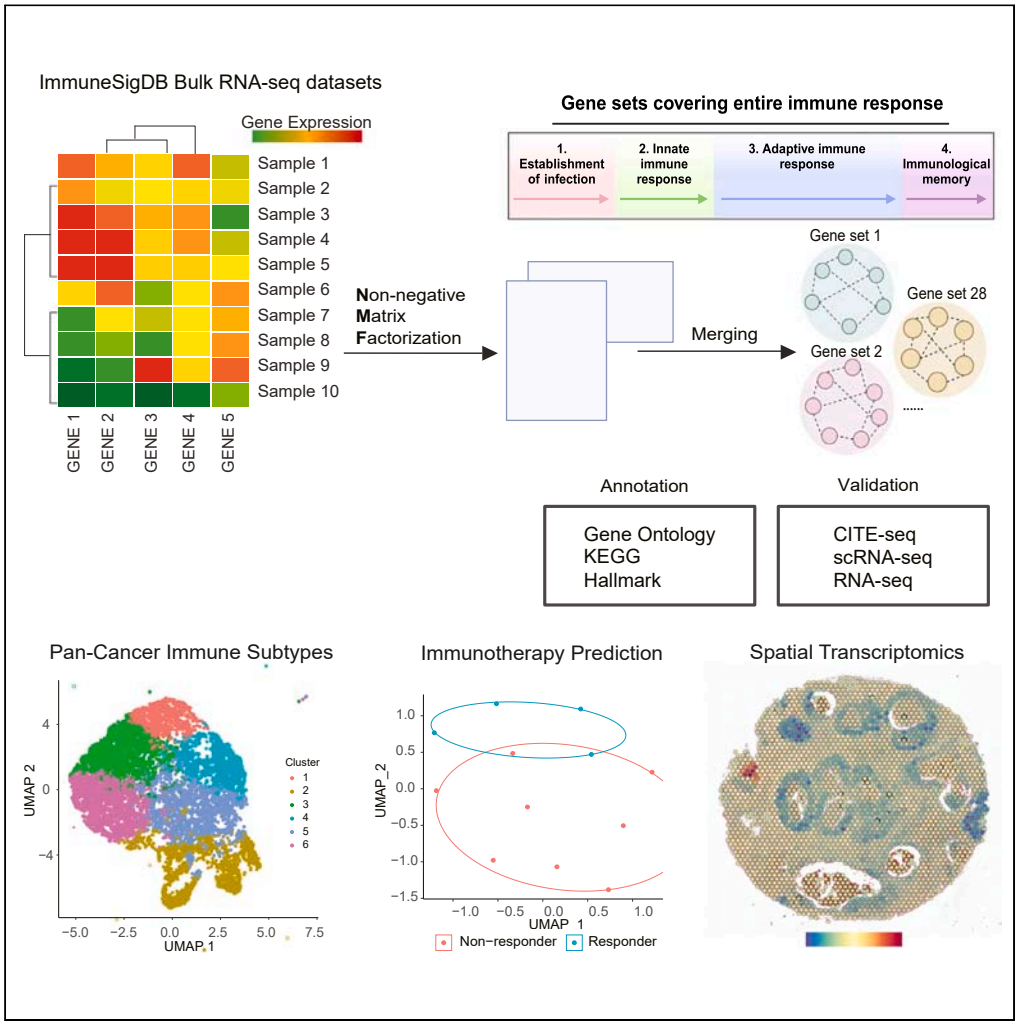
iScience

CellPress
OPEN ACCESS

Article

Elucidating immune-related gene transcriptional programs via factorization of large-scale RNA-profiles



Shan He, Matthew M. Gubin, Hind Rafei, ..., Luisa M. Solis, Vakul Mohanty, Ken Chen

vmohanty@mdanderson.org (V.M.)
kchen3@mdanderson.org (K.C.)

Highlights

Immune-related gene sets (irGS) identified via factorization of RNA-seq database

The irGS refines pan-cancer immune subtypes

The irGS enhances immunotherapy response classification

The irGS improves functional annotation of spatial transcriptomic data

He et al., iScience *27*, 110096
June 21, 2024 Published by Elsevier Inc.
https://doi.org/10.1016/j.isci.2024.110096

iScience

CellPress
OPEN ACCESS

Article

Elucidating immune-related gene transcriptional programs via factorization of large-scale RNA-profiles

Shan He,[1] Matthew M. Gubin,[2] Hind Rafei,[3] Rafet Basar,[3] Merve Dede,[1] Xianli Jiang,[1] Qingnan Liang,[1] Yukun Tan,[1] Kunhee Kim,[1] Maura L. Gillison,[4] Katayoun Rezvani,[3] Weiyi Peng,[5] Cara Haymaker,[6] Sharia Hernandez,[6] Luisa M. Solis,[6] Vakul Mohanty,[1,*] and Ken Chen[1,7,*]

SUMMARY

Recent developments in immunotherapy, including immune checkpoint blockade (ICB) and adoptive cell therapy (ACT), have encountered challenges such as immune-related adverse events and resistance, especially in solid tumors. To advance the field, a deeper understanding of the molecular mechanisms behind treatment responses and resistance is essential. However, the lack of functionally characterized immune-related gene sets has limited data-driven immunological research. To address this gap, we adopted non-negative matrix factorization on 83 human bulk RNA sequencing (RNA-seq) datasets and constructed 28 immune-specific gene sets. After rigorous immunologist-led manual annotations and orthogonal validations across immunological contexts and functional omics data, we demonstrated that these gene sets can be applied to refine pan-cancer immune subtypes, improve ICB response prediction and functionally annotate spatial transcriptomic data. These functional gene sets, informing diverse immune states, will advance our understanding of immunology and cancer research.

INTRODUCTION

Despite recent breakthroughs in immunotherapy such as immune checkpoint blockade (ICB)[1] and adoptive cell therapy (ACT),[2] only a limited fraction of patients benefits from these biomedical advancements. Challenges related to immune-related adverse events and resistance remain and hinder positive outcomes, especially in solid tumors.[3] It is essential for the field of immunotherapy to understand molecular mechanisms underlying the treatment resistance and responses better predict treatment outcomes and develop targeted therapeutics. A systematic way to address these challenges is through statistical analysis of transcriptional programs of human patient samples,[4] which can uncover useful gene expression patterns associated with immunotherapeutic responses. Interpretation of transcriptional programs and their functional significance, however, relies on utilizing previously established gene signature knowledgebases[5–8] such as Molecular Signature Database (MSigDB) Hallmarks (Hallmark),[9] Kyoto Encyclopedia of Genes and Genomes (KEGG),[10] Gene Ontology (GO),[11] etc. Unfortunately, these knowledgebases are largely cell-type agnostic and lack granularity or immune specificity.[7] Such biases severely hamper data interpretation in immunological studies utilizing high-throughput RNA profiling, resulting in many missed opportunities. To further advance cancer immunotherapy, there is an urgent need to discover and expand immune-related gene signatures/sets (irGSs) in current knowledgebases to comprehensively represent diverse immune-specific transcriptional states and to apply them to interpret molecular functions.

A closer inspection of the aforementioned knowledgebases reveals important limitations that are inherent to each source. For example, Hallmark[9] contains only 7 immune-related pathways, with an average of 160 member genes each. The large gene sets in this database can potentially inflate biological significance, without genuine and specific relevance to the designated functions. On the other hand, KEGG[10] contains mostly metabolism related pathways that are not specific to immune systems. The GO[11] terms, with its hierarchical nature, result in collective enrichment of general terms with overlapping functions, diminishing statistical power to detect specific terms. Moreover, although there are some immune-related pathways in the aforementioned knowledgebases, they are too broadly defined. For example, in KEGG, the term T cell receptor (TCR) signaling does not distinguish receptor formation by T cells from antigen presentation by

[1]Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA
[2]Department of Immunology, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA
[3]Department of Stem Cell Transplantation and Cellular Therapy, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA
[4]Department of Thoracic/Head and Neck Medical Oncology, University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA
[5]Department of Biology and Biochemistry, The University of Houston, Houston, TX, USA
[6]Department of Translational Molecular Pathology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA
[7]Lead contact
*Correspondence: vmohanty@mdanderson.org (V.M.), kchen3@mdanderson.org (K.C.)
https://doi.org/10.1016/j.isci.2024.110096

antigen-presenting cells, two complimentary functions in TCR signaling. In Hallmark, "inflammatory response" is too broad to inform detailed molecular programs or mechanisms. Therefore, there is a need to define more specific gene sets and refined annotations to achieve higher granularity.

To address this knowledge gap, some immune specific knowledgebases have been constructed. In particular, ImmuneSigDB,[12] consisting of over 5,000 irGSs, was derived from RNA-expression datasets. These RNA expression datasets are very rich, derived from nearly 400 non-cancer-specific immunological experiments, containing samples challenged with pathogen infections, cytokines or immunological perturbations of different kinds and magnitudes, thus possessing rich immunological states. Some immunological states may be shared by a variety of conditions. For example, the severe immune dysregulation and inflammation,[13,14] observed in sepsis samples may inform immune cell proliferation, inflammatory signaling, and cellular dysfunction and exhaustion in tumor immune microenvironment (TIME).

However, the ImmuneSigDB study, conducted nearly a decade ago, applied empirical approaches that may have resulted in some limitations. First, the study examined sets of genes that are differentially expressed (DEGs) across different comparisons of interest, and these comparison groups were subjectively decided by researchers as important and immune-relevant.[12] Second, differential gene expression analysis tends to output large numbers of DEGs (200 $\pm$ 70, max = 1,992), with 50% of the gene set size ranges from 190 to 196 member genes. Third, the derived irGSs tend to have limited translational implications as they can only be used to differentiate major states (such as cell lineages) instead of more granular substates. Fourth, given that these comparison groups are defined based on empirical knowledge, ImmuneSigDB irGSs are likely incomplete. Lastly, the ImmuneSigDB irGSs are simply annotated using the GEO ID of the dataset and the comparison groups, thus containing limited annotations of immune functions, hampering results interpretation.

In this paper, we aim to construct concise and well-annotated immune-specific gene sets using non-negative matrix factorization (NMF),[15] a powerful factor analysis technique[16–20] that has not yet been applied to ImmuneSigDB data. By decomposing gene expression of samples from ImmuneSigDB into composite programs, followed by addition statistical analysis to consolidate and annotate these programs, we have identified 28 irGSs describing various immunological processes. These irGSs can be applied to study and characterize various immune-related transcriptional states such as those in play in TIMEs. We have systematically validated gene set annotations on 12 different datasets and explored their translational implications on 15 different datasets (see STAR Methods Data Sources). These datasets include various modalities, CITE-seq, BulkRNAseq, scRNA-seq, and spatial transcriptomics, and were profiled from a variety of immunological contexts, including but not limited to cytokine perturbation experiments, infectious disease, and treatment naive or immunotherapy-treated cancer patients. Through these exploratory analyses, we showed that these gene sets can better characterize pan-cancer tumor-immune subtypes, separate ICB response in melanoma, liver, and lung cancer and reveal spatial heterogeneity in ovarian, intestinal, and breast cancer samples.

RESULTS

Twenty eight gene sets covering diverse immunological functions in the TIME

The ImmuneSigDB datasets comprise 83 human bulk gene expression profiles with 1,826 RNA samples from studies published in leading immunology journals (Figure S1A). The datasets included in the ImmuneSigDB were derived from patients affected by severe infections and sepsis or from cell lines challenged with experimental manipulations[12]: gamma delta T cells activated by *IL2*,[21] and *CD14*$^{+}$ monocytes[22] cultured with IFN-gamma and stimulated with Toll-like receptor 2 ligand. Such a large collection of datasets entails transcriptional profiles from diverse cell states, making them suitable to derive generalizable irGSs.

The datasets encompass gene expression data obtained from human and mouse samples (Figure S1B) that originated from samples deriving from different cell types (Figure S1C). Among datasets profiled from *Homo sapiens*, the number of genes and samples profiled varies among datasets (Figures S1D and S1E). We partitioned the 83 ImmuneSigDB human datasets into 47 lymphoid-derived and 36 myeloid-derived datasets.

To obtain unique irGSs, we carried out the analytic workflow, illustrated in Figure 1A and detailed in Figure S1F and STAR Methods, separately for the lymphoid and the myeloid data. We applied NMF on each qualifying dataset (see STAR Methods) and obtained the top 50 genes from each latent factor based on the NMF loadings, producing a total of 529 gene sets. To ensure generalizability, gene sets derived from one dataset were kept only if they overlap at least 20% with at least one gene set derived from another dataset. To ensure non-redundancy, we excluded gene sets with over 20% overlap with gene sets from the same dataset, resulting in 115 gene sets. To further reduce redundancy, we employed the algorithm proposed by Tirosh et al.,[16] whereby gene sets are progressively merged into meta programs (MP) based on Jaccard distances (see STAR Methods).

Using the workflow described previously, we derived 19 immune-related MPs (irMPs) from lymphoid samples (L_MP1-19) and 9 irMPs from myeloid samples (M_MP1-9) (Table 1). To annotate each irMP with immune-relevant functions, we performed over-representation analysis (ORA) using biological processes from GO, KEGG, and Hallmark and sorted the enrichment results first by counts of core enrichment then by ascending adjusted *p* values. Enrichment results and constituent genes for each L_MP and M_MP were presented in Tables S1, S2, S3, and S4. We named the irMPs based on top enriched terms and refined the annotations by going over the constituent genes and the stringDB[23] protein-protein interaction (PPI) network, followed by meticulous scrutiny by multiple immunologists (Matthew Gubin Ph.D.; Hind Rafei M.D.; Rafet Basar M.D.; Katy Rezvani M.D/Ph.D.; and Weiyi Peng M.D/Ph.D.) from MD Anderson Cancer Center. Rationales for each gene set annotation have been documented in supplementary notes.

Overall, the L_MPs are highly specific to immune functions, consisting of pathways such as lymphocyte activation, antigen processing, TCR anchoring, and cytotoxic functions. M_MPs are less diverse, including pathways related to cell migration, adhesion, and antiviral response (Figure 1B; Table 1). A few MPs are related to core cellular processes, such as cell cycle and metabolism.

