers two non-overlapping gene annotations.

To calculate disjoint sets efficiently, we utilize a Union-Find data structure (Aho and Ullman, 1983), which finds disjoint sets in $O(log(n))$ time. Pseudo-code is as follows:

```
summarize alignments as data = {cell=c, umi=u, gene=g}
sort data by c > u > g
for fragment_set in data[{c, u}]:
        for disjoint_set in union_find(fragment_set):
                find number_shared_genes in disjoint_set
                if number_shared_genes == 1:
                        resolve disjoint_set
```

We note that this allows us to more accurately identify the alignment rates for each gene, build better error models for barcode correction, and recover cases where reads align multiply to the same gene. More critically, it gives us the ability to recover fragments that would otherwise not be resolvable due to sequence homology, and these improved fragment counts per molecules act as significant predictors of molecule likelihood and UMI quality. We note that a similar strategy has since been published by (Klein et al., 2015) and a comparable logic underlies the concept of transcript compatibility in Kallisto (Bray et al., 2016).

We had previously created a model wherein disjoint sets with more than 1 common gene could also be disambiguated by calculating the probability of gene-gene multi-alignments from their homology, by comparing gene sequences using a Suffix Array built from the final 1000 bases of each gene. With this strategy, we could estimate the relative probability that genes were generated from each potential candidate molecule shared across all reads in the fragment set. However, the relative rarity of such events (<1% of data) combined with the additional run-time complexity of this method caused us to omit it from the production version of SEQC. This module typically resolves approximately 1M reads per hiseq lane. The result of this module is a bam or h5 file containing only unique alignments to gene features.

Molecular Identifier Correction

Errors in molecular identifiers are well-known to introduce noise in sequencing experiments (Jaitin et al., 2014), since undetected errors induce spurious increases in molecule counts. This module takes an hdf5 read store, identifies errors in UMIs, and replaces them with their corrected value. The most common approach, published in (Jaitin et al., 2014) for MARS-seq, does a very good job of detecting and removing molecule errors in inDrop (due to similarity in the Cell-seq protocol used in both technologies). This approach deletes any UMI for which a higher-abundance donor UMIs can be identified that (1) lies within a single base error (2) has higher count (3) and contain all observed alignment positions of the recipient RNA. This results in removal of approximately 20% of observed UMIs. However, we observed that this model can be overly stringent, correcting UMIs when the donor molecule has as few as one read count higher than the recipient.

We apply a modified version of the (Jaitin et al., 2014) approach, where we replace errors with corrected barcodes instead of deleting them, and where we only eliminate errors when we have adequate statistical evidence. To accomplish this, we utilize the spacer and cell-barcode whitelist to empirically estimate a per-base error UMI error rate of approximately 0.2% per base. e.g. to calculate the probability of a conversion from adenosine to cytosine, where $A \rightarrow C$ denotes this nucleotide conversion:

$$P_{A \rightarrow C} = \frac{1}{n \cdot m} \sum_{j=1}^{m} \sum_{i=1}^{n} \{1 \; if \; x_{ij \,:\, A \rightarrow C} \; else \; 0 \}$$

where x_j is a barcode, $j \in \{1, ..., m\}$ and each barcode has n bases.

To calculate the probability a target read was generated in error from a specific donor molecule, we calculate the product of the errors that could potentially convert a donor into the observed molecule. To convert, for example, ACGTACGT into TTGTACGT, having one $A \rightarrow T$ and one $C \rightarrow T$ conversion:

$$e = \{P_{A \rightarrow T}, P_{C \rightarrow T}\}$$

The probability of the above conversion is

$$P_{ACGTACGT \rightarrow TTGTACGT} = \prod_{i}^{n} e_i$$

Because there are multiple potential donors for each molecule, we calculate the conversion probability for each molecule. Assuming errors are randomly distributed, they can be modeled by a Poisson process, and Poisson rate term can be estimated from the data:

$$\lambda = n_{donor} \times P_{conversion}$$

where n_{donor} is the number of observations (reads) attributed to the donor molecule in the data. Since the sum of multiple Poisson processes is itself Poisson, the rate of conversion from each donor can be combined into a single rate λ_{agg} for each target molecule. The set of conversions s that we consider for each target molecule are all conversions that can occur with two or fewer nucleotide substitutions, in other words, all molecules within a Hamming distance $D_h \leq 2$, where D_h is a matrix of pairwise Hamming distances between barcodes.

$$s = \{\lambda_{j \rightarrow i} \; if \; D_{h, (i,j)} \leq 2\}$$

$$\lambda_{agg} = \sum_{i \in s}^{n} \lambda_i$$

Finally, given the probability of a molecule being observed via the substitution errors that are corrected by the Jaitin method, we can calculate the probability that n observations of a specific molecule x were generated via the Poisson process with rate λ_{agg} :

$$P = \frac{{\lambda_{agg}}^{x} e^{-\lambda}}{x!}$$

Cases that are very unlikely (p < 0.05) are *not* corrected. For inDrop experiments, this results in a recovery of an additional 3-5% of molecules in the data that are otherwise error-corrected. We note that this model is not applicable to all data; It is useful in this instance because we have relatively high coverage (10 reads / molecule) that allows us to evaluate our confidence in molecule observations. For lower-coverage data, it may be appropriate to err towards removing molecules instead of retaining them.

Raw Digital Expression Matrix Construction

To create a digital expression matrix, the uniquely-aligned, error-corrected hdf5 read store is de-duplicated by counting unique groups of reads with the same UMI, cell barcode, and gene annotation. A single molecule then replaces each set, and those molecules are summed to create a cells x genes matrix. scRNA-seq count matrices are often over 95% sparse, and thus are stored in matrix market format and operated on as coordinate sparse or compressed sparse row matrices. We call these count matrices “raw” count matrices because they contain all barcodes observed in an experiment.

Cell Selection and Filtering

Size Selection: Barcoding beads are loaded into inDrop at higher rates than cells in order to ensure that a high fraction of cells are encapsulated with exactly one bead. As a result, the raw count matrix contains a mixture of cell barcodes that were encapsulated with cells and cell barcodes that were encapsulated alone, but may nevertheless capture some ambient mRNA molecules that float in solution due to premature lysis or cell death in the cell solution. We separate these by finding the saddle point in the distribution of total molecule counts per barcode and excluding the mode with lower mean. In practice, we accomplish this by constructing the empirical cumulative density function of cell sizes and finding the minimum of the second derivative (Figure M1E) of the distribution. For typical inDrop runs, this results in the elimination of over 95% of the cell barcodes, but retains as many as 95% of the molecules.

Coverage Selection: Molecule size alone is not adequate to remove all barcodes that were not paired with real cells. Some barcodes appear to aggregate higher numbers of errors, and as such we often see a bimodal distribution of molecule coverage: a higher mode that represents real cells, and a smaller mode that represents aggregated errors (Figure M1F). We remove the low-count density by fitting 2-component and 1-component Gaussian mixture models to each axis and comparing their relative fits using the Bayesian information criterion. When the 2-component model’s difference in likelihood is at least 5% larger than the 1-component model, we exclude the densities with the smaller mean (Figure M1G).

Filtration of dead or dying cells: We score cells for mitochondrial RNA content, which is widely used as a proxy for cell death in scRNA-seq. We observe that a small fraction of cells contain a higher abundance of molecules annotated by this signature, as much as 20−95% of their RNA. Since inDrop does not lyse mitochondria, we reason that these are likely to be cells dying due to stress imposed on them by the inDrop procedure or prior sorting, and remove them from further analysis. This filter may be turned off for studies where apoptosis is a relevant phenotype.

Low-complexity cell filtration: Finally, we regress the number of genes detected per cell against the number of molecules contained in that cell. We observe that there are sometimes cells whose residuals are significantly negative, indicating a cell which detects many fewer genes than would be expected given its number of molecules. We exclude these cells whose residual genes per cell are more than 3 standard deviations below the mean (Figure M1H).

Information Storage & Run Time

scRNA-seq generates large volumes of data whose storage can be costly and onerous, thus we store only aligned, barcode-tagged bam files which losslessly retain all information from the original multiplex fastq files in small storage space. SEQC supports reprocessing of these files, and backwards conversion into fastq files, if users desire the ability to process their data on other platforms or reprocess with updated versions of SEQC. Additional metadata files take up nominal space, and generated count matrices are stored in matrix market sparse format in light of the sparsity of the data. SEQC requires approximately 8 hours to run on a standard 32 GB / 16 core Amazon c4.4xlarge, and costs $5.84 on on-demand or $0.88 on preemptible (spot) instances to process an inDrop, drop-seq or 10x genomics experiment. In contrast, 10x genomics commercial Cell Ranger alternative costs approximately $20, in large part due to high RAM requirements.

Data Quality Analysis of Breast Leukocytes

We applied SEQC to each of the 14 samples and 61 replicates in our data. Each sample had a minimum of 2 replicates (Table S1). Samples were sequenced such that each cell was covered by an average of 22,000 reads. Cells contained on average 15 reads per molecule, and cell saturation was 91% across all samples and replicates. On average, 20% of cells were excluded due to high mitochondrial content, a proxy for cell death and stress. Samples obtained from tissue requiring dissociation had significantly higher mitochondrial transcripts (25%) than those obtained from blood (13%) (t=2.42, p=0.018). Small numbers of low-complexity cells displaying fewer than expected genes for their molecule count were detected and removed, removing 1.2 ∓ 2.3% of total molecules. A complete summary of read abundances, genes detected, and number of cells can be found in Table S6. Having excluded non-viable cells from downstream analysis, we shifted to examining within-sample consistency across technical replicates.

Library Consistency and Quality Control

The inDrop encapsulation procedure runs 10,000 cell aliquots in series, leaving open the possibility of batch-to-batch variation within technical replicates of patient samples. To determine the magnitude of batch-to-batch variation, we compared the variation within patient replicates to between-patient comparison. Each sample was collapsed into a pseudo-bulk by summing over the cells of the digital expression matrix. We determined intra-patient consistency by determining the average pairwise Pearson correlation between pairs of samples, and compared that against the inter-patient correlations. After excluding one aberrant sample with a sample-sample correlation of 0.6, we observed Pearson correlations with a minimum of r2=0.92, a mean of 0.97, and a standard deviation of 0.02 (full pairwise correlation matrix in Figure M1I). This is in contrast to the complete pairwise correlation matrix, where the average correlation between patients is 0.72. This suggests that patient-to-patient variation is primarily biological, and that we have low technical variation between inDrop runs. These comparisons are generated automatically by SEQC, and serve as an internal control for batch variation.

Individual Sample Normalization and Clustering

To characterize the immune cells extracted from patients in this study, we began by analyzing samples independently to identify their cellular composition and cell type abundances (Figures 1C, S1C). Cells were first normalized by median library size.

The normalized data was then decomposed using randomized principle component analysis (Halko et al., 2009). We selected the number of principal components to retain using the knee point (Valle et al., 1999). This resulted in 6-10 principal components per sample. The dimension-reduced PCA projection was used as the input to PhenoGraph (Levine et al., 2015), which was used to cluster the data with default parameters (k=30 nearest neighbors). The same principle components were used to generate tSNE projections (Maaten and Hinton, 2008), which were generated with barnes-hut tSNE, implemented in the bhtsne package https://github.com/lvdmaaten/bhtsne. (Figure 1C, S1D)

Cluster Cell Type Annotation

As reference data, to facilitate the interpretation and labeling of clusters derived from PhenoGraph and Biscuit, we collected previously generated bulk gene expression profiles of sorted cells from a number of sources. To label our immune clusters we collected profiles from several published datasets on sorted immune populations (Novershtern et al., 2011) (Jeffrey et al., 2006), which provided 37 and 32 sorted populations, respectively. In addition, samples from the ENCODE consortium were added as negative controls to identify and subsequently remove contaminating stroma and tumor cells which may have infiltrated the sample due to low-level CD45 expression or auto-fluorescence. To determine the gross cell types of the clustered immune cells extracted in this experiment, we examined the correspondence between cluster centroids for PhenoGraph clusters or mean parameters for Biscuit clusters and the collected cohort of bulk profiles.

Bulk samples were library-size normalized and pre-processed by mean centering, and in the case of microarray data from (Novershtern et al., 2011) (Jeffrey et al., 2006), the data was also scaled by standard deviations. In the case of PhenoGraph clusters (Figures 1C, S1D), cells were library size normalized, clusters were mean centered, and low-expression genes with an average of fewer than 1 count per cell were removed to ensure the correlations were based on the highly expressed genes in each cell type. Then, cluster centroids were correlated with bulk profiles. In the case of Biscuit clusters (Figures 2E,F), mean parameters for each cluster were correlated with bulk profiles.

Each of the bulk profiles was marked as having derived from one of several major cell types: B-cells; T-cells (naive, central memory, cytotoxic, Treg); Monocytic cells (monocytes, dendritic cells, macrophages); Mast cells; Neutrophils; or NK-cells. The highest-scoring bulk profile for each centroid was used to categorize each cluster by its type, and types were split for downstream analysis.

Cells were also typed by examining expression of known marker genes. In this analysis, cells were scored as detecting a marker gene if the cell contained a non-zero molecule count for that gene. Each cell was corrected for its detection rate (the fraction of total genes detected in that cell) and the marker detection rate was then averaged across cells of a cluster. Markers used to type cells included NCAM1, NCR1, NKG2 (NK-cells), GNLY, PFN1, GZMA, GZMB, GMZM, GZMH (cytotoxic T, NK), FOXP3, CTLA4, TIGIT, TNFRSF4, LAG3, PDCD1 (Exhausted T-cell, T-regulatory Cell), CD8, CD3, CD4 (T-cells), IL7R (Naive T-cells), CD19 (B-cells), ENPP3, KIT (Mast cells), IL3RA, LILRA4 (plasmacytoid DC), HLA-DR, FCGR3A, CD68, ANPEP, ITGAX, CD14, ITGAM, CD33 (Monocytic Lineage). For all retained clusters, the two typing methods agreed (Figure 2H).

Gene Signature Summarization

While attempting to interpret biological signals that were observed in the studied immune cells, we realized that it was important to consider several facets of gene signature enrichment: (1) the classically studied mean value of the signature across cells in the cluster, but also (2) the marginal distribution of cell loadings across the signature and (3) the relative contribution of each gene. Therefore, for plots of signature expression by patient we began by constructing a barplot of the counts for each gene in the signature, corrected for cellular observation rate (the total number of genes observed with molecule count > 1). This displays the contribution of each gene to the signature (top panel in Figure 1E, S1E-G and 2B). The normalized values for each signature, per cell, are then summarized as a box plot to display the variation of cells in each patient (left panels). Finally, the cluster median of each gene is taken per patient, and the cluster medians are z-scored across patients. The z-scored values are plotted as a heatmap (center-right panel in Figure 1E, S1E-G and 2B), facilitating a comparison of signatures across patients.

The full lists of compiled gene signatures used throughout this paper can be found in Table S4 and may serve as a valuable resource for additional investigations. To create these lists we broadly surveyed the extant literature and manually curated consensus lists of genes to be included. The relevant literature surveyed to form these signatures includes the following:

For the M1 and M2 macrophage polarization signatures we merged gene lists from (Sica and Mantovani, 2012); (Biswas and Mantovani, 2010); (Bronte et al., 2016) (Ugel et al., 2015) (Gabrilovich, 2017). For other myeloid-specific signatures we used (Villani et al., 2017) (pDCs, AXL/SIGLEC6 DCs, CD141/CLEC9A DCs, CD11C_A DCs, CD1C-/CD141- DCs, CD1C_B DCs, New Monocytes 1, New Monocytes 2, CD14+CD16- Classical Monocytes, and CD14+CD16+ Non-Classical Monocytes); and (Gesta et al., 2007), (Perera et al., 2006), (Farmer, 2006; Lefterova and Lazar, 2009) (Lipid Mediators).

For T Cell-specific signatures we used (Wherry and Kurachi, 2015), (Wherry, 2011), and (Schietinger et al., 2012) (Exhaustion and Anergy); (Glimcher et al., 2004) (Cytolytic Effector Pathway); and (Smith-Garvin et al., 2009), (Chtanova et al., 2005), and (Adam Best et al., 2013) (T Cell Activation).

For gene signatures used across cell types we used (Mantovani et al., 2008), and (Grivennikov et al., 2010) (Pro and Anti-Inflammatory); (Platanias, 2005) (Type I and II Interferon Responses); (Ho et al., 2015) (glucose deprivation); (Benita et al., 2009; Makino et al., 2003) (Hypoxia/HIF Regulated); (Moreno-Sánchez et al., 2009), (Caton et al., 2010; Funes et al., 2007; Mues et al., 2009), (Beale et al., 2007) (Glycolysis, Gluconeogenesis, TCA Cycle, Pentose Phosphate Pathway, and Glycogen Metabolism), and (Whitfield et al., 2002) (G1/S).

Biscuit Clustering and Normalization for Merging Samples

When attempting to merge immune systems from multiple patients, we observed that in some cases, cells are more similar to others from the same patient than they are to those of the cell type. Figure 2A (left) shows scRNA-seq data from 9K immune cells from 4 breast cancer patients after normalization of cells to median library size, suggesting large differences between patients. Moreover, the tSNE projection does not suggest diverse subpopulations and structure beyond two main cell types of lymphoid and myeloid cells (Figure 2A left).

The differences across patients are likely caused by a large number of factors, both technical and biological. Technical factors include differences in machine, enzyme activity, lysis efficiency or experimental protocol. These samples were also subject to operational variation during the clinical resection, transport, and handling. These factors all impact cell viability, which in turn affects the single cell RNA-seq library preparation, in particular molecular capture rate and sampling. Because molecular capture is a binary event, and the capture rates are very low, these technical variations often determine whether a given gene feature is observed in the data for a given cell.

These more technical artifacts, particularly in capture rate, are confounded with biological differences. This is particularly challenging in the case of immune cells, where activated cells have higher transcription rates. Thus, the number of captured molecules may be higher in activated immune cells and therefore activation of immune cells can be convolved with total counts (Blackinton and Keene, 2016; Cheadle, 2005; Singer et al., 2016). As a result, methods that normalize all cells together may remove critical cell type-specific biological variation in the systems of interest such as immune activation and environmental response. Indeed, we see large differences in the number of activated T-cells across patients (Figure 2B), with more activated T-cells in the Triple Negative subtype as expected (Dushyanthen et al., 2015). Hence, normalizing by library size will likely remove these biological variations.

To solve this problem we developed and applied the method “Biscuit” (short for Bayesian Inference for Single-cell ClUstering and ImpuTing) to simultaneously cluster cells and normalize according to their assigned clusters (Figure M2). This is done through incorporating parameters denoting cell-specific technical variation into a Hierarchical Dirichlet Process mixture model (HDPMM) (Görür and Rasmussen, 2010) (Figure M2A). This allows for inference of cell clusters based on similarity in gene expression as well as in co-expression patterns, while identifying and accounting for technical variation per cell (Figure M2B,C). Two key ideas that power Biscuit are the use of gene co-expression as a more robust means to identify cell types, and the normalization of each cell type separately to better account for cell type-specific effects on technical variation. The main idea behind the use of co-expression is that cell types not only share similar mean expression, but also share similar co-expression patterns (covariance) between genes. While mean expression can be more sensitive to capture efficiency, covariation is more robust to such effects. This similarity in co-variation can be used to improve normalization and in turn improve the clustering, through the learning of cluster specific parameters.

