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

79 datasets, and to reveal differential signatures across immunotherapy responses. Finally, by
80 incorporating transcription factor binding data, MetaTiME revealed and prioritized putative
81 transcriptional regulators that may modulate tumor immunity.
82
83 ## Results
84 **MetaTiME as a general framework to discover consensus transcriptomic programs**
85 The MetaTiME framework consists of three stages: meta-component (MeC) discovery,
86 interpretation of MeCs, and application of cell state annotations (**Fig.1a**). The MeC discovery
87 stage detects repeatable sources of variation from multiple single-cell measurements sharing
88 similar cellular properties. The MeC interpretation step involves a one-time curation effort using
89 biomarker databases, pathway information and Cistrome DB chromatin profiling data[27]. In the
90 third step, users map MeCs onto their new tumor scRNA-seq datasets using MetaTiME
91 application tools, to obtain annotated cell states and signature continuums.
92
93 To train MeCs for the TME context, we collected and curated 2,157,387 cells from 76 studies
94 ranging across 27 cancer types, using publicly available tumor scRNA-Seq data mostly from
95 TISCH[28]. After removing the TISCH annotated malignant cells using MAESTRO[17], 102,703
96 stromal cells and 1,617,110 immune cells were retained for downstream training (**Fig.S1**). The
97 76 studies were further partitioned according to cancer type, resulting in 93 datasets, including 7
98 datasets with immune checkpoint blockade (ICB) treatment and 3 10x Genomics provided
99 datasets representing peripheral blood mononuclear cells (PBMC) sampled from healthy
100 donors.
101
102 In the MeC discovery stage, MetaTiME first decomposes the log-transformed expression matrix
103 of each single dataset using Independent Component Analysis (ICA)[25]. We adopted ICA to
104 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

131 as *TNFRSF4* (OX40), *TNFRSF18* (GITR) *TNFRSF9* (4-1BB), and *ICOS* (**Fig.2c**, left). The
132 “CXCL13, exhausted CD8 T cell" MeC features receptors characterizing the exhausted CD8 T
133 cell state[8], including *HAVCR2* (TIM3), *LAG3*, *TIGIT*, and *PDCD1* (PD1) (**Fig.S5**), each being
134 potential ICB targets[31]. In addition, this MeC is characterized by a high level of *CXCL13*
135 (**Fig.2c**, second panel), a cytokine mediating immune cell trafficking to tertiary lymphoid
136 structures[32]. In contrast, a related MeC representing T cell co-signaling receptors in regulatory
137 CD4T cells (Treg) has a different ranking, including *TNFRSF18*, *TNFRSF4*, *TIGIT, TNFRSF1B*,
138 *CTLA4*, *CD27* among the top 20 genes, along with the regulatory T cell-specific marker *FOXP3*
139 (**Fig.2c**, right). Though ICB has been an extremely successful therapy for some patients, it has
140 not yet had an impact on the majority of patients[33]. Investigating the top members in the MeCs
141 involving T cell receptor pathways may help identify new ICB targets.
142
143 **MetaTiME depicts the functional landscape of transcriptomic variation and cell states in**
144 **the tumor microenvironment**
145 We provided functional annotations of all MeCs by examining top z-weight genes and compared
146 these with functional gene sets, such as immune cell type markers[15,18] and gene ontology
147 databases[34]. We found that 75 MeCs clearly mirror gene expression patterns corresponding to
148 cell types, cell states and signaling pathway activities, depicting a landscape of non-cancer cell
149 states in the TME (**Fig.2, Table S1**: MeC annotation). The top genes of the cell type MeCs
150 match well-known lineage-specific markers[15,18]. Examples include *CD74*, *CD79A*, *MS4A1* for B
151 cells (“B cell" MeC), *CD3D*, *CD8A*, *CD8B* for T cells (“CD3 - CD8 T cell” MeC), and *LYZ*, *VCAN*,
152 *S100A9* for CD14+ Monocytes (“CD14 monocyte” MeC) (**Fig.2b, Fig.2d**). The majority of MeCs
153 define high resolution lineage-specific cell states (**Fig.2b, Fig.S4b**). Taking the B cell lineage as
154 an example, multiple MeCs harbor genes specific to B cell developmental stages[35], ranging
155 from a progenitor B cell state (*CD69* and *PAX5* in the “PAX5 B cell"), to a mature B cell state
156 (*CD79A* in the “B cell" MeC), an antibody-secreting plasma cell state (*XBP1* in the “plasma B

157 cell" MeC, and *JCHAIN* in the “alternate plasma B cell” MeC (“Bplasma_1”), and
158 immunoglobulin secretion states (IGK and IGH in the “immunoglobin kappa B cell" MeC, IGL
159 and IGH in the “immunoglobin lambda B cell” MeC) (**Fig.2b**). Lastly, like the interferon
160 responsive MeC mentioned above, we found a subset of MeCs that are more accurately
161 interpreted as signaling pathways because their top genes are more related to pathways or
162 molecular functions than to cell identities.
163
164 We organized the 75 annotated MeCs into six cell lineage-focused categories and one signaling
165 pathway-focused category (**Fig.2a,b** and **Table S1**: MeC annotation, MeC enrichment). Among
166 these, there are 6 B lineage-related MeCs for B cells; 19 T cell lineage MeCs covering CD8 T
167 cells, CD4 T cells, and natural killer (NK) cells; 4 dendritic cell (DC) lineage MeCs; 12 monocyte
168 and macrophage-related MeCs; 3 platelet, erythrocyte, and mast cell MeCs; 6 stromal cell-
169 related MeCs for fibroblasts, myofibroblasts and endothelial cells; and 25 MeCs in the signaling
170 category (**Fig.2b**). We demonstrated that the MeCs are of high specificity, visualizing the z-
171 weights of known cell subtype markers and pathway biomarkers (**Fig.2d**). Correlating MeCs
172 with the comprehensive immune cell type database Azimuth[15] validated the lineage-specificity of
173 several MeCs, while most MeCs reflect cell states that appear specific to the tumor context
174 (**Fig.S4b**).
175
176 **MetaTiME annotates cell states and signature continuums when applied to the tumor**
177 **microenvironment single-cell data**
178 As MetaTiME MeCs provides a highly interpretable basis for the TME in single cells, we
179 provided a toolkit to reveal MeC signature continuums and enriched cell states in scRNA-seq
180 TME data (code deposited in https://github.com/yi-zhang/MetaTiME). The MetaTiME annotation
181 toolkit takes as input the scRNA-seq expression matrix after depth normalization and log
182 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

208 multiple tumor types including colon cancer, breast cancer, and skin cancer, suggesting that
209 immune infiltration is sample-dependent and that cancer treatments should be personalized[38].
210
211 **Differential MetaTiME analysis detects alterations of transcriptional programs in**
212 **immunotherapy.**
213 Single-cell data derived from ICB trials is invaluable for identifying cell types associated with ICB
214 treatment or response[8]. However, the detection of differential cell types abundances in ICB
215 cohorts has been challenging due to the heterogeneity of cell type proportions and to the limited
216 numbers of patients in each cohort[39]. We compared differences in MeC signatures instead of
217 cell count proportions, to understand immune response during ICB. We analyzed two ICB
218 cohorts, a basal cell carcinoma (BCC) cohort with samples from pre- or post-ICB treatment[8],
219 and a bladder cancer (BLCA) cohort with samples from ICB responders and non-responders[40].
220 We applied MetaTiME for per-cluster cell state annotation and per-cell MeC signature
221 evaluation. For each cell state cluster, we tested all MeC signatures passing significance
222 (average z-weight>2) between conditions using the Wilcoxon rank-sum test. We plotted cluster-
223 wise signatures in the significance – effect size scatterplot to highlight the most significant
224 differential MeCs (**Methods**). In a comparison of pre- and post-ICB treatment, we observed
225 higher expression of cytotoxic T cell and B cell MeCs in the post-ICB samples. Moreover,
226 several monocyte and macrophage states are also suppressed after ICB treatment (**Fig.4a**).
227 Notably, the IL1B-positive macrophage signature is also found to be elevated in non-responders
228 compared to responders in the BLCA ICB cohort (**Fig.4b**). Since activation of the IL1B pathway
229 is a known regulator of inflammatory processes[41], we sought to investigate whether the IL1B-
230 positive macrophage signature is associated with tumor survival prognosis in bulk RNA-seq
231 data from The Cancer Genome Atlas Program (TCGA). We evaluated TCGA tumors using the
232 averaged expression of the top 20 genes from the “Macrophage IL1-NFkB” MeC which ranks
233 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",

260 “MHCII-high”, and “RNASE1+,C1Q+” MeCs) (**Fig.S9**). However, MetaTiME reveals additional
261 distinct components such as the “SPP1+” and “C1Q+” MeCs, which were detected as separate
262 myeloid types in the *Cheng et al.* multi-cohort study (**Fig.S10**). While the manual reconciliation
263 of cell types from multi-cohort scRNA data shows many marker genes to be consistent with the
264 top genes in the MetaTiME MeCs, the myeloid cell population is not neatly partitioned into cell
265 clusters and might be better represented in terms of expression signature continuums. For
266 example, when mapping myeloid MeCs onto the kidney myeloid cells, the “IL1B+” MeC
267 signature is distributed across the “Macro_IL1B” cluster as well as the CD14 monocyte cluster
268 (**Fig.S10**).
269
270 To investigate functional differences among the different macrophage states, we applied gene
271 set enrichment (GSEA)[34] analysis using the top MeC genes. Interestingly, the different
272 macrophage states have different metabolic preferences (**Fig.4d**). Glucose metabolism and the
273 glycosylation pathway are highly active in SPP1+ macrophages, while lysosome and
274 phagosome activity are the most highly enriched in C1Q+ macrophages. Lysosome and
275 cholesterol metabolism, including *PPARG* signaling, are enriched in the lipid-rich state. The
276 inflammatory IL1B and NFkB pathways are highly active in IL1B+ macrophages. Several
277 macrophage states are related to cell signaling. *SPP1* for example, encodes Osteopontin, which
278 has been found to foster an environment that promotes cancer metastasis[44]. The C1Q+ MeC
279 features *C1QA*, *C1QB*, and *C1QC*, members of the family of complement molecules that could
280 play dual roles in chronic inflammation[45]. The IL1B+ meta-components features cytokines co-
281 expressed with *IL1B*, including *CXCL8*, *CXCL2*, and *CXCL3*, all of which can interact with other
282 cells in the TME by binding to cytokine receptors[46] (**Fig.2b**).
283
284 **Incorporation of epigenetic data prioritizes transcriptional regulators of tumor immunity**

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

311

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

326

327 Discussion

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

333

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

336 the related MeCs revealed heterogeneity and plasticity of tumor-associated macrophages
337 (TAMs). We thus propose that TAMs, especially for solid tumors, should be classified based on
338 the major states with different metabolic preferences instead of the canonical M1 and M2
339 classification[42]. Similar states that do not fit well into the M1 versus M2 classification scheme
340 were also observed in previous studies analyzing myeloid cells in the TME, where single cells
341 were clustered and labeled using differential markers[10,12,43]. *Cheng et al.*[10] defined several TAM
342 types by clustering myeloid cells separately for each study and naming the TAMs with manually
343 selected top marker, chosen based on consistency across cohorts. Similarly, *Bi et al.*[43] defined
344 TAM types by harmonizing patients and naming the TAMs with top genes in each cluster for an
345 advanced renal cell carcinoma (ccRCC) cohort. Cell type definitions in the previous studies
346 were based on representative genes, which were chosen differently in the respective studies.
347 We propose that the MetaTiME derived monocyte and macrophage MeCs could be used to
348 define macrophage states and functional co-expressed gene modules more consistently for the
349 TME. Reexamining the marker genes from the previous studies: *NLRP3* is highly ranked in the
350 "IL1B+" MeC; in fact, the NLRP3 inflammasome mediates interleukin-1β production. *GPNMB* is
351 weighted among the top 20 genes in both the "C1Q+ macrophage" and "SPP1+ macrophage"
352 MeCs; it encodes a membrane glycoprotein which is typically highly expressed in macrophages.
353 *FOLR2* is ranked 29th in the "C1Q+ macrophage" MeC, indicating this macrophage state also
354 encodes a high folate-activated pathway. Finally, the "SPP1/C1Q macrophage" MeC features
355 an intermediate state with both SPP1 and C1QA, indicating the plasticity and mixed nature of
356 pathways activated in TAMs that could not be defined using exclusive markers. Thus, the
357 myeloid MeCs may provide a consistent definition of TAM states corresponding to different
358 metabolic processes.
359
360 MetaTiME provides a toolkit for analyzing independent TME scRNA-seq datasets by mapping
361 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

