MetaTiME: Meta-components of the Tumor Immune Microenvironment

Yi Zhang[1,2], Guanjue Xiang[1,2], Alva Yijia Jiang[1], Allen Lynch[1,2], Zexian Zeng[1,2], Chenfei Wang[1,2], Wubing Zhang[1,2], Jingyu Fan[1,2], Jiajinlong Kang[1], Shengqing Stan Gu[4], Changxin Wan[1,2], Boning Zhang[1,2], X. Shirley Liu[1,2,3*], Myles Brown[3,4*], Clifford A Meyer[1,2,3*]

1. Department of Data Science, Dana-Farber Cancer Institute, Boston, MA 02215, USA

2. Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA 02215 USA.

3. Center for Functional Cancer Epigenetics, Dana-Farber Cancer Institute, Boston, MA, USA

4. Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA 02215, USA.

*Corresponding author. Email: cliff_meyer@ds.dfci.harvard.edu

*Co-corresponding author. Email: myles_brown@dfci.harvard.edu

*Co-corresponding author. Email: xsliu.res@gmail.com

Abstract

Recent advances in single-cell RNA sequencing have revealed heterogeneous cell types and gene expression states in the non-cancerous cells in tumors. The integration of multiple scRNA-seq datasets across tumors can reveal common cell types and states in the tumor microenvironment (TME). We developed a data driven framework, MetaTiME, to overcome the limitations in resolution and consistency that result from manual labelling using known gene markers. Using millions of TME single cells, MetaTiME learns meta-components that encode independent components of gene expression observed across cancer types. The meta-components are biologically interpretable as cell types, cell states, and signaling activities. By projecting onto the MetaTiME space, we provide a tool to annotate cell states and signature continuums for TME scRNA-seq data. Leveraging epigenetics data, MetaTiME reveals critical transcriptional regulators for the cell states. Overall, MetaTiME learns data-driven meta-components that depict cellular states and gene regulators for tumor immunity and cancer immunotherapy.

Keywords

Tumor microenvironment, cancer immunity, single-cell RNA sequencing, transcriptional regulators, gene expression cell state, cell state representation.

Introduction

Recent advances in cancer research have revealed the integral role of the tumor microenvironment (TME) in tumor progression and therapy responses[1–6]. Understanding interactions between cancer cells and the non-cancer compartments, including immune cells, fibroblasts, and endothelial cells, has revealed potential targets for cancer immunotherapy. Specifically, single-cell RNA-sequencing (scRNA-Seq) applied on multiple patient tumors has enabled the high-resolution identification of TME constituents that interfere with the elimination of cancer cells. For example, exhausted tumor-infiltrating lymphocytes (TILs)[4,7,8], and certain tumor-associated macrophages subtypes[9–11], have been associated with tumor development. However, the definition of cell types and cell states in tumor scRNA analyses still relies on manual labeling by experts using known exclusive biomarkers following unsupervised clustering[12,13], which lacks consistency and varies between different cohorts.

As single-cell data accumulate, integrating a large collection of cells from multiple cohorts can help unify the definition of cell types and states to facilitate the automatic annotation of new scRNA-seq data[14,15]. One approach to cell annotation is to use predefined biomarker lists. However, these biomarkers might not cover domain-specific cellular states, for example, well-defined immune cell markers derived from blood immune cells may not fully cover the TME disease context[14]. Moreover, although cell type definitions in reference databases such as CIBERSORT, Azimuth, and Human Primary Cell Atlas[16–18] can be useful, the granularity of

these definitions varies between databases. Several efforts integrating pan-cancer scRNA data have revealed subtypes in the TME through the manual annotation of clusters using a shortlist of exclusive gene markers[13,19,20].

Another approach is use structure inferred from the data to map cell states from unannotated datasets onto annotated ones. Methods to obtain such representations include canonical correlation analysis (CCA)[21], adjusted principal components (Harmony)[22], or generative deep learning models using variational autoencoders (scVI)[23]. These methods use dimension reduction onto a common latent space to align cells with similar states between datasets, without ascribing meaning to the latent space representations. An alternative data driven approach is to identify low dimensional latent space representations in which a biological meaning can be ascribed to each latent dimension. Several matrix factorization algorithms have been developed to represent high dimensional data in a low-dimensional space with interpretable components, including non-negative matrix factorization (NMF)[24] and independent component analysis (ICA)[25,26].

In this study, we developed a computational framework for mapping millions of single cells from multiple cohorts onto a comprehensive and interpretable latent space, learnt from the data. The framework, MetaTiME (Meta-components of the Tumor immune MicroEnvironment), identifies reproducible low-dimensional meta-components that reflect independent components of gene expression variation across cohorts and cancer types. MetaTiME adopted ICA for dimensional reduction to maximize the mutual independence among components. We used MetaTiME to obtain meta-components (MeCs) from 1.7 million single cells across 79 tumor datasets. These MeCs represent the TME landscape along 75 data-driven transcriptional directions mirroring lineage-specific cell states and signaling activities. Furthermore, we developed a MetaTiME toolkit for using the MeCs to annotate cellular states and signature continuums in tumor scRNA

datasets, and to reveal differential signatures across immunotherapy responses. Finally, by incorporating transcription factor binding data, MetaTiME revealed and prioritized putative transcriptional regulators that may modulate tumor immunity.

Results

MetaTiME as a general framework to discover consensus transcriptomic programs

The MetaTiME framework consists of three stages: meta-component (MeC) discovery, interpretation of MeCs, and application of cell state annotations (**Fig.1a**). The MeC discovery stage detects repeatable sources of variation from multiple single-cell measurements sharing similar cellular properties. The MeC interpretation step involves a one-time curation effort using biomarker databases, pathway information and Cistrome DB chromatin profiling data[27]. In the third step, users map MeCs onto their new tumor scRNA-seq datasets using MetaTiME application tools, to obtain annotated cell states and signature continuums.

To train MeCs for the TME context, we collected and curated 2,157,387 cells from 76 studies ranging across 27 cancer types, using publicly available tumor scRNA-Seq data mostly from TISCH[28]. After removing the TISCH annotated malignant cells using MAESTRO[17], 102,703 stromal cells and 1,617,110 immune cells were retained for downstream training (**Fig.S1**). The 76 studies were further partitioned according to cancer type, resulting in 93 datasets, including 7 datasets with immune checkpoint blockade (ICB) treatment and 3 10x Genomics provided datasets representing peripheral blood mononuclear cells (PBMC) sampled from healthy donors.

In the MeC discovery stage, MetaTiME first decomposes the log-transformed expression matrix of each single dataset using Independent Component Analysis (ICA)[25]. We adopted ICA to maximize mutual independence among gene expression components. In simulations ICA

105 performed slightly better than Non-negative Matrix Factorization (NMF) in simulated single-cell
106 data with pre-embedded transcriptional signatures (**Fig.S2**). The feature weight distribution of
107 each Independent Component (IC) also enables normalization of the gene contribution scores
108 for measuring similarity among components. MetaTiME then applies two transformations to the
109 IC vectors, z-weight normalization and skewness alignment, to ensure the scales of gene
110 representation scores are comparable among components (**Methods**). Next, MetaTiME filters
111 ICs to retain ones that are reproducible across multiple cohorts. These are passed to a graph
112 clustering algorithm to merge IC groups into MeCs (**Methods**). Lastly, MetaTiME computes
113 averaged profiles of gene z-weights within each IC cluster, yielding 86 MeCs trained for the
114 TME (**Fig.1a, Fig.2a**). The number of MeCs was automatically determined by simultaneously
115 optimizing granularity and independence in IC clustering (**Fig.S3**). Importantly, the MeC
116 clustering does not depend on cohort source (**Fig.S4**). This integration after decomposition
117 approach overcomes batch effects, which are often a challenge in single cell RNA-seq data
118 analysis.
119
120 **MetaTiME defines interpretable meta-components**
121 In principle, each MeC represents one independent source of transcriptional variation commonly
122 present in the TME. We investigated top ranked genes in MeCs and found MeCs are highly
123 interpretable, reflecting common biological processes in the TME. For instance, the MeC
124 derived from the largest IC cluster is highly enriched in interferon response genes, such as
125 *ISG15*, *IFI6*, *LY6E*, and *MX1*, indicating that the underlying interferon response is among the
126 most common source of transcriptional variation shared across tumor samples and cohorts
127 (**Fig.2a, b**). Intriguingly, top genes of each MeC are enriched in known biomarkers or regulators.
128 For example, several T cell-related MeCs identify different gene modules co-expressed in T
129 cells reflecting activation of different T-cell related processes (**Fig.2c, Fig.S5**). The “T cell co-
130 signaling” MeC features T cell receptors in co-stimulatory and co-inhibitory pathways[29,30], such

as *TNFRSF4* (OX40), *TNFRSF18* (GITR) *TNFRSF9* (4-1BB), and *ICOS* (**Fig.2c**, left). The "CXCL13, exhausted CD8 T cell" MeC features receptors characterizing the exhausted CD8 T cell state[8], including *HAVCR2* (TIM3), *LAG3*, *TIGIT*, and *PDCD1* (PD1) (**Fig.S5**), each being potential ICB targets[31]. In addition, this MeC is characterized by a high level of *CXCL13* (**Fig.2c**, second panel), a cytokine mediating immune cell trafficking to tertiary lymphoid structures[32]. In contrast, a related MeC representing T cell co-signaling receptors in regulatory CD4T cells (Treg) has a different ranking, including *TNFRSF18*, *TNFRSF4*, *TIGIT, TNFRSF1B*, *CTLA4*, *CD27* among the top 20 genes, along with the regulatory T cell-specific marker *FOXP3* (**Fig.2c**, right). Though ICB has been an extremely successful therapy for some patients, it has not yet had an impact on the majority of patients[33]. Investigating the top members in the MeCs involving T cell receptor pathways may help identify new ICB targets.

MetaTiME depicts the functional landscape of transcriptomic variation and cell states in the tumor microenvironment

We provided functional annotations of all MeCs by examining top z-weight genes and compared these with functional gene sets, such as immune cell type markers[15,18] and gene ontology databases[34]. We found that 75 MeCs clearly mirror gene expression patterns corresponding to cell types, cell states and signaling pathway activities, depicting a landscape of non-cancer cell states in the TME (**Fig.2, Table S1**: MeC annotation). The top genes of the cell type MeCs match well-known lineage-specific markers[15,18]. Examples include *CD74*, *CD79A*, *MS4A1* for B cells ("B cell" MeC), *CD3D*, *CD8A*, *CD8B* for T cells ("CD3 - CD8 T cell" MeC), and *LYZ*, *VCAN*, *S100A9* for CD14+ Monocytes ("CD14 monocyte" MeC) (**Fig.2b, Fig.2d**). The majority of MeCs define high resolution lineage-specific cell states (**Fig.2b, Fig.S4b**). Taking the B cell lineage as an example, multiple MeCs harbor genes specific to B cell developmental stages[35], ranging from a progenitor B cell state (*CD69* and *PAX5* in the "PAX5 B cell"), to a mature B cell state (*CD79A* in the "B cell" MeC), an antibody-secreting plasma cell state (*XBP1* in the "plasma B

cell" MeC, and *JCHAIN* in the “alternate plasma B cell” MeC (“Bplasma_1”), and immunoglobulin secretion states (IGK and IGH in the “immunoglobin kappa B cell" MeC, IGL and IGH in the “immunoglobin lambda B cell” MeC) (**Fig.2b**). Lastly, like the interferon responsive MeC mentioned above, we found a subset of MeCs that are more accurately interpreted as signaling pathways because their top genes are more related to pathways or molecular functions than to cell identities.

We organized the 75 annotated MeCs into six cell lineage-focused categories and one signaling pathway-focused category (**Fig.2a,b** and **Table S1**: MeC annotation, MeC enrichment). Among these, there are 6 B lineage-related MeCs for B cells; 19 T cell lineage MeCs covering CD8 T cells, CD4 T cells, and natural killer (NK) cells; 4 dendritic cell (DC) lineage MeCs; 12 monocyte and macrophage-related MeCs; 3 platelet, erythrocyte, and mast cell MeCs; 6 stromal cell-related MeCs for fibroblasts, myofibroblasts and endothelial cells; and 25 MeCs in the signaling category (**Fig.2b**). We demonstrated that the MeCs are of high specificity, visualizing the z-weights of known cell subtype markers and pathway biomarkers (**Fig.2d**). Correlating MeCs with the comprehensive immune cell type database Azimuth[15] validated the lineage-specificity of several MeCs, while most MeCs reflect cell states that appear specific to the tumor context (**Fig.S4b**).