By jointly performing normalization and clustering, we retain biological heterogeneity and avoid biases that result from independent clustering and normalization, and instead are able to match cells to clusters of the same cell type from different patients which may have very different sampling rates. Figure 2A (right) shows the same data from 4 tumors after normalization with Biscuit. Note that Biscuit does not use any information on sample IDs in the normalization, and normalization is only driven by cluster assignments. The Biscuit-normalized data shows that the differences in library-size normalized data were largely artifacts of normalization and batch effects. Furthermore, data from Biscuit shows richer structure suggesting diversity in cell types. Hence, we then applied Biscuit to data from all 8 tumors to infer the full diversity of immune cell types in the breast tumors, which identified 67 clusters indicating significant diversity in both lymphoid and myeloid cell types (Figure 2C).

Summary of Biscuit model: Starting with the count matrix $X = [\vec{x_1}, ... , \vec{x_n}]$ where each column $\vec{x_j}^{(1,,,,d)}$ contains the expression (number of unique mRNA molecules) of d genes in cell j, the model assumes the log of counts $\vec{l_j} = log(0.1 + \vec{x_j})$ for each cell $j = (1, ..., n)$ follow a multivariate Normal distribution: $\vec{l_j} \mid z_j = k \sim N(\alpha_j \vec{\mu_k}, \beta_j \Sigma_k)$ where z_j denotes the assignment of cell j to cluster k, and $\vec{\mu_k}, \Sigma_k$ are the mean and covariance respectively of the k-th mixture component (cluster), and scalars α_j, β_j are cell-dependent scaling factors used for normalization. Note that the assumption of log Normality has been verified using model mismatch and Lilliefors tests (Prabhakaran, S., Azizi, E., Carr, A., & Pe'er, D, 2016). Within a Bayesian model setting, we assign conjugate prior distributions to the parameters, namely a symmetric Dirichlet prior of the order K, a conjugate-family prior over μ_k as Normal and for Σ_k as Wishart (Díaz-García et al., 1997). A noninformative Normal prior is set for α_j and Inverse-gamma for β_j. The full model specification is thus as below:

$$\{\vec{l}\}_j^{(1,...,d)} \mid z_j = k \sim N(\alpha_j \vec{\mu_k}, \beta_j \Sigma_k)$$

$$\vec{y_j} \sim N(\vec{\mu_k}, \Sigma_k)$$

$$\vec{\mu_k} \sim N(\vec{\mu'}, \Sigma')$$

$$\Sigma_k{}^{-1} \sim Wish\,(H'^{-1}, \sigma')$$

$$\vec{\mu}' \sim N(\vec{\mu}'', \Sigma'')$$

$$\Sigma'^{-1} \sim Wish\,(d, \tfrac{1}{d\Sigma''})$$

$$H' \sim Wish\,(d, \tfrac{1}{d}\Sigma'')$$

$$\sigma' \sim InvGamma(1, \tfrac{1}{d}) - 1 + d$$

$$z_j|\vec{\pi} \sim Mult\,(z_j|\vec{\pi})$$

$$\vec{\pi}|\ \phi,\ K \sim Dir(\vec{\pi}|\phi/K, ..., \phi/K)$$

$$\phi^{-1} \sim Gamma(1, 1)$$

$$\alpha_j \sim N(\nu, \delta^2)$$

$$\beta_j \sim InvGamma\,(\omega, \theta)$$

where $\vec{\mu}'', \Sigma''$ are the empirical mean and covariance across all cells, which are used to allow priors to adapt to different datasets (Figure M2B).

Parameters are inferred through a scalable Gibbs algorithm using the Chinese restaurant process (CRP) (Pitman, 2006) which also infers the number of clusters (K) (Figure M2C). The conditional posterior distributions for model parameters $\{\vec{\pi}, \vec{\mu}_k, \Sigma_k, \vec{\alpha}, \vec{\beta}, \vec{z}, \vec{\mu}', \Sigma', H'\}$ have analytical forms which we derived in (Prabhakaran, S., Azizi, E., Carr, A., & Pe'er, D, 2016).

Final cluster assignments were estimated from the mode of inferred distribution for assignment of cells to clusters $(z_j;\ \forall j)$ after removing a burn-in period. The mean of Gibbs samples for cluster-specific parameters $(\mu_k, \Sigma_k;\ \forall\, k = 1, ..., K)$ and cell-specific parameters $(\alpha_j, \beta_j;\ \forall j)$ were used for further analysis and interpretation.

The goal of normalization is to transform the data from $\vec{l}_j$ to $\vec{y}_j$ in which the expression is corrected for cell-specific factors α_j, β_j using a linear transformation $\vec{y}_j = A\,\vec{l}_j + b$ such that imputed expression for cell j follows $N(\mu_k, \Sigma_k)$ and hence all cells assigned to the same cluster follow the same distribution after correction. One transformation satisfying the above distributions for $\vec{l}_j$ and $\vec{y}_j$ is $A = \frac{I}{\sqrt{\beta_j}}$, $b = (1 - \alpha_j A)\,\vec{\mu}_k$.

This transformation also imputes dropouts in a cell by using the parameters of the cluster it has been assigned to. The use of covariance parameter in the model ensures that intra-cluster heterogeneity is preserved after imputing. We show a systematic evaluation of the algorithm performance (on synthetic and real single cell data), its robustness, as well as the ability of this method to impute dropouts in (Prabhakaran, S., Azizi, E., Carr, A., & Pe'er, D, 2016).

Biscuit Implementation: The joint distribution of Biscuit with cluster- and cell-specific parameters is non-conjugate. Although inference of these parameters via MCMC-Gibbs is possible as presented in (Prabhakaran, S., Azizi, E., Carr, A., & Pe'er, D, 2016), overall runtime of the algorithm was much slower due to multiple covariance matrix inversions leading to slower chain convergence. In order to reduce the number of matrix inversions and enable faster chain mixing, we make use of the conjugate prior for the multivariate Gaussian, namely the Normal-inverse Wishart distribution, where the cluster means and covariances can be jointly inferred (Murphy, 2007).

The Normal-inverse-Wishart (NIW) distribution is the conjugate prior to the multivariate Gaussian distribution with unknown mean and covariance. It is a multivariate four-parameter continuous distribution with probability density function defined as follows:

$$f(\vec{\mu_k}, \Sigma_k \mid \vec{l_j}, \vec{\mu_0}, \Lambda_0, \rho_0, \kappa_0, \alpha_j, \beta_j) \sim NIW(\vec{\mu_k}, \Sigma_k \mid \vec{\mu}', \Lambda', \rho', \kappa', \alpha_j, \beta_j)$$

The inference equations for parameters based on NIW also have closed-forms:

$$\vec{\mu}' = \kappa_0 \vec{\mu_k} + n_k \,, \quad \kappa' = \kappa_0 + n \,, \quad \rho' = \rho_0 + n$$

$$\Lambda' = \Lambda_0 + \sum_{j=1}^{N} (\vec{l_j} - \vec{\bar{l}})(\vec{l_j} - \vec{\bar{l}})^T + \kappa_0 \vec{\mu_0} \vec{\mu_0}^T + \kappa' \vec{\mu}' \vec{\mu}'^T$$

$$f(\Sigma_k \mid \vec{l_j}) \sim median\,(\beta_{j\,:\,z_j \in k}) * InverseWish\,(\Lambda'^{-1}, \rho')$$

$$f(\vec{\mu}_k \mid \vec{l_j}) \sim median\,(\alpha_{j\,:\,z_j \in k}) * Student\ t\,(\vec{\mu}', \tfrac{\Lambda'}{\kappa'(\rho' - d + 1)})$$

where $\vec{\mu_0} \in R^d$, $\lambda > 0$, $\Sigma_0 \in R^{d \times d}$ and is positive semi-definite and $\nu > (d-1)$. This scalable implementation of Biscuit in R can be found in:

https://github.com/sandhya212/BISCUIT_SingleCell_IMM_ICML_2016. The code is parallelized at multiple stages - the NIW-based inference engine works in parallel across blocks of genes ordered through the Fiedler vector. Next, cell assignments per block are consolidated into a confusion matrix via further parallelisation and the overall cluster means and covariances are computed using the law of total expectation, also performed in parallel. A dataset of 3000 cells and 1000 genes takes under two hours on a Cloud platform such as Amazon Web Services (AWS), which is at least 6 times faster than the implementation of the initial model. For the full dataset, the DPMM dispersion parameter was set to 100, gene batches were set to 50, and the number of iterations was set to 150.

Entropy Metric to Evaluate Batch Effect Correction

To evaluate Biscuit's ability to correct batch effects across data from all eight tumors (Figure 2C) and match immune subtypes across the tumors, we devised an entropy-based metric that quantifies the mixing of the normalized data across samples. The entropy-based metric is computed as follows: We constructed a k-NN graph (k=30) on the normalized data using Euclidean distance and computed the distribution of patients (tumors) $m = 1, ..., 8$ in the neighborhood of each cell j, denoted as q_j^m. Then we computed Shannon entropy

$H_j = -\sum_{m=1}^{8} q_j^{\ m} \, log \, q_j^{\ m}$ as a measure of mixing between patients, resulting in one entropy value H_j per cell j. High entropy indicates that the most similar cells come from a well mixed set of additional tumors, whereas low entropy indicates that the most similar cells largely come from the same tumor. Prior to Biscuit, most cells in the data had low entropy values, with 40% of the cells residing in a neighborhood of cells purely from the same tumor. We compare the distribution of entropies across all cells from all 8 tumors, before and after Biscuit, which reveals that the median of entropy shifts significantly towards higher mixing of samples after processing with Biscuit (Mann-Whitney U-test: U=1.7721e+09, p=0; Figure 2D). Thus, we conclude that Biscuit substantially corrected batch effects in this data.

Quantification of Cell Type Enrichment in Tissues

To calculate cell type enrichments in each tissue (Figure 3B), we normalized each tissue to have equal cell count and created a 2-factor contingency table of cell types versus tissues. We then calculated χ^2 enrichments for each tissue type and reported the test statistic and p-value.

Next, we wished to identify which pair of tissues had the greatest phenotypic overlap. To assess similarity in phenotype, we constructed a kk-nearest neighbor graph (with k=10) in a uniformly selected subset of n=3000 cells from each tissue, reasoning that a cell's closest neighbors in this high-dimensional embedding are the cells with the closest phenotypes. We then examined the overlap between each pair of tissues u and v, where n is the number of cells in the subset, k is the number of neighbors, and $u, \ v \ \in \{tumor, \ normal, \ lymph\ node, \ blood\}$:

$$o_{u,v} = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{k} \{1 \ if \ \omega_i = u \ \ and \ \omega_j = \ v \qquad else \ \ 0\}$$

with ω_i denoting the tissue for cell i and $j = 1, ..., k$ denotes the neighbors of cell i . We build a null distribution from all overlaps between all pairs of tissues, and identified tumor and normal as having the highest frequency of being co-identified as neighbors. To test if this enrichment was significantly higher than the other shared-neighbor overlaps, we calculated the z-score of $o_{tumor, \ normal}$ compared to the population of all pairwise overlaps (z=2.68), for which a z-test confers a p-value of p=1.4e-4. Thus, we conclude that tumor and normal have the highest phenotypic overlap between any pair of tissues in our data. Both z-scores and p-values are reported. The above statistics are reported under the section “Environmental impact on the diversity of immune phenotypic states”.

Creating a Global Immune Atlas using Biscuit

To generate a global atlas of immune cell types, we combined samples from all patients and tissues by applying Biscuit to the full set of *n*=62024 cells and *d*=14875 genes, resulting in a global atlas of *K*=95 clusters (Table S2) in which *n*=57143 cells had statistically significant cluster assignments. The remainder of cells had low library size and were hence removed from further analysis.

A subset of these clusters were identified as probable cancer or stromal populations through correlation with bulk gene expression datasets and marker gene expression. While these non-immune clusters may be of significant interest in their own right, they were beyond the scope of this paper and were therefore excluded from downstream analysis, leaving 47,016 cells in 83 clusters (Table S2).

Biscuit is a probabilistic model; hence, we inferred a probability distribution for all parameters. Figure S2A shows the posterior of assignment of cells to clusters, which shows a broader distribution for most naïve T cell clusters, suggesting that these clusters are less distinct as compared to other clusters. Other cell types exhibited a distinct, well separated mode. For simplicity, all further analysis is based on the mode of the probability of assignment of cells to clusters, such that each cell is assigned to one cluster only (Figure 2E).

Biscuit captures cell-type specific scaling factors using parameters α_j, β_j. Figure S2G shows that α_j values capture variation in library size (sum of total counts per cell j). Moreover, we observe different relationships with library size based on assigned clusters, highlighting the cluster-specific normalization that is done in Biscuit. We also observe prominent differences in distributions of α_j, β_j parameters across cells from different patients (Figure S2H,I), emphasizing the differences resulting from technical factors across patients that are captured in Biscuit. The α_j, β_j parameters represent each cell's translational and rotational shifts within the Multivariate Normal distribution associated with the cluster assigned to that cell, by scaling the moments. By correcting these shifts, cells assigned to the same cluster will be normalized and follow the same distribution $N(\vec{\mu_k}, \Sigma_k)$.

Cluster Robustness

To evaluate cluster robustness, we performed 10-fold cross-validation, independently clustering and normalizing on random subsets of data. For each of 10 subsets, we ran Biscuit to obtain a set of clusters. To compare the results across the 10 subsets, we computed the confusion matrix, which indicates the probability of each pair of cells j, j' being assigned to the same cluster: $P(z_j = z_{j'})$. Figure S2C illustrates box plots for the probabilities of co-clustering (across 10 subsets) for every pair of cells that are assigned to the same cluster in the analysis of the full dataset. The average co-clustering probability in each cluster ranges between 92%-100%, showing remarkable robustness of clusters.

Mixing of Samples in Clusters

We observed that clusters displayed differing amounts of mixing between samples (Figure 2G). To further quantify the exact degree of mixing (between patients) in each cluster, we defined an entropy-based metric. We used bootstrapping to correct for cluster size (which ranged from over 8900 cells to just over 30 cells), such that we uniformly sampled 100 cells from each cluster and computed the distribution of patients across these cells, and then computed the Shannon entropy for this distribution. We repeated this procedure 100 times for each cluster, to achieve a range of entropy values per cluster. Figure S2D shows box plots for entropy values for each cluster, with the order of clusters based on their mean entropy. Clusters with entropy of 0 denote entirely patient-specific clusters. Figure S2D shows that there is a continuous range of entropies, and thus a full range of sample specificity versus mixing, across clusters.

Distances between Clusters

Biscuit provides a parametric characterization of each cluster using a multivariate Normal distribution, which allows us to directly compare the distributions that define these clusters. While Euclidean distances are defined for vectorial objects and operate under a Cartesian coordinate system, Euclidean distance with non-vectorial objects such as probability distributions requires embedding them in Euclidean space. Such embeddings are non-unique and lead to loss of information. It is therefore advisable to use the non-vectorial objects as is and to work with the objects' pairwise similarities or distances instead. One such distance metric, which is effective at comparing pairwise probability distributions, is the Bhattacharyya distance (BD) (Bhattacharyya, 1990).

We defined distances between each pair of clusters k, k' with distributions p_k and $p_{k'}$ as $BD = -log(BC(p_k, p_{k'}))$ where BC is the Bhattacharyya coefficient measuring similarity (overlap) of the distributions. We use the BD to compute distances between pairs of inferred clusters' moments to create the Bhattacharyya kernel. The Bhattacharyya kernel has closed forms for any exponential distribution including the (multivariate) Gaussian distribution (Jebara, T., Kondor, R., & Howard, A., 2004), which is Biscuit's underlying data-generation distribution. For the case of multivariate normal distributions: $p_k \sim N(\vec{\mu}_k, \Sigma_k)$ and $p_{k'} \sim N(\vec{\mu}_{k'}, \Sigma_{k'})$:

$$D_B = \tfrac{1}{8}(\vec{\mu}_k - \vec{\mu}_{k'})^T\, \Sigma^{-1}(\vec{\mu}_k - \vec{\mu}_{k'}) \; + \; \tfrac{1}{2}log(\frac{det\,\Sigma}{\sqrt{det\,\Sigma_k det\,\Sigma_{k'}}}) \text{ where } \Sigma = (\Sigma_k + \Sigma_{k'})/2 .$$

Figure S2B shows the heatmap of pairwise distances between all pairs of clusters.

A geometric interpretation of BD is that, via its cosine formulation, the distance subsumes a full hypersphere and the centre of the hypersphere is the centroid (mean) of the cluster, whereas the Euclidean distance only covers a quarter of the hypersphere with the center at the origin.

Contribution of Covariance in Defining Clusters

We used the above Bhattacharyya distance (BD) metric to study the contribution of Biscuit's covariance parameters to characterizing clusters of different cell types. First, we computed the BD distance between pairs of clusters of the same type (T, monocytic, NK, B) and compared these to distances between pairs of clusters of different cell types (e.g. a T cell cluster and a monocytic cluster). Figure S2E shows violin plots for distances between pairs of clusters with dots (overlaid on violins) representing cluster pairs; violins are sorted based on median distance. As reference, we also split each cluster into two halves and computed the empirical BD between two splits (shown at the left end in Figure S2E). We observe that, overall, pairs of clusters of different types are more distant than pairs of clusters from the same type, as expected.

We then computed these same pairwise distances while removing the contribution of mean parameters for each cluster, via setting $\vec{\mu}_k - \vec{\mu}_{k'} = 0$ and computing the distance only based on covariance parameters of the pair of clusters $\Sigma_k, \Sigma_{k'}$ (Figure S2F). We observed that pairs of T cell clusters or monocytic clusters still show prominent distances, and therefore covariance parameters have a crucial role in defining these clusters.

Defining Phenotypic Volume

One of the main aims of this paper was to compare immune phenotypes across tissues, including peripheral blood and lymph node tissue as references in which the immune phenotypes typically present are better understood. The global atlas from merged tissues revealed that clusters differ appreciably in their proportions across tissues (Figure 3A-C). For example, we observed 3 blood-specific clusters with CD4+ naive characteristics, as well as a lymph node-specific T cell cluster. Conversely, we observed a large range of shared clusters between normal and tumor tissues (Figure 3B; Table S2). Interestingly, the clusters observed in normal tissue were a subset of those observed in tumor tissue. We also observed an increase in the variance of gene expression in tumor compared to normal tissue (Figure 3D,E; S3A,B), which prompted us to explore further the range of phenotypic diversity between normal and tumor tissue. We were especially interested to find whether the observed increase in variance of expression is due to activation of additional processes in tumor that are independent of those active in normal.