387 of the identity and function of cells in the TME in future studies and will facilitate the
388 identification of potential new therapeutic targets for immune modulation.
389
390
391 Methods
392
393 **Tumor single-cell RNA-seq data collection and processing**
394 For an extensive collection of single cells from tumor microenvironment, we utilized the public
395 tumor scRNA-seq collection from TISCH[28]. The TISCH collection uniformly processed each
396 dataset with MAESTRO[17] and isolated non-malignant environmental cells from malignant cells.
397 Overall, we collected 2,157,387 cells from 76 studies ranging 27 cancer types. The MAESTRO
398 annotation was labeled using CIBERSORT gene panels[18] followed by curation, enabling
399 selection of 1,719,813 environmental cells, including 1,617,110 immune cells and 102,703
400 stromal cells, were retained for integrative analysis in this study. For studies with data measured
401 from multiple cancer types, cells different cancer types were split into independent datasets,
402 resulting in 93 datasets; it includes 3 PBMC datasets from healthy donors from 10X Genomics
403 as baseline and 7 datasets with ICB treatment.
404
405 For an unsupervised component analysis, each dataset was re-analyzed. For datasets with raw
406 count matrices available, gene expression was normalized towards per-cell read depth 10,000
407 followed by log transformation. For datasets with only TPM or FPKM values available, including
408 Smart-seq data or studies with only normalized matrix available, gene expression underwent log
409 transformation. Cells were filtered based on minimum library size 1000, gene number 500, and
410 maximum mitochondrial read proportion 5%.
411
412 **Decomposing individual studies and denoising low-dimensional components.**

413 We then decomposed the expression matrix of each scRNA-seq dataset using fastICA[51] into an
414 independent component (IC) vector matrix and a projection weight matrix. We tested different
415 values for the number of components (k) and chosen k to be 100 uniformly for each dataset,
416 given it could cover more variations than the number of cell types in the TME, which is around
417 twenty. We applied two denoising approaches to deal with sparsity and the potential noise of
418 ICs. First, we performed a z-normalization of the gene loadings in the component, scaling all
419 gene loading values by the standard deviation of each IC. The gene loadings indicate the
420 degree of contribution to the component as a “metagene” from each gene, and we observed
421 that most genes contribute neutrally to the metagene. Thus, genes with significant contributions
422 are selected using the two-standard deviation threshold from either the positive or negative side.
423 Second, we aligned the positive skewness of components since the sign of an independent
424 component is randomly assigned in fastICA optimization. We observed that asymmetrically
425 extreme gene loading values highlight genes representative of the component's function; thus,
426 we computed each component's skewness statistics and flipped the sign of component loadings
427 if the skewness is negative. We excluded genes with a low contribution (gene weight not
428 passing two standard deviations) to any reproducible components and kept 6623 genes with
429 potentials in driving the reproducible components. The post-decomposition steps ensured the
430 attitude and sign of gene weights are comparable across cohorts, depicting degree of
431 contribution from each gene in the genome-wide background.
432

433 **Meta-components calling and functional annotation**

434 We then aim to discover reproducible patterns from all components from each dataset. We
435 evaluated similarity between pairs of components using cosine distance and retained a set of
436 1043 candidate reproducible components from 69 datasets, each with a minimum Pearson
437 correlation coefficient 0.3 with at least one different IC. We then clustered ICs using Louvain

438 clustering, a graph-based community detection algorithm where the resolution parameter
439 controls segmentation granularity. Clusters with at least five ICs were retained as reproducible
440 IC clusters. The number of clusters is determined by optimizing both Silhouette's score for
441 optimal within-cluster similarity compared to inter-cluster similarity and number of reproducible
442 clusters. The final resolution parameter was chosen to be 1.25 resulting in 86 clusters for meta-
443 component (MeC) calling. The consensus gene z-weights in each MeC were then calculated by
444 averaging ICs in each cluster. Genes of outlier z-weights passing two standard deviations were
445 highlighted as significant, and the ones with positive largest z-weights were considered
446 representative of the MeC.
447
448 MetaTiME MeCs were assigned curated annotation by matching top z-weighted genes to
449 functional biological information including cell type markers, pathway databases from GSEA,
450 cell types expressing top MeC genes, and high-rank transcription factors. In details, GSEA
451 enrichment analyses utilized top 100 highest z-weighted genes and TF database was obtained
452 from AnimalTFDB[49]. The 86 MeCs were first ordered by MeC cluster size, and then organized
453 into seven functional categories and one undefined category. The 11 MeCs in the undefined
454 category are of smaller size, and harbor top genes that are related to cell stress. The remaining
455 75 functional MeCs were assigned six lineage-related categories and one signaling category.
456
457 **Simulating multi-cohort single-cell RNA data with expression programs**
458 To benchmark dimensional reduction methods, we built upon previous effort from *Kotliar et al.*[57]
459 to use the scsim package to simulate multiple count matrices with built-in transcriptional
460 programs. In principle, the built-in gene expression programs (GEP) were sampled as random
461 scaling factors on a subset of genes mimicking overexpression or suppression of a pathway.
462 For testing whether a higher number of cohorts facilitate GEP recovery, we simulated 20 single-

463 cell datasets and tested usage of 5 cohorts, 10 cohorts, and 20 cohorts. Each dataset was
464 embedded with a subset of 14 pre-defined GEPs, since the real tumor scRNA data may not
465 cover every possible cell types or gene programs in every dataset. The 14 GEPs contain 13 cell
466 type-specific programs with distinct cell type-specific genes, and one signaling gene expression
467 program that is randomly active in multiple cell types. Two low-dimensional reduction method
468 are benchmarked using simulated scRNA data: independent component analysis (ICA) and
469 non-negative matrix factorization (NMF). Decomposition was performed on each single cohort
470 separately, and meta-component calling was done as similar in MetaTiME: components are
471 filtered, clustered into meta-components, followed by averaging gene z-weights per cluster as
472 predicted gene expression programs (GEP). The predicted GEPs were compared with pre-
473 defined True GEPs using Pearson correlation. Overall, both ICA and NMF can recover GEPs,
474 while the ICA-based GEPs are more mutually independent and performs slightly better. Since
475 the GEP recovered in the 20 cohorts case matched true GEP better than 5 cohorts and 10
476 cohorts, the increased number of cohorts also improves GEP recovery. Thus, we chose ICA for
477 component integration and use all available datasets for GEP discovery for tumor
478 microenvironmental cells.
479
480 **The MetaTiME annotator for analyzing new tumor scRNA-seq data.**
481 MetaTiME provides an analytical toolkit for annotating cell states and signature activities for
482 tumor scRNA-seq data (https://github.com/yi-zhang/MetaTiME). The scRNA-seq data is first
483 processed following standard procedure, which includes cell depth normalization, log-
484 transformation, batch effect removal using Harmony[22], neighboring graph construction, graph
485 clustering, and UMAP embedding for visualization. Specifically, the clustering step uses an
486 over-clustering strategy, which sets a high-resolution parameter (default 8) that generates a
487 larger number of clusters and help reveal fine structures among the cells. Then, the MetaTiME
488 annotator tool takes as input a single log-transformed expression matrix for TME cells from the

489 dataset. The outputs include both per-cell MetaTiME MeC signature scores and per-cluster
490 enriched MeC state. For the per-cell score, MetaTiME projects each cell onto the MeC space by
491 calculating dot product between the expression vector and the z-weight vector of each MeC,
492 using genes passing significant z-weight criterion (z-weight >2). The projection matrix is then
493 scaled across all cells to ensure normally distributed scores within each MeC, outputting the
494 cell-by-MeC score matrix. Meanwhile, the UMAP view of the projection score shows the
495 signature gradient across the cells positioned by similarity. Lastly, the cluster-wise MeC
496 enrichment results are also generated. The per-cluster MeC enrichment score is calculated by
497 averaging profile of cells along each MeC; MeCs with mean score passing the significant cutoff
498 (2 in the z-weight scale) are called as the set of enriched MeCs. Each cluster may enrich
499 multiple number of MeCs, and the top enriched MeC with highest score is used in UMAP
500 visualization.
501
502 **Differential MeC signature analysis**
503 For tumor scRNA-seq data with different conditions, a differential signature analysis can be
504 carried out following MetaTiME annotation, which provides enriched MeCs for each cluster and
505 names each cluster with the top enriched MeC. Thus, for each cell cluster, the MeC signature
506 strength can be compared across conditions, for all enriched MeC in the current cluster. In
507 details, a simple Wilcoxon rank-sum test is adopted to compare MeC scores of cells in one
508 condition with another. The log-fold change of MeC scores were calculated by the ratio between
509 cell means from the two conditions in comparison. To plot the cluster-specific differential
510 signature plot, the signatures are marked using “EnrichedMeC@ClusterName”, where the
511 “ClusterName” is the top first enriched MeC used as cell state as current cluster. Thus, when
512 the enriched MeC signature is the same with the cluster name, the differential signature is
513 named as “ClusterName” on the Significance-Effect size plot.
514