MetaTiME annotates cell states and signature continuums when applied to the tumor microenvironment single-cell data

As MetaTiME MeCs provides a highly interpretable basis for the TME in single cells, we provided a toolkit to reveal MeC signature continuums and enriched cell states in scRNA-seq TME data (code deposited in https://github.com/yi-zhang/MetaTiME). The MetaTiME annotation toolkit takes as input the scRNA-seq expression matrix after depth normalization and log transformation, maps each single cell onto the pre-trained MeC space, and annotates the most

183 highly enriched cell states for pre-defined cell clusters. The cell clusters are by default
184 calculated using graph clustering with high resolution after an optional batch effect correction
185 with Harmony[22]. We demonstrate usage of MetaTiME on basal cell carcinoma (BCC) single-
186 cells from *Yost et al.*[8], with enriched cell states annotations (**Fig.3a**) highlighting gradients of
187 exhausted CD8 T cells and follicular helper T cells (Tfh) (**Fig.3c**). The most enriched cell states
188 consistently match the manual labelling from the original study with improved resolution
189 (**Fig.3b**). In addition, compared to the Seurat's[14] automated CIBERSORT marker-based
190 annotations, MetaTiME provides higher resolution (**Fig.S6a**). A few other automatic annotation
191 gene panels were also tested, including the human primary cell atlas (HPCA) panel and
192 Blueprint-ENCODE panel used in SingleR[16], where macrophages and plasma cells appear to be
193 mislabeled as subclusters within T cell clusters (**Fig.S6b, c**). Interestingly, the MetaTiME
194 annotation not only indicates the CD8 T cell and CD4 T cell subtypes, but also splits cells further
195 into cell states with polarized expression in proliferation, cytotoxicity, exhausted level, heat
196 stress, co-signaling pathways, *etc.* (**Fig.3a, Fig.S6a**). The B cell group is further partitioned into
197 distinct B cell developmental states including a B cell cluster with cell cycle and MYC activities
198 (**Fig.3a,b, Fig.S6**), which possibly represent germinal center (GC) B cells undergoing active
199 expansion and maturation[36].
200
201 We thus re-annotated all tumor scRNA cohorts using MetaTiME and investigated the distribution
202 of cell state compositions across cancer cohorts. As shown in **Fig. 3e** and **Fig.S7**, tumors are
203 highly heterogenous and the TME cellular composition is only partially determined by cancer
204 type. For example, Cholangiocarcinoma (CHOL) is highly enriched in stromal cells including
205 collagen-secreting fibroblast, as expected[37], while other samples including ovarian cancer (OV),
206 pancreatic adenocarcinoma (PAAD), and multiple myeloma (MM) are also stromal-rich.
207 Furthermore, tumors with high infiltration of the "GZMK+ CCL5+ CD8 T cell" state include

multiple tumor types including colon cancer, breast cancer, and skin cancer, suggesting that immune infiltration is sample-dependent and that cancer treatments should be personalized[38].

Differential MetaTiME analysis detects alterations of transcriptional programs in immunotherapy.

Single-cell data derived from ICB trials is invaluable for identifying cell types associated with ICB treatment or response[8]. However, the detection of differential cell types abundances in ICB cohorts has been challenging due to the heterogeneity of cell type proportions and to the limited numbers of patients in each cohort[39]. We compared differences in MeC signatures instead of cell count proportions, to understand immune response during ICB. We analyzed two ICB cohorts, a basal cell carcinoma (BCC) cohort with samples from pre- or post-ICB treatment[8], and a bladder cancer (BLCA) cohort with samples from ICB responders and non-responders[40]. We applied MetaTiME for per-cluster cell state annotation and per-cell MeC signature evaluation. For each cell state cluster, we tested all MeC signatures passing significance (average z-weight>2) between conditions using the Wilcoxon rank-sum test. We plotted cluster-wise signatures in the significance – effect size scatterplot to highlight the most significant differential MeCs (**Methods**). In a comparison of pre- and post-ICB treatment, we observed higher expression of cytotoxic T cell and B cell MeCs in the post-ICB samples. Moreover, several monocyte and macrophage states are also suppressed after ICB treatment (**Fig.4a**). Notably, the IL1B-positive macrophage signature is also found to be elevated in non-responders compared to responders in the BLCA ICB cohort (**Fig.4b**). Since activation of the IL1B pathway is a known regulator of inflammatory processes[41], we sought to investigate whether the IL1B-positive macrophage signature is associated with tumor survival prognosis in bulk RNA-seq data from The Cancer Genome Atlas Program (TCGA). We evaluated TCGA tumors using the averaged expression of the top 20 genes from the “Macrophage IL1-NFkB” MeC which ranks first in elevated MeCs in non-responders (**Fig.4a**). We found that higher expression of the *IL1B*

signature is associated with lower survival rate in multiple cancer types, especially in Low Grade Glioma (LGG) and in Kidney renal cell carcinoma (KIRC) (**Fig.S8**). This suggests the macrophage state with *IL1B* pathway activation is associated with poor prognosis and lower ICB efficacy.

MetaTiME delineates myeloid cells in different metabolic states

As specific myeloid cell states have been associated with cancer survival and treatment response, we sought to systematically characterize MeCs related to monocytes and macrophages. Although the canonical definition of M1 and M2 macrophages is derived from cytokine polarized macrophages *in vitro*[42], MetaTiME's myeloid-related MeCs represent a more complex framework for understanding tumor-infiltrating macrophages. MetaTiME's 12 monocyte and macrophages related MeCs can be summarized into six central monocyte or macrophage states for the TME, after merging similar states such as "Macrophage IL1-NFkB" and "Macrophage IL1- JUN" due to similarity among top genes (**Fig.4c**). Monocytes are classified as two categories, CD14+ and CD16+. For macrophages, four MeCs define common states of intra-tumor macrophages: C1Q+, SPP1+, lipid-rich, and IL1B+ macrophages, and two MeCs, representing interferon and MHC-II signaling pathways, are less frequently observed among macrophages (**Fig.2b, Fig.4c**). In comparison, previous studies defined different TAMs in terms of manually selected representative genes after clustering myeloid cells. For example, *Cheng et al.*[10] defined several TAM types including *ISG15*+, *SPP1*+, *INHBA*+, *VCAN*+, *NLRP3*+, and *FN1*+ TAMs, while *Bi et al.*[43] defined *CXCL10*-high, *GPNMB*-high, *FOLR2*-high, *VSIR*-high, and cycling TAMs for advanced renal cell carcinoma (ccRCC). We find that the MetaTiME-defined myeloid MeCs reflect co-expression relationships with the selected marker genes. For example, TAM markers from *Bi et al*. rank high in several myeloid MeCs (**Fig.S9**), and the expression pattern of the marker genes picked by *Bi et al.* (*CXCL10*, *GPNMB*, *VSIR*, *FOLR2*, Cycling marker *MKI67*) correspond to several MeCs ("interferon responsive", "PPARG+ lipid-rich",

"MHCII-high", and "RNASE1+,C1Q+" MeCs) (**Fig.S9**). However, MetaTiME reveals additional distinct components such as the "SPP1+" and "C1Q+" MeCs, which were detected as separate myeloid types in the *Cheng et al.* multi-cohort study (**Fig.S10**). While the manual reconciliation of cell types from multi-cohort scRNA data shows many marker genes to be consistent with the top genes in the MetaTiME MeCs, the myeloid cell population is not neatly partitioned into cell clusters and might be better represented in terms of expression signature continuums. For example, when mapping myeloid MeCs onto the kidney myeloid cells, the "IL1B+" MeC signature is distributed across the "Macro_IL1B" cluster as well as the CD14 monocyte cluster (**Fig.S10**).

To investigate functional differences among the different macrophage states, we applied gene set enrichment (GSEA)[34] analysis using the top MeC genes. Interestingly, the different macrophage states have different metabolic preferences (**Fig.4d**). Glucose metabolism and the glycosylation pathway are highly active in SPP1+ macrophages, while lysosome and phagosome activity are the most highly enriched in C1Q+ macrophages. Lysosome and cholesterol metabolism, including *PPARG* signaling, are enriched in the lipid-rich state. The inflammatory IL1B and NFkB pathways are highly active in IL1B+ macrophages. Several macrophage states are related to cell signaling. *SPP1* for example, encodes Osteopontin, which has been found to foster an environment that promotes cancer metastasis[44]. The C1Q+ MeC features *C1QA*, *C1QB*, and *C1QC*, members of the family of complement molecules that could play dual roles in chronic inflammation[45]. The IL1B+ meta-components features cytokines co-expressed with *IL1B*, including *CXCL8*, *CXCL2*, and *CXCL3*, all of which can interact with other cells in the TME by binding to cytokine receptors[46] (**Fig.2b**).

Incorporation of epigenetic data prioritizes transcriptional regulators of tumor immunity

We next investigated the transcription factors (TFs) that regulate the MeCs, hypothesizing that the co-expression of genes in a subset of MeCs is determined through TF regulatory events. Our group previously developed the Cistrome Data Browser and Lisa to predict transcriptional regulators of gene sets based on chromatin immunoprecipitation with sequencing (ChIP-seq) data[27,47]. Thus, we used Lisa to predict the TFs that regulate the top genes of each MeC, and compared these Lisa regulatory prediction scores with the MeC z-weights across TFs. We found that, for many MeCs, the same TFs were predicted to be both regulators of the MeC and were highly expressed in the MeC itself, indicating an autoregulatory control scheme. Often, however, TFs that were predicted by Lisa to be MeC regulators were not represented by high MeC z-weights, and TFs with high MeC z-weight were not always found to have high Lisa scores (**Fig.5, Table S1**: MeC regulators). TFs predicted by Lisa but not represented by high MeC z-weight could be the result of TF activities being regulated through non-transcriptional mechanisms[48] or multiple TFs in a family having similar binding patterns but only a subset being the regulators[49]. TFs that have high MeC z-weights but low Lisa scores are most likely not well represented in the relevant cell types in available ChIP-seq data. In the “interferon response” MeC, *STAT1* is highly represented in the MeC z-weight and Lisa ranks *STAT1* as the top regulator, consistent with *STAT1* being known as the master regulator of the interferon response (**Fig.5a**). Several lineage-defining TFs display the autoregulatory pattern, including *TCF4* in plasmacytoid dendritic cells (pDC) (**Fig.5b**) and *XBP1* in B plasma cells (**Table S1:** MeC regulators**)**. The macrophage related MeCs are regulated by myeloid lineage TFs like *CEBPB*, and TFs related to immune stimulus responses, including NFkB complex TFs. In the "lipid-rich macrophage" MeC, although PPARG ranks among the top Lisa-predicted regulators (**Fig.5c**), *PPARG* expression is not highly represented as a MeC z-weight. In this MeC, the top co-expressed genes are indeed enriched in the PPARG signaling pathway (**Fig.4d**); this result can be accounted for by PPARG being regulated through its ligands, which include a variety of lipophilic acids[48].

We found glucocorticoid receptor (GR) signaling to be implicated in the regulation of the "CXCL13+ Tfh" MeC, with GR being most highly ranked TF in both MeC and Lisa scores (**Fig.5d**). Top genes in the "CXCL13+ Tfh" MeC include several direct target genes of GR, including *SRGN* and *FKBP5* (**Fig.S11b**). We investigated whether the *CXCL13* cytokine itself could be a direct target of GR in *CXCL13* secreting Tfh cells. Since GR ChIP-seq data is not available for the exact Tfh cell state, we collected GR ChIP-seq data from several other cell types. Direct binding of GR is observed at the *CXCL13* gene promoter and nearby the gene locus at putative enhancers, which are conserved across multiple cell lines (**Fig.S11a**), including the B cell line Nalm6, the monocyte cell line THP1, and cancer cell lines. Moreover, in another CXCL13 secreting cell state, the "CXCL13 exhausted CD8 T cell" MeC, GR is also highly ranked in both Lisa and MeC scores (**Fig.5d**). Thus, we hypothesize that GR is likely to be a transcriptional driver of the CXCL13-secreting cell states in exhausted CD8 T cells[50] as well as in CD4 T follicular helper T cells. Thus, the GR pathway could be a candidate target in tumor immunity modulation.

Discussion

We developed the MetaTiME (Meta-components of the tumor immune Microenvironment) framework and performed a large-scale and pan-cancer integration of tumor single cell datasets using ICA to optimize information independence among components[51]. We identified 75 interpretable meta-components (MeCs) that describe common aspects of TME gene expression variation across multiple tumors.