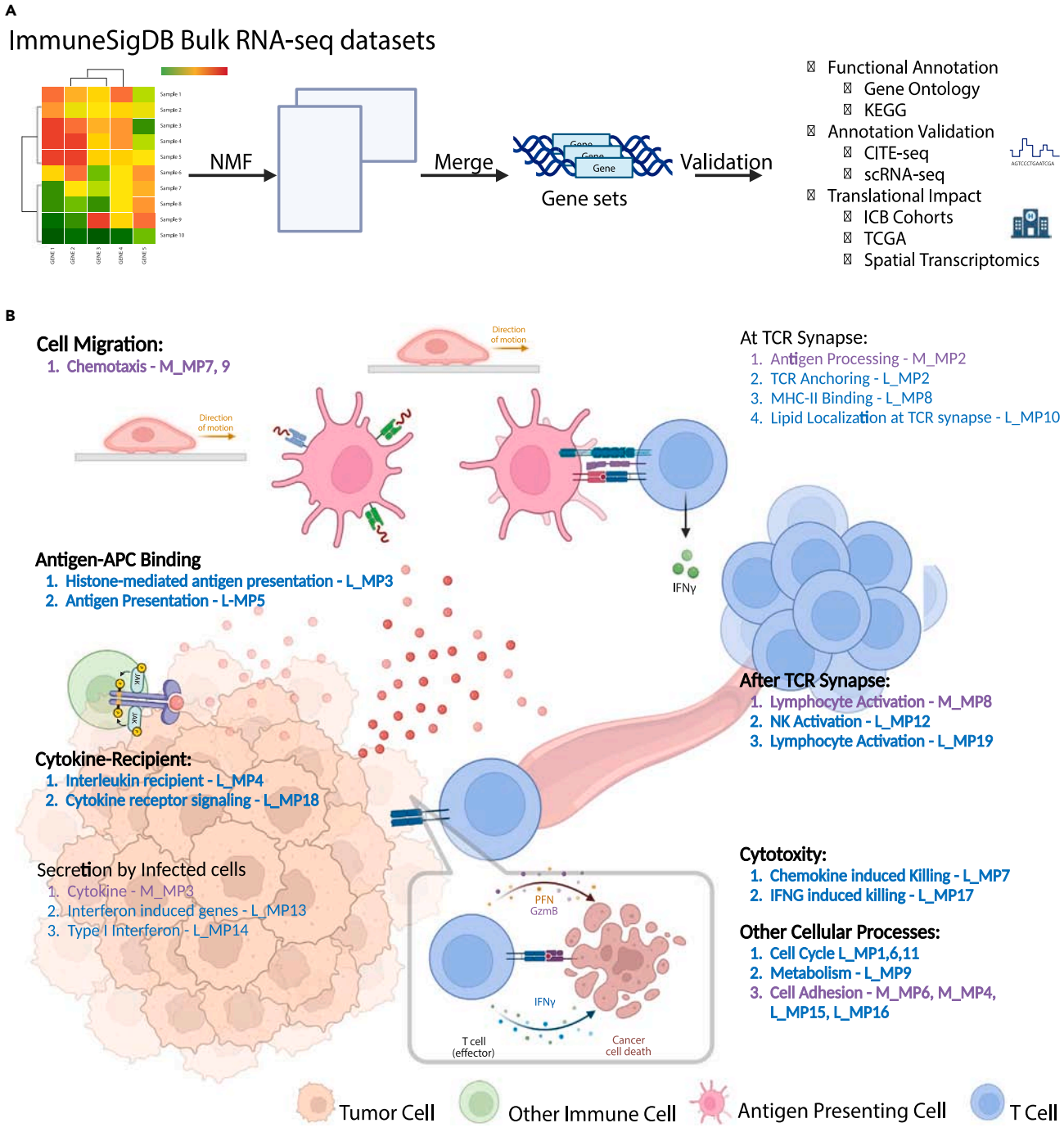


Figure 1. Analytic workflow and constructed gene sets

(A) Gene set construction analysis workflow: we downloaded 83 immunology relevant human studies and performed NMF on each of the qualifying dataset. We curated robust NMF programs and merged them into meta gene sets. We performed over-representation enrichment analysis for each meta gene set using KEGG pathways, Hallmarks, and biological process terms from GO. The name and annotation of each meta gene set was determined based on counts and false discovery rate (FDR) adjusted *p* values of the highly enriched terms. We seek relevant multimodal datasets to functionally name and validate each gene set, and further examined the translational utilities of these gene sets.

(B) The roles of the constructed gene sets in central immunity: constructed gene sets overlaid near the most relevant processes during central immune activities in tumor-immune microenvironment. Created with BioRender.com.

Table 1. Annotations for the 9 myeloid-derived irMPs and 19 lymphoid-derived irMPs

	Myeloid-derived gene sets	Lymphoid-derived gene sets
1	Antiviral defense network	Cell cycle progression
2	Antigen processing	TCR anchoring
3	Cytokine production	Histone-associated lipid antigen presentation
4	Cell adhesion 1	Interleukin-induced Tregs
5	Modulation of cell migration	Antigen presentation
6	Cell adhesion 2	Cell cycle mitosis
7	Eosinophil chemotaxis	Chemokine-mediated T cell activation
8	MHC2-mediated lymphocyte activation	MHC-mediated immunity
9	Chemokine activity	Metabolic process
10		Lipid localization TCR synapse
11		Cell cycle immune response
12		Lymphocyte activation
13		Interferon induced antiviral defense
14		Type-I interferon
15		Cell adhesion 1
16		Cell adhesion 2 ($CD52^{+}$)
17		IFN-gamma-induced cytotoxicity
18		Cytokine receptor signaling
19		T cell activation

To benchmark our irMPs with existing database, we compared and quantified their constituent genes overlap (see STAR Methods) with 231 manually curated Spectra immune gene sets,[24] 5,000 ImmuneSigDB pathways, 50 Hallmark pathways, and 29 immune-related KEGG (irKEGG). We observed significant distinction (false discovery rate [FDR]-adjusted p value < 0.05) in 99.5%, 98%, 99%, and 87% of the comparison with Spectra gene sets (Figure S2), ImmuneSigDB (Figures S3 and S4), Hallmark (Figure S5), and irKEGGs (Figure S6), respectively.

Functional characterization of the irMPs

To validate the functional annotation of irMPs and demonstrate their general applicability, we compiled 12 publicly accessible gene expression datasets encompassing a diverse range of immune contexts, captured through various sequencing techniques (see STAR Methods Data Sources). These datasets were derived from well-designed immunological experiments with well-defined phenotypes, providing independent and orthogonal evidence required to validate annotations of irMPs.

Compared to RNA expression, protein expression is often a more direct indicator of biological activity and reflects real physiological processes. We, therefore, leveraged two CITE-seq datasets,[25,26] in which the whole transcriptome and ~200 surface antibodies were simultaneously profiled in the peripheral blood mononuclear cells (PBMCs) of healthy human and COVID-19 patients, respectively. irMPs were scored for each cell in the RNA level using single sample gene set enrichment analysis (ssGSEA)[27] (see STAR Methods). For each irMP of interest, cells were dichotomized into MP-high or MP-low (see STAR Methods) and the abundance of proteins indicative of the function described by the irMP was compared between MP-high and MP-low groups to show consistency between the irMP and proteomic measurement. For irMP without relevant proteins in the CITE-seq data, we defined highly relevant phenotypes in appropriate scRNA-seq data or The Cancer Genome Atlas (TCGA) pan-cancer data and checked for gene set enrichment in a pre-defined phenotype. In addition, we utilized PPI network constructed by the *StringDB* database[28] to elucidate the activity cascade of each irMP (Figures S7 and S8).

Interestingly, the irMPs we obtained can be broadly classified into categories that align with the sequential progression of a typical immune response: cytokine/chemokine production, cell adhesion, cell chemotaxis, antigen processing, antigen presentation, lymphocyte activation, and cytotoxicity (Figure 1B), indicating the key importance of these functional components and the rich molecular programs that have been under-explored.

In the following, we describe the function of each irMP and evidence from orthogonal validations.

Cytokine/chemokine production

There are two cytokine/chemokine related irMPs from myeloid data (M_MP3 and M_MP9) and two from lymphoid data (L_MP7 and L_MP18). To validate their annotations, we examined scRNA-seq data profiled from two severe COVID-19 patients, with symptoms consistent with inflammatory cytokine storms (ICS).[29] If the irMPs are in fact related to pathogenic cytokine/chemokine activities, they should be enriched in cells deriving from patients rather than health controls. As myeloid cells were concluded to be the source of ICS,[29] we extracted myeloid cells

and performed gene set enrichment analysis (GSEA) on fold-changes of genes expression between ICS and healthy controls. Indeed, we found that all four irMPs were significantly enriched in ICS patients relative to healthy controls (Figure 2A).

Upon further examining the PPI network constructed from the genes in M_MP9 (Figure S7.9), we found that M_MP9 has a central module of chemokine genes (*CXCL10*, *CXCL9*, *CCL4*, *CCL2*, *CCRL2*, *CCL7*, and *CCL8*), connected with antigen presenting cell (APC)-specific human leukocyte antigen (HLA) class II genes (*HLA-DPA1*, *HLA-DMA*, and *HLA-DRA*) by *CD74*, a chaperone responsible for antigen presentation.[30] This further supports the hypothesis that M_MP9 captures the signaling cascade via chemokine secretion leading to the activation of APCs.

Interferon-signaling pathways

There is one antiviral response irMP (M_MP1) from myeloid data and two interferon-mediated viral defense irMPs (L_MP13 and L_MP14) from lymphoid data. To validate these annotations, we used cytokine dictionary curated by Cui et al.,[31] in which the authors have profiled scRNA-seq on 17 different immune cell types extracted from lymph nodes (LNs) of mice treated with 86 different cytokines *in vivo* or phosphate buffered saline (PBS) as control. As these gene sets are associated with interferon signaling, they should be induced upon stimulation by interferon rather than other cytokines. For each cytokine treatment in each cell-type, we performed differential expression relative to corresponding PBS controls followed by GSEA. Across cell-types, we observed that all three irMPs show a general trend of significantly higher enrichment under interferon treatments (IFN-α1, IFN-β, IFN-ε, IFN-κ, IFN-γ, and IFN-λ2) relative to other cytokine treatments (Figure 2B), indicating that these three irMPs are all signaling pathways downstream of interferon stimulation.

Cell adhesion/cell migration

There are two cell adhesion (M_MP4 and M_MP6) and two chemotaxis related (M_MP5 and M_MP7) irMPs from myeloid data, and two cell adhesion irMPs (L_MP15 and L_MP16) from lymphoid data. Cell adhesion and migration are essential steps in the immune response, as they facilitate downstream immune cell trafficking, activation, and effector function.[32] In general, the three steps in cell adhesion are rolling, weak, and strong adhesion.[33] Rolling involves selectin molecules (L-selectin, E-selectin, and P-selectin) loosely moving immune cells on endothelial surface. Weak adhesion occurs when integrin molecules (CD49d, CD29, CD11a, CD11b, and CD11c) weakly bind to ligands (ICAM) on endothelial cells, slowing down the immune cell movement. Strong adhesion occurs when integrin molecules firmly anchor the immune cells to the endothelial surface, allowing them to extravasate through the endothelium and into surrounding tissues. If the irMPs are associated with cell-adhesion and chemotaxis, their activity should be concordant with the expression of relevant protein markers discussed previously.

To validate the functional annotation of these irMPs, we dichotomized cells in the PBMC CITE-seq data[34] by their RNA-based MP activity levels and examined the abundance levels of relevant proteins (see STAR Methods). We found that, among myeloid cells, the activity of M_MP4 and M_MP6 are associated with the abundance of cell adhesion proteins related to both rolling and weak adhesion (Figure S9A). In particular, myeloid cells with higher M_MP4 have the highest CD62L abundance, suggesting that M_MP4 is driven by L-selectin, a cell adhesion molecule on the surfaces of leukocytes. In contrast, the activity levels of M_MP5 and M_MP7 that are associated with cellular migration are associated with lower abundance of cell adhesion marker proteins (Figure S9A), consistent with their contrasting functional annotation. Among lymphoid cells, we found that L_MP15 shows strong association with integrins (CD49d, CD29, CD11a, CD11b, and CD11c), and L_MP16 with selectins (CD62P and CD62L) (Figure S9B). These two irMPs differ from myeloid-derived cell adhesion irMPs in that L_MP15 and L_MP16 are enriched in HLA genes, suggesting possible cell adhesion through antigen presentation.

Core cellular process

Besides highly immune-specific gene sets, there are also gene sets related to core cellular processes. In particular, there are three irMPs (L_MP1, 6, and 11) annotated as cell cycle irMPs from lymphoid data. To validate their functions, we used a COVID-19 single cell gene expression atlas,[35] where author identified a cluster of proliferative $CD8^+$ effector T cells. Since proliferative lymphocytes undergo clonal expansion, we expect cell cycle pathways to be upregulated in this cluster.[28] Indeed, we observed significantly higher gene set activities[26] for all three cell-cycle irMPs in the proliferative lymphocyte cluster (Figure 2C) than in the other clusters.

There is one metabolism-related gene set (L_MP9), enriched in glycolysis genes (*ENO1*, *PGK1*, *ALDOA*, *PGAM1*, *TPI1*, *STMN1*, *LDHA*, *HMMR*, *ENO2*, and *PPIA*). To confirm its relevancy to metabolism, we utilized a study in which authors perform *in vitro* stimulation of naive $CD8^+$ T cells and characterized metabolic programs at different activation stages,[36] showing activation of T cells is accompanied with increased metabolism. We observed that the activity of L_MP9 increases as T cells become more activated, which also coincided with the trend of glycolysis and oxidative phosphorylation (Figure 2D).

Besides the aforementioned gene sets that represent shared functions in both myeloid and lymphoid compartments, we have also discovered irMPs that are highly specific to a cell type or a contact interface between two immune cells.

Leukocyte activation

Two of the lymphoid irMPs (L_MP12 and L_MP19) were annotated as lymphocyte activation and T cell activation, respectively. The activation of these pathways should coincide with high expression of protein markers reflective of lymphocyte activation. As COVID-19 induces immune cell activation, we leveraged the COVID-19 PBMC CITE-seq data.[26] We extracted natural killer (NK) and CD8 T cells based on SingleR[37] annotation (see STAR Methods) and observed that NK/T cells with higher L_MP12 activity have higher abundance of CD16, CD56, KLRG1, CD8, and CD69 (Figure 3A), which are known to be expressed by activated T and NK cells. It is also known that CD8 can be expressed by cytotoxic

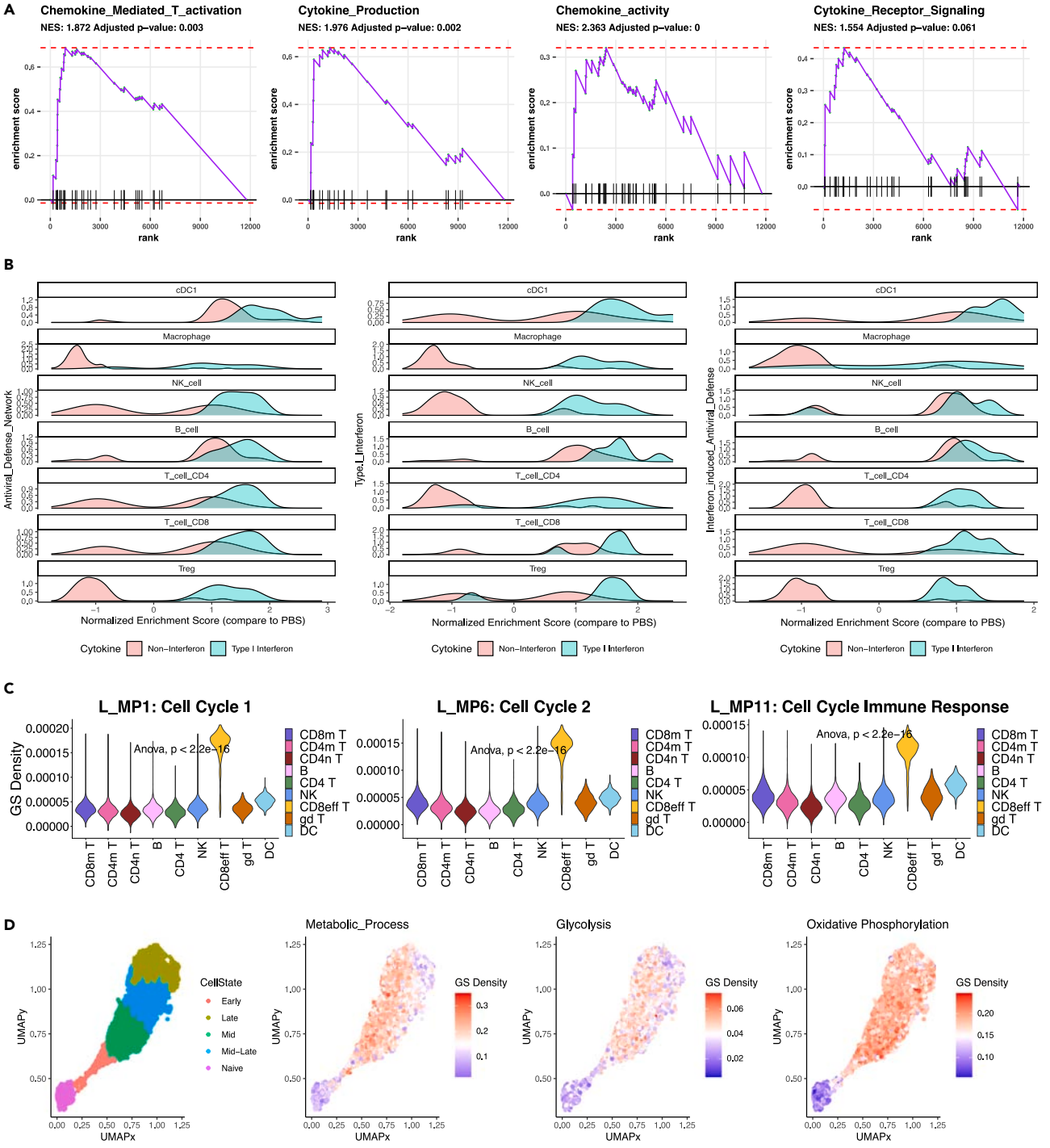


Figure 2. Functional characterization of the generic immune gene sets

(A) Barcode enrichment plot of 4 cytokine/chemokine release irMPs; all genes from the data are ranked along x axis according to enrichment score (y axis). There is a significant upregulation in all 4 irMPs in samples with severe COVID-19 patients with cytokine storm, and the ranked position of each gene within a signature is shown in the x axis.