515 **Incorporation of epigenetic data using Lisa**

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

530

531

532 Grant information

533 NIH grants P01CA163222, U24CA237617, and R01HG011139.

534

535 Acknowledgement

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

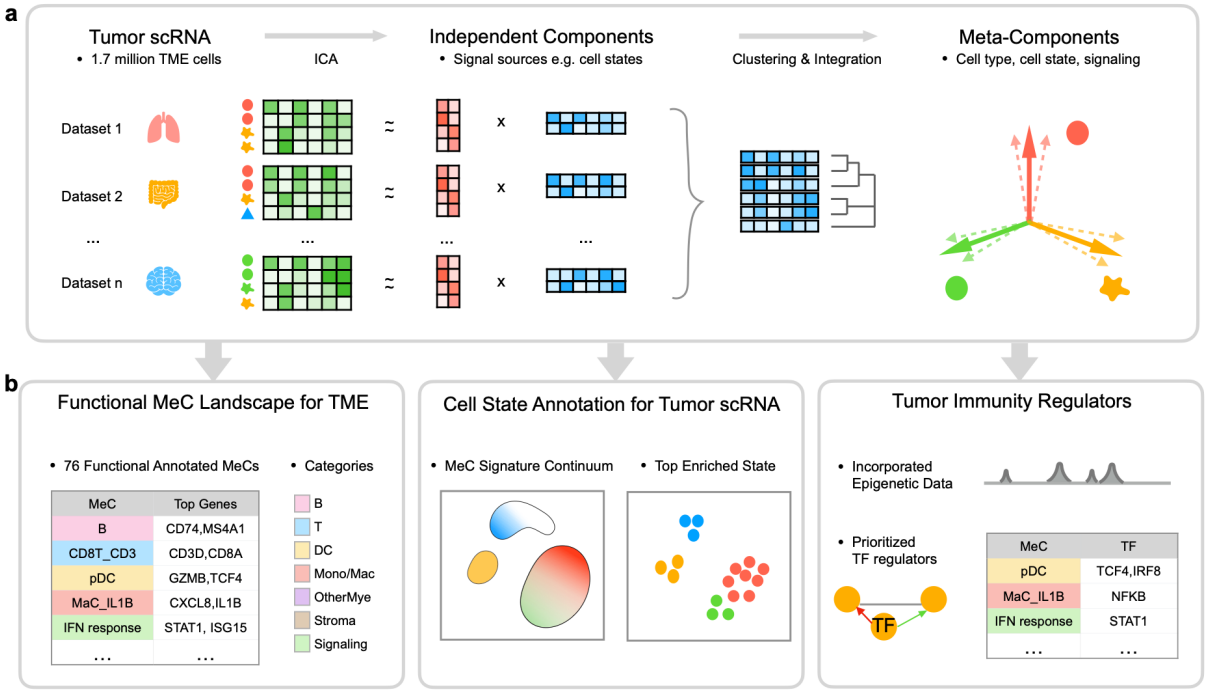
539 Conflict of Interest

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

545

546 Figures

547 **Fig.1.**



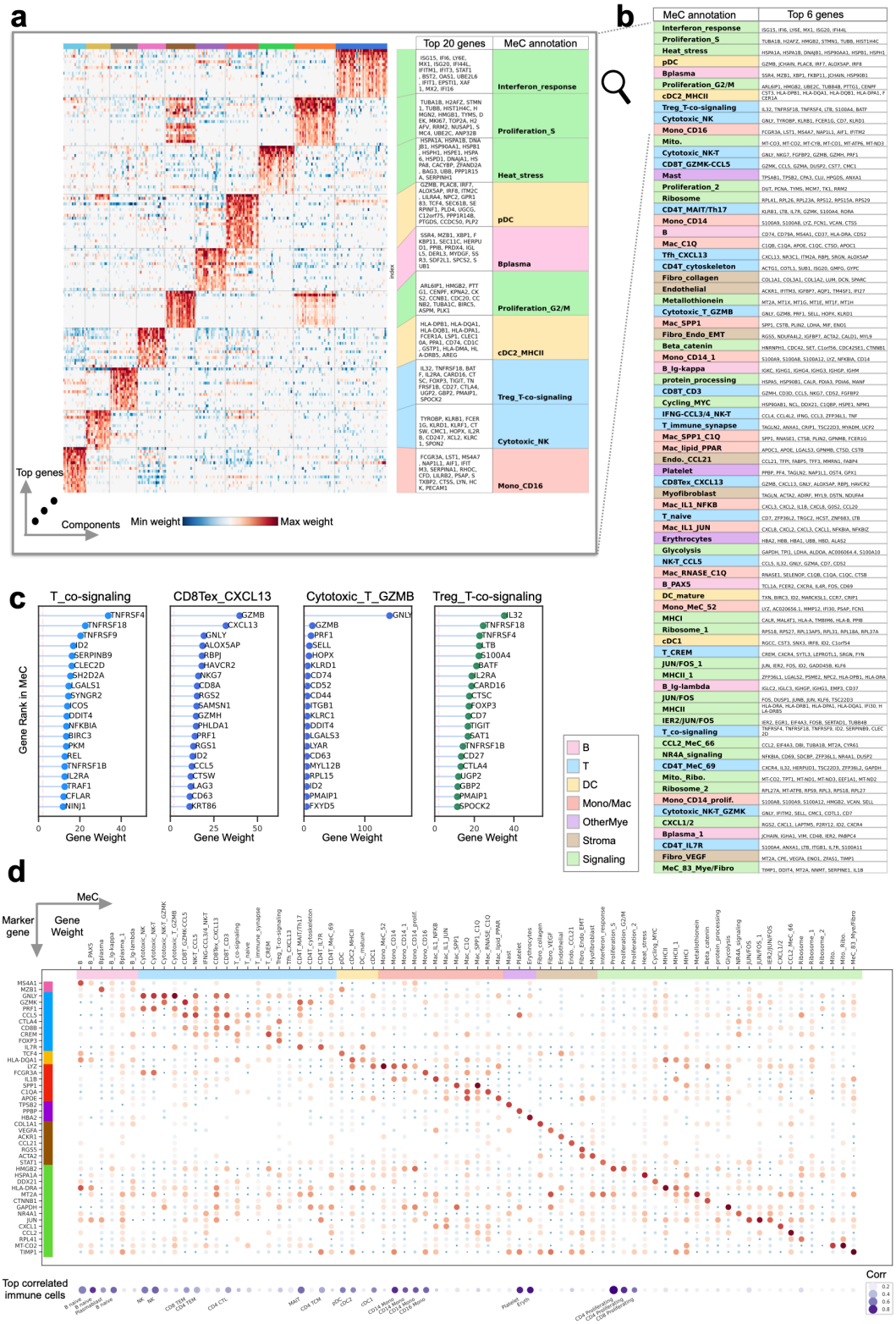
548

549

550 **Fig.1. Overview of MetaTiME.** MetaTiME integrates 1.7 million single cells to learn common
551 transcriptional programs in the tumor microenvironment (TME). **(a)** Steps for Meta-components
552 (MeCs) discovery. For each scRNA dataset, the expression matrix of TME cells is decomposed
553 into a loading matrix (red) and an independent component (IC) matrix through independent
554 component analysis (ICA). The ICs represent mutually independent sources of transcriptional
555 variation. ICs from each dataset are concatenated and clustered into groups of ICs with high
556 similarity, representing transcriptional programs shared across TME. MeCs are then calculated
557 as averaged profiles of ICs from each cluster. Each MeC is interpretable representing gene
558 signatures of cell type, cell states, or signaling pathway activities. **(b)** Left: MetaTiME provides
559 75 functionally annotated MeCs that depict the TME transcriptional landscape. They are
560 grouped into six lineage-related categories and one category reflecting signaling activities.
561 Middle: the MetaTiME annotation tool facilitates automatic annotation of cell states for new

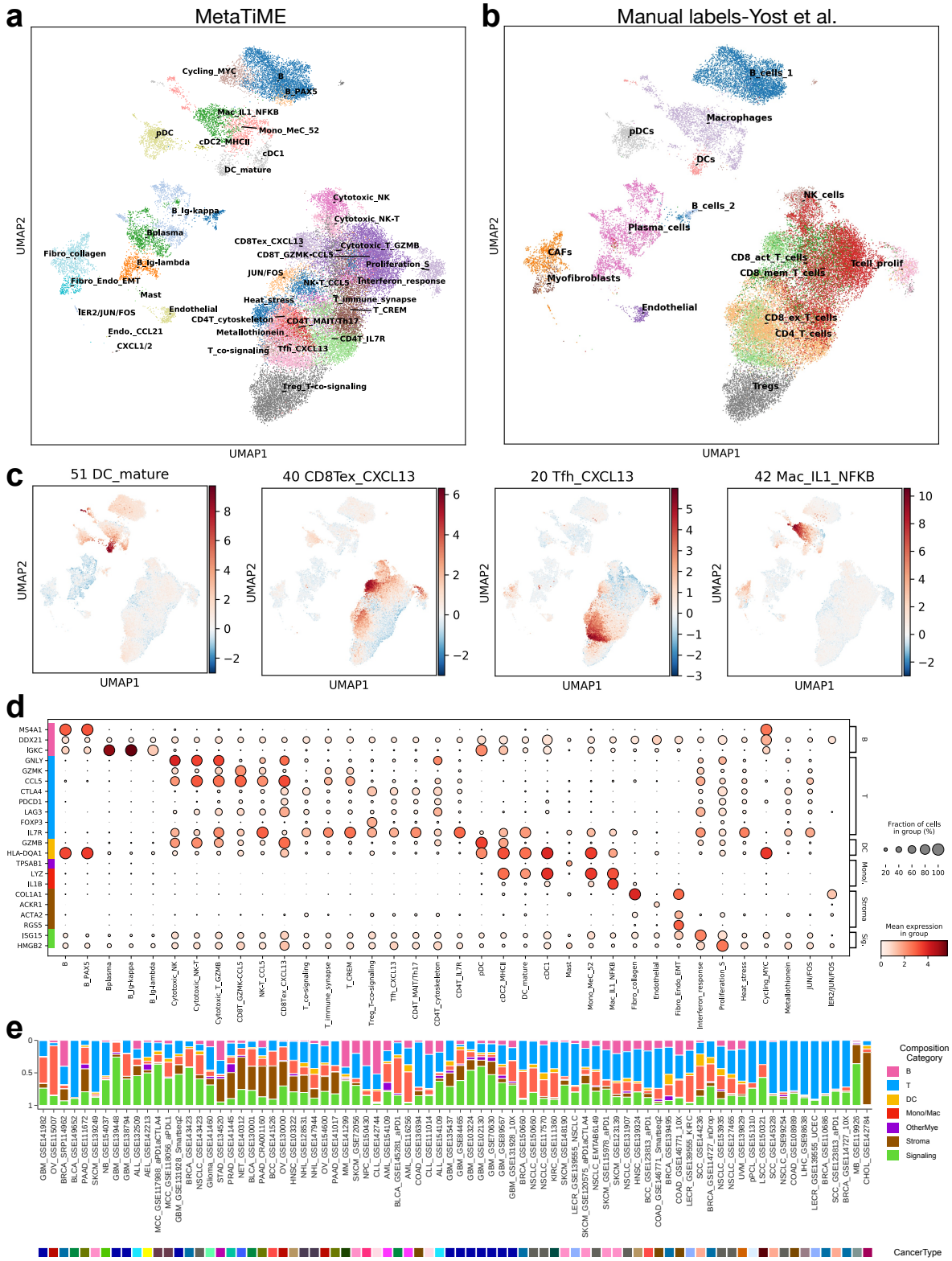
562 tumor scRNA data. Right: candidate regulators of each MeC are prioritized by combining MeC
563 gene weights with epigenetics data. MeC: meta-components, TME: tumor microenvironment,
564 ICA: independent component analysis, MeC: meta-component, TF: transcription factor.
565

Fig. 2.


a
Top 20 genes
MeC annotation
Interferon_response
Proliferation_S
Heat_stress
pDC
Bplasma
Proliferation_G2/M
cDC2_MHCII
Treg_T-co-signaling
Cytotoxic_NK
Mono_CD16
Top genes
Components
Min weight
Max weight
b
MeC annotation
Top 6 genes
c
T_co-signaling
CD8Tex_CXCL13
Cytotoxic_T_GZMB
Treg_T-co-signaling
Gene Rank in MeC
Gene Weight
B
T
DC
Mono/Mac
OtherMye
Stroma
Signaling
d
MeC
Marker gene
Gene Weight
Top correlated immune cells
Corr