The MetaTiME MeCs serve as comprehensive transcriptional signatures that depict a functional landscape of TME transcriptional programs and cell states. For monocytes and macrophages,

the related MeCs revealed heterogeneity and plasticity of tumor-associated macrophages (TAMs). We thus propose that TAMs, especially for solid tumors, should be classified based on the major states with different metabolic preferences instead of the canonical M1 and M2 classification[42]. Similar states that do not fit well into the M1 versus M2 classification scheme were also observed in previous studies analyzing myeloid cells in the TME, where single cells were clustered and labeled using differential markers[10,12,43]. *Cheng et al.*[10] defined several TAM types by clustering myeloid cells separately for each study and naming the TAMs with manually selected top marker, chosen based on consistency across cohorts. Similarly, *Bi et al.*[43] defined TAM types by harmonizing patients and naming the TAMs with top genes in each cluster for an advanced renal cell carcinoma (ccRCC) cohort. Cell type definitions in the previous studies were based on representative genes, which were chosen differently in the respective studies. We propose that the MetaTiME derived monocyte and macrophage MeCs could be used to define macrophage states and functional co-expressed gene modules more consistently for the TME. Reexamining the marker genes from the previous studies: *NLRP3* is highly ranked in the "IL1B+" MeC; in fact, the NLRP3 inflammasome mediates interleukin-1β production. *GPNMB* is weighted among the top 20 genes in both the "C1Q+ macrophage" and "SPP1+ macrophage" MeCs; it encodes a membrane glycoprotein which is typically highly expressed in macrophages. *FOLR2* is ranked 29th in the "C1Q+ macrophage" MeC, indicating this macrophage state also encodes a high folate-activated pathway. Finally, the "SPP1/C1Q macrophage" MeC features an intermediate state with both SPP1 and C1QA, indicating the plasticity and mixed nature of pathways activated in TAMs that could not be defined using exclusive markers. Thus, the myeloid MeCs may provide a consistent definition of TAM states corresponding to different metabolic processes.

MetaTiME provides a toolkit for analyzing independent TME scRNA-seq datasets by mapping gene expression onto the MeC space. The outputs include signature continuums and the most

362 highly enriched cell states. Recent useful single-cell dataset integration algorithms such as
363 Harmony[22] and scArches[52] infer a joint low-dimensional representation among data. In these
364 approaches the shared space is re-computed every time a new dataset is incorporated. The
365 MetaTiME strategy builds upon previous approaches that transfer latent representations from
366 large datasets, but provides a stable and interpretable representation specialized for the TME.
367
368 By leveraging ChIP-seq data, MetaTiME reveals critical transcriptional regulators in tumor
369 immunity. In many cases, we found the joint consideration of MeC-specific co-expression
370 patterns and TF binding enrichments reveals the roles of TFs in defining cellular states and
371 gene expression programs. MetaTiME captured multiple known TFs critical to tumor immunity
372 and could serve as immune modulation targets; this includes *TOX* in the "CXCL13-secreting
373 exhausted CD8 T cell" MeC, a recently discovered regulator of T cell exhaustion[53] (**Table S1**).
374 The MeCs further implicated the glucocorticoid receptor pathway in the regulation of several T
375 cell states. Glucocorticoids are a class of steroid hormones essential to the modulation of
376 multiple biological processes, including immune related ones[54], although the role of the GR
377 pathway in different immune cell types is not fully understood. Since GR is broadly expressed in
378 many cell types, and is regulated through ligand binding, differential analysis of GR expression
379 is unlikely to fully capture GR regulation in single-cell data analysis. Though GR ChIP-seq is not
380 available in the contexts of the relevant T cell states, GR ChIP-seq in other cell lines
381 demonstrate robust binding nearby the top gene *CXCL13*. *CXCL13* is crucial to T follicular
382 helper cell communication with germinal center B cells, through interaction with its receptor
383 CXCR5[55,56].
384
385 Overall, MetaTiME depicts the functional landscape of transcriptomic variation and cell states in
386 the tumor microenvironment. It provides a computational framework to facilitate the elucidation

of the identity and function of cells in the TME in future studies and will facilitate the identification of potential new therapeutic targets for immune modulation.

Methods

Tumor single-cell RNA-seq data collection and processing

For an extensive collection of single cells from tumor microenvironment, we utilized the public tumor scRNA-seq collection from TISCH[28]. The TISCH collection uniformly processed each dataset with MAESTRO[17] and isolated non-malignant environmental cells from malignant cells. Overall, we collected 2,157,387 cells from 76 studies ranging 27 cancer types. The MAESTRO annotation was labeled using CIBERSORT gene panels[18] followed by curation, enabling selection of 1,719,813 environmental cells, including 1,617,110 immune cells and 102,703 stromal cells, were retained for integrative analysis in this study. For studies with data measured from multiple cancer types, cells different cancer types were split into independent datasets, resulting in 93 datasets; it includes 3 PBMC datasets from healthy donors from 10X Genomics as baseline and 7 datasets with ICB treatment.

For an unsupervised component analysis, each dataset was re-analyzed. For datasets with raw count matrices available, gene expression was normalized towards per-cell read depth 10,000 followed by log transformation. For datasets with only TPM or FPKM values available, including Smart-seq data or studies with only normalized matrix available, gene expression underwent log transformation. Cells were filtered based on minimum library size 1000, gene number 500, and maximum mitochondrial read proportion 5%.

Decomposing individual studies and denoising low-dimensional components.

We then decomposed the expression matrix of each scRNA-seq dataset using fastICA[51] into an independent component (IC) vector matrix and a projection weight matrix. We tested different values for the number of components (k) and chosen k to be 100 uniformly for each dataset, given it could cover more variations than the number of cell types in the TME, which is around twenty. We applied two denoising approaches to deal with sparsity and the potential noise of ICs. First, we performed a z-normalization of the gene loadings in the component, scaling all gene loading values by the standard deviation of each IC. The gene loadings indicate the degree of contribution to the component as a “metagene” from each gene, and we observed that most genes contribute neutrally to the metagene. Thus, genes with significant contributions are selected using the two-standard deviation threshold from either the positive or negative side. Second, we aligned the positive skewness of components since the sign of an independent component is randomly assigned in fastICA optimization. We observed that asymmetrically extreme gene loading values highlight genes representative of the component's function; thus, we computed each component's skewness statistics and flipped the sign of component loadings if the skewness is negative. We excluded genes with a low contribution (gene weight not passing two standard deviations) to any reproducible components and kept 6623 genes with potentials in driving the reproducible components. The post-decomposition steps ensured the attitude and sign of gene weights are comparable across cohorts, depicting degree of contribution from each gene in the genome-wide background.

Meta-components calling and functional annotation

We then aim to discover reproducible patterns from all components from each dataset. We evaluated similarity between pairs of components using cosine distance and retained a set of 1043 candidate reproducible components from 69 datasets, each with a minimum Pearson correlation coefficient 0.3 with at least one different IC. We then clustered ICs using Louvain

clustering, a graph-based community detection algorithm where the resolution parameter controls segmentation granularity. Clusters with at least five ICs were retained as reproducible IC clusters. The number of clusters is determined by optimizing both Silhouette's score for optimal within-cluster similarity compared to inter-cluster similarity and number of reproducible clusters. The final resolution parameter was chosen to be 1.25 resulting in 86 clusters for meta-component (MeC) calling. The consensus gene z-weights in each MeC were then calculated by averaging ICs in each cluster. Genes of outlier z-weights passing two standard deviations were highlighted as significant, and the ones with positive largest z-weights were considered representative of the MeC.

MetaTiME MeCs were assigned curated annotation by matching top z-weighted genes to functional biological information including cell type markers, pathway databases from GSEA, cell types expressing top MeC genes, and high-rank transcription factors. In details, GSEA enrichment analyses utilized top 100 highest z-weighted genes and TF database was obtained from AnimalTFDB[49]. The 86 MeCs were first ordered by MeC cluster size, and then organized into seven functional categories and one undefined category. The 11 MeCs in the undefined category are of smaller size, and harbor top genes that are related to cell stress. The remaining 75 functional MeCs were assigned six lineage-related categories and one signaling category.

Simulating multi-cohort single-cell RNA data with expression programs

To benchmark dimensional reduction methods, we built upon previous effort from *Kotliar et al.*[57] to use the scsim package to simulate multiple count matrices with built-in transcriptional programs. In principle, the built-in gene expression programs (GEP) were sampled as random scaling factors on a subset of genes mimicking overexpression or suppression of a pathway. For testing whether a higher number of cohorts facilitate GEP recovery, we simulated 20 single-

cell datasets and tested usage of 5 cohorts, 10 cohorts, and 20 cohorts. Each dataset was embedded with a subset of 14 pre-defined GEPs, since the real tumor scRNA data may not cover every possible cell types or gene programs in every dataset. The 14 GEPs contain 13 cell type-specific programs with distinct cell type-specific genes, and one signaling gene expression program that is randomly active in multiple cell types. Two low-dimensional reduction method are benchmarked using simulated scRNA data: independent component analysis (ICA) and non-negative matrix factorization (NMF). Decomposition was performed on each single cohort separately, and meta-component calling was done as similar in MetaTiME: components are filtered, clustered into meta-components, followed by averaging gene z-weights per cluster as predicted gene expression programs (GEP). The predicted GEPs were compared with pre-defined True GEPs using Pearson correlation. Overall, both ICA and NMF can recover GEPs, while the ICA-based GEPs are more mutually independent and performs slightly better. Since the GEP recovered in the 20 cohorts case matched true GEP better than 5 cohorts and 10 cohorts, the increased number of cohorts also improves GEP recovery. Thus, we chose ICA for component integration and use all available datasets for GEP discovery for tumor microenvironmental cells.

The MetaTiME annotator for analyzing new tumor scRNA-seq data.

MetaTiME provides an analytical toolkit for annotating cell states and signature activities for tumor scRNA-seq data (https://github.com/yi-zhang/MetaTiME). The scRNA-seq data is first processed following standard procedure, which includes cell depth normalization, log-transformation, batch effect removal using Harmony[22], neighboring graph construction, graph clustering, and UMAP embedding for visualization. Specifically, the clustering step uses an over-clustering strategy, which sets a high-resolution parameter (default 8) that generates a larger number of clusters and help reveal fine structures among the cells. Then, the MetaTiME annotator tool takes as input a single log-transformed expression matrix for TME cells from the

dataset. The outputs include both per-cell MetaTiME MeC signature scores and per-cluster enriched MeC state. For the per-cell score, MetaTiME projects each cell onto the MeC space by calculating dot product between the expression vector and the z-weight vector of each MeC, using genes passing significant z-weight criterion (z-weight >2). The projection matrix is then scaled across all cells to ensure normally distributed scores within each MeC, outputting the cell-by-MeC score matrix. Meanwhile, the UMAP view of the projection score shows the signature gradient across the cells positioned by similarity. Lastly, the cluster-wise MeC enrichment results are also generated. The per-cluster MeC enrichment score is calculated by averaging profile of cells along each MeC; MeCs with mean score passing the significant cutoff (2 in the z-weight scale) are called as the set of enriched MeCs. Each cluster may enrich multiple number of MeCs, and the top enriched MeC with highest score is used in UMAP visualization.

Differential MeC signature analysis

For tumor scRNA-seq data with different conditions, a differential signature analysis can be carried out following MetaTiME annotation, which provides enriched MeCs for each cluster and names each cluster with the top enriched MeC. Thus, for each cell cluster, the MeC signature strength can be compared across conditions, for all enriched MeC in the current cluster. In details, a simple Wilcoxon rank-sum test is adopted to compare MeC scores of cells in one condition with another. The log-fold change of MeC scores were calculated by the ratio between cell means from the two conditions in comparison. To plot the cluster-specific differential signature plot, the signatures are marked using “EnrichedMeC@ClusterName”, where the “ClusterName” is the top first enriched MeC used as cell state as current cluster. Thus, when the enriched MeC signature is the same with the cluster name, the differential signature is named as “ClusterName” on the Significance-Effect size plot.

Incorporation of epigenetic data using Lisa

Our group previously developed Lisa (Landscape In-Silico deletion Analysis) that predict influence of TFs on a set of genes. Lisa models public chromatin accessibility and TF binding profiles to score TFs in gene regulation from an epigenetic perspective. We developed Lisa2 that improves on running speed and pipeline integration, which is applied on each MeC to score TFs in regulation potential on top 100 high z-weighted genes. The impact scores of TFs are thus from two sources: MeC z-weights for expression representation, and Lisa scores for binding potential. The TFs are grouped into three classes, TFs highly ranked based on both MeC gene weights and Lisa significance, TFs representative only in MeC, and TFs based on binding information only. In TableS1, we marked TFs from different classes in different colored columns. Significant TFs based on both MeC and Lisa (MeC z-weight >=2, Lisa score -log p-value >=2) are marked as orange color; furthermore, the TFs ranking among top 40 (aggregated rank of MeC and Lisa) compared to all genes are further marked red. TFs ranking among top 10 only in MeC z-weight are colored green, and TFs ranking among top 10 only in Lisa score are colored blue.

Grant information

NIH grants P01CA163222, U24CA237617, and R01HG011139.

Acknowledgement

We acknowledge all authors of the original tumor single-cell studies that generated publicly available data. The results appearing here are in part based upon the data generated by the TCGA Research Network (http://cancergenome.nih.gov/).

Conflict of Interest

MB is a consultant to and receives sponsored research support from Novartis. MB serves on the SAB of H3 Biomedicine, Kronos Bio, and GV20 Oncotherapy. XSL is a cofounder, board member, SAB member, and employee of GV20 Oncotherapy and its subsidiaries; is a stockholder of BMY, TMO, WBA, ABT, ABBV, and JNJ; and received research funding from Takeda, Sanofi, and Novartis. The remaining authors declare no competing interests.