Hence, to quantify the diversity of phenotypes, we defined a metric of phenotypic volume or space occupied by a cell type (e.g. T, NK, or monocytic), such that if more independent phenotypes are active, the total volume spanned by the phenotypes would be larger than an alternative case wherein phenotypes are dependent.

We therefore defined “phenotypic volume” (V) for a subpopulation of cells as the determinant of the gene expression covariance matrix in that subpopulation, which considers covariance between all gene pairs in addition to their variance. The (symmetric) covariance matrix can be written as $\Sigma = [\vec{s_1}, ..., \vec{s_d}]$ where $\vec{s_i}$ for $i = 1, ..., d$ is a vector containing covariance between gene i and all other genes. Its determinant $det(\Sigma)$ is equal to the volume of a parallelepiped spanned by vectors of the covariance matrix (Tao and Vu, 2005).

For example, if the covariance values between a gene i and other genes is very similar to that of another gene i' and other genes, such that $\vec{s_i}, \vec{s_{i'}}$ are dependent, gene i' does not add to the volume. Extending this to all genes, we sought to evaluate whether the increase in expression variances (Figure 3D,E) are associated with phenotypes activated in tumor that are independent from those in normal tissue, i.e. are novel independent phenotypes observed in tumor that suggest additional mechanisms and pathways being activated in tumor.

For example, in a simplified case with only two phenotypes the determinant, which is equal to the area of the parallelogram spanned by two vectors representing the phenotypes, is larger if the phenotypes are independent, but would be equal to zero if they are dependent. With more than two phenotypes, we are then interested in measuring the volume of the parallelepiped spanned by these phenotypes. The (pseudo-)determinant can also be computed as the product of nonzero eigenvalues of the covariance matrix:

$$V = det(\Sigma) = \prod_{e=1}^{E} \lambda_e = \lambda_1 \lambda_2 \ldots \lambda_E$$

To quantify the change in phenotypic volume from normal to tumor, we computed this volume metric for each major cell type of T, monocytic, and NK cells. To correct for the effect of differences in the number of cells across cell types and tissues, we uniformly sampled 1000 cells with replacement from each cell type per tissue and computed the empirical covariance between genes based on imputed expression values for that subset of cells. This was followed by singular-value decomposition (SVD) of each empirical covariance matrix and computation of the product of nonzero eigenvalues as stated in the equation above. B cells were not included in this comparison due to the very small number of B cells in normal tissue. Given the high number of dimensions (genes), the volumes were normalized by the total number of genes (d). For robustness, this process was repeated 20 times to achieve a range of computed volumes for each cluster in each tissue, which are summarized with box plots (Figure 3F) showing statistically significant expansions of volume in tumor compared to normal in all three cell types (Mann-Whitney U-test p=0 for all three cell types).

Diffusion Component Analysis

We used diffusion maps (Coifman et al., 2010) as a nonlinear dimensionality reduction technique to find the major non-linear components of variation across cells. We computed diffusion components in each cell type separately using the Charlotte Python package, which implements diffusion maps as described in (Coifman et al., 2005). To account for differences in cell density and cluster size, we used a fixed perplexity Gaussian kernel with perplexity 30, with symmetric Markov normalization and t=1 diffusion steps. We selected t=1 diffusion steps because this approximates diffusion of information for each cell through its 20 nearest neighbors in our data. This is a conservative value, but we wanted to ensure that information did not diffuse beyond the borders of our smallest cluster (30 cells). To focus on processes explaining variation specifically within major cell types, we performed diffusion map analyses separately on all cells labeled as T cells and myeloid cells, respectively.

In the case of T cells, the first two diffusion components identified two isolated clusters (Figure 3A): The first was cluster 9, which is quite distinct from other T cell clusters as measured by Bhattacharyya distance (Figure S2B) and shows characteristics similar to NKT cells (Table S3); the second was cluster 20, which is a blood-specific naive T cell cluster predominantly from one patient (Table S2). Since these two clusters were very distant from other T cell clusters according to a variety of comparison metrics, the two components corresponding to them were removed from further analysis, as we wished to focus on studying components that are informative of heterogeneity across multiple clusters.

The next components were labeled as T cell activation, Terminal Exhaustion, and Hypoxia, respectively, (Figure 4A) as they were most highly correlated with the corresponding gene signatures (Table S4). The subsequent component is labeled as Tissue Specificity, as it separates cells primarily on the basis of their tissue of origin and helps explain heterogeneity in T cells across tissues.

In the separate analysis of myeloid cells, several isolated clusters, namely neutrophils and mast cells, were not considered in the analysis, with the goal of identifying components explaining variation across multiple monocytic clusters (Figure 6B, S6A). The first component identified a trajectory of macrophage (TAM) activation. Though a remaining component segregated cluster 41 (labeled plasmacytoid dendritic cells - pDCs), it was retained in our analysis due to several generally illuminating facets of the component, including the stark contrast between pDCs primarily found in the lymph node and mDCs found in other tissues. The third diffusion component explained heterogeneity across tissue resident monocytes, and the fourth component further separated monocytes by tissue, delineating a trajectory from monocytes in blood to activated monocytes in tissue.

Significance of Differences in Covariances in Raw Data

To verify that the differing covariance patterns in Figures 5 and 7 were not the result of computational modeling decisions, we tested the difference in covariance in raw unnormalized data (prior to Biscuit). As the raw data involves significant amount of dropouts, co-expression patterns and their signs cannot be easily visualized or inferred.

Hence, we performed hypothesis testing accounting for the level of dropouts by comparing the observed empirical covariance between a pair of genes $i,\ i'$ to a null distribution for the gene pair in which co-expression patterns are removed. We assume the null hypothesis to be the case where covariance between a specific gene pair for a given cluster is the same across all clusters.

Specifically, to test whether $cov(\vec{x_i}, \vec{x_{i'}})$ in a cluster k, with $\vec{x_i}, \vec{x_{i'}}$ being expressions of genes i, i' across cells assigned to cluster k, is significantly different from that in all other clusters, we used bootstrapping and permutation testing as follows: We started by generating a null distribution for the covariance between a pair of genes by first uniformly sampling a subset of cells from all clusters, with the subset being the same size as cluster k. Then, to further remove existing structures of co-expression in cells, we permuted the cell labels for gene i' (while retaining labels for gene i) and computed empirical covariance between the two genes in this subset of "scrambled" cells. We repeated this on 10,000 subsets to achieve a null distribution of $cov(\vec{w_i},\ \vec{w_{i'}})$ where $w_i, w_{i'}$ are the expressions of gene i, i' in the sets of scrambled cells. We then compared the observed $cov(\vec{x_i}, \vec{x_{i'}})$ (marked with a star in Figure S5A, S7A) to the null distribution, which was rejected for that pair of genes if p-value<0.05 indicating that the covariance is significantly different in cluster k compared to all other clusters.

We concluded that the signal is also apparent in raw unnormalized data for all the aforementioned clusters and we observe a range of covariance values with different signs between GITR and CTLA4 across Treg clusters (Figure S5A), and similarly different values in covariance between MARCO and CD276 in TAM clusters (Figure S7A).

Continuity of Cells along Components

We were interested to know whether cells show continuity as opposed to defined cell states along various diffusion components. For example, we were interested to know whether T cells exhibit defined states with different activation levels. For this, we computed the distribution of cells projected on each diffusion component and then used Hartigan's Dip Test (Hartigan and Hartigan, 1985) to test whether the distribution of cells is unimodal (broad continuum of cells) or alternatively multimodal (representative of multiple defined states) with p<0.01. In Figure S4A, we observe that in the case of the T Cell Activation component, the null hypothesis of unimodality is not rejected, indicating that the distribution of cells is similar to a broad unimodal distribution as opposed to a multimodal distribution with defined states. In contrast, other components (such as the Tissue Specificity Component) exhibit multimodal distributions with distinct modes implying distinct states (in this case corresponding to various tissues).

In the case of myeloid cells, the null hypothesis of unimodality is rejected in all diffusion components, indicating that myeloid cells lie in distinct states along all major components explaining variation across cells that were analyzed (Figure S6A).

DATA AND SOFTWARE AVAILABILITY

The sequencing data presented in this paper will be available for download on GEO data repository.

SEQC is available on https://github.com/ambrosejcarr/seqc.git and Biscuit is available on https://github.com/sandhya212/BISCUIT_SingleCell_IMM_ICML_2016.

References

Adam Best, J., Blair, D.A., Knell, J., Yang, E., Mayya, V., Doedens, A., Dustin, M.L., Goldrath, A.W., and The Immunological Genome Project Consortium (2013). Transcriptional insights into the CD8+ T cell response to infection and memory T cell formation. Nat. Immunol. *14*, 404.

Aho, A.V., and Ullman, J.D. (1983). Data structures and algorithms.

Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data.

Beale, E.G., Harvey, B.J., and Forest, C. (2007). PCK1 and PCK2 as candidate diabetes and obesity genes. Cell Biochem. Biophys. *48*, 89–95.

Benita, Y., Kikuchi, H., Smith, A.D., Zhang, M.Q., Chung, D.C., and Xavier, R.J. (2009). An integrative genomics approach identifies Hypoxia Inducible Factor-1 (HIF-1)-target genes that form the core response to hypoxia. Nucleic Acids Res. *37*, 4587–4602.

Bhattacharyya, A. (1990). On a Geometrical Representation of Probability Distributions and its use in Statistical Inference. Calcutta Statist. Assoc. Bull. *40*, 23–49.

Biswas, S.K., and Mantovani, A. (2010). Macrophage plasticity and interaction with lymphocyte subsets: cancer as a paradigm. Nat. Immunol. *11*, ni.1937.

Blackinton, J.G., and Keene, J.D. (2016). Functional coordination and HuR-mediated regulation of mRNA stability during T cell activation. Nucleic Acids Res. *44*, 426–436.

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

Bronte, V., Brandau, S., Chen, S.-H., Colombo, M.P., Frey, A.B., Greten, T.F., Mandruzzato, S., Murray, P.J., Ochoa, A., Ostrand-Rosenberg, S., et al. (2016). Recommendations for myeloid-derived suppressor cell nomenclature and characterization standards. Nat. Commun. *7*, 12150.

Caton, P.W., Nayuni, N.K., Kieswich, J., Khan, N.Q., Yaqoob, M.M., and Corder, R. (2010). Metformin suppresses hepatic gluconeogenesis through induction of SIRT1 and GCN5. J. Endocrinol. *205*, 97–106.

Cheadle, C. (2005). Stability Regulation of mRNA and the Control of Gene Expression. Ann. N. Y. Acad. Sci. *1058*, 196–204.

Chtanova, T., Newton, R., Liu, S.M., Weininger, L., Young, T.R., Silva, D.G., Bertoni, F., Rinaldi, A., Chappaz, S., Sallusto, F., et al. (2005). Identification of T Cell-Restricted Genes, and Signatures for Different T Cell Responses, Using a Comprehensive Collection of Microarray Datasets. The Journal of Immunology *175*, 7837–7847.

Coifman, R., Coppi, A., Hirn, M., and Warner, F. (2010). Diffusion Geometry Based Nonlinear Methods for Hyperspectral Change Detection.

Coifman, R.R., Lafon, S., Lee, A.B., Maggioni, M., Nadler, B., Warner, F., and Zucker, S.W. (2005). Geometric diffusions as a tool for harmonic analysis and structure definition of data: diffusion maps. Proc. Natl. Acad. Sci. U. S. A. *102*, 7426–7431.

Díaz-García, J.A., Jáimez, R.G., and Mardia, K.V. (1997). Wishart and Pseudo-Wishart Distributions and Some Applications to Shape Theory. J. Multivar. Anal. *63*, 73–87.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21.

Dushyanthen, S., Beavis, P.A., Savas, P., Teo, Z.L., Zhou, C., Mansour, M., Darcy, P.K., and Loi, S. (2015). Relevance of tumor-infiltrating lymphocytes in breast cancer. BMC Med. *13*.

Faircloth, B.C., and Glenn, T.C. (2012). Not all sequence tags are created equal: designing and validating sequence identification tags robust to indels. PLoS One *7*, e42543.

Farmer, S.R. (2006). Transcriptional control of adipocyte formation. Cell Metab. *4*, 263–273.

Funes, J.M., Quintero, M., Henderson, S., Martinez, D., Qureshi, U., Westwood, C., Clements, M.O., Bourboulia, D., Pedley, R.B., Moncada, S., et al. (2007). Transformation of human mesenchymal stem cells increases their dependency on oxidative phosphorylation for energy production. Proc. Natl. Acad. Sci. U. S. A. *104*, 6223–6228.

Gabrilovich, D.I. (2017). Myeloid-Derived Suppressor Cells. Cancer Immunol Res *5*, 3–8.

Gesta, S., Tseng, Y.-H., and Ronald Kahn, C. (2007). Developmental Origin of Fat: Tracking Obesity to Its Source. Cell *131*, 242–256.

Glimcher, L.H., Townsend, M.J., Sullivan, B.M., and Lord, G.M. (2004). Recent developments in the transcriptional regulation of cytolytic effector cells. Nat. Rev. Immunol. *4*, 900–911.

Görür, D., and Rasmussen, C.E. (2010). Dirichlet Process Gaussian Mixture Models: Choice of the Base Distribution. J. Comput. Sci. Technol. *25*, 653–664.

Grivennikov, S.I., Greten, F.R., and Karin, M. (2010). Immunity, inflammation, and cancer. Cell *140*, 883–899.

Halko, N., Martinsson, P.-G., and Tropp, J.A. (2009). Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions.

Hartigan, J.A., and Hartigan, P.M. (1985). The Dip Test of Unimodality. Ann. Stat. *13*, 70–84.

Ho, P.-C., Bihuniak, J.D., Macintyre, A.N., Staron, M., Liu, X., Amezquita, R., Tsui, Y.-C., Cui, G., Micevic, G., Perales, J.C., et al. (2015). Phosphoenolpyruvate Is a Metabolic Checkpoint of Anti-tumor T Cell Responses. Cell *162*, 1217–1228.

Jaitin, D.A., Kenigsberg, E., Keren-Shaul, H., Elefant, N., Paul, F., Zaretsky, I., Mildner, A., Cohen, N., Jung, S., Tanay, A., et al. (2014). Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. Science *343*, 776–779.

Jebara, T., Kondor, R., & Howard, A. (2004). Probability product kernels. J. Mach. Learn. Res. *5*, 819–844.

Jeffrey, K.L., Brummer, T., Rolph, M.S., Liu, S.M., Callejas, N.A., Grumont, R.J., Gillieron, C., Mackay, F., Grey, S., Camps, M., et al. (2006). Positive regulation of immune cell function and inflammatory responses by phosphatase PAC-1. Nat. Immunol. *7*, 274–283.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S.L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. *14*, R36.

Klein, A.M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D.A., and Kirschner, M.W. (2015). Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell *161*, 1187–1201.

Lefterova, M.I., and Lazar, M.A. (2009). New developments in adipogenesis. Trends Endocrinol. Metab. *20*, 107–114.

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

Maaten, L. van der, and Hinton, G. (2008). Visualizing Data using t-SNE. J. Mach. Learn. Res.

9, 2579–2605.

Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A.R., Kamitaki, N., Martersteck, E.M., et al. (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell *161*, 1202–1214.

Makino, Y., Nakamura, H., Ikeda, E., Ohnuma, K., Yamauchi, K., Yabe, Y., Poellinger, L., Okada, Y., Morimoto, C., and Tanaka, H. (2003). Hypoxia-inducible factor regulates survival of antigen receptor-driven T cells. J. Immunol. *171*, 6534–6540.

Mamedov, T.G., Pienaar, E., Whitney, S.E., TerMaat, J.R., Carvill, G., Goliath, R., Subramanian, A., and Viljoen, H.J. (2008). A fundamental study of the PCR amplification of GC-rich DNA templates. Comput. Biol. Chem. *32*, 452–457.

Manley, L.J., Ma, D., and Levine, S.S. (2016). Monitoring Error Rates In Illumina Sequencing. J. Biomol. Tech. *27*, 125–128.

Mantovani, A., Allavena, P., Sica, A., and Balkwill, F. (2008). Cancer-related inflammation. Nature *454*, 436–444.

Moreno-Sánchez, R., Rodríguez-Enríquez, S., Saavedra, E., Marín-Hernández, A., and Gallardo-Pérez, J.C. (2009). The bioenergetics of cancer: Is glycolysis the main ATP supplier in all tumor cells? Biofactors *35*, 209–225.

Mues, C., Zhou, J., Manolopoulos, K.N., Korsten, P., Schmoll, D., Klotz, L.-O., Bornstein, S.R., Klein, H.H., and Barthel, A. (2009). Regulation of glucose-6-phosphatase gene expression by insulin and metformin. Horm. Metab. Res. *41*, 730–735.

Murphy., K.P. (2007). Conjugate bayesian analysis of the gaussian distribution. Tech. Rep.

Novershtern, N., Subramanian, A., Lawton, L.N., Mak, R.H., Haining, W.N., McConkey, M.E., Habib, N., Yosef, N., Chang, C.Y., Shay, T., et al. (2011). Densely interconnected transcriptional circuits control cell states in human hematopoiesis. Cell *144*, 296–309.

Patro, R., Duggal, G., Love, M.I., Irizarry, R.A., and Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. Nat. Methods *14*, 417–419.

Perera, R.J., Marcusson, E.G., Koo, S., Kang, X., Kim, Y., White, N., and Dean, N.M. (2006). Identification of novel PPARγ target genes in primary human adipocytes. Gene *369*, 90–99.

Pitman, J. (2006). Combinatorial Stochastic Processes (Springer Science & Business Media).

Platanias, L.C. (2005). Mechanisms of type-I- and type-II-interferon-mediated signalling. Nat. Rev. Immunol. *5*, 375–386.

Prabhakaran, S., Azizi, E., Carr, A., & Pe'er, D (2016). Dirichlet Process Mixture Model for Correcting Technical Variation in Single-Cell Gene Expression Data. Journal of Machine Learning Research, W&CP (ICML) *48*, 1070–1079.

Schietinger, A., Delrow, J.J., Basom, R.S., Blattman, J.N., and Greenberg, P.D. (2012). Rescued Tolerant CD8 T Cells Are Preprogrammed to Reestablish the Tolerant State. Science

335, 723–727.

Sica, A., and Mantovani, A. (2012). Macrophage plasticity and polarization: in vivo veritas. J. Clin. Invest. *122*, 787.

Singer, M., Wang, C., Cong, L., Marjanovic, N.D., Kowalczyk, M.S., Zhang, H., Nyman, J., Sakuishi, K., Kurtulus, S., Gennert, D., et al. (2016). A Distinct Gene Module for Dysfunction Uncoupled from Activation in Tumor-Infiltrating T Cells. Cell *166*, 1500–1511.e9.

Smith-Garvin, J.E., Koretzky, G.A., and Jordan, M.S. (2009). T Cell Activation. Annu. Rev. Immunol. *27*, 591–619.