(B) Density plot comparing the normalized enrichment score of three antiviral irMPs in different immune cell types when stimulated with interferon vs. non-interferon.

(C) Violin plot comparing cell cycle MP activities calculated using GS density across different cell types identified from COVID-19 single cell atlas. ANOVA test was performed for multi-group comparison.

(D) UMAPs labeled with T cell activation stages, L_MP9 activity, glycolysis, and oxidative phosphorylation activities from Hallmark using GS density.

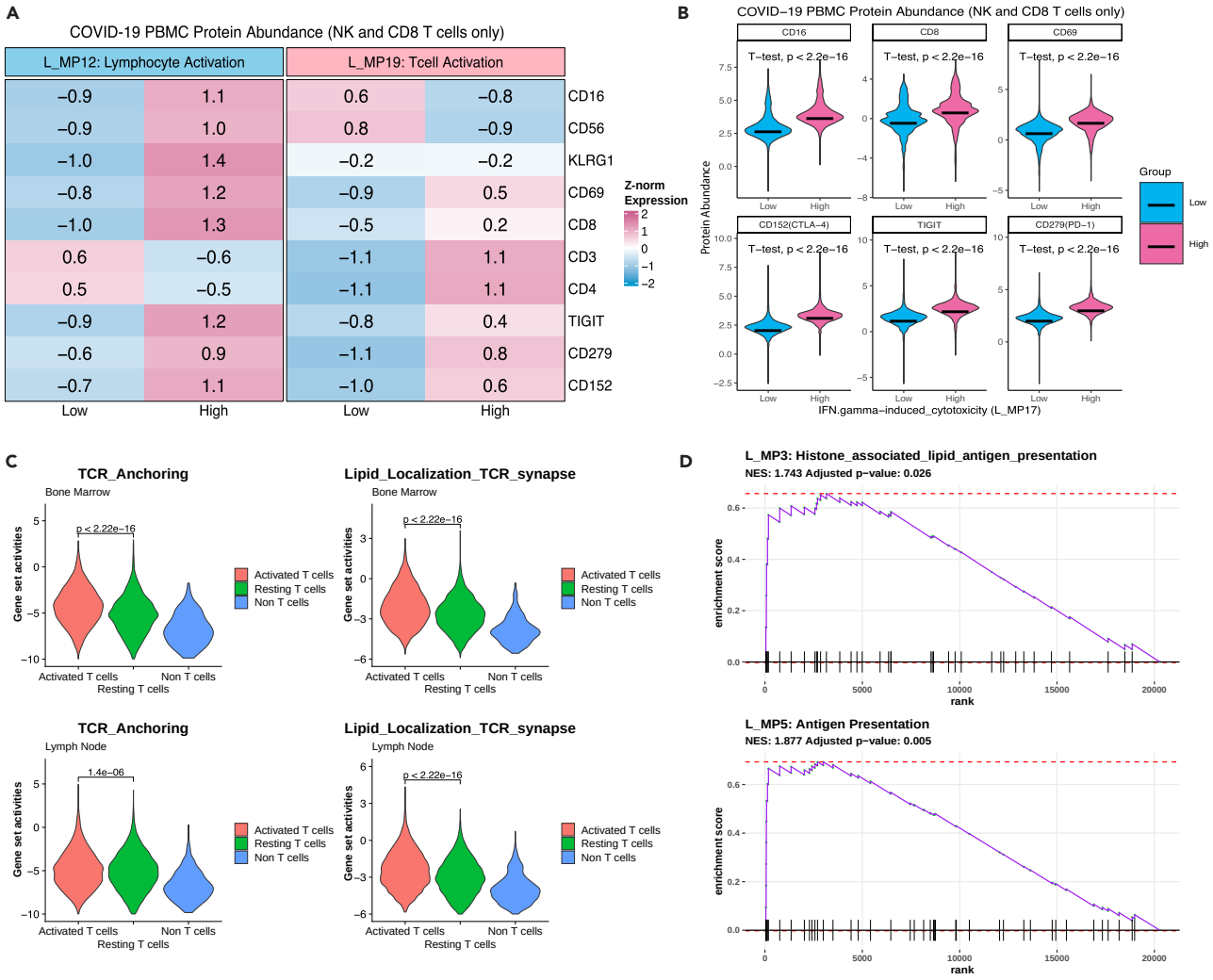


Figure 3. Functional characterization for the specific irMPs

(A) COVID-19 PBMC cells annotated as NK or CD8$^+$ T cells were dichotomized into two groups (high vs. low) based on the activity levels of 2 lymphoid-derived gene sets: L_MP12 and 19, estimated respectively from the RNA expression data. Shown in the heatmap table (and colored accordingly) are the average Z-transformed protein abundance of surface proteins: CD16, CD56, KLRG1, CD8, CD3, CD4, CD69, TIGIT, CD279, and CD152 in each of the respective cell groups.

(B) Shown in violin plots are the abundance of 6 surface proteins: CD16, CD8, CD69, CD152, CD279, TIGIT in two dichotomized COVID-19 PBMC cells (annotated as NK or CD8$^+$ T cells) groups with respectively low and high L_MP17 activity levels, determined from the RNA expression data. In all the cases, dichotomization was performed using Gaussian mixture models. Activities between two groups were compared using t test with significance level 0.05.

(C) Violin plot comparing two TCR-related irMPs across different T cell activation states using ssGSEA. Activities between two groups were compared using t test with significance level of 0.05.

(D) Barcode enrichment plot of L_MP3 and L_MP5; All genes from the TCGA are ranked along x axis according to enrichment score (y axis). There is a significant upregulation in both L_MP3 and L_MP5 in samples with a high persistent tumor mutational burden, and the ranked position of each gene within a signature is shown in the x axis. Statistical significance was determined by a rank-based test.

NK cells.[38] Cells with higher L_MP19 activity have higher abundance of CD3, CD4, CD8, and CD69 (Figure 3A), which are known to be expressed by activated T cells. In addition, L_MP12 and L_MP19 appear to be associated with checkpoint proteins, TIGIT, CD152 (CTLA4), and CD279 (PD-1), which also indicate activation and regulation of immune homeostasis.[39]

Interferon-induced cytotoxicity pathway

One lymphoid irMP (L_MP17) was annotated as IFN-gamma induced cytotoxicity. As COVID-19 induces cytotoxicity, we leveraged the COVID-19 PBMC CITE-seq data.[26] Extracting NK and CD8 T cells based on SingleR annotation, we observed that cells with higher

L_MP17 activity have significantly higher abundance of CD8, CD16, and CD69 (FDR adjusted p value < 0.05), which are protein markers for cytotoxic CD8 T cells and cytotoxic NK cells (Figure 3B). In addition, higher abundance of TIGIT, CD152 (CTLA4), and CD279 (PD-1) in L_MP17 high cells indicate activation, required for achieving cytotoxicity and regulating immune response to avoid over-activation.[39] The PPI network for this MP suggests the following activity cascade (Figure S7.17): Activated NK cells (*NKG7* and *GNLY*) secrete *IFNG*, inducing the activation of granzyme-producing (*GZMH*, *GZMB*, *GZMK*, and *GZMA*) cytotoxic macrophages (*LYZ*). Moreover, downstream targets of leukocyte activation are also captured by L_MP17, specifically *NFKBIA*, *KLF6*, and nuclear receptor genes (*NR4A2* and *NR4A3*), which work harmoniously to ensure T cell homeostasis and proper proliferation.[40,41]

To further validate the cytotoxicity nature of this irMP, we used a COVID-19 single cell atlas data.[35] We observed that lymphoid cells with higher L_MP17 activities correspond to cells with higher cytotoxicity scores (Figure S9C), calculated using GSDensity with the following genes: *PRF1*, *GZMA*, *GZMB*, *GZMH*, and *GNLY*.[35] In addition to cytotoxicity, L_MP17 scored high in the CD8 effector T cell cluster, a cluster defined as highly proliferating T cells in the paper, suggesting that L_MP17 describes not only a cytotoxic but also a proliferative cell state. Indeed, immediately connected to the granzyme modules (Figure S7.17) are *TUBA4A* and *TUBA1C*, which are known to be involved in forming and stabilizing microtubules in proliferating T cells during various stages of cell cycling, contributing to spindle formation during mitosis and ensuring proper chromosome segregation and accurate cell division.[42]

TCRs

Two lymphoid irMPs (L_MP2 and L_MP10) were annotated as TCR-anchoring and lipid localization at TCR synapse, respectively. L_MP2 shows significant enrichment in antigen receptor-mediated signaling pathways (Table S1.2). In addition, L_MP2 is enriched in genes related to T cell surface glycoprotein and transmembrane adapter (*CD247*, *CD2*, *CD28*, *CD3D*, *TRAT1*, *CD52*, and *HLA-E*) and genes that have key roles in T cell antigen receptor-linked signal transduction pathways (*ICOS*, *SLAMF1*, *LCK*, and *LAT*)[43–46] (Figure S7.2). As we expect TCR signaling to be upregulated when T cells are activated, we examined scRNA-seq from T cells isolated from bone marrow (BM) and LNs of healthy donors that were cultured in resting condition vs. activated by anti-CD3/anti-CD28 antibodies.[47] As expected, T cells showed higher L_MP2 and L_MP10 activities, indicating cell-type specificity (Figure 3C). Further, the activities of these gene sets were significantly higher in activated T cells indicating their relevance to TCR engagement upon T cell activation (Figure 3C).

Antigen presentation and processing pathway

Two of the 19 lymphoid irMPs (L_MP3 and L_MP5) are related to antigen-presentation mediating immunity. We named L_MP3 as histone-associated lipid antigen presentation and L_MP5 as antigen presentation. L_MP3 is enriched in chromatin organization histones genes with lipid antigen presentation genes connected by a V(D)J recombination gene RAG1. Histone genes play an important factor in V(D)J recombination and formation of antigen receptors on lymphocytes.[48] The PPI network of L_MP3 (Figure S7.3) shows a HISTONE gene module and a *CD1* gene module connected by *RAG1*. L_MP5 is enriched in genes responsible for cell cycle regulation (*DLGAP5*, *BUB1B*, and *CCNB1*) and genome stability (*TUBB*, *HMGB4*, *HNRNPA2B1*, *PTTG1*, and *HMGB1*), each of which is an important element for V(D)J recombination fidelity[49] (*RAG1* and *RAG2*) (Figure S7.5).

To validate the antigen presentation function of L_MP3 and L_MP5, we used pan-cancer mutational and gene-expression data from TCGA.[50] Persistent tumor mutational burden (pTMB), defined as the number of single copy and multiple copy mutations,[51] informs immune activation. Tumors with higher pTMB burdens are more likely to be visible to the immune system and are associated with sustained anti-tumor immune responses and improved response to ICB.[51] These tumors should therefore have higher activity of antigen presentation pathways. We performed differential gene expression analysis (DGE) between TCGA samples with high vs. low pTMB, regardless of cancer type and stage (see STAR Methods). GSEA with 19 lymphoid irMPs was performed. Both L_MP3 and L_MP5 are significantly upregulated among samples with high pTMB (Figure 3D), confirming the antigen presentation function of these two irMPs. Notably, L_MP3 and L_MP5 are the only two irMPs (out of the 19) that are significantly enriched among high pTMB samples in the full GSEA profiles (FDR adjusted p value < 0.05) (Figure S9D).

pMHC-TCR contact interface

Among myeloid irMPs, M_MP2 was annotated as antigen processing. Figure S10A depicts a typical antigen uptake by an APC, such as, dendritic cell (DC), in which an immature DC (iDC) transforms into a mature DC (mDC) upon detecting pathogen-associated molecular patterns (PAMPs). This activation process involves the DC recognizing PAMPs, leading to upregulation of major histocompatibility complex (MHC) molecules for antigen presentation and the co-stimulatory molecules *CD80*/*CD86*, crucial for T cell activation.[52,53] Simultaneously, there is a change in cytoskeleton organization, notably the F-actin, to facilitate processing and presentation of the peptide-MHC (p-MHC) complexes on their surface for downstream T cell activation.

Concordant with Figure S10A, we observed that APCs (dendritic cells and macrophages, identified by *SingleR*[37]) with higher ssGSEA in M_MP2 also have significantly (p value < 0.05) higher abundance in MHC proteins in the COVID-19 CITE-seq data (Figure S10A), such as HLA-F, HLA-A-B-C, and HLA-DR, suggesting APC activation. In addition, these cells with higher M_MP2 activity also have significantly (p value < 0.05) higher abundance in CD11a/CD18, which facilitate downstream T cell binding, and in CD86, a co-stimulatory molecule for T cells activation.