Figures

Fig.1.

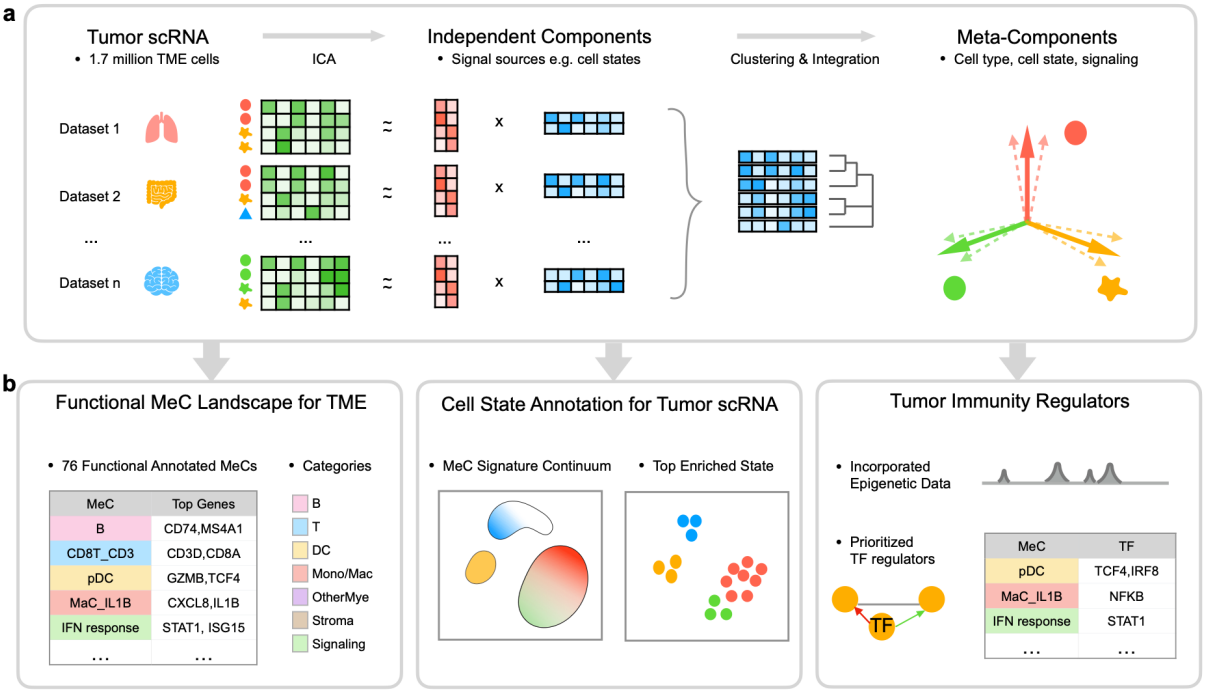
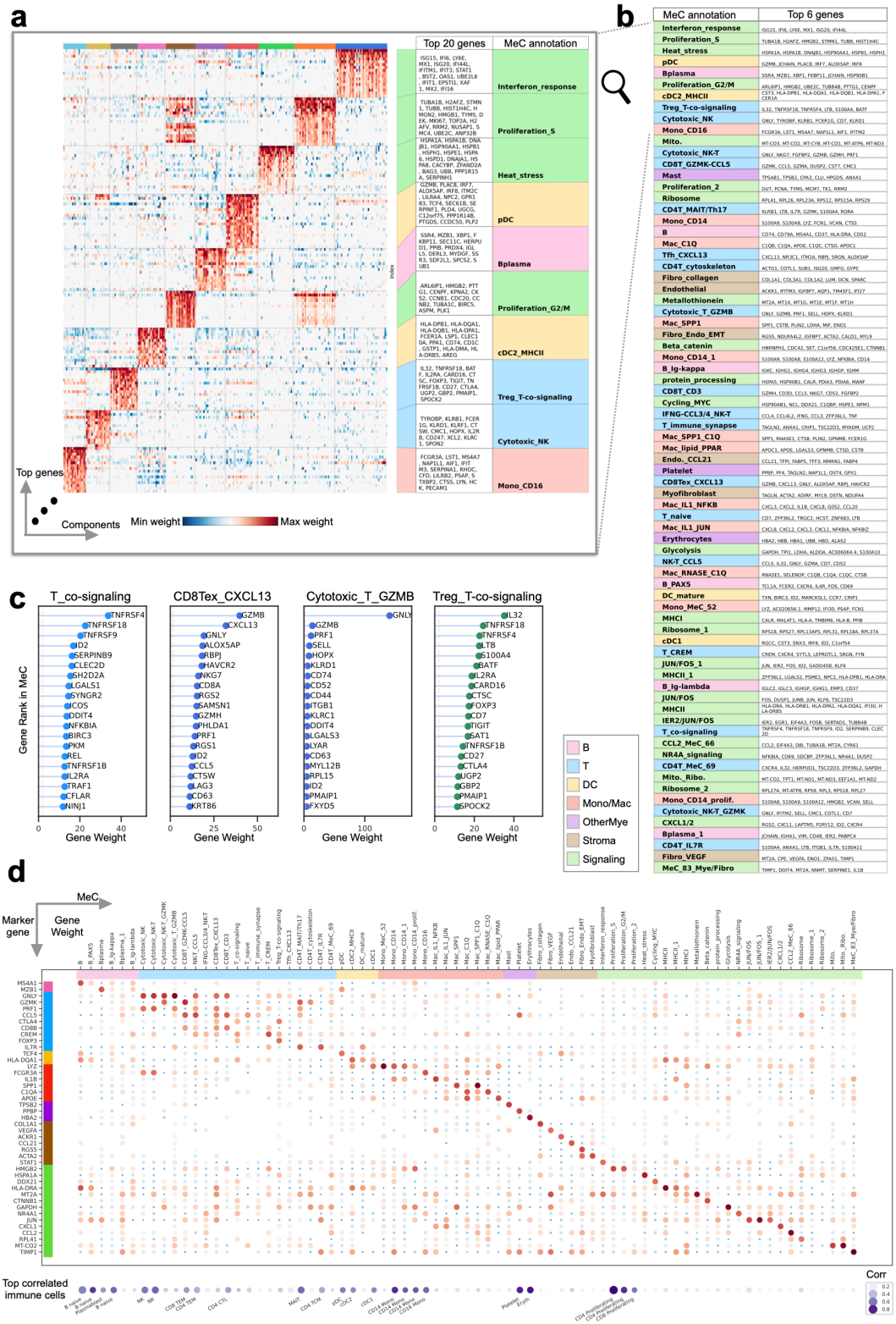


Fig.1. Overview of MetaTiME. MetaTiME integrates 1.7 million single cells to learn common transcriptional programs in the tumor microenvironment (TME). **(a)** Steps for Meta-components (MeCs) discovery. For each scRNA dataset, the expression matrix of TME cells is decomposed into a loading matrix (red) and an independent component (IC) matrix through independent component analysis (ICA). The ICs represent mutually independent sources of transcriptional variation. ICs from each dataset are concatenated and clustered into groups of ICs with high similarity, representing transcriptional programs shared across TME. MeCs are then calculated as averaged profiles of ICs from each cluster. Each MeC is interpretable representing gene signatures of cell type, cell states, or signaling pathway activities. **(b)** Left: MetaTiME provides 75 functionally annotated MeCs that depict the TME transcriptional landscape. They are grouped into six lineage-related categories and one category reflecting signaling activities. Middle: the MetaTiME annotation tool facilitates automatic annotation of cell states for new

562 tumor scRNA data. Right: candidate regulators of each MeC are prioritized by combining MeC
563 gene weights with epigenetics data. MeC: meta-components, TME: tumor microenvironment,
564 ICA: independent component analysis, MeC: meta-component, TF: transcription factor.
565

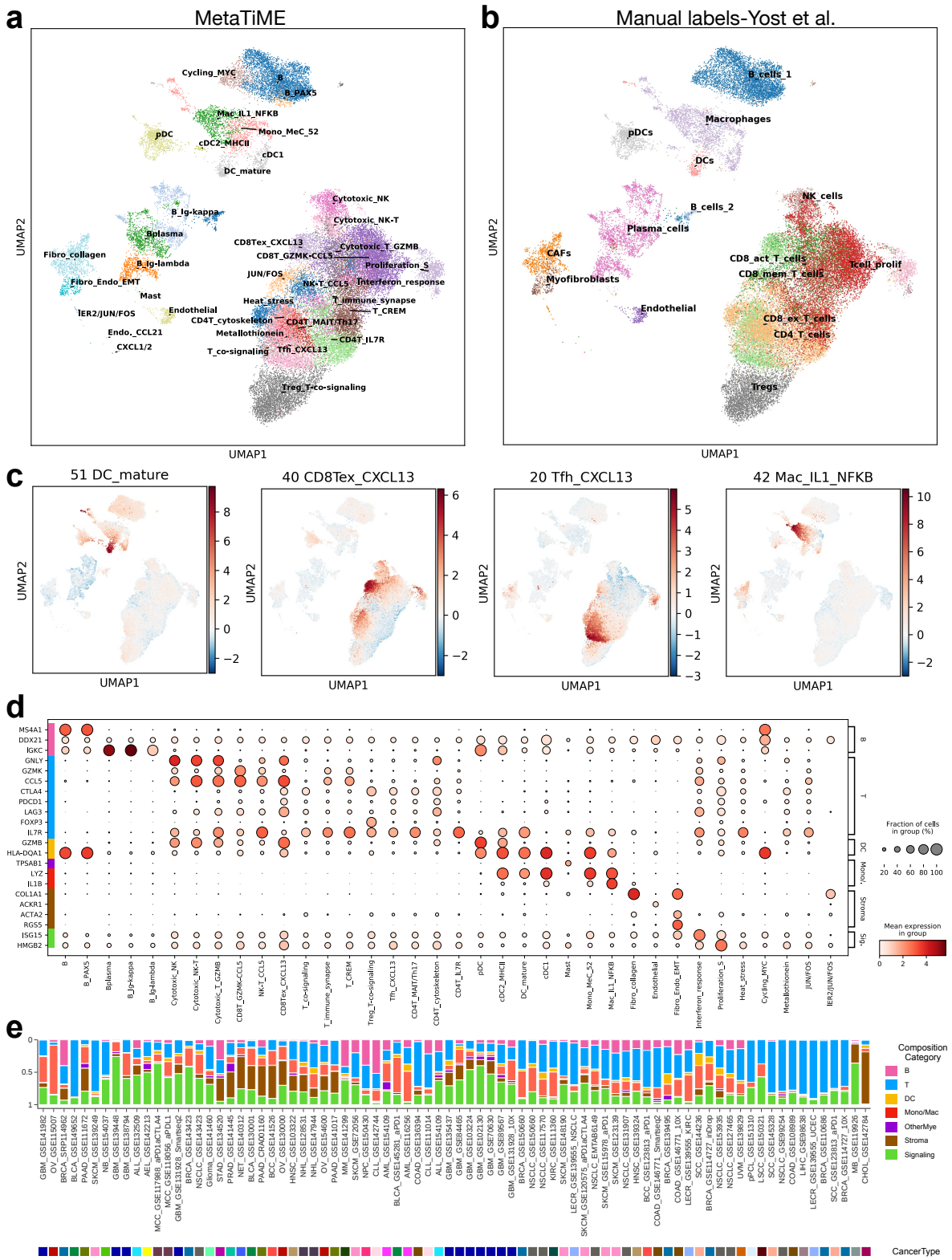
566 **Fig. 2.**



567

568 **Fig.2. MetaTiME meta-components are biologically interpretable with top genes. (a)**
569 Heatmap of top ten most recurrent clusters of MeCs showing normalized gene weights. **(b)**
570 Biological characterization of each MeC with top genes. To facilitate biological interpretation,
571 MeCs are categorized into six lineage-associated (B cells, T cells for CD4T, CD8T and NK cells,
572 dendritic cells, monocyte and macrophages, other myeloid cell types, and stroma cells) and one
573 signaling pathway-associated class. **(c)** Examples of T cell related MeCs with top 20 genes with
574 largest weights. **(d)** Gene contribution of known lineage-related biomarkers for each MeC, and
575 correlation with known immune markers from Azimuth. In the top dot plot, size and color
576 represents MeC z-weights of each gene in each MeC. In the bottom dot plot, size and color
577 represents the maximum correlation coefficient between MeC and Azimuth defined marker
578 genes per cell type. MeC: meta-component, DC: dendritic cell, Mono/Mac: monocytes and
579 macrophages.
580