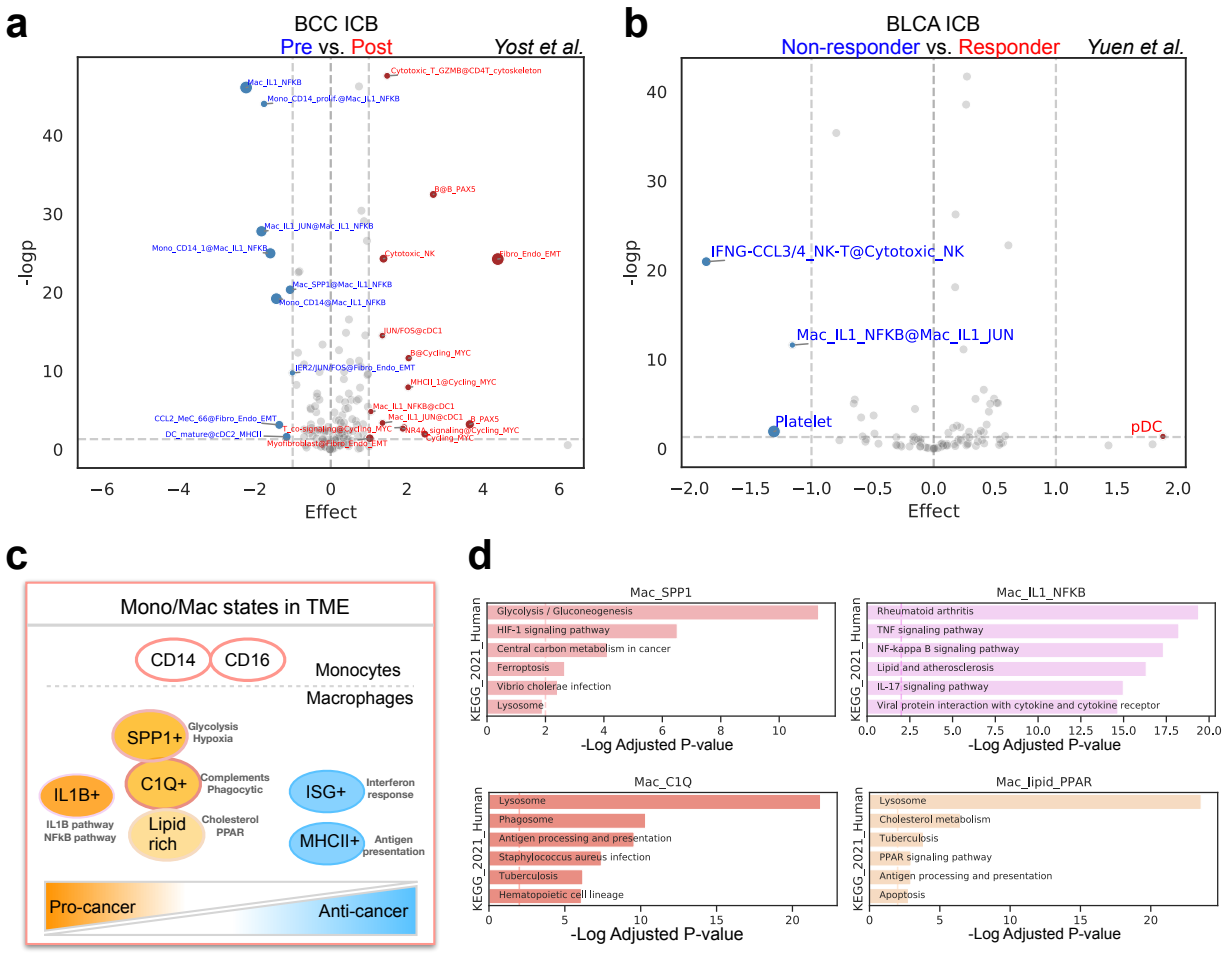
568 **Fig.2. MetaTiME meta-components are biologically interpretable with top genes. (a)**
569 Heatmap of top ten most recurrent clusters of MeCs showing normalized gene weights. **(b)**
570 Biological characterization of each MeC with top genes. To facilitate biological interpretation,
571 MeCs are 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

Fig.3.


a
MetaTiME
Cycling_MYC
B
B_PAX5
Mac_IL1_NFKB
Mono_MeC_52
pDC
cDC2_MHCII
cDC1
DC_mature
Cytotoxic_NK
B_Ig-kappa
Cytotoxic_NK-T
Bplasma
CD8Tex_CXCL13
Cytotoxic_T_GZMB
Fibro_collagen
CD8T_GZMK-CCL5
Proliferation_S
B_Ig-lambda
JUN/FOS
NK-T_CCL5
Interferon_response
Fibro_Endo_EMT
Heat_stress
T_immune_synapse
Mast
T_CREM
Endothelial
IER2/JUN/FOS
CD4T_cytoskeleton
CD4T_MAIT/Th17
Endo._CCL21
Metallothionein
CD4T_IL7R
CXCL1/2
T_co-signaling
Tfh_CXCL13
Treg_T-co-signaling
UMAP2
UMAP1
b
Manual labels-Yost et al.
B_cells_1
Macrophages
pDCs
DCs
NK_cells
B_cells_2
Plasma_cells
CAFs
CD8_act_T_cells
Tcell_prolif
Myofibroblasts
CD8_mem_T_cells
Endothelial
CD8_ex_T_cells
CD4_T_cells
Tregs
UMAP2
UMAP1
c
51 DC_mature
40 CD8Tex_CXCL13
20 Tfh_CXCL13
42 Mac_IL1_NFKB
UMAP2
UMAP1
d
MS4A1
DDX21
IGKC
GNLY
GZMK
CCL5
CTLA4
PDCD1
LAG3
FOXP3
IL7R
GZMB
HLA-DQA1
TPSAB1
LYZ
IL1B
COL1A1
ACKR1
ACTA2
RGS5
ISG15
HMGB2
B
T
DC
Mono/
Stroma
Sig.
Fraction of cells in group (%)
20 40 60 80 100
Mean expression in group
0 2 4
e
Composition Category
B
T
DC
Mono/Mac
OtherMye
Stroma
Signaling
CancerType

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.

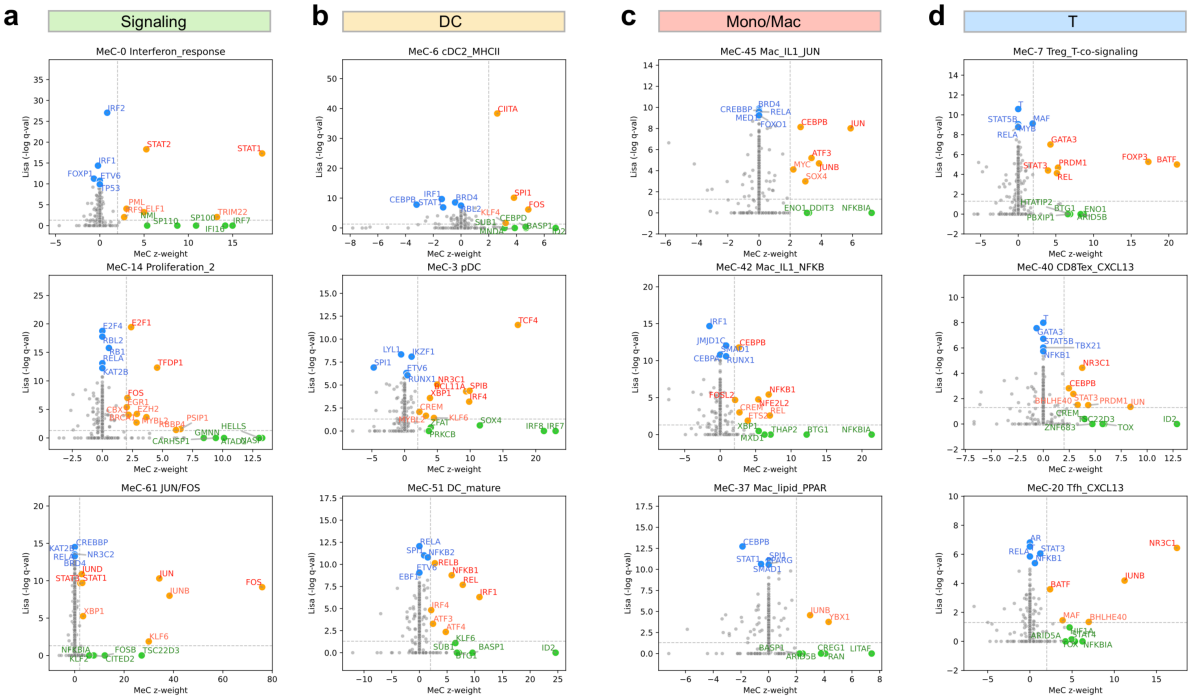


Fig.5. MetaTiME prioritizes tumor immunity transcriptional regulators. For selected MeCs, TFs are prioritized by their MeC expression representation and Lisa, ChIP-seq based, regulatory potentials. X-axis: gene z-weight of the TF for the current MeC. Y-axis: Lisa-based regulatory potential significance 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).

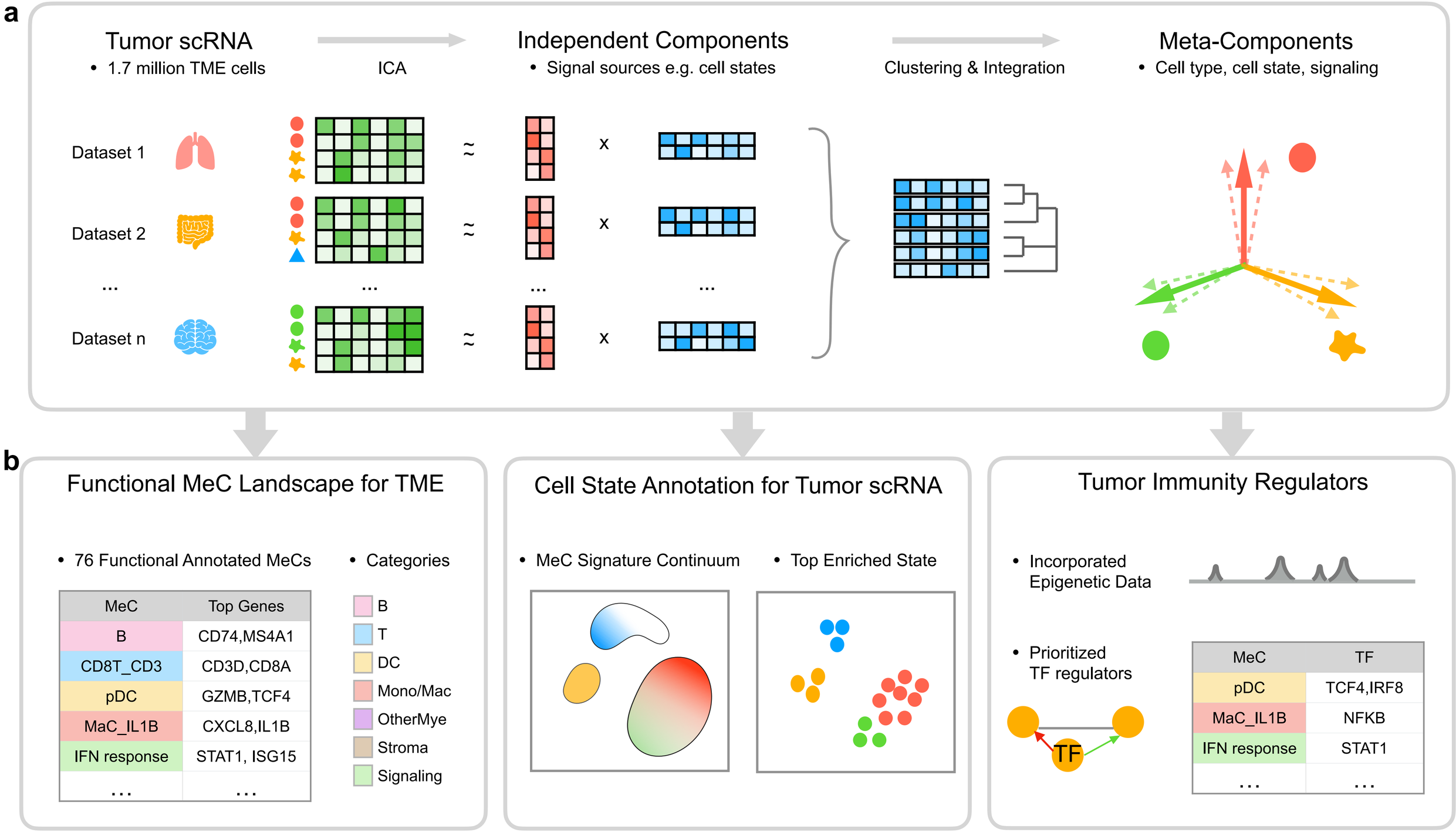
649 12. Wu, S. Z. *et al.* A single-cell and spatially resolved atlas of human breast cancers. *Nat.*
650 *Genet.* **53**, 1334–1347 (2021).
651 13. Nieto, P. *et al.* A single-cell tumor immune atlas for precision oncology. *Genome Res.* **31**,
652 1913–1926 (2021).
653 14. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902.e21
654 (2019).
655 15. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587.e29
656 (2021).
657 16. Aran, D. *et al.* Reference-based analysis of lung single-cell sequencing reveals a
658 transitional profibrotic macrophage. *Nat. Immunol.* **20**, 163–172 (2019).
659 17. Wang, C. *et al.* Integrative analyses of single-cell transcriptome and regulome using
660 MAESTRO. *Genome Biol.* **21**, (2020).
661 18. Newman, A. M. *et al.* Robust enumeration of cell subsets from tissue expression profiles.
662 *Nat. Methods 2015 125* **12**, 453–457 (2015).
663 19. Andreatta, M. *et al.* Interpretation of T cell states from single-cell transcriptomics data
664 using reference atlases. *Nat. Commun.* **12**, 1–19 (2021).
665 20. Zheng, A. L., Qin, S., Wang, A., Xing, B. & Gao, R. Pan-Cancer Single Cell Landscape of
666 Tumor-Infiltrating T Cells. **6474**, (2021).
667 21. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell
668 transcriptomic data across different conditions, technologies, and species. *Nat.*
669 *Biotechnol. 2018 365* **36**, 411–420 (2018).
670 22. Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with
671 Harmony. *Nat. Methods* **16**, 1289–1296 (2019).
672 23. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for
673 single-cell transcriptomics. *Nat. Methods* (2018). doi:10.1038/s41592-018-0229-2
674 24. Pelka, K. *et al.* Spatially organized multicellular immune hubs in human colorectal cancer.