Among lymphoid irMPs, L_MP8 was annotated as MHC mediated immunity. Figure S10B depicts a typical interaction between an mDC and a T cell at the TCR synapse: Upon recognition of the p-MHC complex by TCR, the T cell upregulates *CD69*, indicating that the T cell has

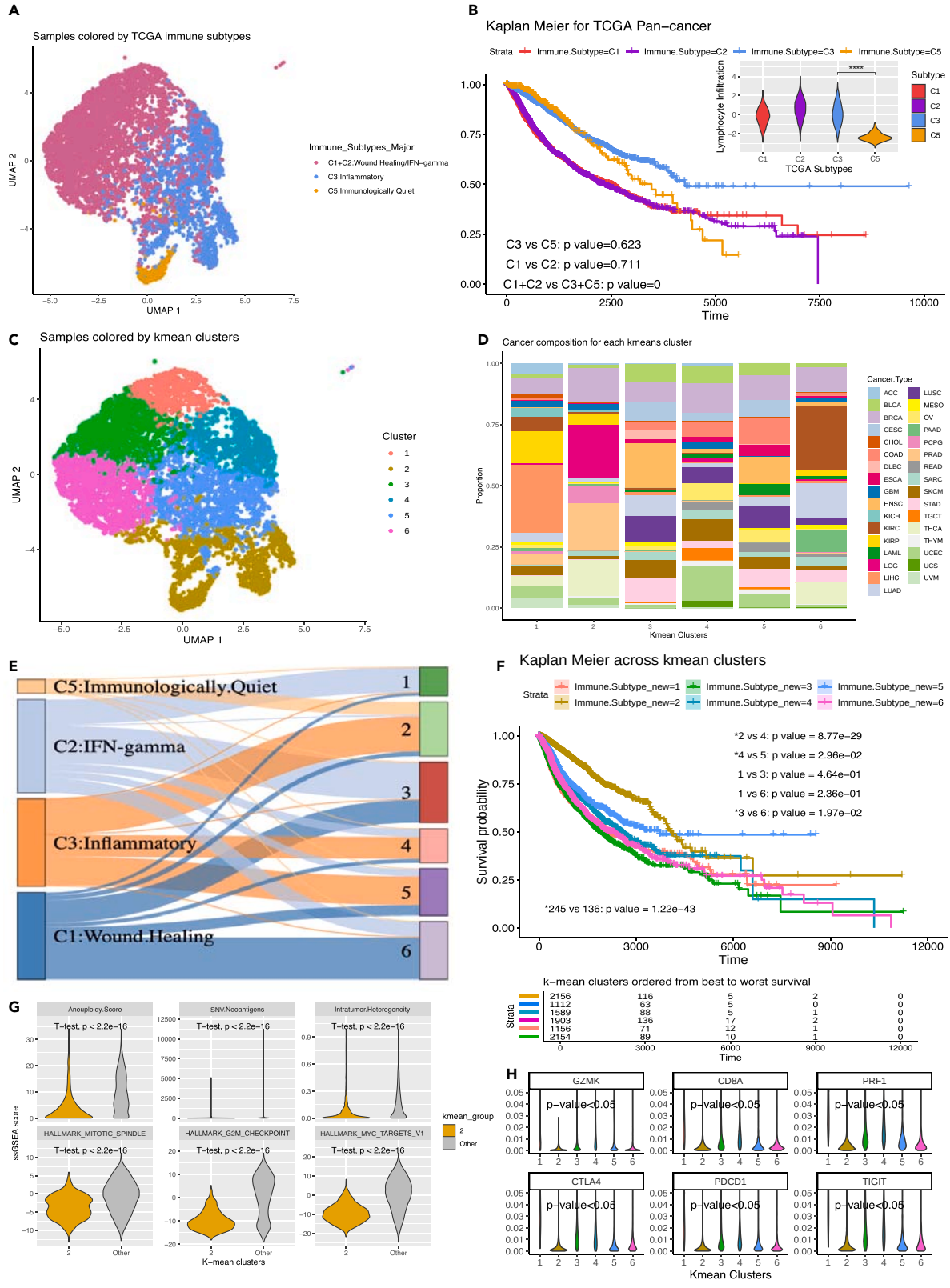

A
Samples colored by TCGA immune subtypes
UMAP 2
UMAP 1
Immune_Subtypes_Major
C1+C2:Wound Healing/IFN−gamma
C3:Inflammatory
C5:Immunologically Quiet
B
Kaplan Meier for TCGA Pan−cancer
Strata
Immune.Subtype=C1
Immune.Subtype=C2
Immune.Subtype=C3
Immune.Subtype=C5
Lymphocyte Infiltration
TCGA Subtypes
Subtype
C3 vs C5: p value=0.623
C1 vs C2: p value=0.711
C1+C2 vs C3+C5: p value=0
Time
C
Samples colored by kmean clusters
UMAP 2
UMAP 1
Cluster
D
Cancer composition for each kmeans cluster
Proportion
Kmean Clusters
Cancer.Type
E
C5:Immunologically.Quiet
C2:IFN-gamma
C3:Inflammatory
C1:Wound.Healing
F
Kaplan Meier across kmean clusters
Strata
Immune.Subtype_new=1
Immune.Subtype_new=2
Immune.Subtype_new=3
Immune.Subtype_new=4
Immune.Subtype_new=5
Immune.Subtype_new=6
*2 vs 4: p value = 8.77e−29
*4 vs 5: p value = 2.96e−02
1 vs 3: p value = 4.64e−01
1 vs 6: p value = 2.36e−01
*3 vs 6: p value = 1.97e−02
*245 vs 136: p value = 1.22e−43
Survival probability
Time
k−mean clusters ordered from best to worst survival
G
Aneuploidy.Score
SNV.Neoantigens
Intratumor.Heterogeneity
HALLMARK_MITOTIC_SPINDLE
HALLMARK_G2M_CHECKPOINT
HALLMARK_MYC_TARGETS_V1
T−test, p < 2.2e−16
ssGSEA score
K−mean clusters
kmean_group
2
Other
H
GZMK
CD8A
PRF1
CTLA4
PDCD1
TIGIT
p−value<0.05
Kmean Clusters

Figure 4. irMPs redefine immune subtypes in TCGA pan-cancer data

(A) UMAP of subsets of TCGA samples (n = 7,768) using single sample gene-set enrichment (ssGSEA) scores calculated from the 28 irMPs. Each dot is a TCGA sample colored by its immune subtypes defined by Thorsson et al.

(B) Kaplan-Meier curves for overall survival between TCGA immune subtypes. Log rank test is performed between comparisons of interests, with corresponding lymphocyte infiltration scores compared on the top right (p value is calculated with one-way ANOVA test).

(C) UMAP of all TCGA samples (n = 10,128) using ssGSEA scores calculated from the 28 irMPs. Each dot is a TCGA sample colored by clusters identified using k-mean clustering.

(D) Stacked plot showing the proportions of cancer types across k-mean clusters.

(E) Sankey River plot showing the shuffling of each TCGA immune subtype into each of the k-mean clusters.

(F) Kaplan-Meier overall survival curves across the different k-mean clusters. Log rank test is performed between pairs of interests (2 vs. 5, 2 vs. 4, 4 vs. 5). Risk table is shown below with clusters ordered from best survival (top) to worst survival (bottom).

(G) Violin plot comparing tumor intrinsic characteristics between cluster 2 and the rest of the clusters by scoring programs with ssGSEA. Activities between two groups were compared using t test with significance level 0.05.

(H) Violin plot comparing the gene expression levels of 3 classical activation and exhaustion markers across the k-mean clusters. Activites across multiple groups were compared using anova with significance level 0.05.

been successfully engaged and activated. Concordant with Figure 3F, we observed that cells with higher M_MP2 activity also have significantly (p value < 0.05) higher abundance in MHC molecules, CD69, and TCRab from COVID-19 CITE-seq data, indicating successful TCR engagement at the immunological synapse.

Among myeloid irMP, M_MP8 was annotated as MHC-II-mediated lymphocyte activation because it is enriched in MHC-II markers (*HLA-DRB4*, *HLA-DQB1*, *HLA-DMA*, and *HLA-DMB*) and TCR complex (*TRAC* and *TRBC1*), suggesting MHC-II dependent CD4 T cell activation (as MHC-II presents to helper T cells).[54] To confirm if this gene set is describing functions at the interface between macrophages and CD4 T cells, we extracted macrophages and CD4 T cells based on SingleR annotation and observed that cells with high activity in M_MP8 have higher protein abundance in M2-like macrophages (CD163 and CD206), T cell activation (CD45RO and TCR), and MHC-II (HLA-DR) (Figure S10C), suggesting CD4 T cell activation via MHC-II on M2-like macrophages.

Interleukin-induced Tregs

There is one lymphoid irMP (L_MP4) annotated as interleukin induced Tregs because it is enriched in Treg activation markers (*CTLA4*, *IL2RA*, *TIGIT*, *TNFRSF9*, and *IL1R2*). Figure S10D depicts one of the mechanisms for Treg survival: binding between *IL2* and *CD25* (*IL2RA*) promotes the growth and suppressive functions of Tregs, upregulating *CD39*[55,56], eliciting a more suppressive Treg phenotype, defined by upregulation of *TNFRSF9* and checkpoint transcripts.[56] Using COVID-19 CITE-seq data, we extracted all Tregs based on *SingleR* annotation and found that Tregs with higher L_MP4 activity do have significantly (p value < 0.05) higher abundance in IL2 receptors (IL2RA and IL2RB), CD39, CTLA4, TIGIT, and TNFRS9 (Figure S10D), which are all protein markers over-expressed in effector Tregs, suggesting possible interleukin-induced Treg activation.

irMPs redefine immune subtypes in TCGA pan-cancer data

In TCGA pan-cancer immunity study, tumor samples were previously categorized into six major immune subtypes (C1: wound healing, C2: IFN-gamma dominant, C3: inflammatory, C4: lymphocyte depleted, C5: immunologically quiet, and C6: TGF-beta dominant), based on 160 immune signatures.[57] We performed Uniform Manifold Approximation and Projection (UMAP) based on ssGSEA[27] scores with the 28 irMPs (Figure S11A). Since C4 and C6 are composed of less than 10% and 2% of TCGA samples, respectively, we did not see a clear separation of these two clusters, but C4 is mostly distributed in the bottom left and C6 is mostly distributed in the top right of the UMAP (Figure S11A). Nevertheless, the UMAP well assigned the tumor samples into C1+C2, C3, or C5 subtypes with significant (log rank p value < 0.0001) differential overall survival between C1+C2 and C3+C5 (Figures 4A and 4B), but the overall survival difference between C3 and C5 is insignificant (p value = 0.623), despite having significantly (p value < 2.2e−16) different levels of lymphocyte infiltration. Through k-mean clustering with k = 6 based on the same UMAP, we re-clustered the same set of TCGA samples into 6 new subtypes (Figure 4C), each of which is enriched in different types of cancer (Figure 4D).

Upon re-clustering, C3 and C5 are resolved into new clusters 2, 4, and 5 (Figure 4E) that well delineate patients by their overall survival (Figure 4F), with cluster 2 having the best overall survival. To further study the molecular characteristics of cluster 2, we assessed selected TCGA signature scores on tumor characteristics[57] (see STAR Methods) across our clusters and found that cluster 2 has significantly (p value < 0.05) lower scores, compared to other k-mean clusters in terms of aneuploidy score, neoantigen loads, intratumor heterogeneity, and hallmark pathways indicative of proliferation (mitotic spindle, G2M checkpoint, and MYC targets V1) (Figure 4G), suggesting that cluster 2 contains more indolent tumors, corroborated with the observed low cell proliferation. In addition, cluster 2 has significantly lower expression in cytotoxic markers (*GZMK*, *PRF1*, and *CD8A*) and exhaustion markers (*CTLA4*, *PDCD1*, and *TIGIT*) (Figure 4H), suggesting less immunogenic environment. Cluster 2 is also enriched in lower grade glioma, prostate adenocarcinoma, and thyroid carcinoma (Figure 4D), which are often lower grade cancers with indolent to slower growth[58] and are less likely to invade nearby tissues or metastasize to distant organs. These tumors generally have better prognosis[59] compared to higher-grade cancers. Therefore, the superior survival of cluster 2 is mainly driven by its simple tumor composition, low proliferation, and low immunogenicity, showing that our irMPs can also discern tumor intrinsic differences via their associated immunological states. Clusters 4 and 5 are both characterized by low tumor proliferation and high tumor development

Table 2. Number of samples contained in each of the melanoma ICB cohort

GEO ID	Disease	Treatment	Pre (*n*)	Post (*n*)	Responder (*n*)	Non-responder (*n*)
91061	Metastatic melanoma	Anti PD-1	51	58	60	49
115821	Metastatic melanoma	Anti PD-1	9	18	1	26
78220	Metastatic melanoma	Anti PD-1	25	0	13	12
145996	Metastatic melanoma	Anti PD-1	14	0	9	5
Total (*n*)			99	76	83	92

(Figure S11B). In term of immune landscape, cluster 4 has significantly (p value < 0.05) higher immune infiltration (Th1 cells and Gamma-Delta T cells) than clusters 2 or 5 and cluster 5 has significantly (p value < 0.05) higher level of activated NK cells compared to clusters 2 or 4 (Figures S11B and S11C).

The k-mean algorithm also redistributed C1 and C2 into clusters 1, 3, and 6 (Figure 4E). Specifically, we identified a transitional subtype between C1 and C2 (cluster 3 in Figure 4C)[30,31] characterized by in-between expression levels of characteristic IFNG-subtype (C2) markers (*IFNG*, *JAK1*, and *STAT1*)[60] and classical wound-healing (C1) markers (*S100A1*, *E2F1*, and *APC*)[61] (Figure S11D). Cluster 3 has significantly (p value < 0.05) worse overall survival compared to clusters 1 and 6, respectively (Figure S11E), and it is characterized by high tumor proliferation, significantly (p value < 0.05) higher tumor glycolysis and oxidative phosphorylation than cluster 6 (Figure S11B). Clusters 1 and 6 are both characterized by high tumor proliferation (Figure S11B). In term of immune landscape, cluster 1 has significantly (p value < 0.05) higher infiltration of CD8 T cells and M1 macrophages (Figure S11C) than the other clusters, suggesting high level of immune activation, which is related to the observed high expression of exhaustion markers[62] (Figure S11F). Cluster 6 is characterized by significantly (p value < 0.05) lower hallmark immune pathway scores than other clusters (Figure S11B).

In summary, our irGSs redefine TCGA subtypes with significantly distinct survival patterns, unveiling not only the intrinsic traits of tumors such as their proliferation and metabolic profiles, but also the dynamic immune activities within the tumor microenvironment. Compared with the immune archetypes defined by Combes et al.,[63] we are defining differential functional states rather than cell type compositions. Figures S11B and S11C contain detailed information on hallmark pathway scores and the relative abundance of immune cell infiltration for each cluster.

irMPs better distinguish ICB response

We investigated if the irMP expression at baseline could be used to differentiate ICB responders and non-responders. We utilized four ICB melanoma cohorts,[64–67] details in Table 2. To account for differences in sequencing depth and coverage, we calculated the fragments per kilobase of transcript per million mapped reads (FPKM) for each of the four RNA-seq data and obtained 104 RNA samples at baseline. We calculated ssGSEA score for each RNA expression sample in the integrated ICB cohort using (1) 50 Hallmark pathways,[9] (2) 29 immune-relevant KEGG (irKEGG) pathways,[10] (3) lymphoid irMPs (L_MPs), (4) myeloid irMPs (M_MPs), or (5) lymphoid-myeloid combined irMPs. We randomly divided 104 baseline samples into 70% training (n_{train} = 73) and 30% testing (n_{test} = 21). To avoid over-fitting, we fitted five generalized linear models (GLM) with LASSO regularization adjusting for the aforementioned gene set variables. Using the best model selected from 10-fold cross-validation, we fitted the model on the test data and evaluated model performance by comparing classification accuracy, which is the average of sensitivity and specificity. To deal with randomness in data splitting, we performed the procedure 1,000 times and computed the range for classification accuracy from each of the five models (see STAR Methods and Figure 5A). Our gene sets resulted in the highest mean accuracy of 0.71, followed by Hallmark (0.69) and irKEGG (0.62), with most of the gene sets being negatively associated with response (Figure S12A). The classification accuracy is comparable (p value = 0.13) between Hallmark and combined irMPs (Figure 5A). However, Hallmark contains 50 gene sets with relatively large (146 $\pm$ 67) gene set sizes, whereas the combined irMPs has only 28 gene sets with 50 genes each, possessing comparable predictive power using fewer parameters. We also compared the classification accuracies with the 18-gene tumor inflammation signature (TIS),[68] which has been approved for use as an investigational use-only (IUO) criteria to stratify patients based on potential to respond to ICB, and the classification accuracy with TIS only on the test data is 52.83%.