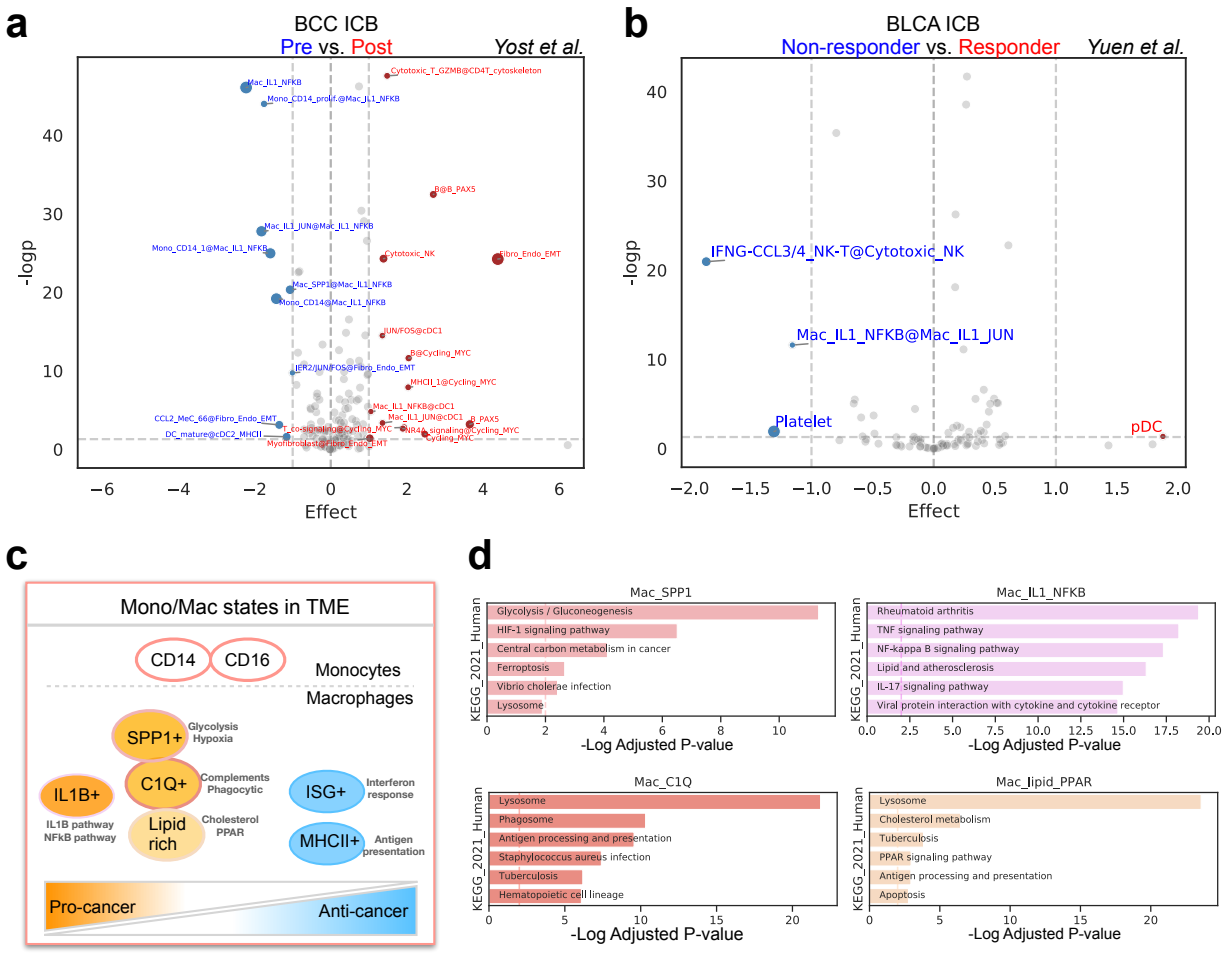
581 **Fig.3.**



582

583 **Fig.3. MetaTiME annotates cell states with high resolution on tumor microenvironment**
584 **single-cell data. (a)** MetaTiME cell state annotation of cell clusters in a melanoma scRNA
585 dataset based on top enriched MeCs. **(b)** Manual annotation labels by experts from the original
586 study shown on the same UMAP space. **(c)** Signature continuum of four MeCs representing
587 mature dendritic cell state, CXCL13-secreting exhausted T cell state, CXCL13-secreting T
588 follicular helper cell state, and IL1B pathway-activated macrophage state. **(d)** Marker gene
589 expression for each annotated cell cluster as in (a)**. (e)** Bar plot showing cell state composition
590 of tumor microenvironment for tumor scRNA dataset cell states. The proportion of cell states
591 from the same MeC category are aggregated.
592
593
594
595

596 **Fig.4.**



597

598 **Fig.4. Differential signature analysis and delineated macrophage states in TME. (a)**
599 Differential MeC signature testing for enriched cell states comparing pre- and post-
600 immunotherapy conditions in Basal Cell Carcinoma (BCC). X-axis: Log odds ratio of mean
601 signature scores between post- and pre-immunotherapy conditions. Y-axis: minus log p-value
602 from Wilcoxon test. **(b)** Differential signature testing for enriched cell states comparing non-
603 responders and responders from pre-treatment condition in Bladder Carcinoma (BLCA). **(c)**
604 Model of monocytes and macrophage states in tumor and their metabolic differences. **(d)** Top
605 pathways enriched in different macrophage MeCs. BCC: Basal Cell Carcinoma, BLCA: Bladder
606 Carcinoma.

Fig.5.

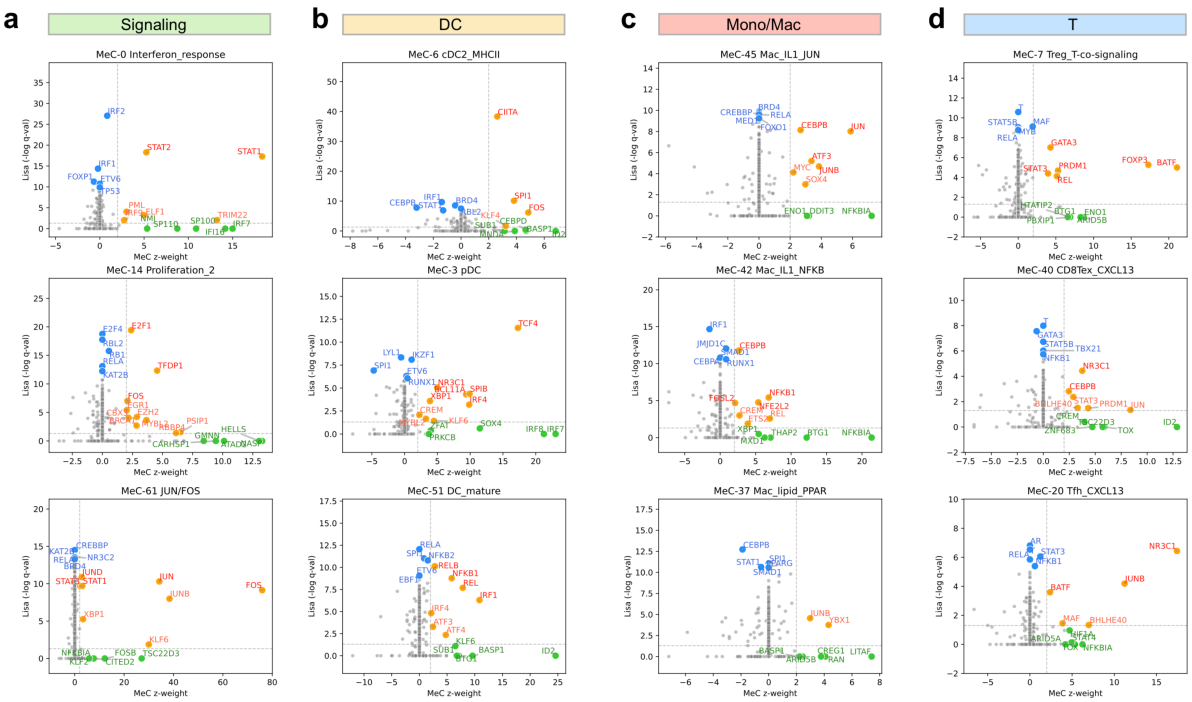


Fig.5. MetaTiME prioritizes tumor immunity transcriptional regulators. For selected MeCs, TFs are prioritized by their MeC expression representation and Lisa, ChIP-seq based, regulatory potentials. X-axis: gene z-weight of the TF for the current MeC. Y-axis: Lisa-based regulatory potential significance for top genes in the current MeC. Red and orange factors: MeC regulators prioritized based on both MeC gene weights and Lisa regulatory potential significance. Green factors: TFs highly weighted in MeCs and not in Lisa analysis. Blue factors: TFs with high Lisa regulatory potential and not highly weighted in MeCs. **(a)** TFs prioritized for three MeCs in the signaling category. **(b)** TFs prioritized for three MeCs in the dendritic cell category. **(c)** TFs prioritized for three MeCs representing different macrophage states. **(d)** TFs prioritized for three MeCs representing different T cell states. TF: transcription factor, MeC: meta-component.

623

624 Reference

625 1. Liu, X. S. & Mardis, E. R. Applications of Immunogenomics to Cancer. *Cell* **168**, 600–612
626 (2017).
627 2. Jin, M. Z. & Jin, W. L. The updated landscape of tumor microenvironment and drug
628 repurposing. *Signal Transduct. Target. Ther.* **5**, (2020).
629 3. Penter, L. *et al.* Local and Systemic Effects of Immune Checkpoint Blockade on Relapsed
630 Myeloid Malignancies Following Allogeneic Hematopoietic Stem Cell Transplantation.
631 *Blood* **136**, 34–35 (2020).
632 4. Zhang, Y. *et al.* The cancer-associated genetic variant rs3903072 modulates immune
633 cells in the tumor microenvironment. *Front. Genet.* (2019). doi:10.3389/fgene.2019.00754
634 5. Gu, S. S. *et al.* Therapeutically Increasing MHC-I Expression Potentiates Immune
635 Checkpoint Blockade. *Cancer Discov.* **11**, 1524–1541 (2021).
636 6. Zeng, Z. *et al.* TISMO: syngeneic mouse tumor database to model tumor immunity and
637 immunotherapy response. *Nucleic Acids Res.* **50**, D1391–D1397 (2022).
638 7. Jiang, P. *et al.* Signatures of T cell dysfunction and exclusion predict cancer
639 immunotherapy response. *Nat. Med.* (2018). doi:10.1038/s41591-018-0136-1
640 8. Yost, K. E. *et al.* Clonal replacement of tumor-specific T cells following PD-1 blockade.
641 *Nat. Med.* (2019). doi:10.1038/s41591-019-0522-3
642 9. Zhang, L. *et al.* Single-Cell Analyses Inform Mechanisms of Myeloid-Targeted Therapies
643 in Colon Cancer. *Cell* **181**, 442-459.e29 (2020).
644 10. Cheng, S. *et al.* A pan-cancer single-cell transcriptional atlas of tumor infiltrating myeloid
645 cells. *Cell* **184**, 792-809.e23 (2021).
646 11. Wang, X. *et al.* In vivo CRISPR screens identify the E3 ligase Cop1 as a modulator of
647 macrophage infiltration and cancer immunotherapy target. *Cell* **184**, 5357-5374.e22
648 (2021).

12. Wu, S. Z. *et al.* A single-cell and spatially resolved atlas of human breast cancers. *Nat. Genet.* **53**, 1334–1347 (2021).

13. Nieto, P. *et al.* A single-cell tumor immune atlas for precision oncology. *Genome Res.* **31**, 1913–1926 (2021).

14. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902.e21 (2019).

15. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587.e29 (2021).

16. Aran, D. *et al.* Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat. Immunol.* **20**, 163–172 (2019).

17. Wang, C. *et al.* Integrative analyses of single-cell transcriptome and regulome using MAESTRO. *Genome Biol.* **21**, (2020).

18. Newman, A. M. *et al.* Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods 2015 125* **12**, 453–457 (2015).

19. Andreatta, M. *et al.* Interpretation of T cell states from single-cell transcriptomics data using reference atlases. *Nat. Commun.* **12**, 1–19 (2021).

20. Zheng, A. L., Qin, S., Wang, A., Xing, B. & Gao, R. Pan-Cancer Single Cell Landscape of Tumor-Infiltrating T Cells. **6474**, (2021).

21. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol. 2018 365* **36**, 411–420 (2018).

22. Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).

23. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nat. Methods* (2018). doi:10.1038/s41592-018-0229-2

24. Pelka, K. *et al.* Spatially organized multicellular immune hubs in human colorectal cancer.

Cell **184**, 4734-4752.e20 (2021).

25. Hyvärinen, A. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Trans. Neural Networks* (1999). doi:10.1109/72.761722

26. Wang, W. *et al.* Independent component analysis based gene co-expression network inference (ICAnet) to decipher functional modules for better single-cell clustering and batch integration. *Nucleic Acids Res.* **49**, E54 (2021).

27. Zheng, R. *et al.* Cistrome Data Browser: expanded datasets and new tools for gene regulatory analysis. *Nucleic Acids Res.* **47**, D729–D735 (2019).