Tao, T., and Vu, V. (2005). On random ±1 matrices: Singularity and determinant. Random Struct. Algorithms *28*, 1–23.

Ugel, S., De Sanctis, F., Mandruzzato, S., and Bronte, V. (2015). Tumor-induced myeloid deviation: when myeloid-derived suppressor cells meet tumor-associated macrophages. J. Clin. Invest. *125*, 3365–3376.

Valle, S., Li, W., and Joe Qin, S. (1999). Selection of the Number of Principal Components: The Variance of the Reconstruction Error Criterion with a Comparison to Other Methods†. Ind. Eng. Chem. Res. *38*, 4389–4401.

Villani, A.-C., Satija, R., Reynolds, G., Sarkizova, S., Shekhar, K., Fletcher, J., Griesbeck, M., Butler, A., Zheng, S., Lazo, S., et al. (2017). Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. Science *356*.

Wherry, J.E. (2011). T cell exhaustion. Nat. Immunol. *12*, ni.2035.

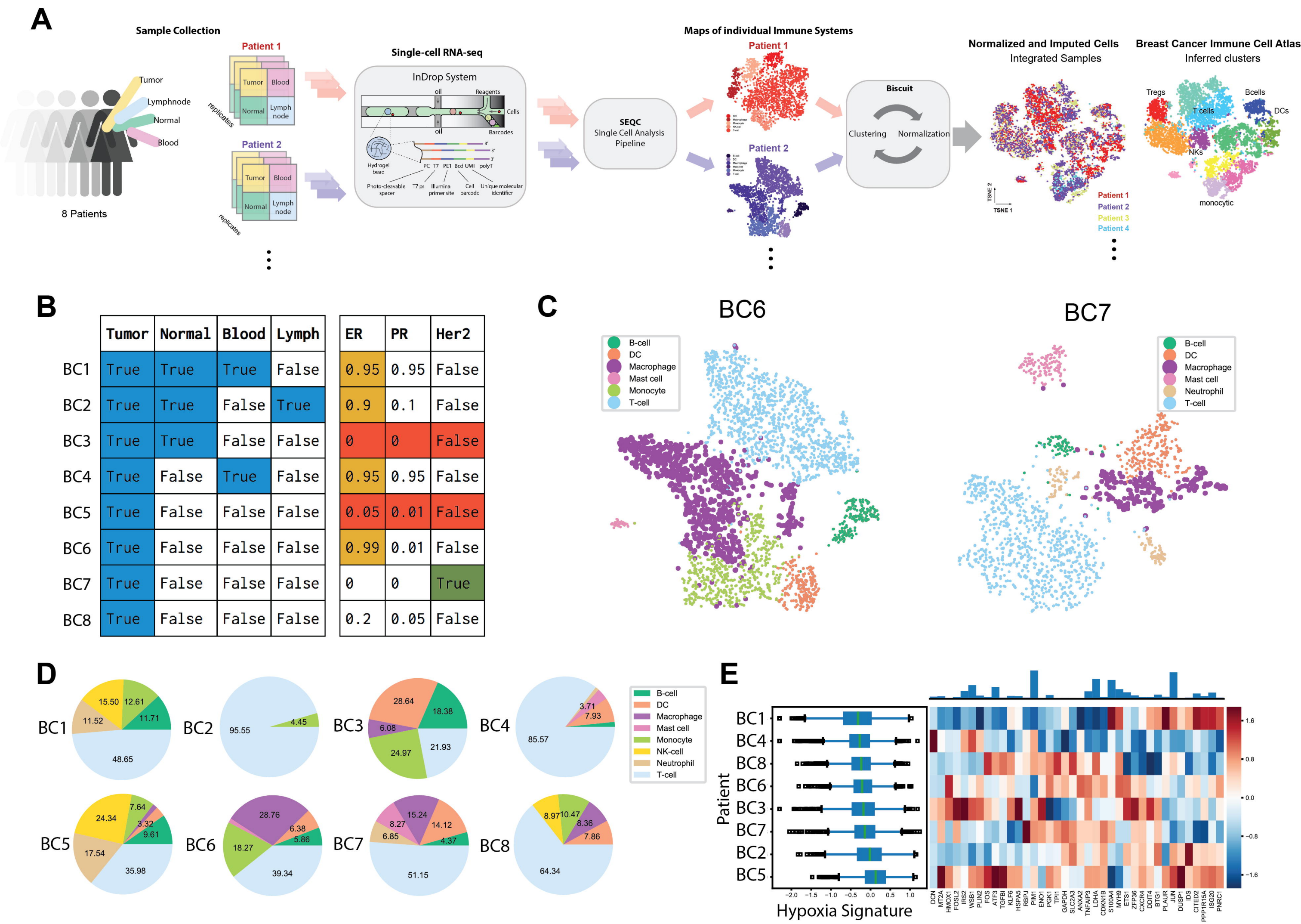
Wherry, J.E., and Kurachi, M. (2015). Molecular and cellular insights into T cell exhaustion. Nat. Rev. Immunol. *15*, nri3862.

Whitfield, M.L., Sherlock, G., Saldanha, A.J., Murray, J.I., Ball, C.A., Alexander, K.E., Matese, J.C., Perou, C.M., Hurt, M.M., Brown, P.O., et al. (2002). Identification of Genes Periodically Expressed in the Human Cell Cycle and Their Expression in Tumors. Mol. Biol. Cell *13*, 1977–2000.

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.

Zilionis, R., Nainys, J., Veres, A., Savova, V., Zemmour, D., Klein, A.M., and Mazutis, L. (2017). Single-cell barcoding and sequencing using droplet microfluidics. Nat. Protoc. *12*, nprot.2016.154.

Figure 1



	Tumor	Normal	Blood	Lymph	ER	PR	Her2
BC1	True	True	True	False	0.95	0.95	False
BC2	True	True	False	True	0.9	0.1	False
BC3	True	True	False	False	0	0	False
BC4	True	False	True	False	0.95	0.95	False
BC5	True	False	False	False	0.05	0.01	False
BC6	True	False	False	False	0.99	0.01	False
BC7	True	False	False	False	0	0	True
BC8	True	False	False	False	0.2	0.05	False

Figure 2

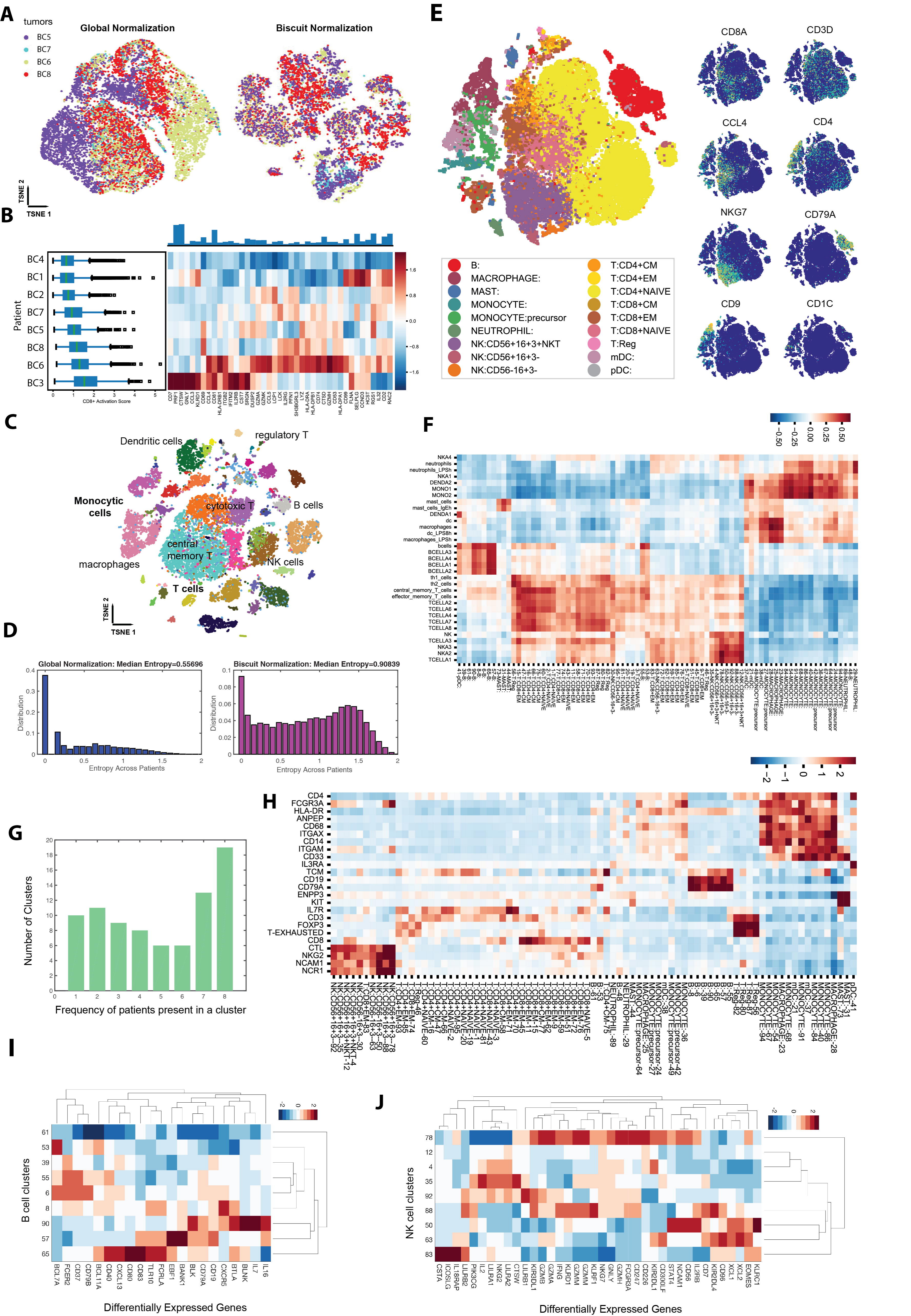
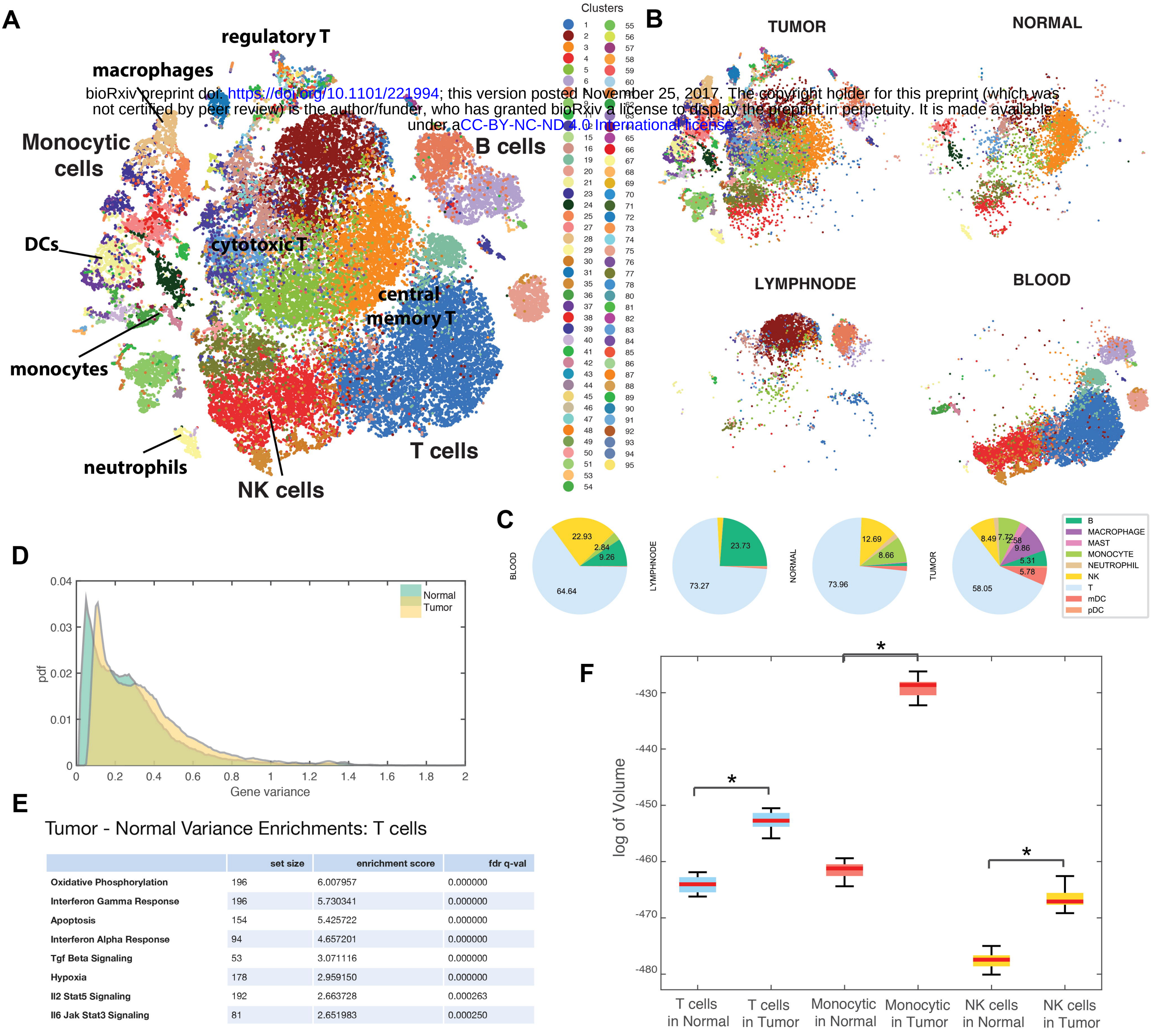

A
tumors
BC5
BC7
BC6
BC8
Global Normalization
Biscuit Normalization
TSNE 2
TSNE 1
B
Patient
BC4
BC1
BC2
BC7
BC5
BC8
BC6
BC3
CD8+ Activation Score
1.6
0.8
0.0
−0.8
−1.6
CD7
PRF1
CTSW
GNLY
CCL3
KLRD1
CD69
CCL4
CD81
HLA-DRB1
ITGB2
IFITM2
IL2RB
CST7
SRGN
DUSP2
GZMA
GZMK
CCL5
LCP1
LCK
IL2RG
PFN1
SH3BGRL3
LYZ
HLA-DRA
HLA-DPB1
CD74
CTSD
GZMH
CD53
HLA-DPA1
CD99
FLNA
SELPLG
CCND3
HCST
RGS1
IL32
CD2
RAC2
C
Dendritic cells
regulatory T
Monocytic cells
cytotoxic T
B cells
central memory T
macrophages
NK cells
T cells
TSNE 2
TSNE 1
D
Global Normalization: Median Entropy=0.55696
Biscuit Normalization: Median Entropy=0.90839
Distribution
Entropy Across Patients
E
CD8A
CD3D
CCL4
CD4
NKG7
CD79A
CD9
CD1C
B:
MACROPHAGE:
MAST:
MONOCYTE:
MONOCYTE:precursor
NEUTROPHIL:
NK:CD56+16+3+NKT
NK:CD56+16+3-
NK:CD56-16+3-
T:CD4+CM
T:CD4+EM
T:CD4+NAIVE
T:CD8+CM
T:CD8+EM
T:CD8+NAIVE
T:Reg
mDC:
pDC:
F
-0.50
-0.25
0.00
0.25
0.50
NKA4
neutrophils
neutrophils_LPSh
NKA1
DENDA2
MONO1
MONO2
mast_cells
mast_cells_IgEh
DENDA1
dc
macrophages
dc_LPS8h
macrophages_LPSh
bcells
BCELLA3
BCELLA4
BCELLA1
BCELLA2
th1_cells
th2_cells
central_memory_T_cells
effector_memory_T_cells
TCELLA2
TCELLA6
TCELLA4
TCELLA7
TCELLA8
NK
TCELLA3
NKA3
NKA2
TCELLA1
G
Number of Clusters
Frequency of patients present in a cluster
H
−2
−1
0
1
2
CD4
FCGR3A
HLA-DR
ANPEP
CD68
ITGAX
CD14
ITGAM
CD33
IL3RA
TCM
CD19
CD79A
ENPP3
KIT
IL7R
CD3
FOXP3
T-EXHAUSTED
CD8
CTL
NKG2
NCAM1
NCR1
I
B cell clusters
61
53
39
55
6
8
90
57
65
BCL7A
FCER2
CD37
CD79B
BCL11A
CD40
CXCL13
CD80
CD83
TLR10
FCRLA
EBF1
BANK1
BLK
CD79A
CD19
CXCR5
BTLA
BLNK
IL7
IL16
Differentially Expressed Genes
J
NK cell clusters
78
12
4
35
92
88
50
63
83
CSTA
ICOSLG
IL18RAP
IL18R2
PIK3CG
IL2
LILRA1
NKG2
LILRA2
CTSW
LILRB1
KIR3DL1
GZMB
GZMA
IFNG
KLRD1
GZMM
GZMM
KLRF1
NKG7
GNLY
GZMH
FCGR3A
CD247
CD226
KIR2DL1
CD300LF
STAT4
NCAM1
CD56
IL2RB
CD7
KIR2DL4
CD96
XCL1
XCL2
EOMES
KLRC1
Differentially Expressed Genes

Figure 3



Tumor - Normal Variance Enrichments: T cells

	set size	enrichment score	fdr q-val
Oxidative Phosphorylation	196	6.007957	0.000000
Interferon Gamma Response	196	5.730341	0.000000
Apoptosis	154	5.425722	0.000000
Interferon Alpha Response	94	4.657201	0.000000
Tgf Beta Signaling	53	3.071116	0.000000
Hypoxia	178	2.959150	0.000000
Il2 Stat5 Signaling	192	2.663728	0.000263
Il6 Jak Stat3 Signaling	81	2.651983	0.000250