To corroborate if our gene sets can better separate ICB response with baseline gene expression only, we extracted the top 10 most selected irMPs across the 1,000 LASSO models (Figure 5B) and calculated their activities at baseline using BulkRNAseq profiled from lung cancer,[69] melanoma,[70] and liver cancer.[71] We again observed that these 10 irMPs can better separate ICB response in comparison to irKEGGs or Hallmarks (Figure 5C). These data indicate that the irMPs can be useful to develop tumor type agnostic general prognostic signatures to predict ICB response.

To understand what these irMPs explain, we calculated the ssGSEA for exhaustion, cytotoxicity, and tissue resident memory (T_{RM}) using their respective marker gene expression (full marker list detailed in STAR Methods). Computing the correlation between irMPs activity and these signature expressions at baseline (Figure S12B), we observed significant (FDR adjusted p value < 0.05) positive correlation between most of the irMPs and exhaustion, cytotoxicity, and T_{RM}, suggesting that at baseline non-responders show phenotype of activation induced exhaustion, a result of prior antigen stimulation. To confirm our observation, we deconvoluted each of the bulk RNA sample into exhausted (T_{EX}), tissue resident memory (T_{RM}), tissue effector memory (T_{EM}), naive (T_N), and effector T cells (T_{EFF}) using CibersortX[72] and pan-cancer T cell

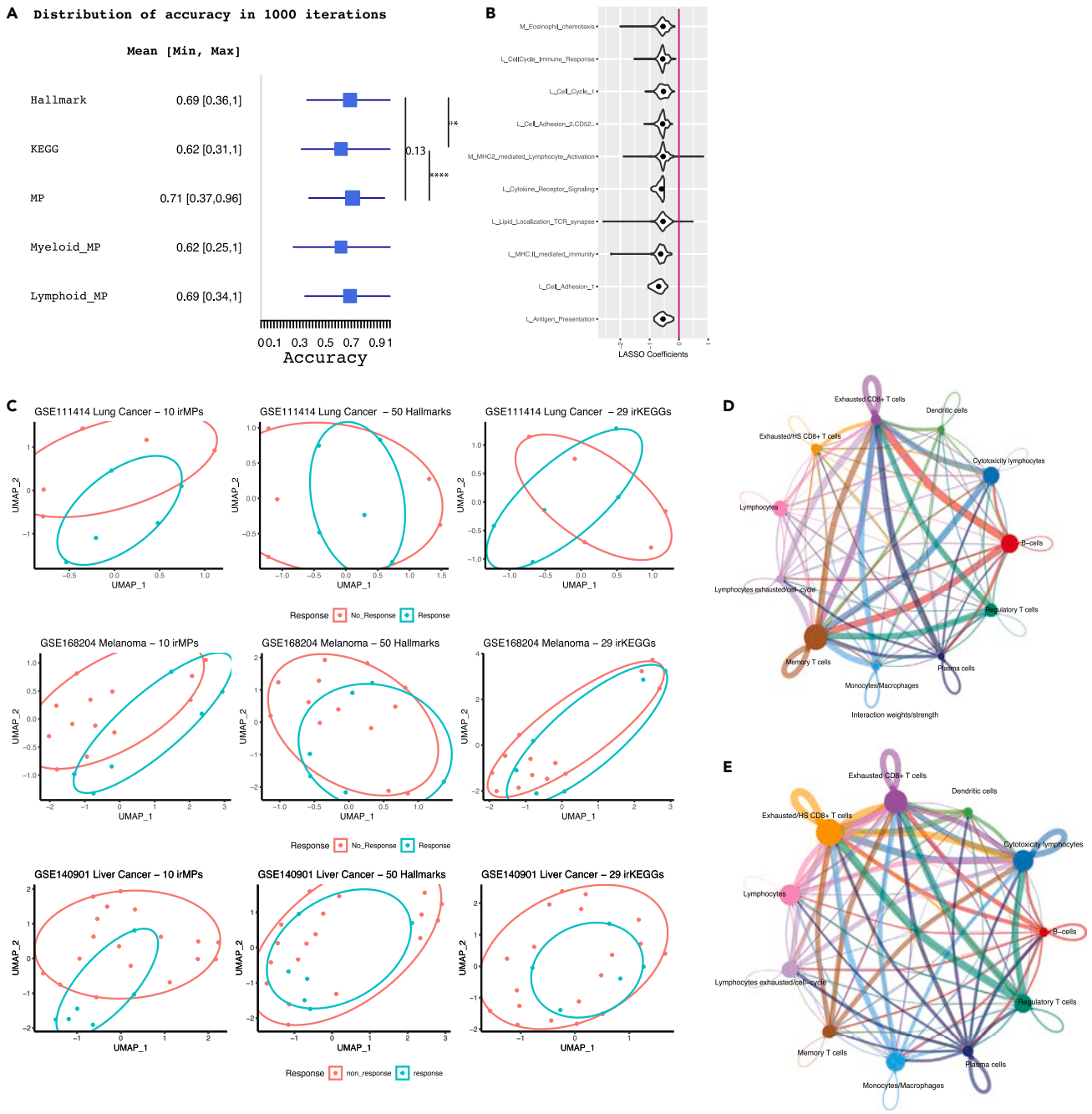


Figure 5. irMPs better separate ICB responses

(A) Distribution of ICB classification accuracy in each of the five model across 1,000 iterations. Distribution of accuracies was compared using t test with significance level 0.05.

(B) The average fitted coefficients for top 10 selected irMPs across 1,000 iterations.

(C) UMAP derived from 10 most selected irMPs activities (left), 50 Hallmarks (middle), and 29 irKEGGs (right) in lung (top), melanoma (middle), and liver (bottom) cancer. Color represents different responses.

(D and E) Baseline cell-cell communications for responders (D) and non-responders (E) inferred from cell chat. Nodes are the 11 cell types annotated in the original single cell data, edges represent significant interactions (adjusted p value < 0.05) and the thickness of the edges represents the probability of interactions.

signature matrix[73] as reference (see STAR Methods). Indeed, non-responders show higher T_{EX} and T_{RM} and lower T_{EM} and T_N at baseline (Figure S12C). However, when we fit the model with only exhaustion, cytotoxicity and T_{RM} score, the classification accuracy is only 54%, suggesting that irMPs have captured addition factors that dictate response to ICB in addition abundance of T cell states.

As cellular communication plays an important role in orchestrating the immune response,[74,75] we hypothesize that irMPs have also captured differential cellular interactions between responders and non-responders. Utilizing an independent scRNA-seq dataset[76] from baseline melanoma patients later treated with anti-PD1, we converted the data into pseudo-bulk expression (see STAR Methods), calculated activity for selected irMPs (top 10 most frequently selected irMPs from 1,000 LASSO regressions in the previous analysis), Hallmarks and irKEGGs pathways, and projected the 12 samples onto a UMAP. Once again, we observed that irMPs more effectively separated out responders and non-responders (Figure S12D) and these 10 irMPs are also associated with exhaustion, cytotoxicity and T_{RM} (Figure S12E). We then performed CellChat[77] analysis (see STAR Methods) to infer cell-cell interactions based on ligand-receptor expression and observed that responders are enriched in interactions surrounding memory T cells, while non-responders are enriched in interactions surrounding exhausted T cells at baseline (Figure 5D). To offer additional evidence of these differential interactions, we used scrublet[78] to identify doublets and double-annotated the doublets (see STAR Methods). Although doublets in droplet-based scRNA-seq data may result from multiple sources, they have been shown to reflect true physical interactions between cell types.[79] Responders were characterized by a limited repertoire of doublet phenotypes with a marked enrichment for doublets with memory T cells. In contrast, non-responders presented with a more varied doublet repertoire with frequent doublets involving various exhausted lymphocytes (Figure S12F). These trends mirror results of the cell-cell communication analysis (Figures 5D and 5E). To confirm whether our irMPs have captured these interactions, we calculated the activities of the 10 important irMPs in each type of doublets at baseline. We observed that the top 10 selected irMPs all have high activities in doublets that are enriched in non-responders at baseline (Figure S13). Taken together, the irMPs may capture interactions between exhausted lymphocytes and other cell types at baseline, which have been associated with poor response in anti-PD1 immunotherapy.[80–82]

irMPs granularly segment spatial transcriptomics data

irMPs can potentially enhance segmentation and phenotyping of spatially resolved transcriptomic (SRT) data. For illustration, we obtained 10× Visium data from an FFPE human breast tissue section.[83] The H&E slide from the tissue section was segmented by pathologists into major cell types (Figure 6A). We calculated the activity scores of the irMPs from the SRT sample and observed that irMPs consistently isolate out the tumor regions (dark blue), aligning with pathologist's annotation (Figures 6B and S14). Besides identifying tumor regions, our irMPs can also be characterized by the infiltrated immunological patterns on the H&E image. We used CellTrek[84] to project single cells from a reference breast cancer patient profiled by scRNA-seq in Tumor Immune Single-cell Hub 2 (TISCH2)[85] onto the spatial spots[84] (see STAR Methods).[83] We then visualized the average irMP activity score for each projected cell type in Figure 6C using a clustered dot plot and observed that clustering of irMPs was largely driven by lineages (3 clusters from top to bottom): myeloid, lymphoid, and cycling. On one hand, some irMPs are strongly lineage specific. For example, the "TCR anchoring" score is prominent among T cells, nearly absent in other cell types (Figure 6C). On the other hand, some irMPs also provide functional insight across the TIME, not limited to a cell type. For example, "IFN-gamma induced cytotoxicity" exhibits the highest score in T cells and a notable score in macrophages (Figure 6C), suggesting a co-enforcement feedback loop between activated T/NK cells and cytotoxic macrophages mediated by potent interferon gamma, a hub gene in the PPI network (Figure S7.17), which aligns with the annotation of this irMP. Benchmarking with immune-related KEGG pathways and Hallmarks, we observed that Hallmark pathways are mostly specific to cancer and endothelial regions (Figure S15A) and irKEGG pathways are myeloid-specific (Figure S15B). To examine the generalizability of our results, we repeated the aforementioned experiments on an FFPE human ovarian tumor tissue and an FFPE human intestine tumor tissue using ovarian cancer patients[86] and colorectal cancer[87] patients as single cell references, respectively. We obtained coherent results (Figure S16). Together, we found the irMPs can be applied to not only segment spatial transcriptomic data aligning with standard pathological segmentation but also reveal granular spot-level cell type information and multicellular processes in TIME.

To further assess how gene sets explain spatial heterogeneity and identify niches with distinct functions, we referred back to Figure 6D and observed that TCR_anchoring shows high activities in regions dominated by B and T cells. We further plotted the activity of TCR_anchoring on intestine, breast, and ovarian cancer Visium data and observed that, in each of the three tumor slides, there is a region with TCR_anchoring activity (Figures 7A–7C, left panel) that is significantly higher than the rest of the spots (Figures 7A–7C, right panel). To further define these regions, a board-certified pathologist (Sharia Hernandez) performed pathological review from respectively matched H&E images and concluded that the identified regions corresponded to a tertiary lymphoid structure (TLS) and two lymphoid aggregates (LAs), respectively (Figures 7A–7C, middle panel). TLS has recently gained attention in cancer research as their presence in the TIME often reflect active and local immune response against the tumor, contributing to favorable prognosis.[88,89,90] LA, on the other hand, is often considered the precursor to TLS.[89] Therefore, the strong association between TCR_anchoring scores and TLS/LA regions demonstrated the potential of applying our irMPs to delineate spatial heterogeneity and identify functionally active regions.

DISCUSSION

Our study aims to address the lack of objectively defined immune-specific gene sets in cancer immunology research. We developed these gene sets based on different studies representing immune cell involvement from bulk samples challenged with infections and immune perturbations. We further utilized CITE-seq and other orthogonal data to validate the immunological functions of the irMPs we identified in our

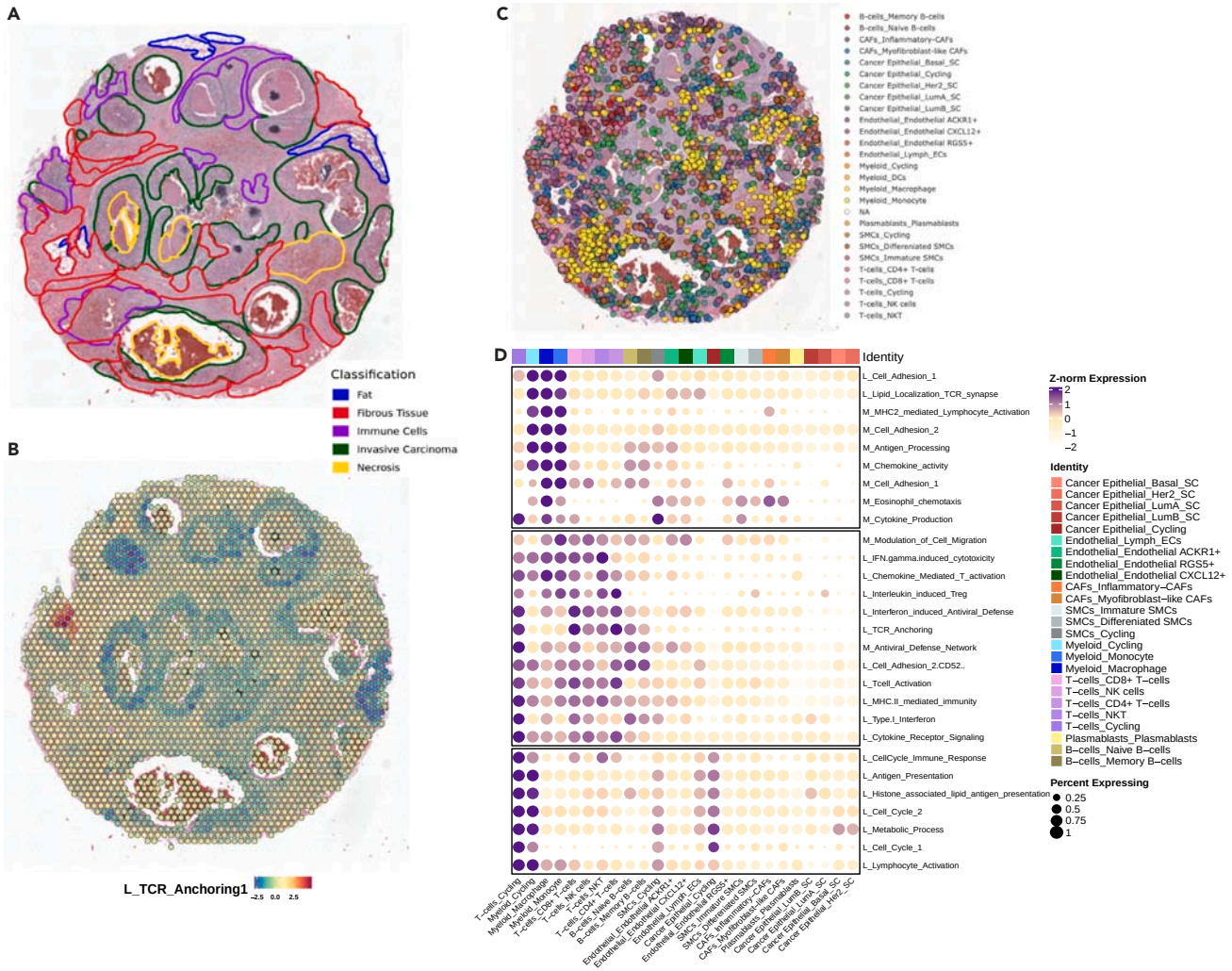


Figure 6. irMPs granularly segment spatial transcriptomics data

(A) Pathologist-annotated 10× Genomics-acquired FFPE human breast tissue spatial image with colored-boundaries separating different cell types.