28. Sun, D. *et al.* TISCH: A comprehensive web resource enabling interactive single-cell transcriptome visualization of tumor microenvironment. *Nucleic Acids Res.* **49**, D1420–D1430 (2021).

29. Wykes, M. N. & Lewin, S. R. Immune checkpoint blockade in infectious diseases. *Nat. Rev. Immunol. 2017 182* **18**, 91–104 (2017).

30. Zhang, Q. & Vignali, D. A. A. Co-stimulatory and co-inhibitory pathways in autoimmunity. *Immunity* **44**, 1034 (2016).

31. Qin, S. *et al.* Novel immune checkpoint targets: moving beyond PD-1 and CTLA-4. *Mol. Cancer 2019 181* **18**, 1–14 (2019).

32. Thommen, D. S. *et al.* A transcriptionally and functionally distinct PD-1+ CD8+ T cell pool with predictive potential in non-small-cell lung cancer treated with PD-1 blockade. *Nat. Med. 2018 247* **24**, 994–1004 (2018).

33. Chowell, D. *et al.* Improved prediction of immune checkpoint blockade efficacy across multiple cancer types. *Nat. Biotechnol. 2021 404* **40**, 499–506 (2021).

34. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 15545–15550 (2005).

35. Morgan, D. & Tergaonkar, V. Unraveling B cell trajectories at single cell resolution.

Trends Immunol. **43**, 210–229 (2022).

36. Finkin, S., Hartweger, H., Oliveira, T. Y., Kara, E. E. & Nussenzweig, M. C. Protein Amounts of the MYC Transcription Factor Determine Germinal Center B Cell Division Capacity. *Immunity* **51**, 324-336.e5 (2019).

37. Vaquero, J., Aoudjehane, L. & Fouassier, L. Cancer-associated fibroblasts in cholangiocarcinoma. *Curr. Opin. Gastroenterol.* **36**, 63–69 (2020).

38. Luca, B. A. *et al.* Atlas of clinically distinct cell states and ecosystems across human solid tumors. *Cell* **184**, 5482-5496.e28 (2021).

39. Zhao, J. *et al.* Detection of differentially abundant cell subpopulations in scrna-seq data. *Proc. Natl. Acad. Sci. U. S. A.* **118**, (2021).

40. Yuen, K. C. *et al.* High systemic and tumor-associated IL-8 correlates with reduced clinical benefit of PD-L1 blockade. *Nat. Med.* **26**, 693–698 (2020).

41. Kaplanov, I. *et al.* Blocking IL-1β reverses the immunosuppression in mouse breast cancer and synergizes with anti-PD-1 for tumor abrogation. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 1361–1369 (2019).

42. Duan, Z. & Luo, Y. Targeting macrophages in cancer immunotherapy. *Signal Transduct. Target. Ther. 2021 61* **6**, 1–21 (2021).

43. Bi, K. *et al.* Tumor and immune reprogramming during immunotherapy in advanced renal cell carcinoma. *Cancer Cell* **39**, 649-661.e5 (2021).

44. Zhao, H. *et al.* The role of osteopontin in the progression of solid organ tumour. *Cell Death Dis.* **9**, (2018).

45. Spivia, W., Magno, P. S., Le, P. & Fraser, D. A. Complement protein C1q promotes macrophage anti-inflammatory M2-like polarization during the clearance of atherogenic lipoproteins. *Inflamm. Res.* **63**, 885–893 (2014).

46. Liu, Q. *et al.* The CXCL8-CXCR1/2 pathways in cancer. *Cytokine Growth Factor Rev.* **31**, 61–71 (2016).

47. Qin, Q. *et al.* Lisa: inferring transcriptional regulators through integrative modeling of public chromatin accessibility and ChIP-seq data. *Genome Biol.* (2020). doi:10.1186/s13059-020-1934-6

48. Peters, J. M., Shah, Y. M. & Gonzalez, F. J. The role of peroxisome proliferator-activated receptors in carcinogenesis and chemoprevention. *Nat. Rev. Cancer 2012 123* **12**, 181–195 (2012).

49. Hu, H. *et al.* AnimalTFDB 3.0: a comprehensive resource for annotation and prediction of animal transcription factors. *Nucleic Acids Res.* **47**, D33–D38 (2019).

50. Li, H. *et al.* Dysfunctional CD8 T Cells Form a Proliferative, Dynamically Regulated Compartment within Human Melanoma. *Cell* **176**, 775-789.e18 (2019).

51. Hyvärinen, A. & Oja, E. Independent component analysis: algorithms and applications. *Neural Networks* **13**, 411–430 (2000).

52. Lotfollahi, M. *et al.* Mapping single-cell data to reference atlases by transfer learning. *Nat. Biotechnol.* (2021). doi:10.1038/s41587-021-01001-7

53. Khan, O. *et al.* TOX transcriptionally and epigenetically programs CD8+ T cell exhaustion. *Nat. 2019 5717764* **571**, 211–218 (2019).

54. Tang, Q. & Brown, M. A. Abstract A218: Modulating glucocorticoid receptor-mediated signaling for enhancement of cancer immunotherapy. *Cancer Immunol. Res.* **7**, A218–A218 (2019).

55. Ohandjo, A. Q. *et al.* Transcriptome Network Analysis Identifies CXCL13-CXCR5 Signaling Modules in the Prostate Tumor Immune Microenvironment. *Sci. Reports 2019 91* **9**, 1–13 (2019).

56. Havenar-Daughton, C. *et al.* CXCL13 is a plasma biomarker of germinal center activity. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 2702–2707 (2016).

57. Kotliar, D. *et al.* Identifying gene expression programs of cell-type identity and cellular activity with single-cell RNA-Seq. *Elife* **8**, 1–26 (2019).

753

a

Tumor scRNA

- 1.7 million TME cells

ICA

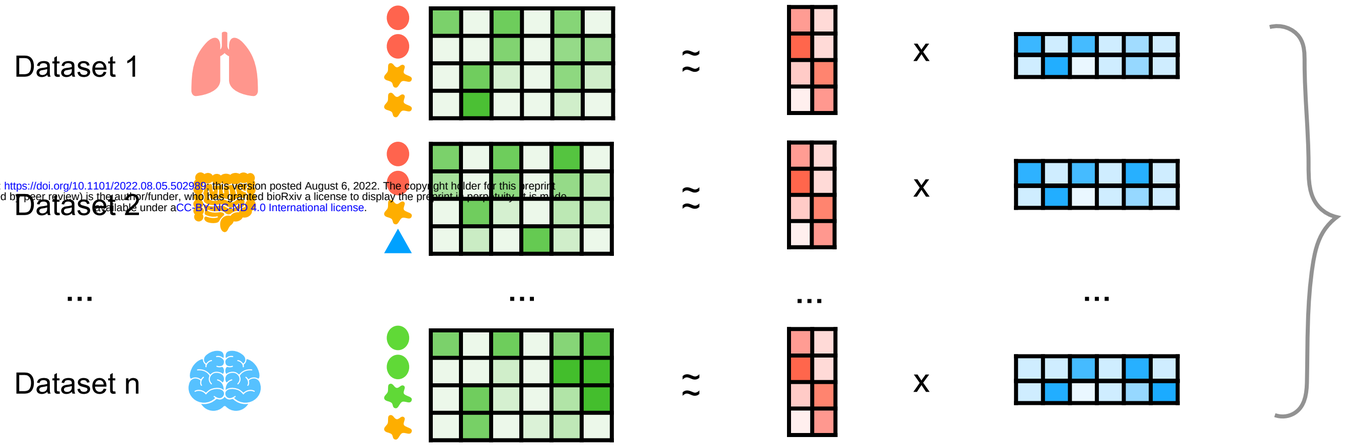
Independent Components

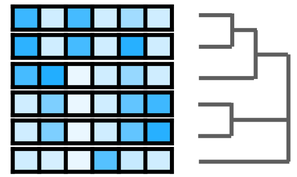
- Signal sources e.g. cell states

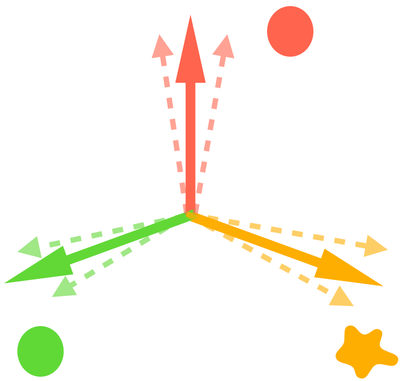
Clustering & Integration

Meta-Components

- Cell type, cell state, signaling







b

Functional MeC Landscape for TME

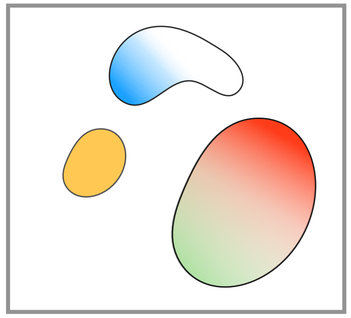
- 76 Functional Annotated MeCs

MeC	Top Genes
B	CD74,MS4A1
CD8T_CD3	CD3D,CD8A
pDC	GZMB,TCF4
MaC_IL1B	CXCL8,IL1B
IFN response	STAT1, ISG15
...	...

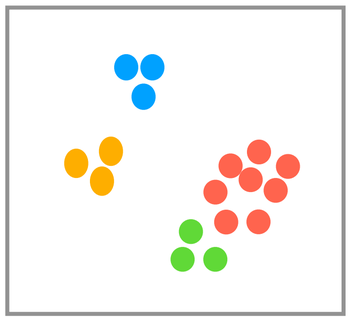
- Categories

Cell State Annotation for Tumor scRNA

- MeC Signature Continuum



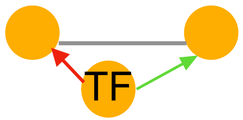
- Top Enriched State



Tumor Immunity Regulators

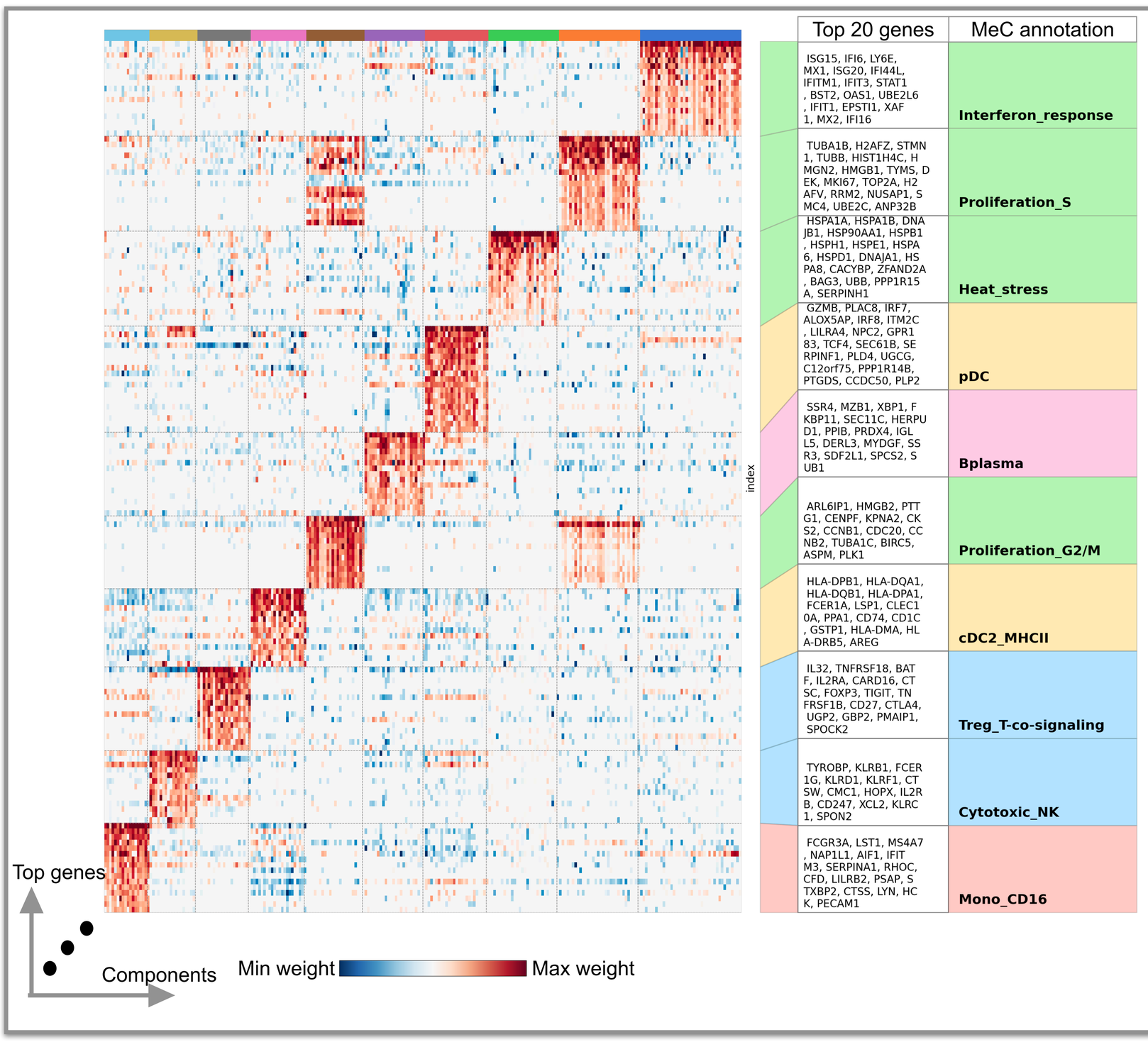
- Incorporated Epigenetic Data

- Prioritized TF regulators



MeC	TF
pDC	TCF4,IRF8
MaC_IL1B	NFKB
IFN response	STAT1
...	...