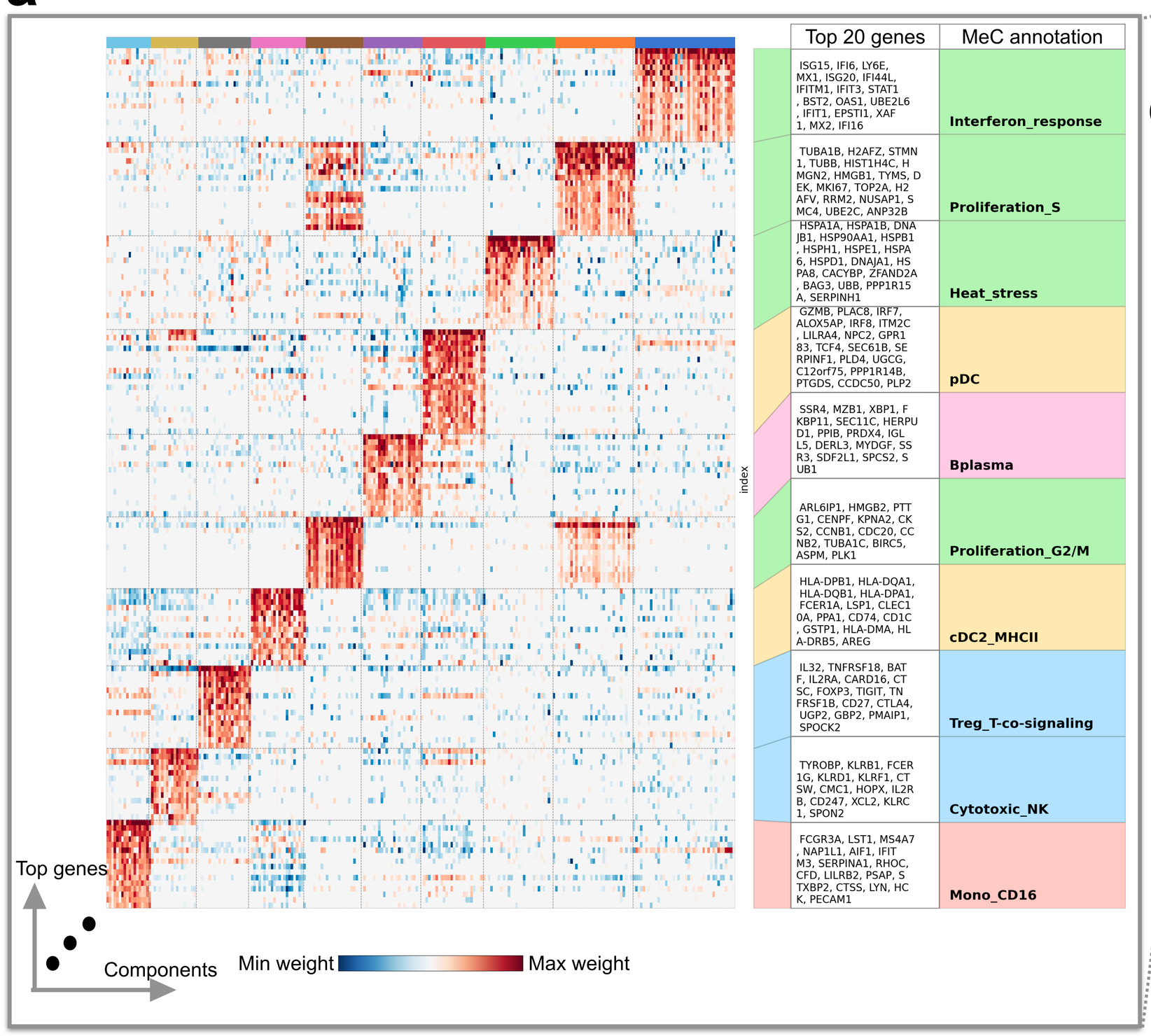
675 *Cell* **184**, 4734-4752.e20 (2021).
676 25. Hyvärinen, A. Fast and robust fixed-point algorithms for independent component analysis.
677 *IEEE Trans. Neural Networks* (1999). doi:10.1109/72.761722
678 26. Wang, W. *et al.* Independent component analysis based gene co-expression network
679 inference (ICAnet) to decipher functional modules for better single-cell clustering and
680 batch integration. *Nucleic Acids Res.* **49**, E54 (2021).
681 27. Zheng, R. *et al.* Cistrome Data Browser: expanded datasets and new tools for gene
682 regulatory analysis. *Nucleic Acids Res.* **47**, D729–D735 (2019).
683 28. Sun, D. *et al.* TISCH: A comprehensive web resource enabling interactive single-cell
684 transcriptome visualization of tumor microenvironment. *Nucleic Acids Res.* **49**, D1420–
685 D1430 (2021).
686 29. Wykes, M. N. & Lewin, S. R. Immune checkpoint blockade in infectious diseases. *Nat.*
687 *Rev. Immunol. 2017 182* **18**, 91–104 (2017).
688 30. Zhang, Q. & Vignali, D. A. A. Co-stimulatory and co-inhibitory pathways in autoimmunity.
689 *Immunity* **44**, 1034 (2016).
690 31. Qin, S. *et al.* Novel immune checkpoint targets: moving beyond PD-1 and CTLA-4. *Mol.*
691 *Cancer 2019 181* **18**, 1–14 (2019).
692 32. Thommen, D. S. *et al.* A transcriptionally and functionally distinct PD-1+ CD8+ T cell pool
693 with predictive potential in non-small-cell lung cancer treated with PD-1 blockade. *Nat.*
694 *Med. 2018 247* **24**, 994–1004 (2018).
695 33. Chowell, D. *et al.* Improved prediction of immune checkpoint blockade efficacy across
696 multiple cancer types. *Nat. Biotechnol. 2021 404* **40**, 499–506 (2021).
697 34. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for
698 interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 15545–
699 15550 (2005).
700 35. Morgan, D. & Tergaonkar, V. Unraveling B cell trajectories at single cell resolution.

701 *Trends Immunol.* **43**, 210–229 (2022).
702 36. Finkin, S., Hartweger, H., Oliveira, T. Y., Kara, E. E. & Nussenzweig, M. C. Protein
703 Amounts of the MYC Transcription Factor Determine Germinal Center B Cell Division
704 Capacity. *Immunity* **51**, 324-336.e5 (2019).
705 37. Vaquero, J., Aoudjehane, L. & Fouassier, L. Cancer-associated fibroblasts in
706 cholangiocarcinoma. *Curr. Opin. Gastroenterol.* **36**, 63–69 (2020).
707 38. Luca, B. A. *et al.* Atlas of clinically distinct cell states and ecosystems across human solid
708 tumors. *Cell* **184**, 5482-5496.e28 (2021).
709 39. Zhao, J. *et al.* Detection of differentially abundant cell subpopulations in scrna-seq data.
710 *Proc. Natl. Acad. Sci. U. S. A.* **118**, (2021).
711 40. Yuen, K. C. *et al.* High systemic and tumor-associated IL-8 correlates with reduced
712 clinical benefit of PD-L1 blockade. *Nat. Med.* **26**, 693–698 (2020).
713 41. Kaplanov, I. *et al.* Blocking IL-1β reverses the immunosuppression in mouse breast
714 cancer and synergizes with anti-PD-1 for tumor abrogation. *Proc. Natl. Acad. Sci. U. S. A.*
715 **116**, 1361–1369 (2019).
716 42. Duan, Z. & Luo, Y. Targeting macrophages in cancer immunotherapy. *Signal Transduct.*
717 *Target. Ther. 2021 61* **6**, 1–21 (2021).
718 43. Bi, K. *et al.* Tumor and immune reprogramming during immunotherapy in advanced renal
719 cell carcinoma. *Cancer Cell* **39**, 649-661.e5 (2021).
720 44. Zhao, H. *et al.* The role of osteopontin in the progression of solid organ tumour. *Cell*
721 *Death Dis.* **9**, (2018).
722 45. Spivia, W., Magno, P. S., Le, P. & Fraser, D. A. Complement protein C1q promotes
723 macrophage anti-inflammatory M2-like polarization during the clearance of atherogenic
724 lipoproteins. *Inflamm. Res.* **63**, 885–893 (2014).
725 46. Liu, Q. *et al.* The CXCL8-CXCR1/2 pathways in cancer. *Cytokine Growth Factor Rev.* **31**,
726 61–71 (2016).

727 47. Qin, Q. *et al.* Lisa: inferring transcriptional regulators through integrative modeling of
728 public chromatin accessibility and ChIP-seq data. *Genome Biol.* (2020).
729 doi:10.1186/s13059-020-1934-6
730 48. Peters, J. M., Shah, Y. M. & Gonzalez, F. J. The role of peroxisome proliferator-activated
731 receptors in carcinogenesis and chemoprevention. *Nat. Rev. Cancer 2012 123* **12**, 181–
732 195 (2012).
733 49. Hu, H. *et al.* AnimalTFDB 3.0: a comprehensive resource for annotation and prediction of
734 animal transcription factors. *Nucleic Acids Res.* **47**, D33–D38 (2019).
735 50. Li, H. *et al.* Dysfunctional CD8 T Cells Form a Proliferative, Dynamically Regulated
736 Compartment within Human Melanoma. *Cell* **176**, 775-789.e18 (2019).
737 51. Hyvärinen, A. & Oja, E. Independent component analysis: algorithms and applications.
738 *Neural Networks* **13**, 411–430 (2000).
739 52. Lotfollahi, M. *et al.* Mapping single-cell data to reference atlases by transfer learning. *Nat.*
740 *Biotechnol.* (2021). doi:10.1038/s41587-021-01001-7
741 53. Khan, O. *et al.* TOX transcriptionally and epigenetically programs CD8+ T cell
742 exhaustion. *Nat. 2019 5717764* **571**, 211–218 (2019).
743 54. Tang, Q. & Brown, M. A. Abstract A218: Modulating glucocorticoid receptor-mediated
744 signaling for enhancement of cancer immunotherapy. *Cancer Immunol. Res.* **7**, A218–
745 A218 (2019).
746 55. Ohandjo, A. Q. *et al.* Transcriptome Network Analysis Identifies CXCL13-CXCR5
747 Signaling Modules in the Prostate Tumor Immune Microenvironment. *Sci. Reports 2019*
748 *91* **9**, 1–13 (2019).
749 56. Havenar-Daughton, C. *et al.* CXCL13 is a plasma biomarker of germinal center activity.
750 *Proc. Natl. Acad. Sci. U. S. A.* **113**, 2702–2707 (2016).
751 57. Kotliar, D. *et al.* Identifying gene expression programs of cell-type identity and cellular
752 activity with single-cell RNA-Seq. *Elife* **8**, 1–26 (2019).