Figure 4

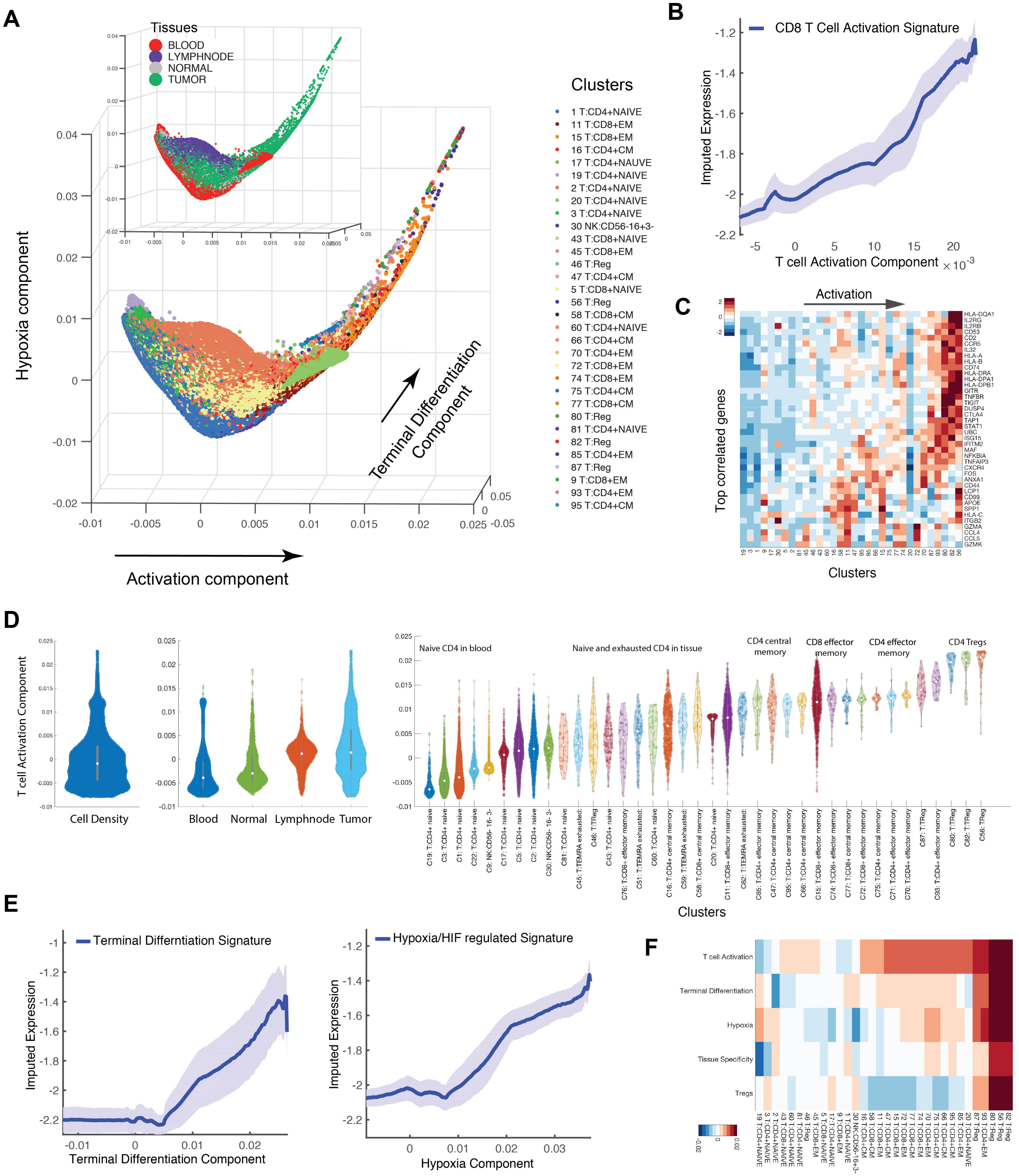

A
Tissues
BLOOD
LYMPHNODE
NORMAL
TUMOR
Hypoxia component
Activation component
Terminal Differentiation Component
Clusters
1 T:CD4+NAIVE
11 T:CD8+EM
15 T:CD8+EM
16 T:CD4+CM
17 T:CD4+NAUVE
19 T:CD4+NAIVE
2 T:CD4+NAIVE
20 T:CD4+NAIVE
3 T:CD4+NAIVE
30 NK:CD56-16+3-
43 T:CD8+NAIVE
45 T:CD8+EM
46 T:Reg
47 T:CD4+CM
5 T:CD8+NAIVE
56 T:Reg
58 T:CD8+CM
60 T:CD4+NAIVE
66 T:CD4+CM
70 T:CD4+EM
72 T:CD8+EM
74 T:CD8+EM
75 T:CD4+CM
77 T:CD8+CM
80 T:Reg
81 T:CD4+NAIVE
82 T:Reg
85 T:CD4+EM
87 T:Reg
9 T:CD8+EM
93 T:CD4+EM
95 T:CD4+CM
B
CD8 T Cell Activation Signature
Imputed Expression
T cell Activation Component ×10^-3
C
Activation
Top correlated genes
HLA-DQA1
IL2RG
IL2RB
CD53
CD2
CCR5
IL32
HLA-A
HLA-B
CD74
HLA-DRA
HLA-DPA1
HLA-DPB1
GITR
TNFBR
TIGIT
DUSP4
CTLA4
TAP1
STAT1
UBC
ISG15
IFITM2
MAF
NFKBIA
TNFAIP3
CXCR4
FOS
ANXA1
CD44
LCP1
CD99
APOE
SPP1
HLA-C
ITGB2
GZMA
CCL4
CCL5
GZMK
19 3 1 9 17 30 5 2 81 45 46 43 60 16 58 11 47 95 85 66 15 75 77 74 20 72 70 87 93 80 82 56
Clusters
D
T cell Activation Component
Cell Density
Blood
Normal
Lymphnode
Tumor
Naive CD4 in blood
Naive and exhausted CD4 in tissue
CD4 central memory
CD8 effector memory
CD4 effector memory
CD4 Tregs
C19: T:CD4+ naive
C3: T:CD4+ naive
C1: T:CD4+ naive
C22: T:CD4+ naive
C9: NK:CD56- 16- 3-
C17: T:CD4+ naive
C5: T:CD4+ naive
C2: T:CD4+ naive
C30: NK:CD56- 16- 3-
C81: T:CD4+ naive
C45: T:TEMRA exhausted:
C46: T:TReg
C43: T:CD4+ naive
C76: T:CD8+ effector memory
C51: T:TEMRA exhausted:
C60: T:CD4+ naive
C16: T:CD4+ central memory
C59: T:TEMRA exhausted:
C58: T:CD8+ central memory
C20: T:CD4+ naive
C11: T:CD8+ effector memory
C62: T:TEMRA exhausted:
C85: T:CD4+ effector memory
C47: T:CD4+ central memory
C95: T:CD4+ central memory
C66: T:CD4+ central memory
C15: T:CD8+ effector memory
C74: T:CD8+ effector memory
C77: T:CD8+ central memory
C72: T:CD8+ effector memory
C75: T:CD4+ central memory
C71: T:CD4+ effector memory
C70: T:CD4+ effector memory
C87: T:TReg
C93: T:CD4+ effector memory
C80: T:TReg
C82: T:TReg
C56: TReg
Clusters
E
Terminal Differntiation Signature
Imputed Expression
Terminal Differentiation Component
Hypoxia/HIF regulated Signature
Imputed Expression
Hypoxia Component
F
T cell Activation
Terminal Differentiation
Hypoxia
Tissue Specificity
Tregs
19 T:CD4+NAIVE
3 T:CD4+NAIVE
2 T:CD4+NAIVE
43 T:CD8+NAIVE
60 T:CD4+NAIVE
81 T:CD4+NAIVE
46 T:Reg
45 T:CD8+EM
5 T:CD8+NAIVE
17:T:CD4+NAIVE
9 T:CD8+EM
1 T:CD4+NAIVE
30 NK:CD56-16+3-
16 T:CD4+CM
58 T:CD8+CM
11 T:CD8+EM
47 T:CD4+CM
15 T:CD8+EM
72 T:CD8+EM
77 T:CD8+CM
74 T:CD8+EM
70 T:CD4+EM
75 T:CD4+CM
66 T:CD4+CM
95 T:CD4+CM
85 T:CD4+EM
20 T:CD4+NAIVE
87 T:Reg
93 T:CD4+EM
80 T:Reg
56 T:Reg
82 T:Reg
-0.02
0
0.02

Figure 5

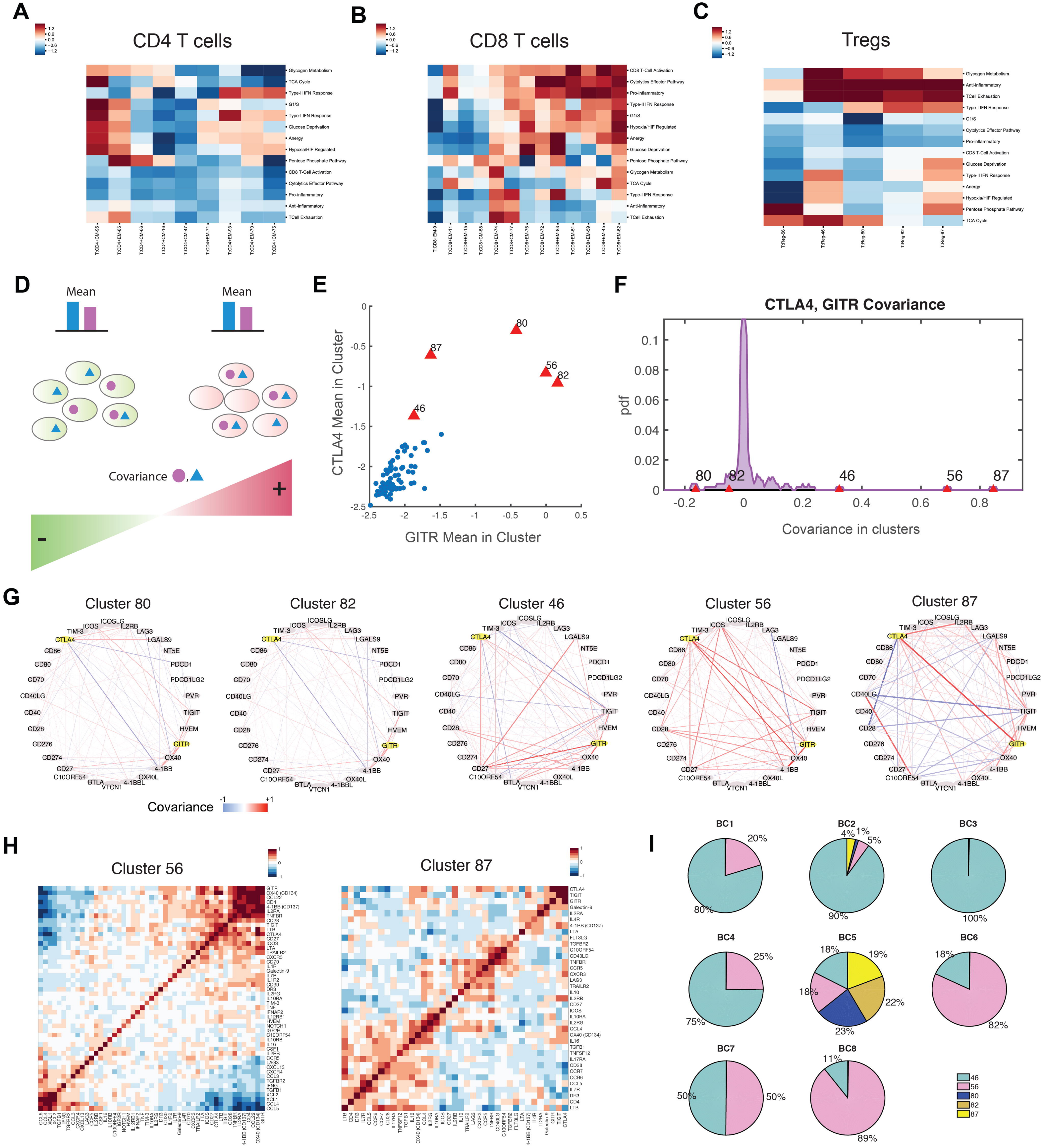

A
1.2
0.6
0.0
−0.6
−1.2
CD4 T cells
Glycogen Metabolism
TCA Cycle
Type-II IFN Response
G1/S
Type-I IFN Response
Glucose Deprivation
Anergy
Hypoxia/HIF Regulated
Pentose Phosphate Pathway
CD8 T-Cell Activation
Cytolytics Effector Pathway
Pro-inflammatory
Anti-inflammatory
TCell Exhaustion
T:CD4+CM-95
T:CD4+EM-85
T:CD4+CM-66
T:CD4+CM-16
T:CD4+CM-47
T:CD4+EM-71
T:CD4+EM-93
T:CD4+EM-70
T:CD4+CM-75
B
CD8 T cells
CD8 T-Cell Activation
Cytolytics Effector Pathway
Pro-inflammatory
Type-II IFN Response
G1/S
Hypoxia/HIF Regulated
Anergy
Glucose Deprivation
Pentose Phosphate Pathway
Glycogen Metabolism
TCA Cycle
Type-I IFN Response
Anti-inflammatory
TCell Exhaustion
T:CD8+EM-9
T:CD8+EM-11
T:CD8+EM-15
T:CD8+CM-58
T:CD8+EM-74
T:CD8+CM-77
T:CD8+EM-76
T:CD8+EM-72
T:CD8+EM-83
T:CD8+EM-51
T:CD8+EM-59
T:CD8+EM-45
T:CD8+EM-62
C
Tregs
Glycogen Metabolism
Anti-inflammatory
TCell Exhaustion
Type-I IFN Response
G1/S
Cytolytics Effector Pathway
Pro-inflammatory
CD8 T-Cell Activation
Glucose Deprivation
Type-II IFN Response
Anergy
Hypoxia/HIF Regulated
Pentose Phosphate Pathway
TCA Cycle
T:Reg-56
T:Reg-46
T:Reg-80
T:Reg-82
T:Reg-87
D
Mean
Mean
Covariance
-
+
E
CTLA4 Mean in Cluster
GITR Mean in Cluster
0
-0.5
-1
-1.5
-2
-2.5
-2.5
-2
-1.5
-1
-0.5
0
0.5
80
87
56
82
46
F
CTLA4, GITR Covariance
pdf
0.1
0.08
0.06
0.04
0.02
0
-0.2
0
0.2
0.4
0.6
0.8
Covariance in clusters
80
82
46
56
87
G
Cluster 80
Cluster 82
Cluster 46
Cluster 56
Cluster 87
ICOSLG
IL2RB
LAG3
LGALS9
NT5E
PDCD1
PDCD1LG2
PVR
TIGIT
HVEM
GITR
OX40
4-1BB
OX40L
4-1BBL
VTCN1
BTLA
C10ORF54
CD27
CD274
CD276
CD28
CD40
CD40LG
CD70
CD80
CD86
CTLA4
TIM-3
ICOS
Covariance
-1
+1
H
Cluster 56
1
0
-1
GITR
OX40 (CD134)
CCL22
CD4
4-1BB (CD137)
IL2RA
TNFBR
CD28
TIGIT
LTB
CTLA4
CD27
ICOS
LTA
TRAILR2
CXCR3
CD70
IL4R
Galectin-9
IL7R
IL1R2
CD30
DR3
IL2RG
IL10RA
TIM-3
TNF
IFNAR2
IL12RB1
HVEM
NOTCH1
IGF2R
C10ORF54
IL10RB
IL16
CSF1
IL2RB
CCR5
LAG3
CXCL13
CXCR4
CCL3
TGFBR2
IFNG
TGFB1
XCL2
XCL1
CCL4
CCL5
Cluster 87
1
0
-1
CTLA4
TIGIT
GITR
Galectin-9
IL2RA
IL4R
4-1BB (CD137)
LTA
FLT3LG
TGFBR2
C10ORF54
CD40LG
TNFBR
CCR5
CXCR3
LAG3
TRAILR2
IL10
IL2RB
CD27
ICOS
IL10RA
IL2RG
CCL4
OX40 (CD134)
IL16
TGFB1
TNFSF12
IL17RA
CD28
CCR7
CCR6
CCL5
IL7R
DR3
CD4
LTB
I
BC1
20%
80%
BC2
4%
1%
5%
90%
BC3
100%
BC4
25%
75%
BC5
18%
19%
18%
22%
23%
BC6
18%
82%
BC7
50%
50%
BC8
11%
89%
46
56
80
82
87