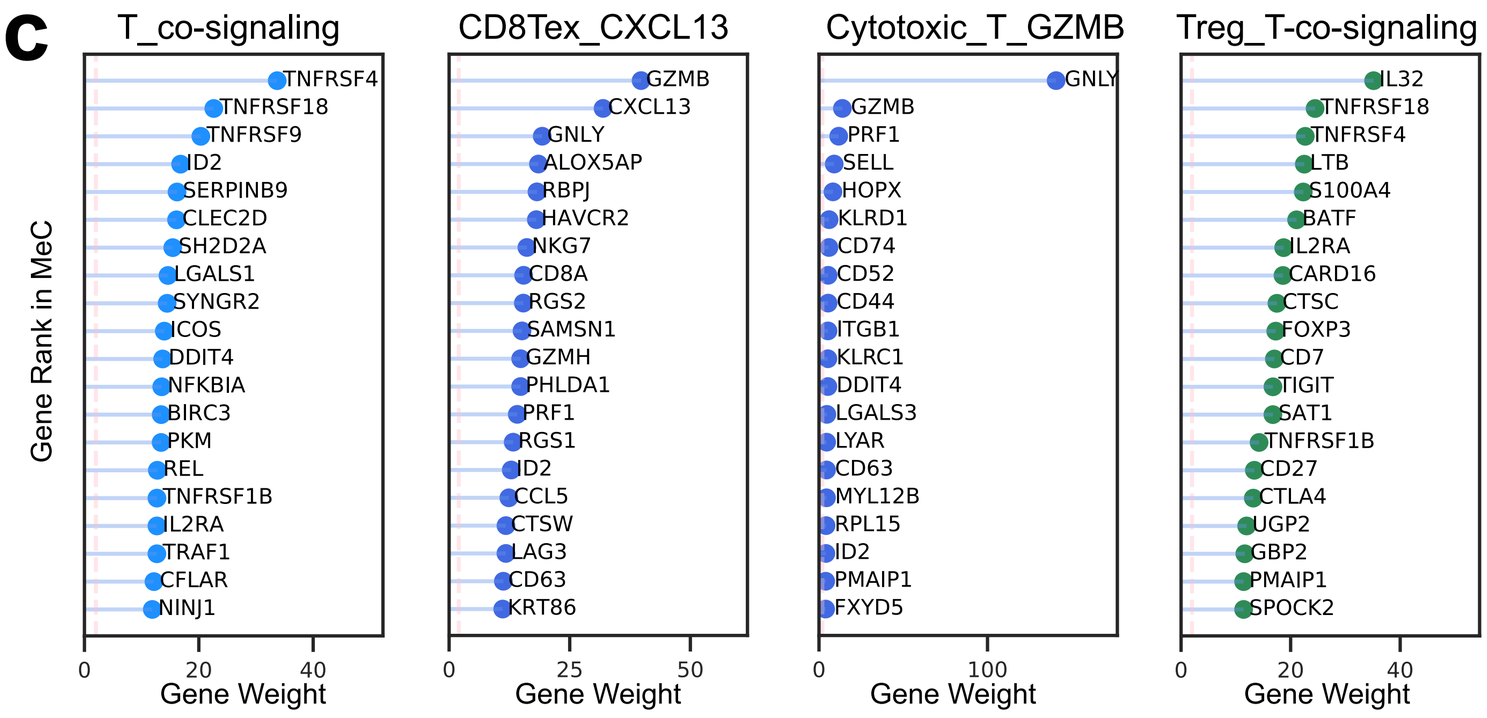
a



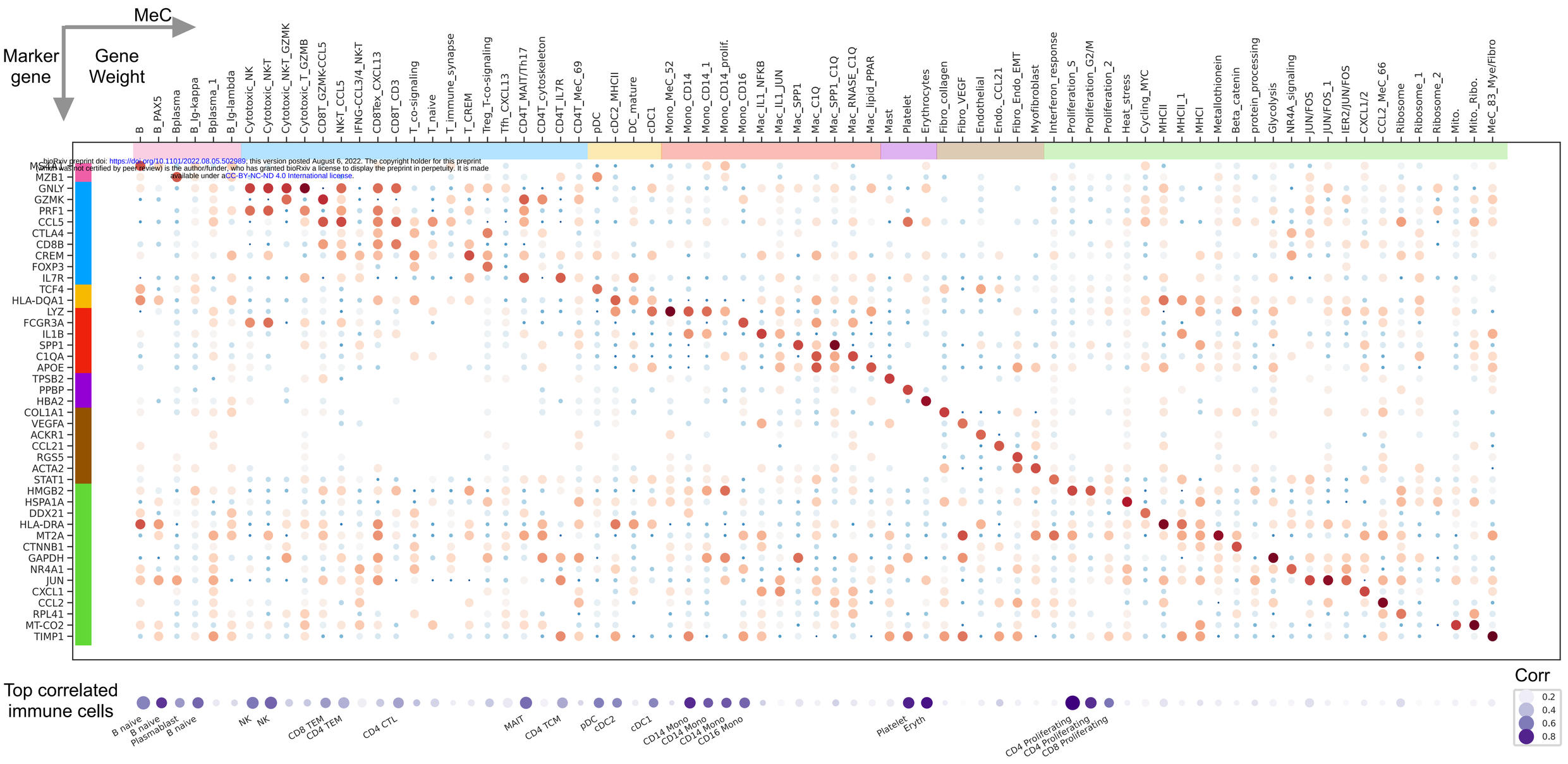
b

MeC annotation	Top 6 genes
Interferon_response	ISG15, IFI6, LY6E, MX1, ISG20, IFI44L
Proliferation_S	TUBA1B, H2AFZ, HMGB2, STMN1, TUBB, HIST1H4C
Heat_stress	HSPA1A, HSPA1B, DNAJB1, HSP90AA1, HSPB1, HSPH1
pDC	GZMB, JCHAIN, PLAC8, IRF7, ALOX5AP, IRF8
Bplasma	SSR4, MZB1, XBP1, FKBP11, JCHAIN, HSP90B1
Proliferation_G2/M	ARL6IP1, HMGB2, UBE2C, TUBB4B, PTTG1, CENPF
cDC2_MHCII	CST3, HLA-DPB1, HLA-DQA1, HLA-DQB1, HLA-DPA1, FCER1A
Treg_T-co-signaling	IL32, TNFRSF18, TNFRSF4, LTB, S100A4, BATF
Cytotoxic_NK	GNLY, TYROBP, KLRB1, FCER1G, CD7, KLRD1
Mono_CD16	FCGR3A, LST1, MS4A7, NAP1L1, AIF1, IFITM2
Mito.	MT-CO3, MT-CO2, MT-CYB, MT-CO1, MT-ATP6, MT-ND3
Cytotoxic_NK-T	GNLY, NKG7, FGFBP2, GZMB, GZMH, PRF1
CD8T_GZMK-CCL5	GZMK, CCL5, GZMA, DUSP2, CST7, CMC1
Mast	TPSAB1, TPSB2, CPA3, CLU, HPGDS, ANXA1
Proliferation_2	DUT, PCNA, TYMS, MCM7, TK1, RRM2
Ribosome	RPL41, RPL26, RPL23A, RPS12, RPS15A, RPS29
CD4T_MAIT/Th17	KLRB1, LTB, IL7R, GZMK, S100A4, RORA
Mono_CD14	S100A9, S100A8, LYZ, FCN1, VCAN, CTSS
B	CD74, CD79A, MS4A1, CD37, HLA-DRA, CD52
Mac_C1Q	C1QB, C1QA, APOE, C1QC, CTSD, APOC1
Tfh_CXCL13	CXCL13, NR3C1, ITM2A, RBPJ, SRGN, ALOX5AP
CD4T_cytoskeleton	ACTG1, COTL1, SUB1, ISG20, GMFG, GYPC
Fibro_collagen	COL1A1, COL3A1, COL1A2, LUM, DCN, SPARC
Endothelial	ACKR1, IFITM3, IGFBP7, AQP1, TM4SF1, IFI27
Metallothionein	MT2A, MT1X, MT1G, MT1E, MT1F, MT1H
Cytotoxic_T_GZMB	GNLY, GZMB, PRF1, SELL, HOPX, KLRD1
Mac_SPP1	SPP1, CSTB, PLIN2, LDHA, MIF, ENO1
Fibro_Endo_EMT	RGS5, NDUFA4L2, IGFBP7, ACTA2, CALD1, MYL9
Beta_catenin	HNRNPH1, CDC42, SET, C1orf56, CDC42SE1, CTNNB1
Mono_CD14_1	S100A9, S100A8, S100A12, LYZ, NFKBIA, CD14
B_Ig-kappa	IGKC, IGHG1, IGHG4, IGHG3, IGHGP, IGHM
protein_processing	HSPA5, HSP90B1, CALR, PDIA3, PDIA6, MANF
CD8T_CD3	GZMH, CD3D, CCL5, NKG7, CD52, FGFBP2
Cycling_MYC	HSP90AB1, NCL, DDX21, C1QBP, HSPE1, NPM1
IFNG-CCL3/4_NK-T	CCL4, CCL4L2, IFNG, CCL3, ZFP36L1, TNF
T_immune_synapse	TAGLN2, ANXA1, CRIP1, TSC22D3, MYADM, UCP2
Mac_SPP1_C1Q	SPP1, RNASE1, CTSB, PLIN2, GPNMB, FCER1G
Mac_lipid_PPAR	APOC1, APOE, LGALS3, GPNMB, CTSD, CSTB
Endo._CCL21	CCL21, TFPI, FABP5, TFF3, MMRN1, FABP4
Platelet	PPBP, PF4, TAGLN2, NAP1L1, OST4, GPX1
CD8Tex_CXCL13	GZMB, CXCL13, GNLY, ALOX5AP, RBPJ, HAVCR2
Myofibroblast	TAGLN, ACTA2, ADIRF, MYL9, DSTN, NDUFA4
Mac_IL1_NFKB	CXCL3, CXCL2, IL1B, CXCL8, G0S2, CCL20
T_naive	CD7, ZFP36L2, TRGC2, HCST, ZNF683, LTB
Mac_IL1_JUN	CXCL8, CXCL2, CXCL3, CXCL1, NFKBIA, NFKBIZ
Erythrocytes	HBA2, HBB, HBA1, UBB, HBD, ALAS2
Glycolysis	GAPDH, TPI1, LDHA, ALDOA, AC006064.4, S100A10
NK-T_CCL5	CCL5, IL32, GNLY, GZMA, CD7, CD52
Mac_RNASE_C1Q	RNASE1, SELENOP, C1QB, C1QA, C1QC, CTSB
B_PAX5	TCL1A, FCER2, CXCR4, IL4R, FOS, CD69
DC_mature	TXN, BIRC3, ID2, MARCKSL1, CCR7, CRIP1
Mono_MeC_52	LYZ, AC020656.1, MMP12, IFI30, PSAP, FCN1
MHCI	CALR, MALAT1, HLA-A, TMBIM6, HLA-B, PPIB
Ribosome_1	RPS18, RPS27, RPL13AP5, RPL31, RPL18A, RPL37A
cDC1	RGCC, CST3, SNX3, IRF8, ID2, C1orf54
T_CREM	CREM, CXCR4, SYTL3, LEPROTL1, SRGN, FYN
JUN/FOS_1	JUN, IER2, FOS, ID2, GADD45B, KLF6
MHCII_1	ZFP36L1, LGALS2, PSME2, NPC2, HLA-DPB1, HLA-DRA
B_Ig-lambda	IGLC2, IGLC3, IGHGP, IGHG1, EMP3, CD37
JUN/FOS	FOS, DUSP1, JUNB, JUN, KLF6, TSC22D3
MHCII	HLA-DRA, HLA-DRB1, HLA-DPA1, HLA-DQA1, IFI30, HLA-DRB5
IER2/JUN/FOS	IER2, EGR1, EIF4A3, FOSB, SERTAD1, TUBB4B
T_co-signaling	TNFRSF4, TNFRSF18, TNFRSF9, ID2, SERPINB9, CLEC2D
CCL2_MeC_66	CCL2, EIF4A3, DBI, TUBA1B, MT2A, CYR61
NR4A_signaling	NFKBIA, CD69, SDCBP, ZFP36L1, NR4A1, DUSP2
CD4T_MeC_69	CXCR4, IL32, HERPUD1, TSC22D3, ZFP36L2, GAPDH
Mito._Ribo.	MT-CO2, TPT1, MT-ND1, MT-ND3, EEF1A1, MT-ND2
Ribosome_2	RPL27A, MT-ATP8, RPS9, RPL3, RPS18, RPL27
Mono_CD14_prolif.	S100A8, S100A9, S100A12, HMGB2, VCAN, SELL
Cytotoxic_NK-T_GZMK	GNLY, IFITM2, SELL, CMC1, COTL1, CD7
CXCL1/2	RGS2, CXCL1, LAPTM5, P2RY12, ID2, CXCR4
Bplasma_1	JCHAIN, IGHA1, VIM, CD48, IER2, PABPC4
CD4T_IL7R	S100A4, ANXA1, LTB, ITGB1, IL7R, S100A11
Fibro_VEGF	MT2A, CPE, VEGFA, ENO1, ZFAS1, TIMP1
MeC_83_Mye/Fibro	TIMP1, DDIT4, MT2A, NNMT, SERPINE1, IL1B