753

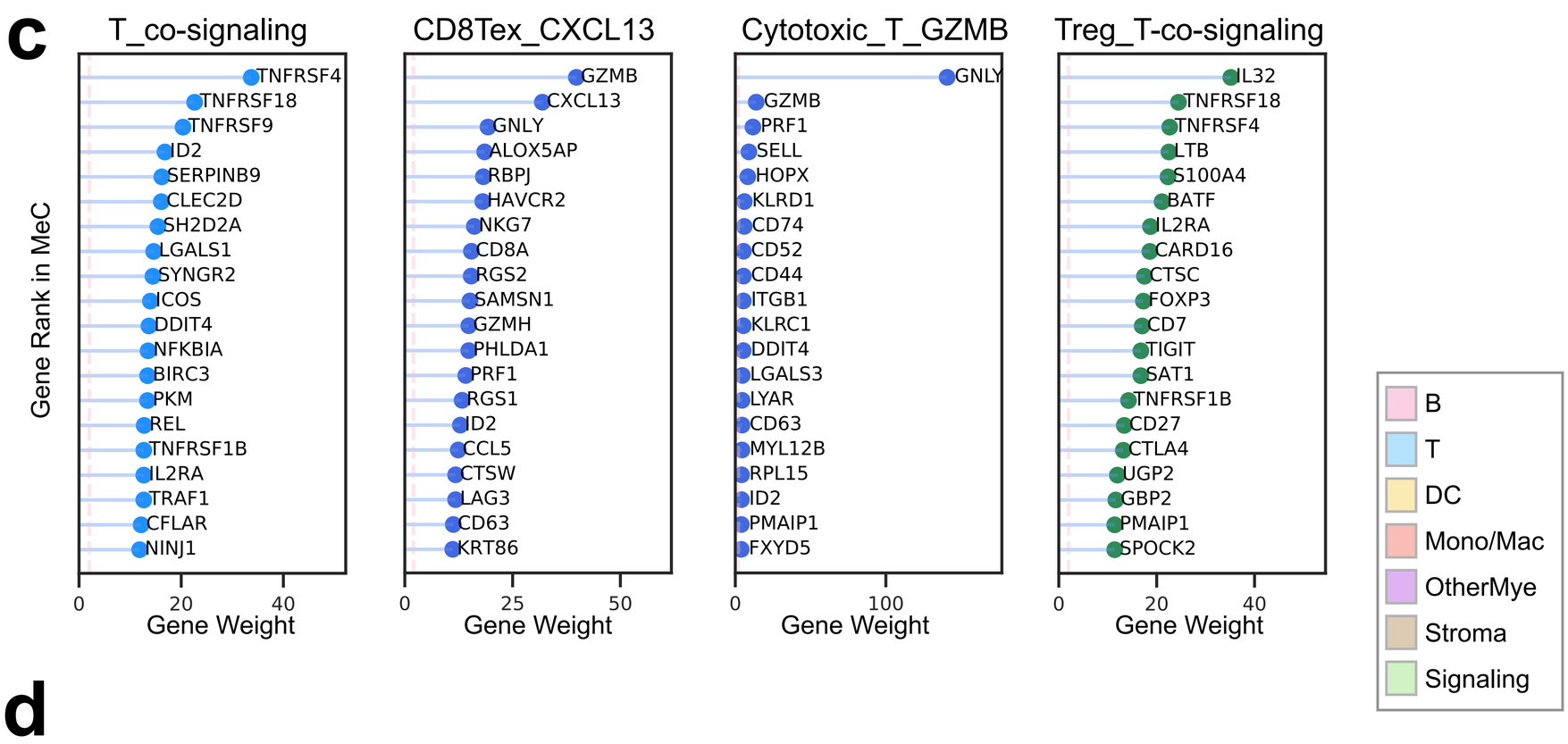

a
Tumor scRNA
• 1.7 million TME cells
ICA
Independent Components
• Signal sources e.g. cell states
Clustering & Integration
Meta-Components
• Cell type, cell state, signaling
Dataset 1
Dataset 2
...
Dataset n
b
Functional MeC Landscape for TME
• 76 Functional Annotated MeCs
• Categories
MeC | Top Genes
B | CD74,MS4A1
CD8T_CD3 | CD3D,CD8A
pDC | GZMB,TCF4
MaC_IL1B | CXCL8,IL1B
IFN response | STAT1, ISG15
B
T
DC
Mono/Mac
OtherMye
Stroma
Signaling
Cell State Annotation for Tumor scRNA
• MeC Signature Continuum
• Top Enriched State
Tumor Immunity Regulators
• Incorporated Epigenetic Data
• Prioritized TF regulators
TF
MeC | TF
pDC | TCF4,IRF8
MaC_IL1B | NFKB
IFN response | STAT1



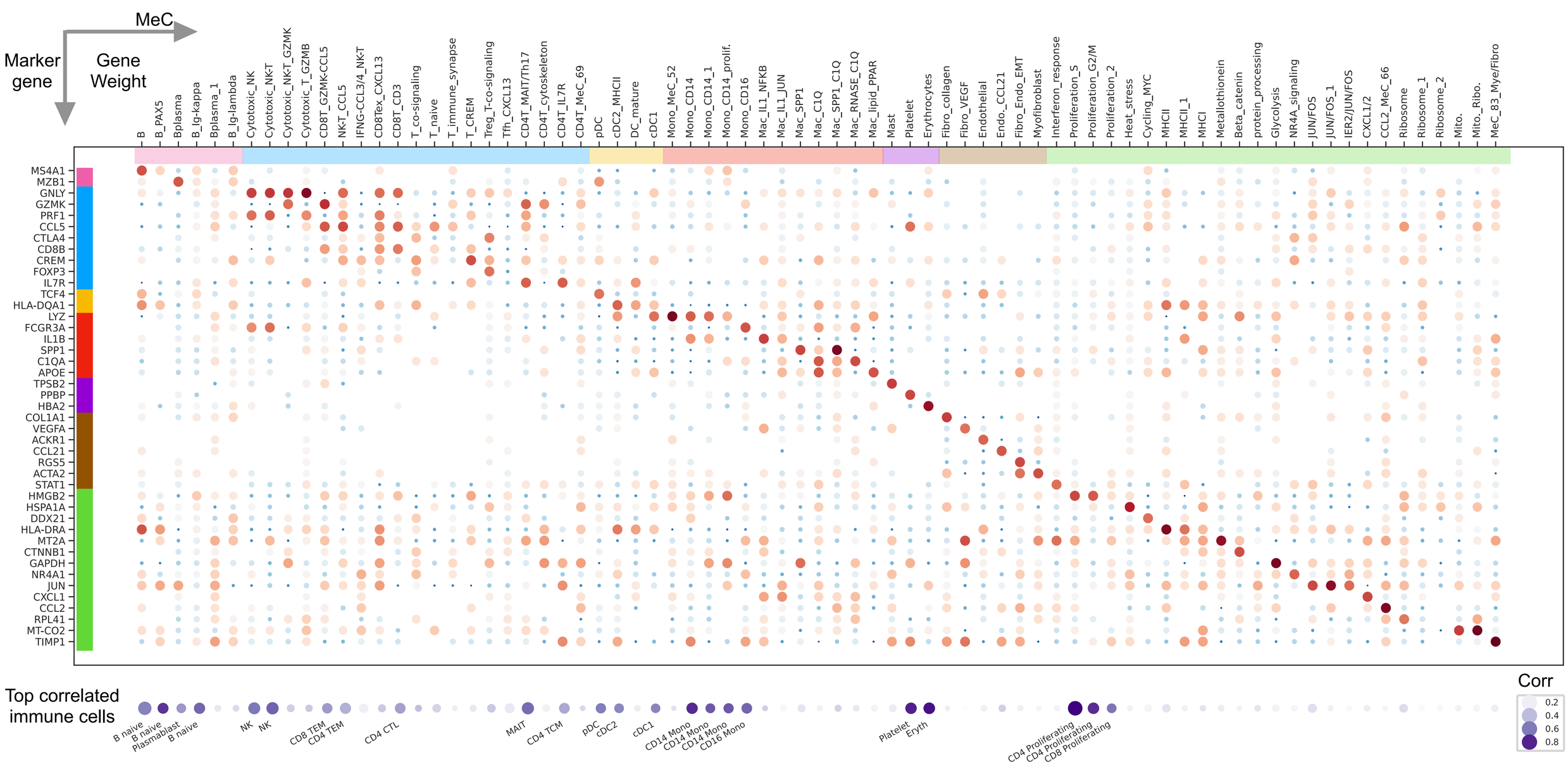
b

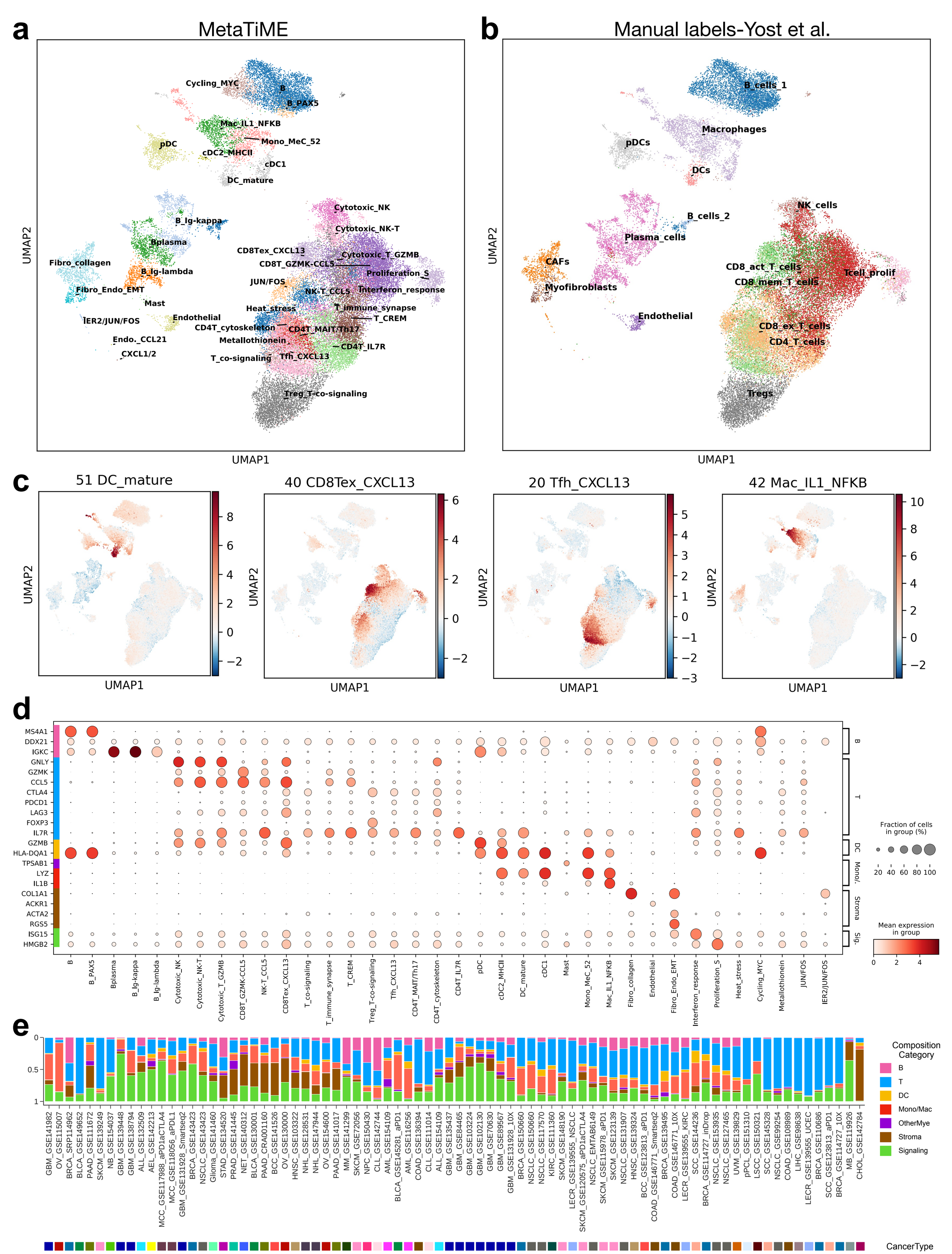
MeC annotation	Top 6 genes
Interferon_response	ISG15, IFI6, LY6E, MX1, ISG20, IFI44L
Proliferation_S	TUBA1B, H2AFZ, HMGB2, STMN1, TUBB, HIST1H4C
Heat_stress	HSPA1A, HSPA1B, DNAJB1, HSP90AA1, HSPB1, HSPH1
pDC	GZMB, JCHAIN, PLAC8, IRF7, ALOX5AP, IRF8
Bplasma	SSR4, MZB1, XBP1, FKBP11, JCHAIN, HSP90B1
Proliferation_G2/M	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