(B) Module score overlaid on spatial spots for L_MP2 (TCR anchoring).

(C) CellTrek projected single cells onto the breast tumor tissue spatial slide. Color represents different cell types (important compartments are myeloid (yellow), lymphoid (pink/red), and tumor (green).

(D) Dot plot comparing the Z-transformed expression level of 28 irMPs across different celltrek-annotated cell types. Color represents the average expression level, and the dot size represents the percentage of cells expressing the irMP activity.

study. Most importantly, we demonstrated that these transcriptional programs, derived from high-throughput experiments, can provide valuable insights into various aspects of cancer immunological research. Specifically, (1) irMPs improved tumor immune subtype clustering and identified new subtypes with distinct survival outcomes, enhancing our understanding of immune heterogeneity; (2) irMPs demonstrated higher accuracy in separating ICB response, highlighting their potential for predicting ICB outcomes by providing insights into baseline cytotoxicity and cellular communications; and (3) irMPs offered enhanced granularity in delineating tumor-immune boundaries and niches in spatial transcriptomic data. These findings suggest promising translational applications in diverse immunological contexts, as these irMPs are originated from non-cancerous experiments, implying common immunological mechanisms shared by cancer and other pathophysiological conditions.

Our work also shows the importance of cross-lineage coordination and the interconnectivity between immune and tumor compartment. By leveraging BulkRNAseq data, we have discovered transcriptional programs that span across multiple cell lineages. These observations are important as they underpin the intricate web of cellular communications, orchestrating functional phenotypes crucial for mounting effective immune responses. By unraveling these shared transcriptional signatures across diverse cell populations, this research enhances our fundamental understanding of signaling cascades within the immune system. Communication between tumor and immune cells was demonstrated in Figures 4F–4H, when samples grouped based on immune activities also show differences in tumor-specific characteristics. There is a

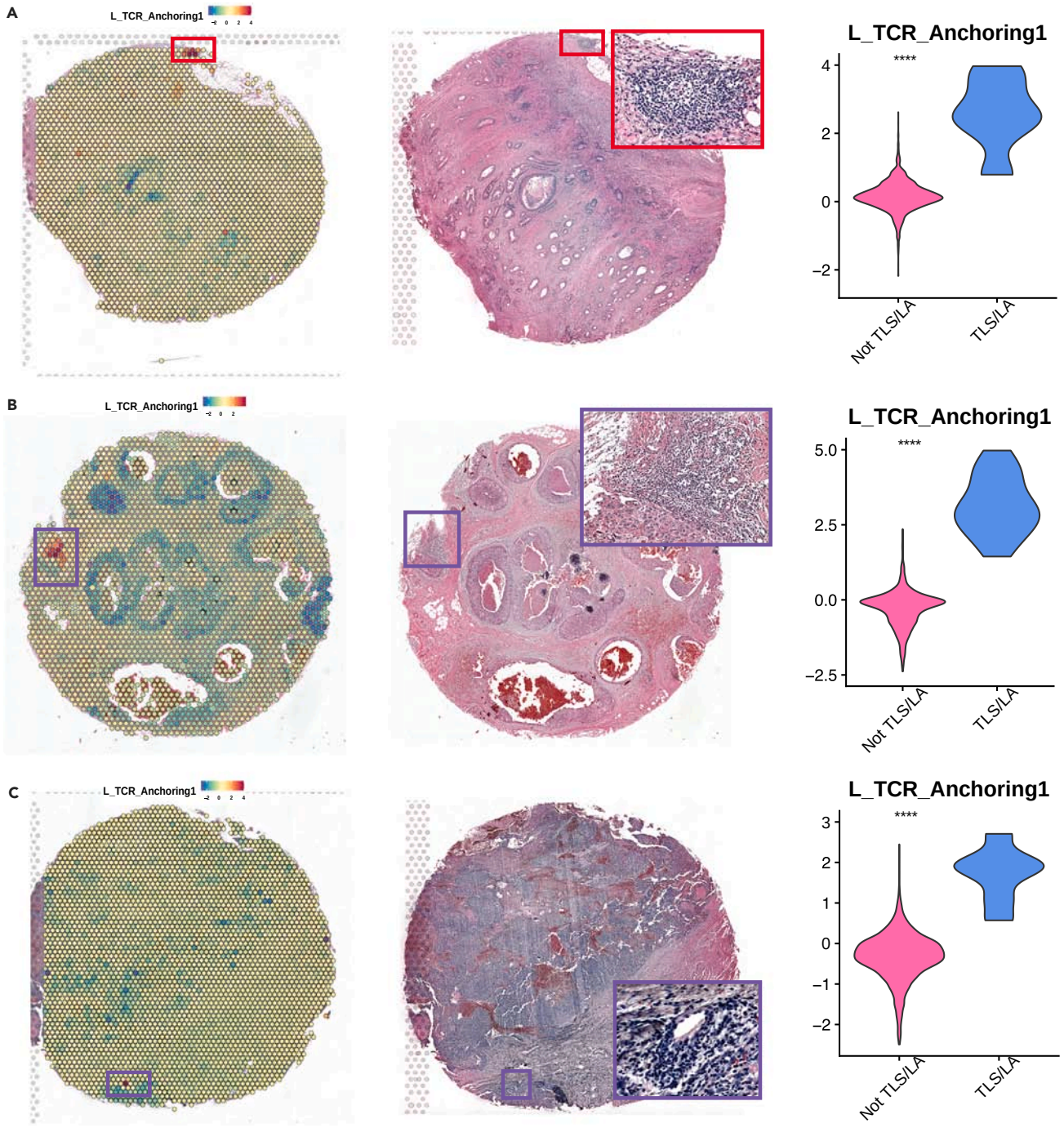


Figure 7. L_MP2 associates with lymphoid aggregates and tertiary lymphoid structure

(A–C) Gene set activity for L_MP2 (left), H&E image with pathologist-confirmed lymphoid aggregates (middle), and statistical test for gene set activities comparing TLS/LA spots vs. other spots (right) for (A) intestine cancer, (B) breast cancer, and (C) ovarian cancer. (Red boxes are confirmed TLS, purple boxes are LA with insufficient evidence to be TLS).

complex bidirectional dialogue between immune cells and cancer cells, thus the activities of irMPs provides valuable insights into the nature of the tumor, its aggressiveness, and even its susceptibility to treatments like immunotherapy. Therefore, studying immune transcriptional programs offers a window into understanding the broader dynamics of tumor biology and devising strategies to harness the immune system's potential in combating cancer.

Our work serves as a parallel effort to the cross-tissue immune atlas by Teichmann group, in which they leveraged scRNA-seq dataset and developed CellTypist[91] to classify immune cells into 101 distinct populations with context-dependent functional states across various human healthy tissues. In a similar vein, our work leveraged BulkRNAseq and concentrate on elucidating immune activities in diseased samples. Future studies are warranted to (1) analyze the immune landscape differences between healthy and diseased microenvironment and (2) leverage scRNA-seq to derive cell type specific transcriptional programs.

Algorithms designed to enhance our comprehension of molecular-level functional states are crucial, offering abundant information unattainable through individual laboratory experiments. By integrating extensive data sources, we often capture a broader spectrum of functional states under varying conditions, offering insights into numerous biological inquiries. For instance, recent research[92] linking metabolic fitness to the resistance of chimeric antigen receptor-engineered natural killer (CAR-NK) cells underscores the importance of comprehending cellular characteristics in deciphering therapeutic resistance. In this study, we demonstrated the NMF can be used to efficiently output functional gene sets with translational implications. Overall, our study constructed 28 immune-specific gene sets, taking a significant step toward advancing our comprehension of TIME. This knowledge can empower cancer immunologists to gain deeper insights into ICB response and cancer survival, potentially expediting the development of immunotherapies and biomarkers, and ultimately improving patient outcomes.

Limitations of the study

In this study, we constructed gene sets using extensive bulk RNA data, offering valuable insights into the molecular landscape. We divided the data by lymphoid and myeloid lineages. While that may have enhanced the discovery of lineage-specific irGSs, it may bias against discovery of irGSs shared between the lineages. Another inherent limitation in our work, as well as studies of similar nature,[9,12,16,93,94] is the challenge of annotating the functional roles of each gene sets. To address this limitation and enhance the reliability of our findings, we emphasize the importance of employing various functional assays for validation. One promising avenue is the utilization of targeted perturb-seq experiments, which can provide experimental evidence to corroborate our gene set annotations. As a matter of fact, there are already *in silico* experiments to predict transcriptional profiles from combinatorial perturbations,[95] and such predictions, if successfully validated in wet lab, can further bolster our functional annotations. Another promising avenue is to apply these gene sets on scRNA data derived in different contexts, thereby enhancing annotations with higher granularity and context-dependent information. By applying gene sets to single cell data, we can precisely assess their functional relevance in distinct cellular contexts and unveil insights into the regulation of biological processes at the single-cell level.

STAR★METHODS

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

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - Bulk RNA-seq data curation
  - "Data-driven" statistical analysis
  - Functional annotation
  - Protein-protein interaction network for each MP
  - Benchmarking
  - Cell type annotation for CITE-seq
  - Cell dichotomization in CITE-seq
  - Gene set scoring
  - Antigen-related pathways validation
  - Cytotoxicity pathway validation
  - TCGA immune subtypes refinement
  - ICB response model
  - Marker genes for immune signatures
  - Pseudo-bulk aggregation
  - Cell type deconvolution
  - Cell chat analysis
  - Doublet detection and annotation
  - Single cell level annotation on spatial transcriptomic
  - scRNAseq data processing
  - Differential gene expression and pathway analysis
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2024.110096.

ACKNOWLEDGMENTS

This project has been made possible in part by grant U01CA247760 and U01CA281902 to K.C. and the Cancer Center Support Grant P30 CA016672 from National Cancer Institute. This project was also partially supported by the MD Anderson Moonshot programs; M.M.G. is a Cancer Prevention and Research Institute of Texas (CPRIT) Scholar in Cancer Research and is supported by CPRIT (Recruitment of First-Time Tenure-Track Faculty Members; RR190017). We also thank Dr. Chloé Villani for their helpful comments.

AUTHOR CONTRIBUTIONS

Writing – original draft: S. He. Conceptualization: S. He, V.M., K.C., and K.R. Methodology: S. He, V.M., H.R., and R.B. Visualization: S. He. Writing – review & editing: S. He, V.M., K.C., M.D., C.H., M.M.G., and W.P. Suggestions: X.J., Q.L., C.H., Y.T., K.K., C.H., K.R., S. Hernandez, L.M.S., and M.L.G. Supervision: V.M and K.C.

DECLARATION OF INTERESTS

H.R. and The University of Texas MD Anderson Cancer Center have an institutional financial conflict of interest with Takeda Pharmaceutical; K.R. and The University of Texas MD Anderson Cancer Center have an institutional financial conflict of interest with Takeda Pharmaceutical and Affimed GmbH. K.R. participates on the Scientific Advisory Board for GemoAb, AvengeBio, Virogin Biotech, GSK, Bayer, Navan Technologies, Caribou Biosciences, BitBio Limited and Innate Pharma. K.R. is the scientific founder of Syena.

Received: December 14, 2023
Revised: April 3, 2024
Accepted: May 21, 2024
Published: May 23, 2024

REFERENCES

1. Dine, J., Gordon, R., Shames, Y., Kasler, M.K., and Barton-Burke, M. (2017). Immune checkpoint inhibitors: An innovation in immunotherapy for the treatment and management of patients with cancer. Asia Pac. J. Oncol. Nurs. *4*, 127–135.
2. Kim, S.P., Vale, N.R., Zacharakis, N., Krishna, S., Yu, Z., Gasmi, B., Gartner, J.J., Sindiri, S., Malekzadeh, P., Deniger, D.C., et al. (2022). Adoptive Cellular Therapy with Autologous Tumor-Infiltrating Lymphocytes and T-cell Receptor- Engineered T Cells Targeting Common p53 Neoantigens in Human Solid Tumors. Cancer Immunol. Res. *10*, 932–946.
3. Koustas, E., Sarantis, P., Papavassiliou, A.G., and Karamouzis, M.V. (2020). The resistance mechanisms of checkpoint inhibitors in solid tumors. Biomolecules *10*, 666. https://doi.org/10.3390/biom10050666.
4. Breschi, A., Muñoz-Aguirre, M., Wucher, V., Davis, C.A., Garrido-Martín, D., Djebali, S., Gillis, J., Pervouchine, D.D., Vlasova, A., Dobin, A., et al. (2020). A limited set of transcriptional programs define major cell types. Genome Res. *30*, 1047–1059.
5. Dai, Y., Hu, R., Liu, A., Cho, K.S., Manuel, A.M., Li, X., Dong, X., Jia, P., and Zhao, Z. (2022). WebCSEA: Web-based cell-type-specific enrichment analysis of genes. Nucleic Acids Res. *50*, W782–W790.
6. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., and Mesirov, J.P. (2005). Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles. Proc. Natl. Acad. Sci. USA *102*, 15545–15550. https://doi.org/10.1073/pnas.0506580102.
7. Ben-Ari Fuchs, S., Lieder, I., Stelzer, G., Mazor, Y., Buzhor, E., Kaplan, S., Bogoch, Y., Plaschkes, I., Shitrit, A., Rappaport, N., et al. (2016). GeneAnalytics: An Integrative Gene Set Analysis Tool for Next Generation Sequencing, RNAseq and Microarray Data. OMICS *20*, 139–151.
8. Mathur, R., Rotroff, D., Ma, J., Shojaie, A., and Motsinger-Reif, A. (2018). Gene set analysis methods: A systematic comparison. BioData Min. *11*, 8.
9. Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J.P., and Tamayo, P. (2015). The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst. *1*, 417–425.
10. Kanehisa, M., and Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. *28*, 27–30. http://www.genome.ad.jp/kegg/.
11. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. *25*, 25–29. https://doi.org/10.1038/75556.
12. Godec, J., Tan, Y., Liberzon, A., Tamayo, P., Bhattacharya, S., Butte, A.J., Mesirov, J.P., and Haining, W.N. (2016). Compendium of Immune Signatures Identifies Conserved and Species-Specific Biology in Response to Inflammation. Immunity *44*, 194–206.
13. Gudiol, C., Albasanz-Puig, A., Cuervo, G., and Carratalà, J. (2021). Understanding and Managing Sepsis in Patients With Cancer in the Era of Antimicrobial Resistance. Front. Med. *8*, 636547. https://doi.org/10.3389/fmed.2021.636547.
14. Tripathi, H., Mukhopadhyay, S., and Mohapatra, S.K. (2020). Sepsis-associated pathways segregate cancer groups. BMC Cancer *20*, 309.
15. Li, Y., and Ngom, A. (2013). The Non-Negative Matrix Factorization Toolbox for Biological Data Mining. http://www.scfbm.org/content/8/1/10.
16. Gavish, A., Tyler, M., Simkin, D., Kovarsky, D., Gonzalez Castro, L.N., Halder, D., Chanoch-Myers, R., Laffy, J., Mints, M., Greenwald, A.R., et al. (2021). The Transcriptional Hallmarks of Intra-tumor Heterogeneity across a Thousand Tumors. https://doi.org/10.1101/2021.12.19.473368.
17. Boccarelli, A., Del Buono, N., and Esposito, F. (2022). Colorectal cancer in Crohn's disease evaluated with genes belonging to fibroblasts of the intestinal mucosa selected by NMF. Pathol. Res. Pract. *229*, 153728.
18. Khan, S.M., Das, T., Chakraborty, S., Choudhury, A.M.A.R., Karim, H.F., Mostofa, M.A., Ahmed, H.U., Hossain, M.A., Le Calvez-Kelm, F., Hosen, M.I., and Shekhar, H.U. (2023). A transcriptome study of p53-pathway related prognostic gene signature set in bladder cancer. Heliyon *9*, e21058.
19. Boccarelli, A., Del Buono, N., and Esposito, F. (2023). Cluster of resistance-inducing genes in MCF-7 cells by estrogen, insulin, methotrexate and tamoxifen extracted via NMF. Pathol. Res. Pract. *242*, 154347.
20. Kim, D., and Cho, K.H. (2023). Hidden patterns of gene expression provide prognostic insight for colorectal cancer. Cancer Gene Ther. *30*, 11–21.
21. Pont, F., Familiades, J., Déjean, S., Fruchon, S., Cendron, D., Poupot, M., Poupot, R., L'faqihi-Olive, F., Prade, N., Ycart, B., and

Fournié, J.J. (2012). The gene expression profile of phosphoantigen-specific human γδ T lymphocytes is a blend of αβ T-cell and NK-cell signatures. Eur. J. Immunol. *42*, 228–240.