c



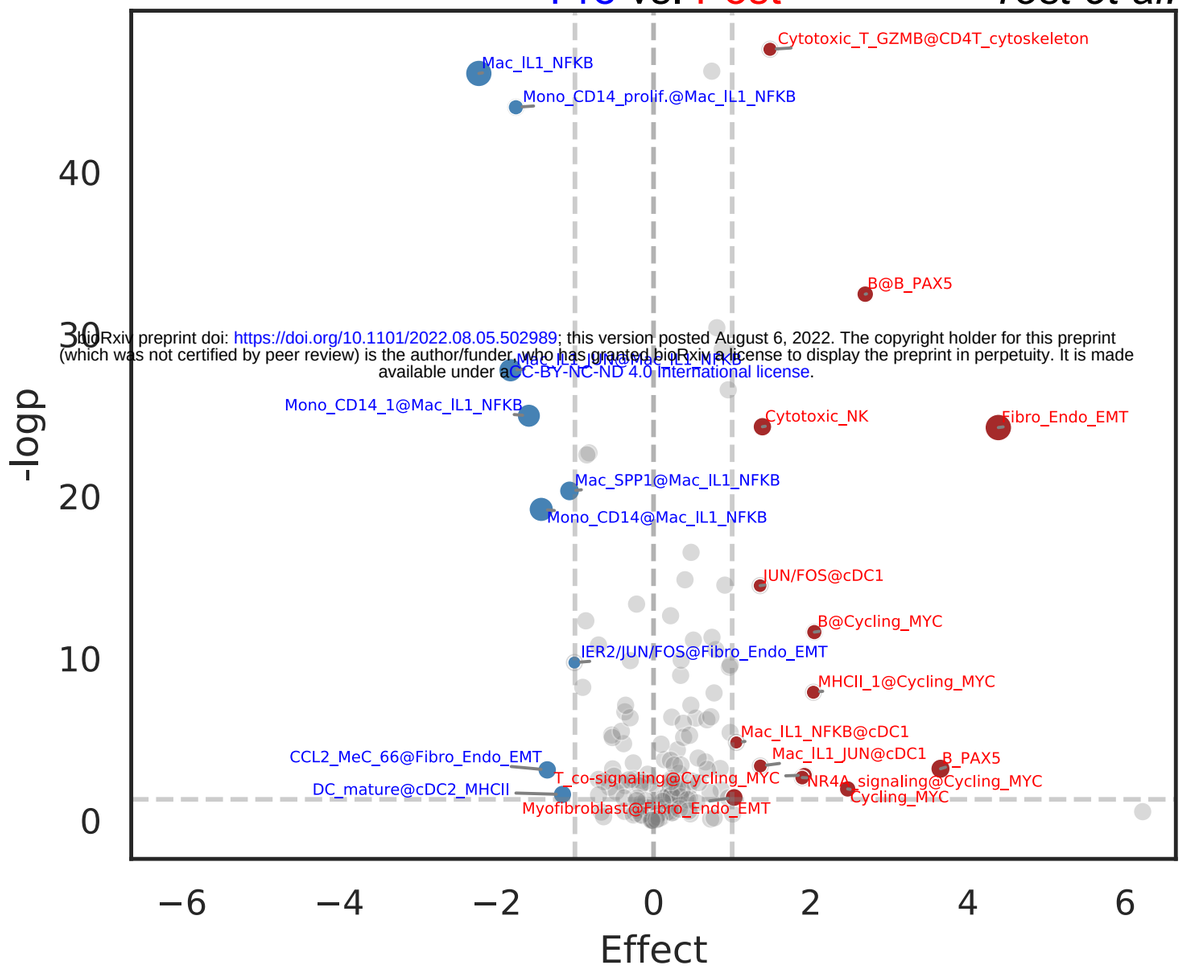
d



a

BCC ICB

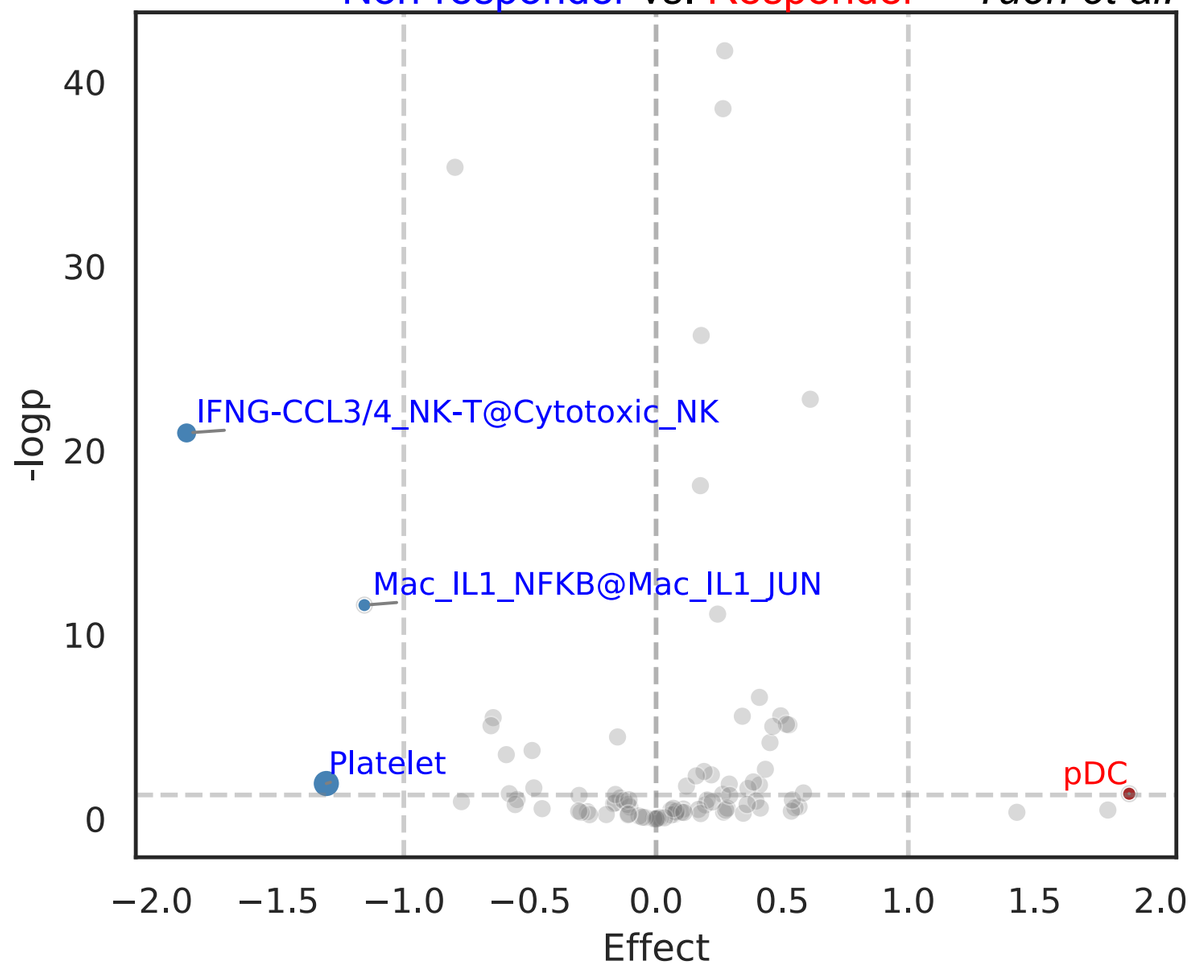
Pre vs. Post *Yost et al.*



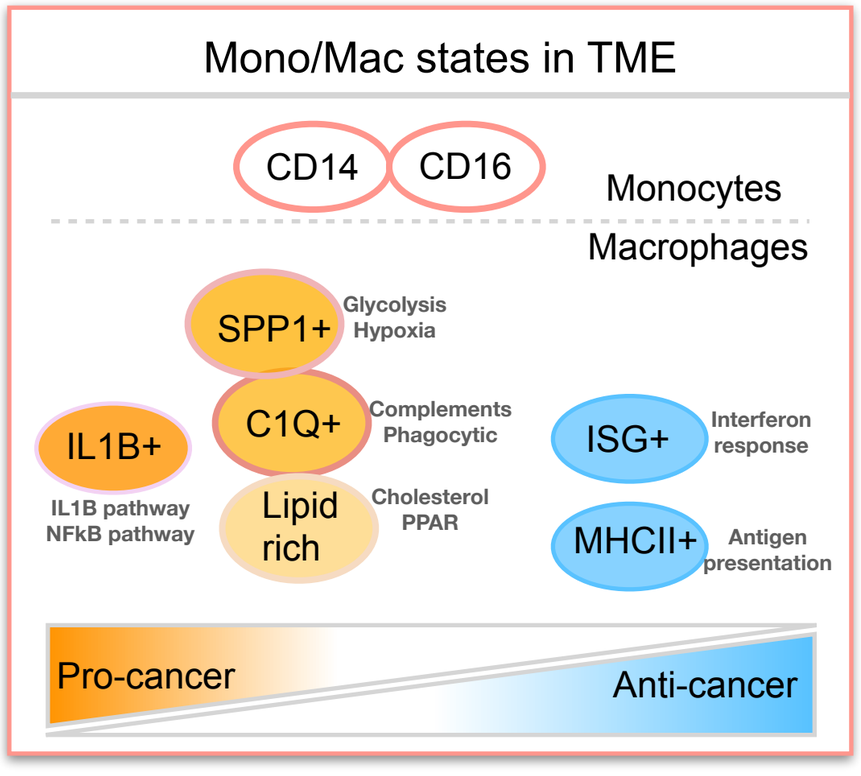
b

BLCA ICB

Non-responder vs. Responder *Yuen et al.*



c



d