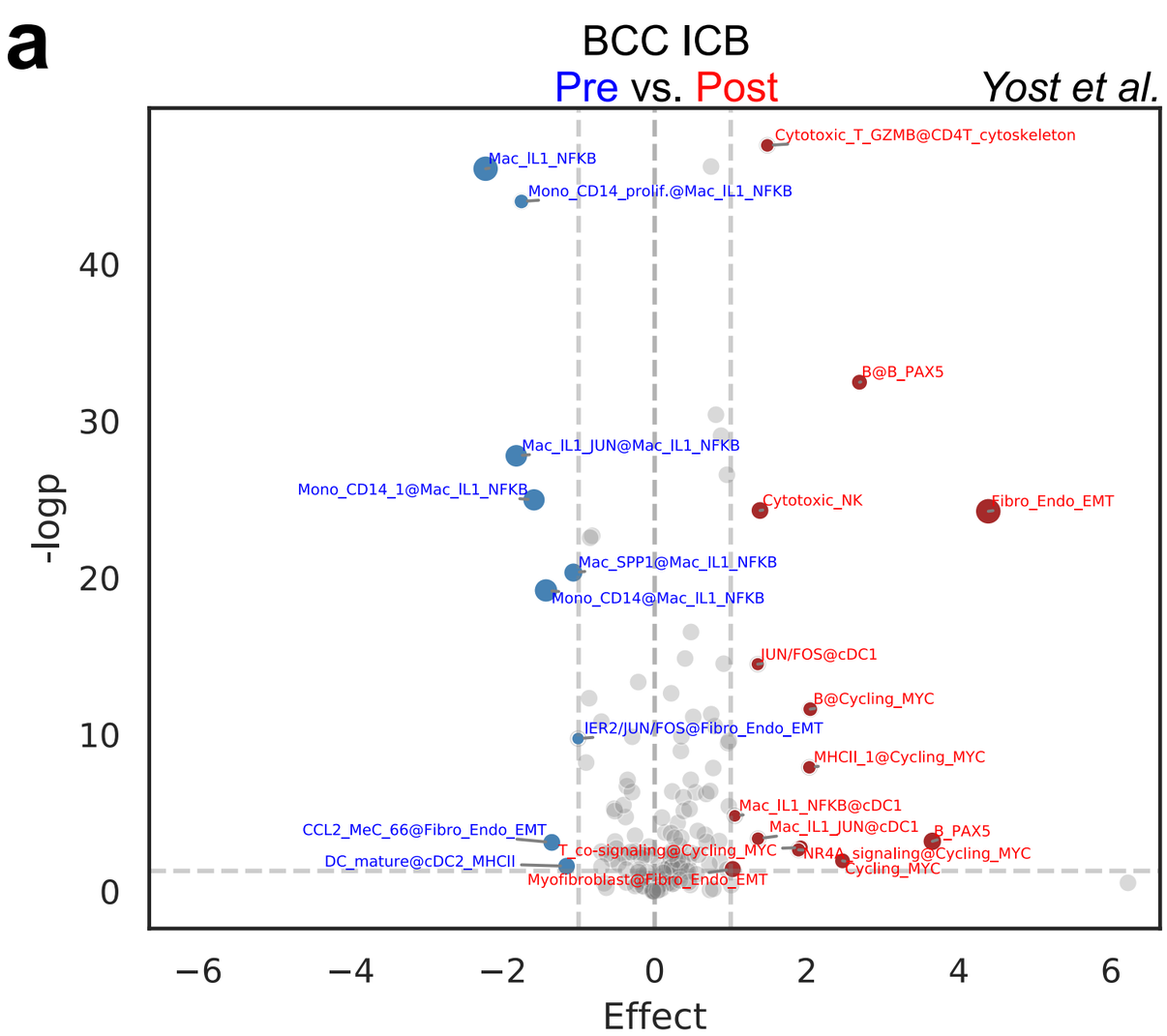
c

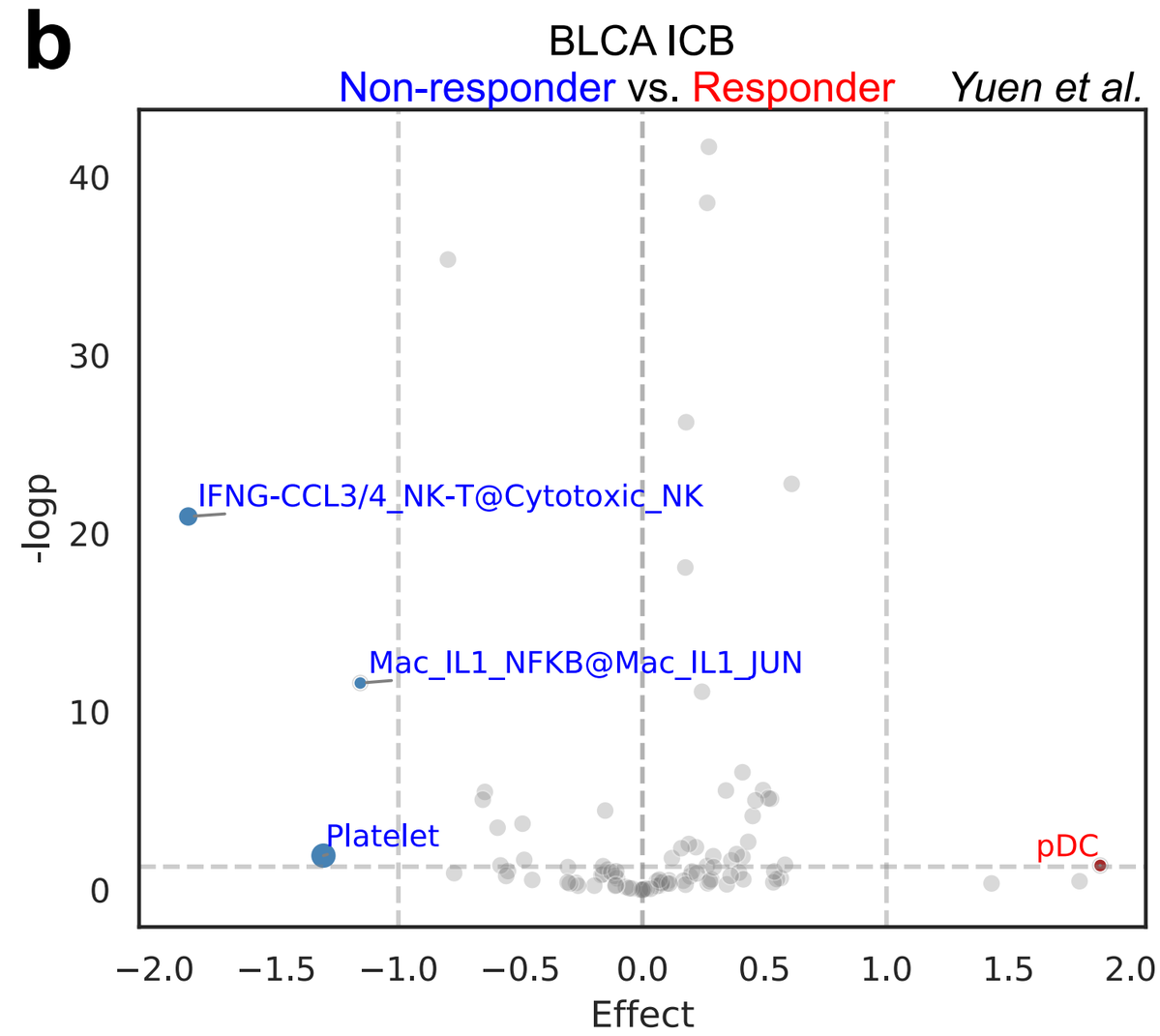


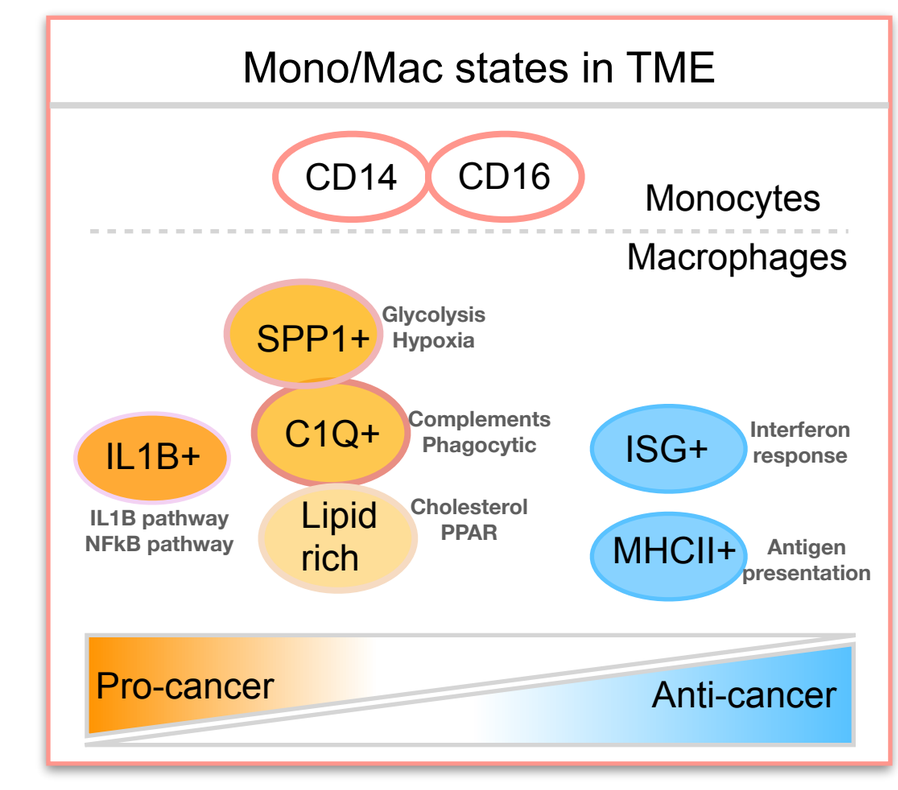
d

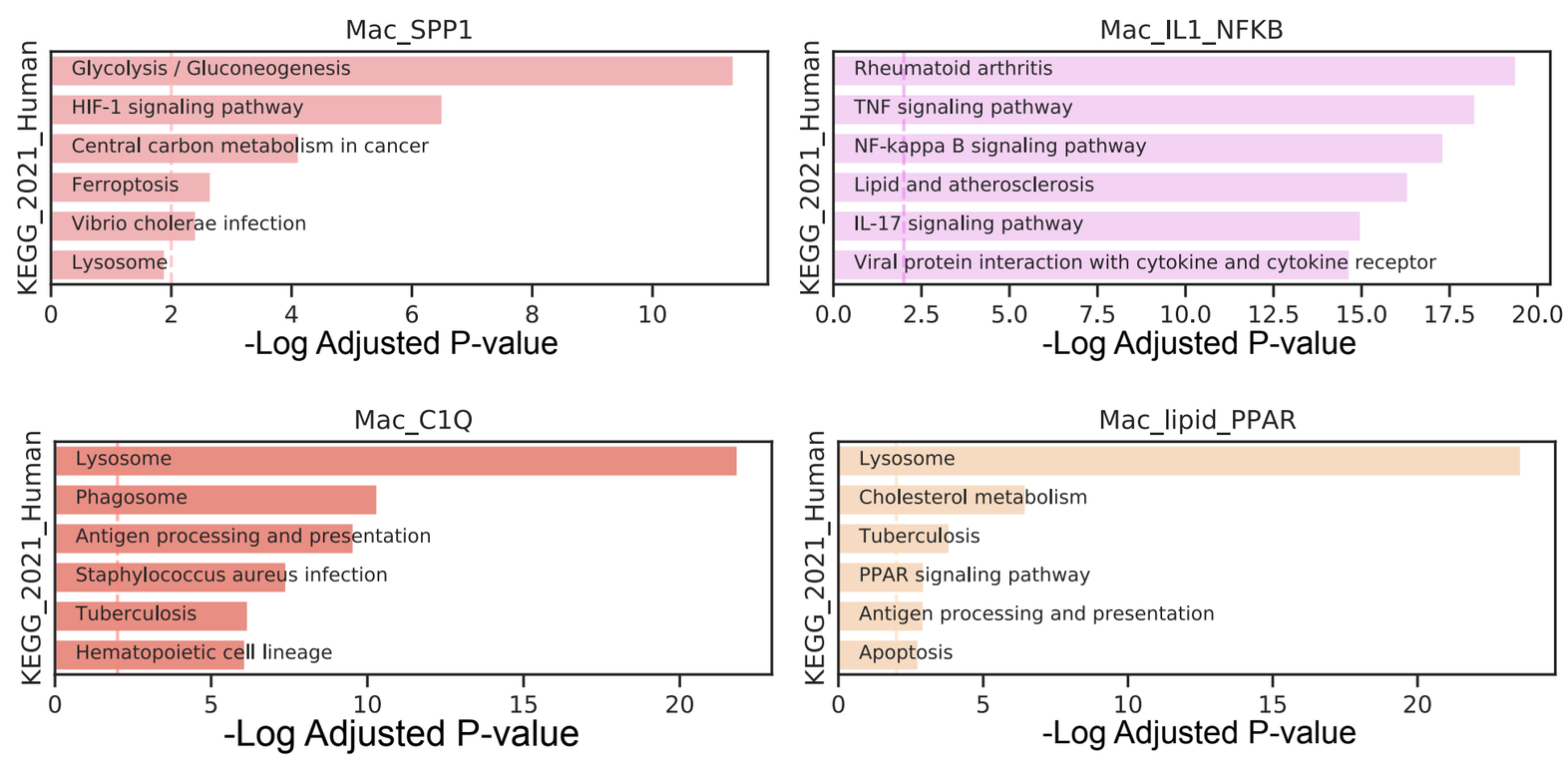


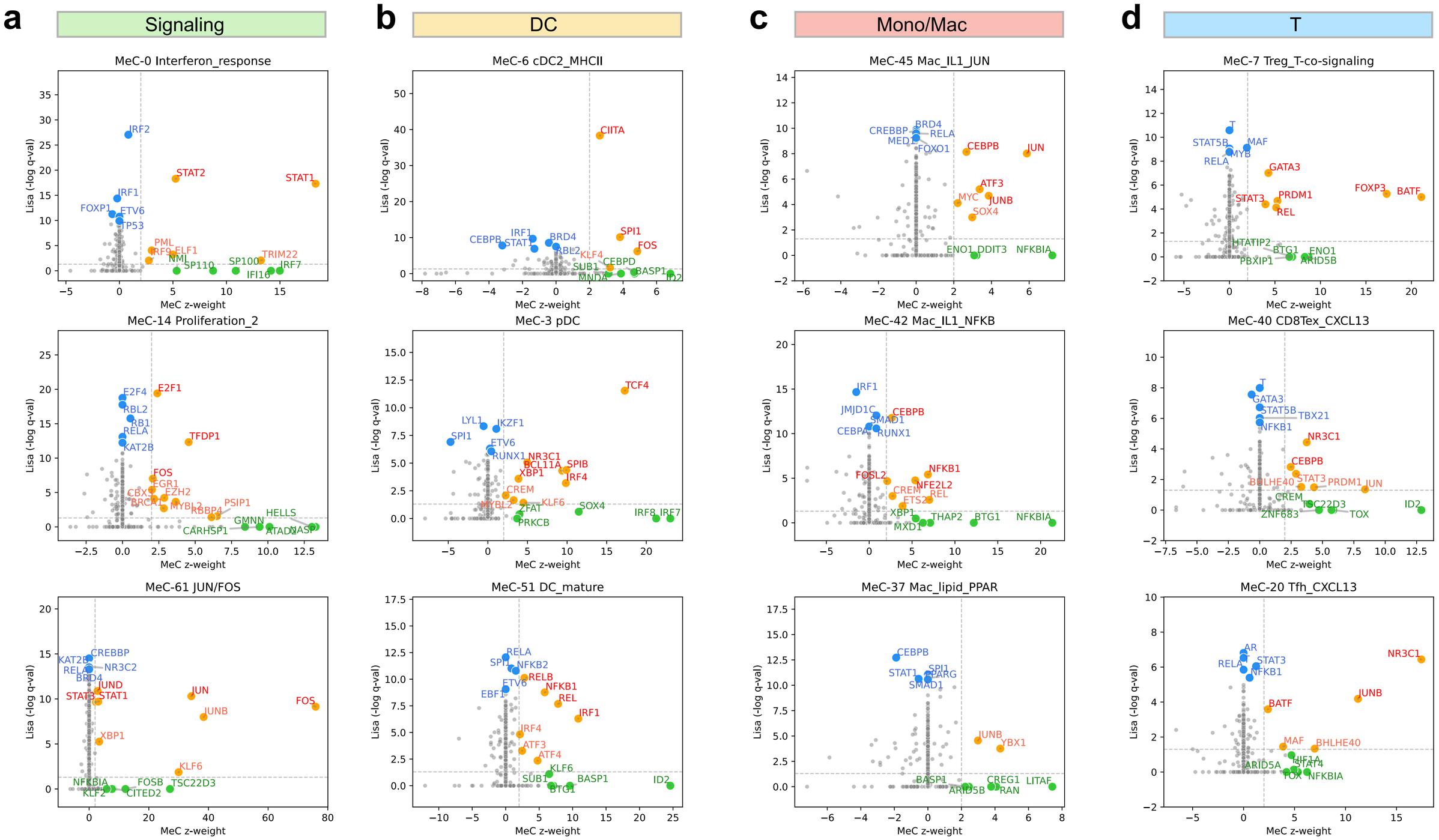

a
MetaTiME
Cycling_MYC
B
B_PAX5
Mac_IL1_NFKB
Mono_MeC_52
pDC
cDC2_MHCII
cDC1
DC_mature
B_Ig-kappa
Bplasma
Fibro_collagen
B_Ig-lambda
Fibro_Endo_EMT
Mast
IER2/JUN/FOS
Endothelial
Endo._CCL21
CXCL1/2
Cytotoxic_NK
Cytotoxic_NK-T
CD8Tex_CXCL13
Cytotoxic_T_GZMB
CD8T_GZMK-CCL5
Proliferation_S
JUN/FOS
NK-T_CCL5
Interferon_response
Heat_stress
T_immune_synapse
T_CREM
CD4T_cytoskeleton
CD4T_MAIT/Th17
Metallothionein
CD4T_IL7R
T_co-signaling
Tfh_CXCL13
Treg_T-co-signaling
UMAP2
UMAP1
b
Manual labels-Yost et al.
B_cells_1
Macrophages
pDCs
DCs
B_cells_2
NK_cells
Plasma_cells
CAFs
CD8_act_T_cells
Tcell_prolif
CD8_mem_T_cells
Myofibroblasts
Endothelial
CD8_ex_T_cells
CD4_T_cells
Tregs
UMAP2
UMAP1
c
51 DC_mature
40 CD8Tex_CXCL13
20 Tfh_CXCL13
42 Mac_IL1_NFKB
UMAP2
UMAP1
d
MS4A1
DDX21
IGKC
GNLY
GZMK
CCL5
CTLA4
PDCD1
LAG3
FOXP3
IL7R
GZMB
HLA-DQA1
TPSAB1
LYZ
IL1B
COL1A1
ACKR1
ACTA2
RGS5
ISG15
HMGB2
B
T
DC
Mono/.
Stroma
Sig.
Fraction of cells in group (%)
20 40 60 80 100
Mean expression in group
0 2 4
e
Composition Category
B
T
DC
Mono/Mac
OtherMye
Stroma
Signaling
CancerType


a
BCC ICB
Pre vs. Post
Yost et al.
Cytotoxic_T_GZMB@CD4T_cytoskeleton
Mac_IL1_NFKB
Mono_CD14_prolif.@Mac_IL1_NFKB
B@B_PAX5
Mac_IL1_JUN@Mac_IL1_NFKB
Mono_CD14_1@Mac_IL1_NFKB
Cytotoxic_NK
Fibro_Endo_EMT
Mac_SPP1@Mac_IL1_NFKB
Mono_CD14@Mac_IL1_NFKB
JUN/FOS@cDC1
B@Cycling_MYC