Figure 6

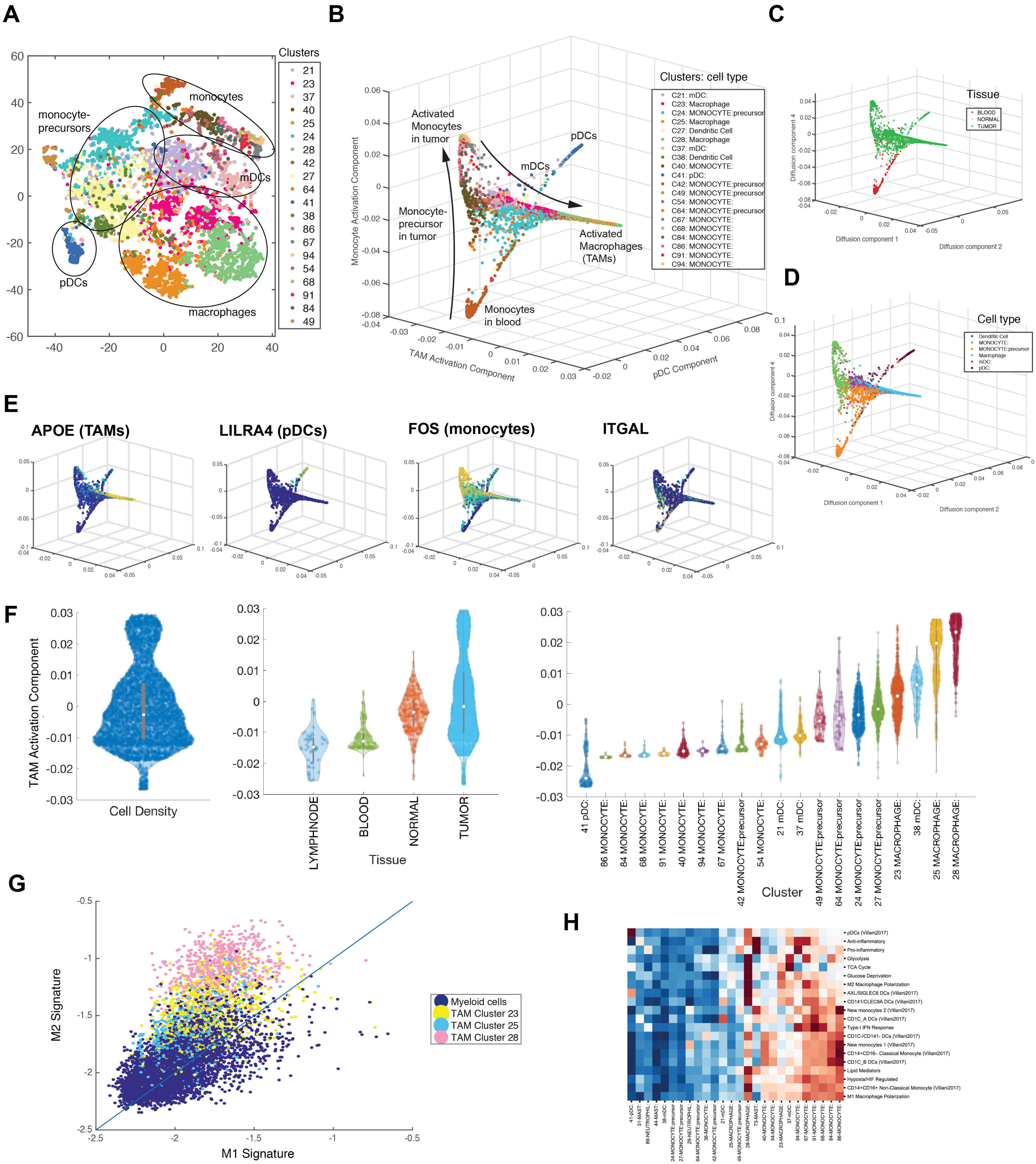

A
Clusters
21
23
37
40
25
24
28
42
27
64
41
38
86
67
94
54
68
91
84
49
monocytes
monocyte-
precursors
mDCs
pDCs
macrophages
B
Clusters: cell type
C21: mDC:
C23: Macrophage
C24: MONOCYTE:precursor
C25: Macrophage
C27: Dendritic Cell
C28: Macrophage
C37: mDC:
C38: Dendritic Cell
C40: MONOCYTE:
C41: pDC:
C42: MONOCYTE:precursor
C49: MONOCYTE:precursor
C54: MONOCYTE:
C64: MONOCYTE:precursor
C67: MONOCYTE:
C68: MONOCYTE:
C84: MONOCYTE:
C86: MONOCYTE:
C91: MONOCYTE:
C94: MONOCYTE:
Activated
Monocytes
in tumor
pDCs
mDCs
Monocyte-
precursor
in tumor
Activated
Macrophages
(TAMs)
Monocytes
in blood
Monocyte Activation Component
TAM Activation Component
pDC Component
C
Tissue
BLOOD
NORMAL
TUMOR
Diffusion component 4
Diffusion component 1
Diffusion component 2
D
Cell type
Dendritic Cell
MONOCYTE:
MONOCYTE:precursor
Macrophage
mDC:
pDC:
Diffusion component 4
Diffusion component 1
Diffusion component 2
E
APOE (TAMs)
LILRA4 (pDCs)
FOS (monocytes)
ITGAL
F
TAM Activation Component
Cell Density
LYMPHNODE
BLOOD
NORMAL
TUMOR
Tissue
41 pDC:
86 MONOCYTE:
84 MONOCYTE:
68 MONOCYTE:
91 MONOCYTE:
40 MONOCYTE:
94 MONOCYTE:
67 MONOCYTE:
42 MONOCYTE:precursor
54 MONOCYTE:
21 mDC:
37 mDC:
49 MONOCYTE:precursor
64 MONOCYTE:precursor
24 MONOCYTE:precursor
27 MONOCYTE:precursor
23 MACROPHAGE:
38 mDC:
25 MACROPHAGE:
28 MACROPHAGE:
Cluster
G
M2 Signature
M1 Signature
Myeloid cells
TAM Cluster 23
TAM Cluster 25
TAM Cluster 28
H
pDCs (Villani2017)
Anti-inflammatory
Pro-inflammatory
Glycolysis
TCA Cycle
Glucose Deprivation
M2 Macrophage Polarization
AXL/SIGLEC6 DCs (Villani2017)
CD141/CLEC9A DCs (Villani2017)
New monocytes 2 (Villani2017)
CD1C_A DCs (Villani2017)
Type-I IFN Response
CD1C-/CD141- DCs (Villani2017)
New monocytes 1 (Villani2017)
CD14+CD16– Classical Monocyte (Villani2017)
CD1C_B DCs (Villani2017)
Lipid Mediators
Hypoxia/HIF Regulated
CD14+CD16+ Non-Classical Monocyte (Villani2017)
M1 Macrophage Polarization
41-pDC:
31-MAST:
89-NEUTROPHIL:
44-MAST:
38-mDC:
24-MONOCYTE:precursor
27-MONOCYTE:precursor
29-NEUTROPHIL:
64-MONOCYTE:precursor
36-MONOCYTE:
42-MONOCYTE:precursor
21-mDC:
25-MACROPHAGE:
49-MONOCYTE:precursor
28-MACROPHAGE:
73-MAST:
40-MONOCYTE:
54-MONOCYTE:
23-MACROPHAGE:
37-mDC:
94-MONOCYTE:
67-MONOCYTE:
91-MONOCYTE:
68-MONOCYTE:
84-MONOCYTE:
86-MONOCYTE:

Figure 7

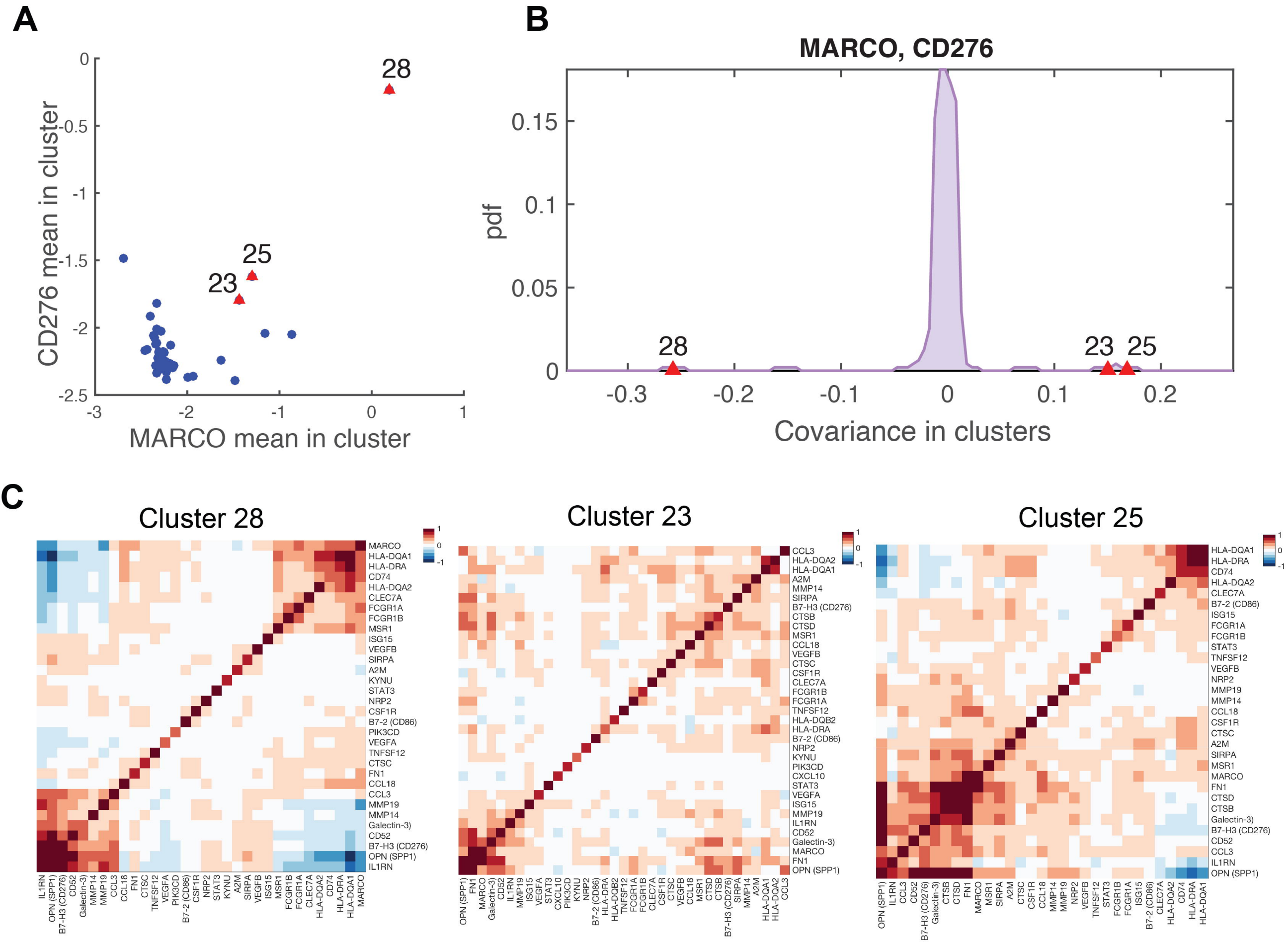
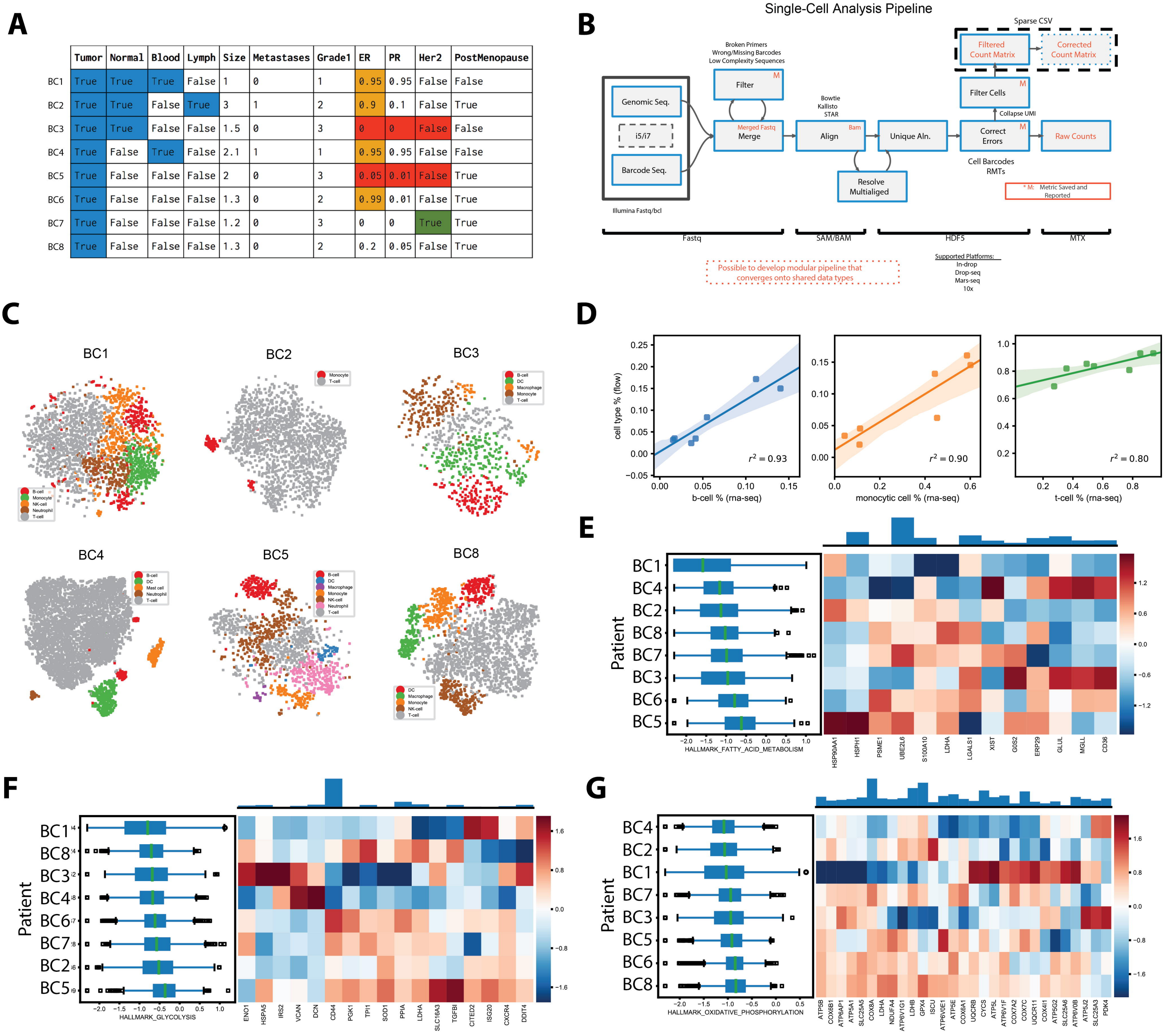

A
0
-0.5
-1
-1.5
-2
-2.5
CD276 mean in cluster
-3
-2
-1
0
1
MARCO mean in cluster
28
25
23
B
MARCO, CD276
pdf
0.15
0.1
0.05
0
-0.3
-0.2
-0.1
0
0.1
0.2
Covariance in clusters
28
23
25
C
Cluster 28
1
0
-1
MARCO
HLA-DQA1
HLA-DRA
CD74
HLA-DQA2
CLEC7A
FCGR1A
FCGR1B
MSR1
ISG15
VEGFB
SIRPA
A2M
KYNU
STAT3
NRP2
CSF1R
B7-2 (CD86)
PIK3CD
VEGFA
TNFSF12
CTSC
FN1
CCL18
CCL3
MMP19
MMP14
Galectin-3)
CD52
B7-H3 (CD276)
OPN (SPP1)
IL1RN
IL1RN
OPN (SPP1)
B7-H3 (CD276)
CD52
Galectin-3)
MMP14
MMP19
CCL3
CCL18
FN1
CTSC
TNFSF12
VEGFA
PIK3CD
B7-2 (CD86)
CSF1R
NRP2
STAT3
KYNU
A2M
SIRPA
VEGFB
ISG15
MSR1
FCGR1B
FCGR1A
CLEC7A
HLA-DQA2
CD74
HLA-DRA
HLA-DQA1
MARCO
Cluster 23
1
0
-1
CCL3
HLA-DQA2
HLA-DQA1
A2M
MMP14
SIRPA
B7-H3 (CD276)
CTSB
CTSD
MSR1
CCL18
VEGFB
CTSC
CSF1R
CLEC7A
FCGR1B
FCGR1A
TNFSF12
HLA-DQB2
HLA-DRA
B7-2 (CD86)
NRP2
KYNU
PIK3CD
CXCL10
STAT3
VEGFA
ISG15
MMP19
IL1RN
CD52
Galectin-3)
MARCO
FN1
OPN (SPP1)
OPN (SPP1)
FN1
MARCO
Galectin-3)
CD52
IL1RN
MMP19
ISG15
VEGFA
STAT3
CXCL10
PIK3CD
KYNU
NRP2
B7-2 (CD86)
HLA-DRA
HLA-DQB2
TNFSF12
FCGR1A
FCGR1B
CLEC7A
CSF1R
CTSC
VEGFB
CCL18
MSR1
CTSD
CTSB
B7-H3 (CD276)
SIRPA
MMP14
A2M
HLA-DQA1
HLA-DQA2
CCL3
Cluster 25
1
0
-1
HLA-DQA1
HLA-DRA
CD74
HLA-DQA2
CLEC7A
B7-2 (CD86)
ISG15
FCGR1A
FCGR1B
STAT3
TNFSF12
VEGFB
NRP2
MMP19
MMP14
CCL18
CSF1R
CTSC
A2M
SIRPA
MSR1
MARCO
FN1
CTSD
CTSB
Galectin-3)
B7-H3 (CD276)
CD52
CCL3
IL1RN
OPN (SPP1)
OPN (SPP1)
IL1RN
CCL3
CD52
B7-H3 (CD276)
Galectin-3)
CTSB
CTSD
FN1
MARCO
MSR1
SIRPA
A2M
CTSC
CSF1R
CCL18
MMP14
MMP19
NRP2
VEGFB
TNFSF12
STAT3
FCGR1B
FCGR1A
ISG15
B7-2 (CD86)
CLEC7A
HLA-DQA2
CD74
HLA-DRA
HLA-DQA1

Figure S1



	Tumor	Normal	Blood	Lymph	Size	Metastases	Grade1	ER	PR	Her2	PostMenopause
BC1	True	True	True	False	1	0	1	0.95	0.95	False	False
BC2	True	True	False	True	3	1	2	0.9	0.1	False	True
BC3	True	True	False	False	1.5	0	3	0	0	False	False
BC4	True	False	True	False	2.1	1	1	0.95	0.95	False	False
BC5	True	False	False	False	2	0	3	0.05	0.01	False	True
BC6	True	False	False	False	1.3	0	2	0.99	0.01	False	True
BC7	True	False	False	False	1.2	0	3	0	0	True	True
BC8	True	False	False	False	1.3	0	2	0.2	0.05	False	True

Figure S2

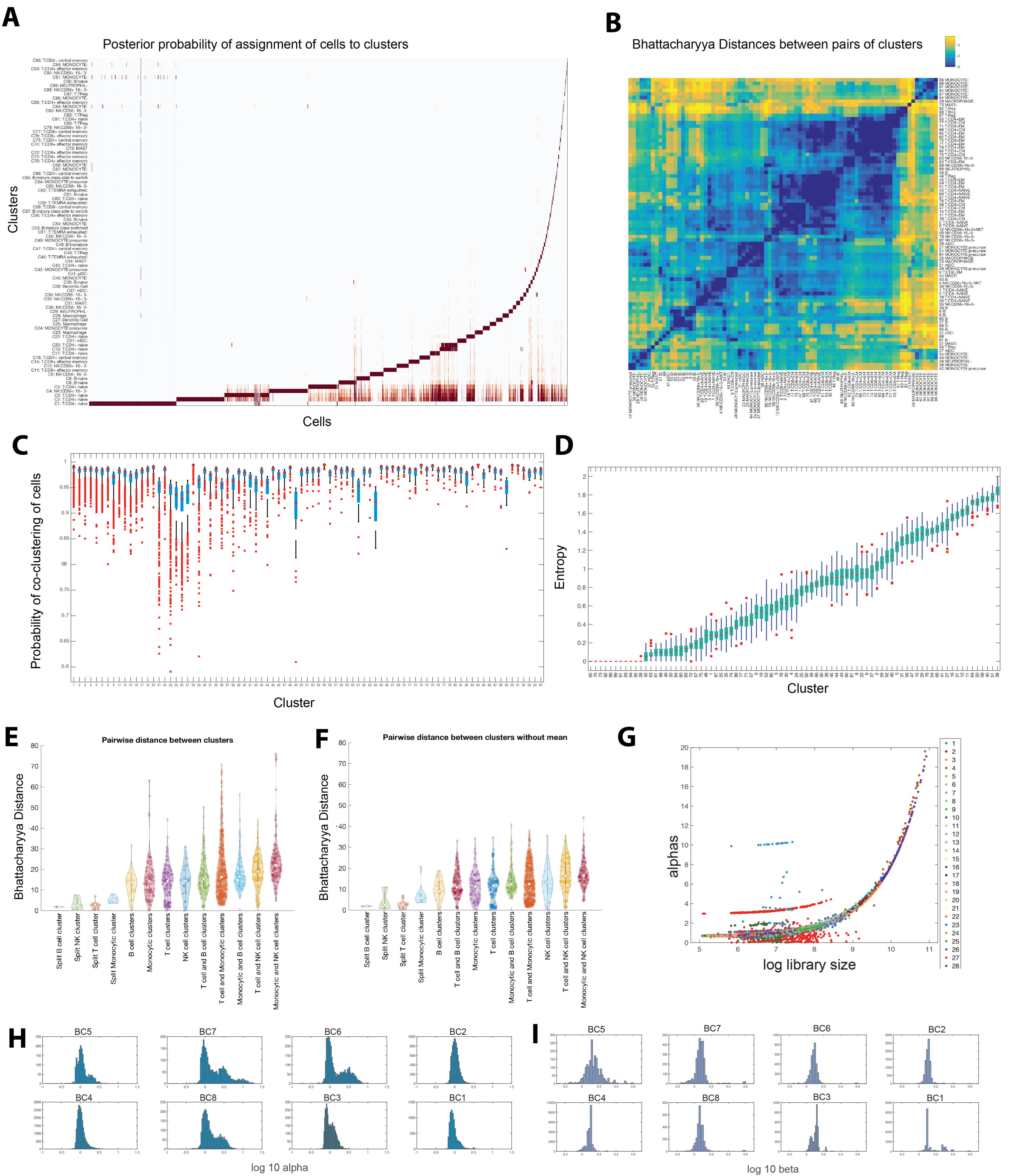
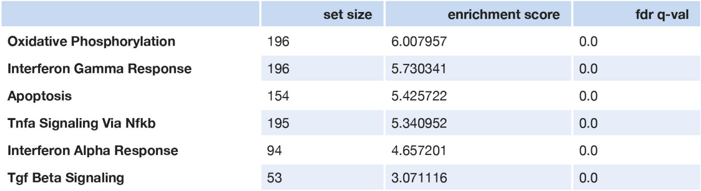