22. Hu, X., Chung, A.Y., Wu, I., Foldi, J., Chen, J., Ji, J.D., Tateya, T., Kang, Y.J., Han, J., Gessler, M., et al. (2008). Integrated Regulation of Toll-like Receptor Responses by Notch and Interferon-γ Pathways. Immunity *29*, 691–703.
23. Szklarczyk, D., Kirsch, R., Koutrouli, M., Nastou, K., Mehryary, F., Hachilif, R., Gable, A.L., Fang, T., Doncheva, N.T., Pyysalo, S., et al. (2023). The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. Nucleic Acids Res. *51*, D638–D646.
24. Kunes, R.Z., Walle, T., Land, M., Nawy, T., and Pe'er, D. (2023). Supervised discovery of interpretable gene programs from single-cell data. Nat Biotechnol. https://doi.org/10.1038/s41587-023-01940-3.
25. Nettersheim, F.S., Armstrong, S.S., Durant, C., Blanco-Dominguez, R., Roy, P., Orecchioni, M., Suryawanshi, V., and Ley, K. (2022). Titration of 124 antibodies using CITE-Seq on human PBMCs. Sci. Rep. *12*, 20817.
26. Unterman, A., Sumida, T.S., Nouri, N., Yan, X., Zhao, A.Y., Gasque, V., Schupp, J.C., Asashima, H., Liu, Y., Cosme, C., Jr., et al. (2022). Single-cell multi-omics reveals dyssynchrony of the innate and adaptive immune system in progressive COVID-19. Nat. Commun. *13*, 440.
27. Foroutan, M., Bhuva, D.D., Lyu, R., Horan, K., Cursons, J., and Davis, M.J. (2018). Single sample scoring of molecular phenotypes. BMC Bioinf. *19*, 404.
28. Szklarczyk, D., Gable, A.L., Nastou, K.C., Lyon, D., Kirsch, R., Pyysalo, S., Doncheva, N.T., Legeay, M., Fang, T., Bork, P., et al. (2021). The STRING database in 2021: Customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucleic Acids Res. *49*, D605–D612.
29. Guo, C., Li, B., Ma, H., Wang, X., Cai, P., Yu, Q., Zhu, L., Jin, L., Jiang, C., Fang, J., et al. (2020). Single-cell analysis of two severe COVID-19 patients reveals a monocyte-associated and tocilizumab-responding cytokine storm. Nat. Commun. *11*, 3924.
30. Fukuda, Y., Bustos, M.A., Cho, S.N., Roszik, J., Ryu, S., Lopez, V.M., Burks, J.K., Lee, J.E., Grimm, E.A., Hoon, D.S.B., and Ekmekcioglu, S. (2022). Interplay between soluble CD74 and macrophage-migration inhibitory factor drives tumor growth and influences patient survival in melanoma. Cell Death Dis. *13*, 117.
31. Cui, A., Huang, T., Li, S., Ma, A., Pérez, J.L., Sander, C., Keskin, D.B., Wu, C.J., Fraenkel, E., and Hacohen, N. (2024). Dictionary of immune responses to cytokines at single-cell resolution. Nature *625*, 377–384. https://doi.org/10.1038/s41586-023-06816-9.
32. Kinashi, T., and Katagiri, K. (2005). Regulation of immune cell adhesion and migration by regulator of adhesion and cell polarization enriched in lymphoid tissues. Immunology *116*, 164–171. https://doi.org/10.1111/j.1365-2567.2005.02214.x.
33. Dustin, M.L. (2019). Integrins and Their Role in Immune Cell Adhesion. Cell *177*, 499–501. https://doi.org/10.1016/j.cell.2019.03.038.
34. Nettersheim, F.S., Armstrong, S.S., Durant, C., Blanco-Dominguez, R., Roy, P., Orecchioni, M., Suryawanshi, V., and Ley, K. (2022). Titration of 124 antibodies using CITE-Seq on human PBMCs. Sci. Rep. *12*, 20817.
35. Wilk, A.J., Rustagi, A., Zhao, N.Q., Roque, J., Martínez-Colón, G.J., McKechnie, J.L., Ivison, G.T., Ranganath, T., Vergara, R., Hollis, T., et al. (2020). A single-cell atlas of the peripheral immune response in patients with severe COVID-19. Nat. Med. *26*, 1070–1076.
36. Fernández-García, J., Franco, F., Parik, S., Altea-Manzano, P., Alejandro Pane, A., Broekaert, D., van Elsen, J., Di Conza, G., Vermeire, I., Schalley, T., et al. (2022). CD8+ T cell metabolic rewiring defined by scRNA-seq identifies a critical role of ASNS expression dynamics in T cell differentiation. Cell Rep. *41*, 111639.
37. Aran, D., Looney, A.P., Liu, L., Wu, E., Fong, V., Hsu, A., Chak, S., Naikawadi, R.P., Wolters, P.J., Abate, A.R., et al. (2019). Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. Nat. Immunol. *20*, 163–172.
38. Addison, E.G., North, J., Bakhsh, I., Marden, C., Haq, S., Al-Sarraj, S., Malayeri, R., Wickremasinghe, R.G., Davies, J.K., and Lowdell, M.W. (2005). Ligation of CD8alpha on human natural killer cells prevents activation-induced apoptosis and enhances cytolytic activity. Immunology *116*, 354–361.
39. Pardoll, D.M. (2012). The blockade of immune checkpoints in cancer immunotherapy. Nat. Rev. Cancer *12*, 252–264. https://doi.org/10.1038/nrc3239.
40. Hiwa, R., Nielsen, H.V., Mueller, J.L., Mandla, R., and Zikherman, J. (2021). NR4A family members regulate T cell tolerance to preserve immune homeostasis and suppress autoimmunity. JCI Insight *6*, e151005. https://doi.org/10.1172/jci.insight.151005.
41. Silv, A., Cornish, G., Ley, S.C., and Seddon, B. (2014). NF-κB signaling mediates homeostatic maturation of new T cells. Proc. Natl. Acad. Sci. USA *111*, E846–E855.
42. Bunning, A.R., and Gupta, M.L. (2023). The importance of microtubule-dependent tension in accurate chromosome segregation. Front. Cell Dev. Biol. *11*, 1096333. https://doi.org/10.3389/fcell.2023.1096333.
43. Rujas, E., Cui, H., Sicard, T., Semesi, A., and Julien, J.P. (2020). Structural characterization of the ICOS/ICOS-L immune complex reveals high molecular mimicry by therapeutic antibodies. Nat. Commun. *11*, 5066.
44. Cantrell, D.A. (2002). T-cell antigen receptor signal transduction. Immunology *105*, 369–374. https://doi.org/10.1046/j.1365-2567.2002.01391.x.
45. Orentas, R.J., Nordlund, J., He, J., Sindiri, S., Mackall, C., Fry, T.J., and Khan, J. (2014). Bioinformatic description of immunotherapy targets for pediatric T-cell leukemia and the impact of normal gene sets used for comparison. Front. Oncol. *4*, 134.
46. Gartshteyn, Y., Askanase, A.D., and Mor, A. (2021). SLAM Associated Protein Signaling in T Cells: Tilting the Balance Toward Autoimmunity. Front. Immunol. *12*, 654839. https://doi.org/10.3389/fimmu.2021.654839.
47. Szabo, P.A., Levitin, H.M., Miron, M., Snyder, M.E., Senda, T., Yuan, J., Cheng, Y.L., Bush, E.C., Dogra, P., Thapa, P., et al. (2019). Single-cell transcriptomics of human T cells reveals tissue and activation signatures in health and disease. Nat. Commun. *10*, 4706.
48. Shimazaki, N., and Lieber, M.R. (2014). Histone methylation and V(D)J recombination. Int. J. Hematol. *100*, 230–237. https://doi.org/10.1007/s12185-014-1637-4.
49. Roth, D.B. V(D)J (2014). Recombination: Mechanism, Errors, and Fidelity. Microbiol Spectr *2*. https://doi.org/10.1128/microbiolspec.MDNA3-0041-2014.
50. Cancer Genome Atlas Research Network, Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R.M., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J.M. (2013). The cancer genome atlas pan-cancer analysis project. Nat. Genet. *45*, 1113–1120.
51. Niknafs, N., Balan, A., Cherry, C., Hummelink, K., Monkhorst, K., Shao, X.M., Belcaid, Z., Marrone, K.A., Murray, J., Smith, K.N., et al. (2023). Persistent mutation burden drives sustained anti-tumor immune responses. Nat. Med. *29*, 440–449. https://doi.org/10.1038/s41591-022-02163-w.
52. Oth, T., Vanderlocht, J., Van Elssen, C.H.M.J., Bos, G.M.J., and Germeraad, W.T.V. (2016). Pathogen-Associated Molecular Patterns Induced Crosstalk between Dendritic Cells, T Helper Cells, and Natural Killer Helper Cells Can Improve Dendritic Cell Vaccination. Mediators Inflamm. *2016*, 5740373. https://doi.org/10.1155/2016/5740373.
53. Li, J.G., DU, Y.M., Yan, Z.D., Yan, J., Zhuansun, Y.X., Chen, R., Zhang, W., Feng, S.L., and Ran, P.X. (2016). CD80 and CD86 knockdown in dendritic cells regulates Th1/Th2 cytokine production in asthmatic mice. Exp. Ther. Med. *11*, 878–884.
54. Furuta, K., Ishido, S., and Roche, P.A. (2012). Encounter with antigen-specific primed CD4 T cells promotes MHC class II degradation in dendritic cells. Proc. Natl. Acad. Sci. USA *109*, 19380–19385.
55. Barron, L., Dooms, H., Hoyer, K.K., Kuswanto, W., Hofmann, J., O'Gorman, W.E., and Abbas, A.K. (2010). Cutting Edge: Mechanisms of IL-2–Dependent Maintenance of Functional Regulatory T Cells. J. Immunol. *185*, 6426–6430.
56. Miragaia, R.J., Gomes, T., Chomka, A., Jardine, L., Riedel, A., Hegazy, A.N., Whibley, N., Tucci, A., Chen, X., Lindeman, I., et al. (2019). Single-Cell Transcriptomics of Regulatory T Cells Reveals Trajectories of Tissue Adaptation. Immunity *50*, 493–504.e7.
57. Thorsson, V., Gibbs, D.L., Brown, S.D., Wolf, D., Bortone, D.S., Ou Yang, T.H., Porta-Pardo, E., Gao, G.F., Plaisier, C.L., Eddy, J.A., et al. (2018). The Immune Landscape of Cancer. Immunity *48*, 812–830.e14.
58. Lowenstein, L.M., Basourakos, S.P., Williams, M.D., Troncoso, P., Gregg, J.R., Thompson, T.C., and Kim, J. (2019). Active surveillance for prostate and thyroid cancers: evolution in clinical paradigms and lessons learned. Nat. Rev. Clin. Oncol. *16*, 168–184. https://doi.org/10.1038/s41571-018-0116-x.
59. Rakha, E.A., Reis-Filho, J.S., Baehner, F., Dabbs, D.J., Decker, T., Eusebi, V., Fox, S.B., Ichihara, S., Jacquemier, J., Lakhani, S.R.,

et al. (2010). Breast cancer prognostic classification in the molecular era: the role of histological grade. Breast Cancer Res *12*, 207. https://doi.org/10.1186/bcr2607.