IER2/JUN/FOS@Fibro_Endo_EMT
MHCII_1@Cycling_MYC
Mac_IL1_NFKB@cDC1
Mac_IL1_JUN@cDC1
B_PAX5
CCL2_MeC_66@Fibro_Endo_EMT
T_co-signaling@Cycling_MYC
NR4A_signaling@Cycling_MYC
DC_mature@cDC2_MHCII
Cycling_MYC
Myofibroblast@Fibro_Endo_EMT
-logp
Effect


b
BLCA ICB
Non-responder vs. Responder
Yuen et al.
IFNG-CCL3/4_NK-T@Cytotoxic_NK
Mac_IL1_NFKB@Mac_IL1_JUN
Platelet
pDC
-logp
Effect


c
Mono/Mac states in TME
CD14
CD16
Monocytes
Macrophages
SPP1+
Glycolysis
Hypoxia
C1Q+
Complements
Phagocytic
IL1B+
IL1B pathway
NFkB pathway
Lipid rich
Cholesterol
PPAR
ISG+
Interferon response
MHCII+
Antigen presentation
Pro-cancer
Anti-cancer


d
Mac_SPP1
Glycolysis / Gluconeogenesis
HIF-1 signaling pathway
Central carbon metabolism in cancer
Ferroptosis
Vibrio cholerae infection
Lysosome
Mac_IL1_NFKB
Rheumatoid arthritis
TNF signaling pathway
NF-kappa B signaling pathway
Lipid and atherosclerosis
IL-17 signaling pathway
Viral protein interaction with cytokine and cytokine receptor
Mac_C1Q
Lysosome
Phagosome
Antigen processing and presentation
Staphylococcus aureus infection
Tuberculosis
Hematopoietic cell lineage
Mac_lipid_PPAR
Lysosome
Cholesterol metabolism
Tuberculosis
PPAR signaling pathway
Antigen processing and presentation
Apoptosis
KEGG_2021_Human
-Log Adjusted P-value


a
Signaling
b
DC
c
Mono/Mac
d
T
MeC-0 Interferon_response
MeC-6 cDC2_MHCII
MeC-45 Mac_IL1_JUN
MeC-7 Treg_T-co-signaling
MeC-14 Proliferation_2
MeC-3 pDC
MeC-42 Mac_IL1_NFKB
MeC-40 CD8Tex_CXCL13
MeC-61 JUN/FOS
MeC-51 DC_mature
MeC-37 Mac_lipid_PPAR
MeC-20 Tfh_CXCL13
Lisa (-log q-val)
MeC z-weight