A
Posterior probability of assignment of cells to clusters
Clusters
Cells
B
Bhattacharyya Distances between pairs of clusters
C
Probability of co-clustering of cells
Cluster
D
Entropy
Cluster
E
Pairwise distance between clusters
Bhattacharyya Distance
Split B cell cluster
Split NK cell cluster
Split T cell cluster
Split Monocytic cluster
B cell clusters
Monocytic clusters
T cell clusters
NK cell clusters
T cell and B cell clusters
T cell and Monocytic clusters
Monocytic and B cell clusters
T cell and NK cell clusters
Monocytic and NK cell clusters
F
Pairwise distance between clusters without mean
Bhattacharyya Distance
Split B cell cluster
Split NK cell cluster
Split T cell cluster
Split Monocytic cluster
B cell clusters
T cell and B cell clusters
Monocytic clusters
T cell clusters
Monocytic and B cell clusters
T cell and Monocytic clusters
NK cell clusters
T cell and NK cell clusters
Monocytic and NK cell clusters
G
alphas
log library size
H
BC5
BC7
BC6
BC2
BC4
BC8
BC3
BC1
log 10 alpha
I
BC5
BC7
BC6
BC2
BC4
BC8
BC3
BC1
log 10 beta

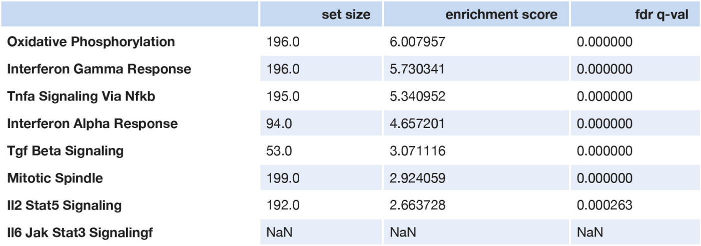
Figure S3

A Tumor - Normal Variance Enrichments: NK cells



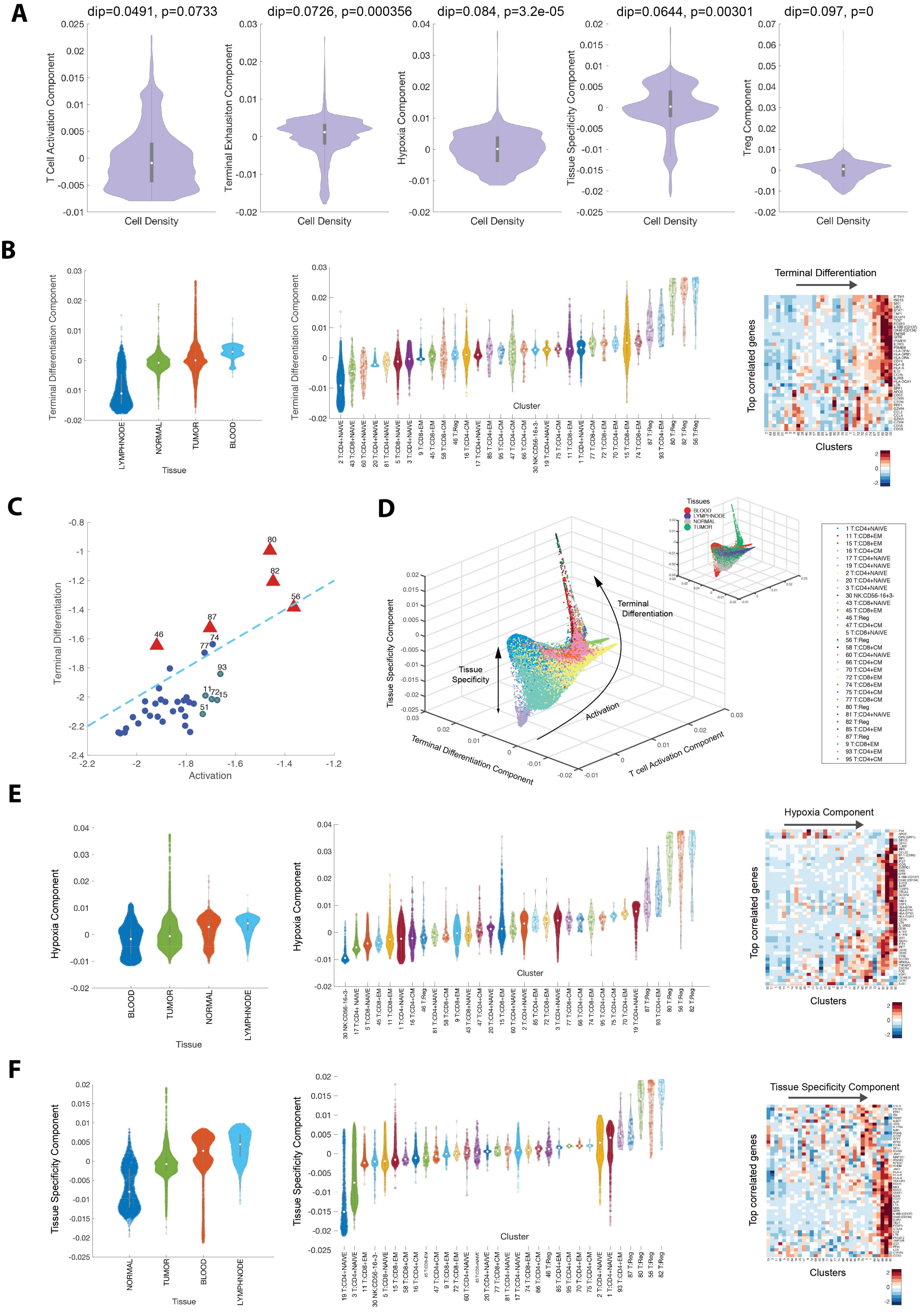
	set size	enrichment score	fdr q-val
Oxidative Phosphorylation	196	6.007957	0.0
Interferon Gamma Response	196	5.730341	0.0
Apoptosis	154	5.425722	0.0
Tnfa Signaling Via Nfkb	195	5.340952	0.0
Interferon Alpha Response	94	4.657201	0.0
Tgf Beta Signaling	53	3.071116	0.0

B Tumor - Normal Variance Enrichments: Monocytic cells



	set size	enrichment score	fdr q-val
Oxidative Phosphorylation	196.0	6.007957	0.000000
Interferon Gamma Response	196.0	5.730341	0.000000
Tnfa Signaling Via Nfkb	195.0	5.340952	0.000000
Interferon Alpha Response	94.0	4.657201	0.000000
Tgf Beta Signaling	53.0	3.071116	0.000000
Mitotic Spindle	199.0	2.924059	0.000000
Il2 Stat5 Signaling	192.0	2.663728	0.000263
Il6 Jak Stat3 Signalingf	NaN	NaN	NaN

Figure S4



A
dip=0.0491, p=0.0733
T Cell Activation Component
Cell Density
dip=0.0726, p=0.000356
Terminal Exhausiton Component
Cell Density
dip=0.084, p=3.2e-05
Hypoxia Component
Cell Density
dip=0.0644, p=0.00301
Tissue Specificity Component
Cell Density
dip=0.097, p=0
Treg Component
Cell Density
B
Terminal Differentiation Component
LYMPHNODE
NORMAL
TUMOR
BLOOD
Tissue
Terminal Differentiation Component
Cluster
2 T:CD4+NAIVE
43 T:CD8+NAIVE
60 T:CD4+NAIVE
20 T:CD4+NAIVE
81 T:CD4+NAIVE
5 T:CD8+NAIVE
3 T:CD4+NAIVE
9 T:CD8+EM
45 T:CD8+EM
58 T:CD8+CM
46 T:Reg
16 T:CD4+CM
17 T:CD4+NAIVE
85 T:CD4+EM
95 T:CD4+CM
47 T:CD4+CM
66 T:CD4+CM
30 NK:CD56-16+3-
19 T:CD4+NAIVE
75 T:CD4+CM
11 T:CD8+EM
1 T:CD4+NAIVE
77 T:CD8+CM
72 T:CD8+EM
70 T:CD4+EM
15 T:CD8+EM
74 T:CD8+EM
87 T:Reg
93 T:CD4+EM
80 T:Reg
82 T:Reg
56 T:Reg
Terminal Differentiation
Top correlated genes
Clusters
2
0
-2
C
Terminal Differentiation
Activation
80
82
56
87
46
77
74
93
11
72
15
51
D
Tissues
BLOOD
LYMPHNODE
NORMAL
TUMOR
Tissue Specificity Component
Terminal Differentiation
Tissue Specificity
Activation
Terminal Differentiation Component
T cell Activation Component
1 T:CD4+NAIVE
11 T:CD8+EM
15 T:CD8+EM
16 T:CD4+CM
17 T:CD4+NAIVE
19 T:CD4+NAIVE
2 T:CD4+NAIVE
20 T:CD4+NAIVE
3 T:CD4+NAIVE
30 NK:CD56-16+3-
43 T:CD8+NAIVE
45 T:CD8+EM
46 T:Reg
47 T:CD4+CM
5 T:CD8+NAIVE
56 T:Reg
58 T:CD8+CM
60 T:CD4+NAIVE
66 T:CD4+CM
70 T:CD4+EM
72 T:CD8+EM
74 T:CD8+EM
75 T:CD4+CM
77 T:CD8+CM
80 T:Reg
81 T:CD4+NAIVE
82 T:Reg
85 T:CD4+EM
87 T:Reg
9 T:CD8+EM
93 T:CD4+EM
95 T:CD4+CM
E
Hypoxia Component
BLOOD
TUMOR
NORMAL
LYMPHNODE
Tissue
Hypoxia Component
Cluster
30 NK:CD56-16+3-
17 T:CD4+ NAIVE
5 T:CD8+NAIVE
45 T:CD8+EM
11 T:CD8+EM
1 T:CD4+NAIVE
16 T:CD4+CM
46 T:Reg
81 T:CD4+NAIVE
58 T:CD8+CM
9 T:CD8+EM
43 T:CD8+NAIVE
47 T:CD4+CM
20 T:CD4+NAIVE
15 T:CD8+EM
60 T:CD4+NAIVE
2 T:CD4+NAIVE
85 T:CD4+EM
72 T:CD8+EM
3 T:CD4+NAIVE
77 T:CD8+CM
66 T:CD4+CM
74 T:CD8+EM
95 T:CD4+CM
75 T:CD4+CM
70 T:CD4+EM
19 T:CD4+NAIVE
87 T:Reg
93 T:CD4+EM
80 T:Reg
56 T:Reg
82 T:Reg
Hypoxia Component
Top correlated genes
Clusters
2
0
-2
F
Tissue Specificity Component
NORMAL
TUMOR
BLOOD
LYMPHNODE
Tissue
Tissue Specificity Component
Cluster
19 T:CD4+NAIVE
3 T:CD4+NAIVE
11 T:CD8+EM
30 NK:CD56-16+3-
5 T:CD8+NAIVE
15 T:CD8+EM
58 T:CD8+CM
16 T:CD4+CM
45 T:CD8+EM
47 T:CD4+CM
9 T:CD8+EM
72 T:CD8+EM
60 T:CD4+NAIVE
43 T:CD8+NAIVE
20 T:CD4+NAIVE
77 T:CD8+CM
81 T:CD4+NAIVE
17 T:CD4+NAIVE
74 T:CD8+EM
66 T:CD4+CM
46 T:Reg
85 T:CD4+EM
95 T:CD4+CM
70 T:CD4+CM
75 T:CD4+CM
2 T:CD4+NAIVE
1 T:CD4+NAIVE
93 T:CD4+EM
87 T:Reg
80 T:Reg
56 T:Reg
82 T:Reg
Tissue Specificity Component
Top correlated genes
Clusters
2
0
-2

Figure S5

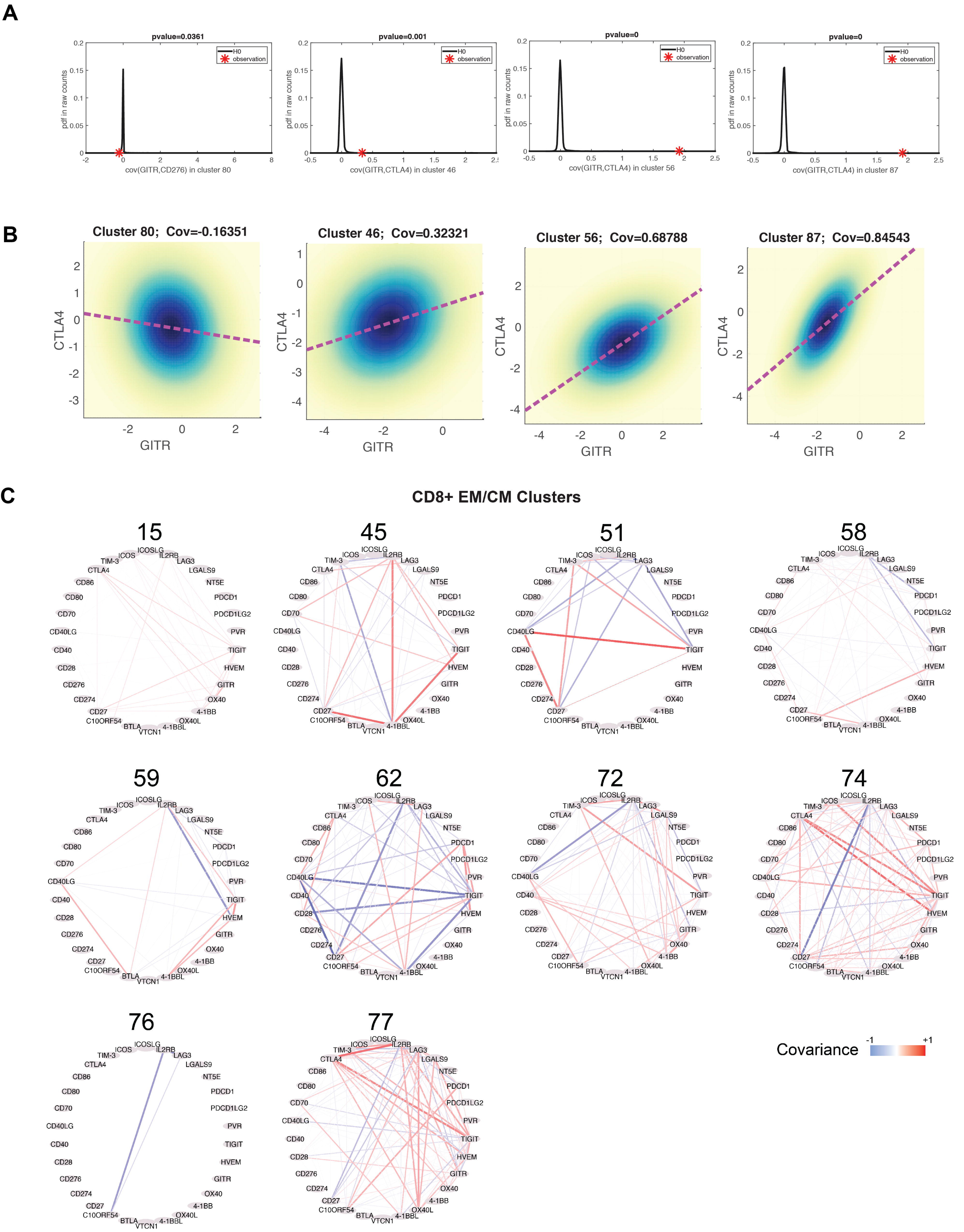

A
pvalue=0.0361
pvalue=0.001
pvalue=0
pvalue=0
H0
observation
pdf in raw counts
cov(GITR,CD276) in cluster 80
cov(GITR,CTLA4) in cluster 46
cov(GITR,CTLA4) in cluster 56
cov(GITR,CTLA4) in cluster 87
B
Cluster 80; Cov=-0.16351
Cluster 46; Cov=0.32321
Cluster 56; Cov=0.68788
Cluster 87; Cov=0.84543
CTLA4
GITR
C
CD8+ EM/CM Clusters
15
45
51
58
59
62
72
74
76
77
ICOSLG
IL2RB
LAG3
LGALS9
NT5E
PDCD1
PDCD1LG2
PVR
TIGIT
HVEM
GITR
OX40
4-1BB
OX40L
4-1BBL
VTCN1
BTLA
C10ORF54
CD27
CD274
CD276
CD28
CD40
CD40LG
CD70
CD80
CD86
CTLA4
TIM-3
ICOS
Covariance
-1
+1

Figure S6

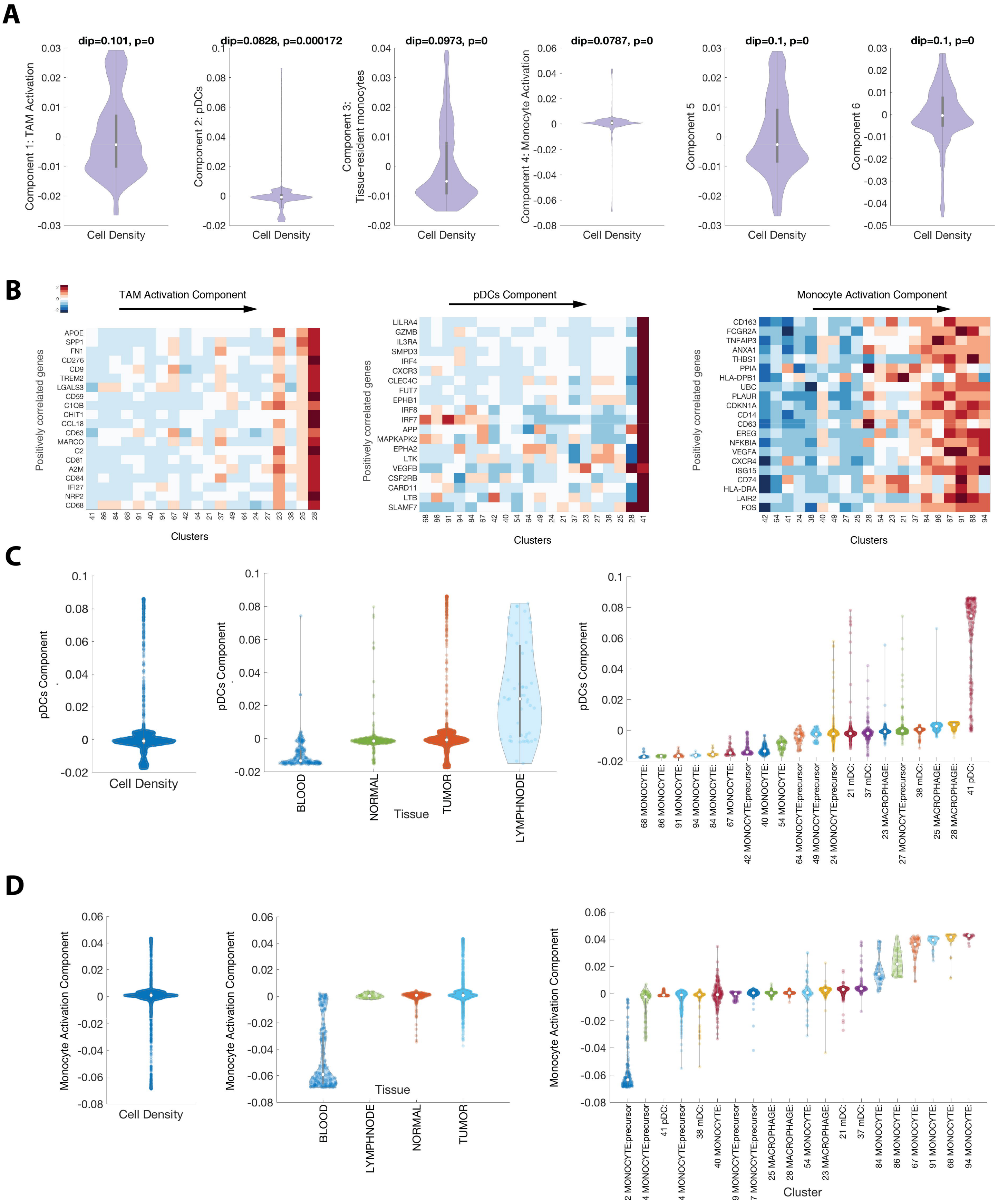

A
dip=0.101, p=0
Component 1: TAM Activation
Cell Density
dip=0.0828, p=0.000172
Component 2: pDCs
Cell Density
dip=0.0973, p=0
Component 3: Tissue-resident monocytes
Cell Density
dip=0.0787, p=0
Component 4: Monocyte Activation
Cell Density
dip=0.1, p=0
Component 5
Cell Density
dip=0.1, p=0
Component 6
Cell Density
B
2
0
-2
TAM Activation Component
Positively correlated genes
APOE
SPP1
FN1
CD276
CD9
TREM2
LGALS3
CD59
C1QB
CHIT1
CCL18
CD63
MARCO
C2
CD81
A2M
CD84
IFI27
NRP2
CD68
41 86 84 68 91 40 94 67 42 54 21 37 49 64 24 27 23 38 25 28
Clusters
pDCs Component
Positively correlated genes
LILRA4
GZMB
IL3RA
SMPD3
IRF4
CXCR3
CLEC4C
FUT7
EPHB1
IRF8
IRF7
APP
MAPKAPK2
EPHA2
LTK
VEGFB
CSF2RB
CARD11
LTB
SLAMF7
68 86 91 94 84 67 42 40 54 64 49 24 21 37 23 27 38 25 28 41
Clusters
Monocyte Activation Component
Positively correlated genes
CD163
FCGR2A
TNFAIP3
ANXA1
THBS1
PPIA
HLA-DPB1
UBC
PLAUR
CDKN1A
CD14
CD63
EREG
NFKBIA
VEGFA
CXCR4
ISG15
CD74
HLA-DRA
LAIR2
FOS
42 64 41 24 38 40 49 27 25 28 54 23 21 37 84 86 67 91 68 94
Clusters
C
pDCs Component
Cell Density
pDCs Component
BLOOD
NORMAL
TUMOR
LYMPHNODE
Tissue
pDCs Component
68 MONOCYTE:
86 MONOCYTE:
91 MONOCYTE:
94 MONOCYTE:
84 MONOCYTE:
67 MONOCYTE:
42 MONOCYTE:precursor
40 MONOCYTE:
54 MONOCYTE:
64 MONOCYTE:precursor
49 MONOCYTE:precursor
24 MONOCYTE:precursor
21 mDC:
37 mDC:
23 MACROPHAGE:
27 MONOCYTE:precursor
38 mDC:
25 MACROPHAGE:
28 MACROPHAGE:
41 pDC:
D
Monocyte Activation Component
Cell Density
Monocyte Activation Component
BLOOD
LYMPHNODE
NORMAL
TUMOR
Tissue
Monocyte Activation Component
2 MONOCYTE:precursor
4 MONOCYTE:precursor
41 pDC:
4 MONOCYTE:precursor
38 mDC:
40 MONOCYTE:
9 MONOCYTE:precursor
7 MONOCYTE:precursor
25 MACROPHAGE:
28 MACROPHAGE:
54 MONOCYTE:
23 MACROPHAGE:
21 mDC:
37 mDC:
84 MONOCYTE:
86 MONOCYTE:
67 MONOCYTE:
91 MONOCYTE:
68 MONOCYTE:
94 MONOCYTE:
Cluster

Figure S7

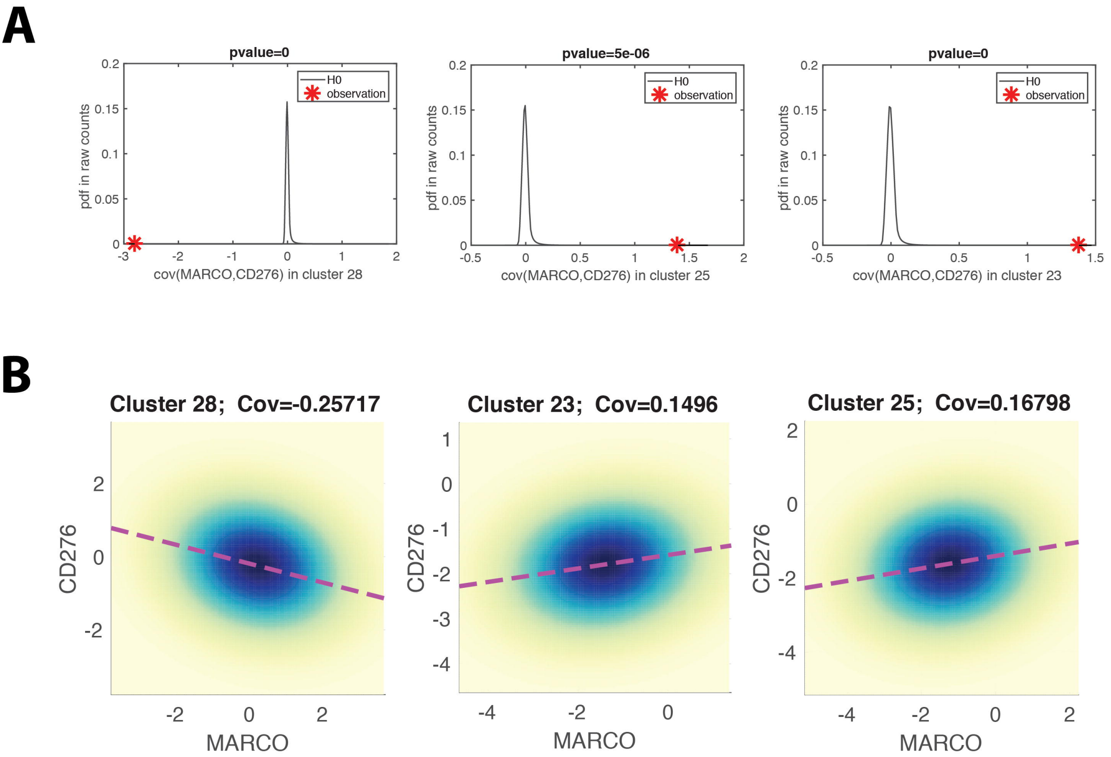
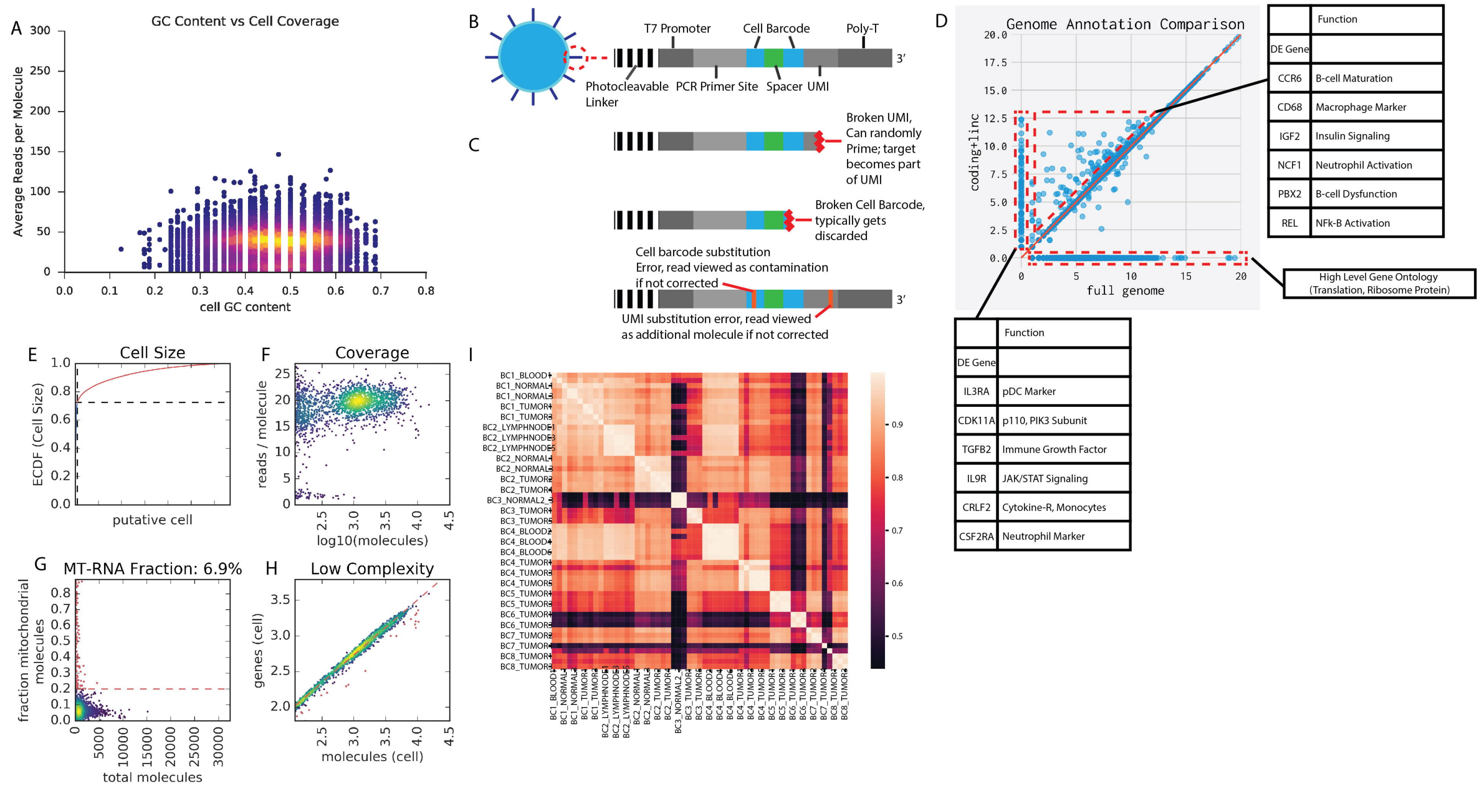

A
pvalue=0
pvalue=5e-06
pvalue=0
H0
observation
pdf in raw counts
0.2
0.15
0.1
0.05
0
-3
-2
-1
0
1
2
cov(MARCO,CD276) in cluster 28
-0.5
0
0.5
1
1.5
2
cov(MARCO,CD276) in cluster 25
-0.5
0
0.5
1
1.5
cov(MARCO,CD276) in cluster 23
B
Cluster 28; Cov=-0.25717
Cluster 23; Cov=0.1496
Cluster 25; Cov=0.16798
CD276
MARCO
2
0
-2
1
0
-1
-2
-3
-4
2
0
-2
-4
-2
0
2
-4
-2
0
-4
-2
0
2

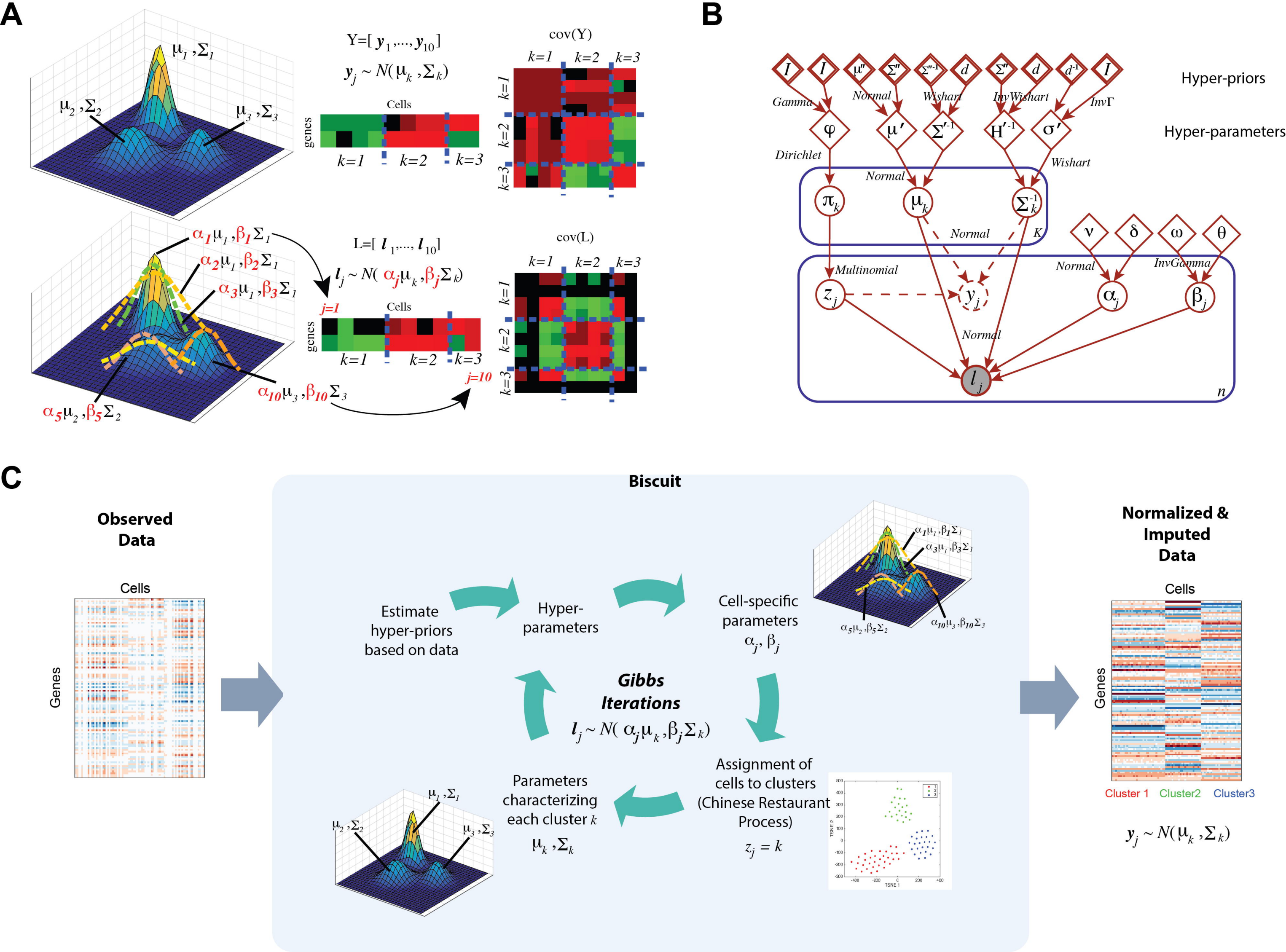
Figure M1



	Function
DE Gene	
CCR6	B-cell Maturation
CD68	Macrophage Marker
IGF2	Insulin Signaling
NCF1	Neutrophil Activation
PBX2	B-cell Dysfunction
REL	NFk-B Activation

	Function
DE Gene	
IL3RA	pDC Marker
CDK11A	p110, PIK3 Subunit
TGFB2	Immune Growth Factor
IL9R	JAK/STAT Signaling
CRLF2	Cytokine-R, Monocytes
CSF2RA	Neutrophil Marker

Figure M2


A
μ_1, Σ_1
μ_2, Σ_2
μ_3, Σ_3
$Y = [\, y_1, \ldots, y_{10}]$
$y_j \sim N(\mu_k, \Sigma_k)$
Cells
genes
k=1
k=2
k=3
cov(Y)
$\alpha_1\mu_1, \beta_1\Sigma_1$
$\alpha_2\mu_1, \beta_2\Sigma_1$
$\alpha_3\mu_1, \beta_3\Sigma_1$
$\alpha_{10}\mu_3, \beta_{10}\Sigma_3$
$\alpha_5\mu_2, \beta_5\Sigma_2$
$L = [\, l_1, \ldots, l_{10}]$
$l_j \sim N(\alpha_j\mu_k, \beta_j\Sigma_k)$
j=1
j=10
cov(L)
B
I
I
μ''
Σ''
Σ''^{-1}
d
Σ''
d
d^{-1}
I
Hyper-priors
Gamma
Normal
Wishart
InvWishart
InvΓ
φ
μ'
Σ'^{-1}
H'^{-1}
σ'
Hyper-parameters
Dirichlet
Wishart
Normal
π_k
μ_k
Σ_k^{-1}
Normal
K
ν
δ
ω
θ
Multinomial
Normal
InvGamma
z_j
y_j
α_j
β_j
Normal
l_j
n
C
Biscuit
Observed Data
Cells
Genes
Estimate hyper-priors based on data
Hyper-parameters
Cell-specific parameters α_j, β_j
$\alpha_1\mu_1, \beta_1\Sigma_1$
$\alpha_3\mu_1, \beta_3\Sigma_1$
$\alpha_5\mu_2, \beta_5\Sigma_2$
$\alpha_{10}\mu_3, \beta_{10}\Sigma_3$
Gibbs Iterations
$l_j \sim N(\alpha_j\mu_k, \beta_j\Sigma_k)$
Assignment of cells to clusters (Chinese Restaurant Process)
$z_j = k$
TSNE 1
TSNE 2
Parameters characterizing each cluster k
μ_k, Σ_k
μ_1, Σ_1
μ_2, Σ_2
μ_3, Σ_3
Normalized & Imputed Data
Cells
Genes
Cluster 1
Cluster2
Cluster3
$y_j \sim N(\mu_k, \Sigma_k)$