60. Wolf, D.M., Lenburg, M.E., Yau, C., Boudreau, A., and Van't Veer, L.J. (2014). Gene co-expression modules as clinically relevant hallmarks of breast cancer diversity. PLoS One *9*, e88309.
61. Chang, H.Y., Sneddon, J.B., Alizadeh, A.A., Sood, R., West, R.B., Montgomery, K., Chi, J.T., van de Rijn, M., Botstein, D., and Brown, P.O. (2004). Gene expression signature of fibroblast serum response predicts human cancer progression: Similarities between tumors and wounds. PLoS Biol. *2*, E7.
62. Mojic, M., Takeda, K., and Hayakawa, Y. (2017). The dark side of IFN-γ: Its role in promoting cancer immunoevasion. Int. J. Mol. Sci. *19*, 89. https://doi.org/10.3390/ijms19010089.
63. Combes, A.J., Samad, B., Tsui, J., Chew, N.W., Yan, P., Reeder, G.C., Kushnoor, D., Shen, A., Davidson, B., Barczak, A.J., et al. (2022). Discovering dominant tumor immune archetypes in a pan-cancer census. Cell *185*, 184–203.e19.
64. Riaz, N., Havel, J.J., Makarov, V., Desrichard, A., Urba, W.J., Sims, J.S., Hodi, F.S., Martín-Algarra, S., Mandal, R., Sharfman, W.H., et al. (2017). Tumor and Microenvironment Evolution during Immunotherapy with Nivolumab. Cell *171*, 934–949.e16.
65. Auslander, N., Zhang, G., Lee, J.S., Frederick, D.T., Miao, B., Moll, T., Tian, T., Wei, Z., Madan, S., Sullivan, R.J., et al. (2018). Robust prediction of response to immune checkpoint blockade therapy in metastatic melanoma. Nat. Med. *24*, 1545–1549.
66. Hugo, W., Zaretsky, J.M., Sun, L., Song, C., Moreno, B.H., Hu-Lieskovan, S., Berent-Maoz, B., Pang, J., Chmielowski, B., Cherry, G., et al. (2016). Genomic and Transcriptomic Features of Response to Anti-PD-1 Therapy in Metastatic Melanoma. Cell *165*, 35–44.
67. Amato, C.M., Hintzsche, J.D., Wells, K., Applegate, A., Gorden, N.T., Vorwald, V.M., Tobin, R.P., Nassar, K., Shellman, Y.G., Kim, J., et al. (2020). Pre-treatment mutational and transcriptomic landscape of responding metastatic melanoma patients to anti-pd1 immunotherapy. Cancers *12*, 1943.
68. Danaher, P., Warren, S., Lu, R., Samayoa, J., Sullivan, A., Pekker, I., Wallden, B., Marincola, F.M., and Cesano, A. (2018). Pan-cancer adaptive immune resistance as defined by the Tumor Inflammation Signature (TIS): Results from The Cancer Genome Atlas (TCGA). J. Immunother. Cancer *6*, 63.
69. Trefny, M.P., Rothschild, S.I., Uhlenbrock, F., Rieder, D., Kasenda, B., Stanczak, M.A., Berner, F., Kashyap, A.S., Kaiser, M., Herzig, P., et al. (2019). A variant of a killer cell immunoglobulin-like receptor is associated with resistance to PD-1 blockade in lung cancer. Clin. Cancer Res. *25*, 3026–3034.
70. Du, K., Wei, S., Wei, Z., Frederick, D.T., Miao, B., Moll, T., Tian, T., Sugarman, E., Gabrilovich, D.I., Sullivan, R.J., et al. (2021). Pathway signatures derived from on-treatment tumor specimens predict response to anti-PD1 blockade in metastatic melanoma. Nat. Commun. *12*, 6023.
71. Hsu, C.L., Ou, D.L., Bai, L.Y., Chen, C.W., Lin, L., Huang, S.F., Cheng, A.L., Jeng, Y.M., and Hsu, C. (2021). Exploring Markers of Exhausted CD8 T Cells to Predict Response to Immune Checkpoint Inhibitor Therapy for Hepatocellular Carcinoma. Liver Cancer *10*, 346–359.
72. Newman, A.M., Steen, C.B., Liu, C.L., Gentles, A.J., Chaudhuri, A.A., Scherer, F., Khodadoust, M.S., Esfahani, M.S., Luca, B.A., Steiner, D., et al. (2019). Determining cell type abundance and expression from bulk tissues with digital cytometry. Nat. Biotechnol. *37*, 773–782.
73. Zheng, L., Qin, S., Si, W., Wang, A., Xing, B., Gao, R., Ren, X., Wang, L., Wu, X., Zhang, J., et al. (2021). Pan-cancer single-cell landscape of tumor-infiltrating T cells. Science *374*, abe6474.
74. Chen, L.X., Zeng, S.J., Liu, X.D., Tang, H.B., Wang, J.W., and Jiang, Q. (2023). Cell–cell communications shape tumor microenvironment and predict clinical outcomes in clear cell renal carcinoma. J. Transl. Med. *21*, 113.
75. Maffuid, K., and Cao, Y. (2023). Decoding the Complexity of Immune–Cancer Cell Interactions: Empowering the Future of Cancer Immunotherapy. Cancers *15*, 4188. https://doi.org/10.3390/cancers15164188.
76. Sade-Feldman, M., Yizhak, K., Bjorgaard, S.L., Ray, J.P., de Boer, C.G., Jenkins, R.W., Lieb, D.J., Chen, J.H., Frederick, D.T., Barzily-Rokni, M., et al. (2018). Defining T Cell States Associated with Response to Checkpoint Immunotherapy in Melanoma. Cell *175*, 998–1013.e20.
77. Jin, S., Guerrero-Juarez, C.F., Zhang, L., Chang, I., Ramos, R., Kuan, C.H., Myung, P., Plikus, M.V., and Nie, Q. (2021). Inference and analysis of cell-cell communication using CellChat. Nat. Commun. *12*, 1088.
78. Wolock, S.L., Lopez, R., and Klein, A.M. (2019). Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. Cell Syst. *8*, 281–291.e9.
79. Barras, D., Ghisoni, E., Chiffelle, J., Orcurto, A., Dagher, J., Fahr, N., Benedetti, F., Crespo, I., Grimm, A.J., Morotti, M., et al. (2024). Response to Tumor-Infiltrating Lymphocyte Adoptive Therapy Is Associated with Preexisting CD8 + T-Myeloid Cell Networks in Melanoma. Sci. Immunol. *9*, eadg7995. https://www.science.org.
80. Du, Y., Lin, Y., Gan, L., Wang, S., Chen, S., Li, C., Hou, S., Hu, B., Wang, B., Ye, Y., and Shen, Z. (2024). Potential crosstalk between SPP1 + TAMs and CD8 + exhausted T cells promotes an immunosuppressive environment in gastric metastatic cancer. J. Transl. Med. *22*, 158.
81. Attias, M., and Piccirillo, C.A. (2024). The impact of $Foxp3^+$ regulatory T-cells on $CD8^+$ T-cell dysfunction in tumour microenvironments and responses to immune checkpoint inhibitors. Br. J. Pharmacol. https://doi.org/10.1111/bph.16313.
82. Bauer, V., Ahmetlić, F., Hömberg, N., Geishauser, A., Röcken, M., and Mocikat, R. (2021). Immune checkpoint blockade impairs immunosuppressive mechanisms of regulatory T cells in B-cell lymphoma. Transl. Oncol. *14*, 101170.
83. Human Breast Cancer. Ductal Carcinoma In Situ, Invasive Carcinoma (FFPE). Spatial Gene Expression Dataset by Space Ranger 1.3.0, 10x Genomics. https://www.10xgenomics.com/datasets/human-breast-cancer-ductal-carcinoma-in-situ-invasive-carcinoma-ffpe-1-standard-1-3-0.
84. Wei, R., He, S., Bai, S., Sei, E., Hu, M., Thompson, A., Chen, K., Krishnamurthy, S., and Navin, N.E. (2022). Spatial charting of single-cell transcriptomes in tissues. Nat. Biotechnol. *40*, 1190–1199.
85. Han, Y., Wang, Y., Dong, X., Sun, D., Liu, Z., Yue, J., Wang, H., Li, T., and Wang, C. (2023). TISCH2: expanded datasets and new tools for single-cell transcriptome analyses of the tumor microenvironment. Nucleic Acids Res. *51*, D1425–D1431.
86. Geistlinger, L., Oh, S., Ramos, M., Schiffer, L., LaRue, R.S., Henzler, C.M., Munro, S.A., Daughters, C., Nelson, A.C., Winterhoff, B.J., et al. (2020). Multiomic analysis of subtype evolution and heterogeneity in high-grade serous ovarian carcinoma. Cancer Res. *80*, 4335–4345.
87. Wu, T.D., Madireddi, S., de Almeida, P.E., Banchereau, R., Chen, Y.J.J., Chitre, A.S., Chiang, E.Y., Iftikhar, H., O'Gorman, W.E., Au-Yeung, A., et al. (2020). Peripheral T cell expansion predicts tumour infiltration and clinical response. Nature *579*, 274–278.
88. Sautès-Fridman, C., Petitprez, F., Calderaro, J., and Fridman, W.H. (2019). Tertiary lymphoid structures in the era of cancer immunotherapy. Nat. Rev. Cancer *19*, 307–325. https://doi.org/10.1038/s41568-019-0144-6.
89. Zou, X., Guan, C., Gao, J., Shi, W., Cui, Y., and Zhong, X. (2023). Tertiary lymphoid structures in pancreatic cancer: a new target for immunotherapy. Front. Immunol. *14*, 1222719. https://doi.org/10.3389/fimmu.2023.1222719.
90. Schumacher, T.N., and Thommen, D.S. (2022). Tertiary lymphoid structures in cancer. Science *375*, eabf9419. https://doi.org/10.1126/science.abf9419.
91. Domínguez Conde, C., Xu, C., Jarvis, L.B., Rainbow, D.B., Wells, S.B., Gomes, T., Howlett, S.K., Suchanek, O., Polanski, K., King, H.W., et al. (2022). Cross-tissue immune cell analysis reveals tissue-specific features in humans. Science *376*, eabl5197.
92. Li, L., Mohanty, V., Dou, J., Huang, Y., Banerjee, P.P., Miao, Q., Lohr, J.G., Vijaykumar, T., Frede, J., Knoechel, B., et al. (2023). Loss of Metabolic Fitness Drives Tumor Resistance after CAR-NK Cell Therapy and Can Be Overcome by Cytokine Engineering. Sci. Adv. *9*, eadd6997. https://www.science.org.
93. Barkley, D., Moncada, R., Pour, M., Liberman, D.A., Dryg, I., Werba, G., Wang, W., Baron, M., Rao, A., Xia, B., et al. (2022). Cancer Cell States Recur across Tumor Types and Form Specific Interactions with the Tumor Microenvironment. Nat. Genet. *54*, 1192–1201. https://doi.org/10.5281/zenodo.6611786.
94. Peng, L., Renauer, P.A., Ye, L., Yang, L., Park, J.J., Chow, R.D., Zhang, Y., Lin, Q., Bai, M., Sanchez, A., et al. (2023). Perturbomics of Tumor-Infiltrating NK Cells. https://doi.org/10.1101/2023.03.14.532653.
95. Roohani, Y., Huang, K., and Leskovec, J. (2023). Predicting transcriptional outcomes of novel multigene perturbations with GEARS. Nat Biotechnol. https://doi.org/10.1038/s41587-023-01905-6.
96. Edgar, R., Domrachev, M., and Lash, A.E. (2002). Gene Expression Omnibus: NCBI Gene Expression and Hybridization Array Data Repository, *30*. http://www.ninds.nih.gov/.

97. Murin, C.D. (2020). Considerations of Antibody Geometric Constraints on NK Cell Antibody Dependent Cellular Cytotoxicity. Front. Immunol. *11*, 1635. https://doi.org/10.3389/fimmu.2020.01635.
98. Maleki, F., Ovens, K., Hogan, D.J., and Kusalik, A.J. (2020). Gene Set Analysis: Challenges, Opportunities, and Future Research. Front. Genet. *11*, 654. https://doi.org/10.3389/fgene.2020.00654.
99. Yu, G., Wang, L.G., Han, Y., and He, Q.Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS *16*, 284–287.
100. Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J. Roy. Stat. Soc. B *57*, 289–300.
101. de Kanter, J.K., Lijnzaad, P., Candelli, T., Margaritis, T., and Holstege, F.C.P. (2019). CHETAH: a selective, hierarchical cell type identification method for single-cell RNA sequencing. Nucleic Acids Res. *47*, E95.
102. Fraley, C., Raftery, A.E., and Murphy, T.B. (2012). Mclust Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation. http://cran.r-project.org/web/packages/mclust/index.html.
103. Anders, S. Analysing RNA-Seq Data with the DESeq Package. http://www-huber.embl.de/users/.
104. Yu, G., Wang, L.G., Han, Y., and He, Q.Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS *16*, 284–287.
105. Liang, Q., Huang, Y., He, S., et al. (2023). Pathway centric analysis for single-cell RNA-seq and spatial transcriptomics data with GSDensity. Nat. Commun *14*, 8416. https://doi.org/10.1038/s41467-023-44206-x.
106. Riaz, N., Havel, J.J., Makarov, V., Desrichard, A., Urba, W.J., Sims, J.S., Hodi, F.S., Martín-Algarra, S., Mandal, R., Sharfman, W.H., et al. (2017). Tumor and Microenvironment Evolution during Immunotherapy with Nivolumab. Cell *171*, 934–949.e16.
107. Prat, A., Navarro, A., Paré, L., Reguart, N., Galván, P., Pascual, T., Martínez, A., Nuciforo, P., Comerma, L., Alos, L., et al. (2017). Immune-related gene expression profiling after PD-1 blockade in non–small cell lung carcinoma, head and neck squamous cell carcinoma, and melanoma. Cancer Res. *77*, 3540–3550.
108. Auslander, N., Zhang, G., Lee, J.S., Frederick, D.T., Miao, B., Moll, T., Tian, T., Wei, Z., Madan, S., Sullivan, R.J., et al. (2018). Robust prediction of response to immune checkpoint blockade therapy in metastatic melanoma. Nat. Med. *24*, 1545–1549.
109. Newman, A.M., Liu, C.L., Green, M.R., Gentles, A.J., Feng, W., Xu, Y., Hoang, C.D., Diehn, M., and Alizadeh, A.A. (2015). Robust enumeration of cell subsets from tissue expression profiles. Nat. Methods *12*, 453–457.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and algorithms		
Seurat	Bioconductor	https://satijalab.org/seurat
NMF	R library	https://cran.r-project.org/web/packages/NMF/index.html
Survival	R library	https://cran.r-project.org/web/packages/survival/index.html
clusterProfiler	Bioconductor	https://bioconductor.org/packages/release/bioc/html/clusterProfiler.html
DEseq2	Bioconductor	https://bioconductor.org/packages/release/bioc/html/DESeq2.html
GSEA	R library	https://cran.r-project.org/web/packages/corto/index.html
STRINGdb	Bioconductor	https://www.bioconductor.org/packages/release/bioc/html/STRINGdb.html
Glmnet	R library	https://cran.r-project.org/web/packages/glmnet/index.html
SingleR	Bioconductor	https://www.bioconductor.org/packages/release/bioc/html/SingleR.html
Gmm	R library	https://cran.r-project.org/web/packages/gmm/index.html
ssGSEA	R library	https://cran.r-project.org/web/packages/corto/index.html
GSDensity	Github	https://github.com/KChen-lab/gsdensity
CIBERSORTx	CIBERSORTx	https://cibersortx.stanford.edu
CellChat	Github	https://github.com/sqjin/CellChat
CellTrek	Github	https://github.com/navinlabcode/CellTrek
Scrublet	Github	https://github.com/swolock/scrublet
Deposited data		
ImmuneSigDB Datasets	ImmuneSigDB	https://data.broadinstitute.org/gsea-msigdb/msigdb/release/7.5.1/
TCGA RNA-seq	TCGA	https://gdc.cancer.gov/about-data/publications/panimmune
TCGA Clinical data	TCGA	https://xenabrowser.net/datapages/?dataset=Survival_SupplementalTable_S1_20171025_xena_sp&host=https%3A%2F%2Fpancanatlas.xenahubs.net&removeHub=http%3A%2F%2F127.0.0.1%3A7222&removeHub=https%3A%2F%2Fxena.treehouse.gi.ucsc.edu%3A443
TCGA Mutation data	TCGA	https://gdc.cancer.gov/about-data/publications/panimmune
TCGA Copy Number data	TCGA	https://xenabrowser.net/datapages/?dataset=broad.mit.edu_PANCAN_Genome_Wide_SNP_6_whitelisted.xena&host=https%3A%2F%2Fpancanatlas.xenahubs.net&removeHub=http%3A%2F%2F127.0.0.1%3A7222&removeHub=https%3A%2F%2Fxena.treehouse.gi.ucsc.edu%3A443
TCGA Immune Signature Score	TCGA	https://gdc.cancer.gov/about-data/publications/panimmune
PBMC CITE-seq	GEO	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE213282
COVID PBMC CITE-seq	GEO	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE155224
COVID-19 single cell atlas	GEO	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE150728
T cell stimulation	GEO	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE126030
COVID-19 cytokine storm	GEO	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE150861
Cytokine Dictionary	GEO	https://www.immune-dictionary.org/app/home
CD8 T cell stimulation - metabolism	GEO	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE211602
ICB Cohort 1	GEO	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE91061
ICB Cohort 2	GEO	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE115821

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
ICB Cohort 3	GEO	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=gse78220
ICB Cohort 4	GEO	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE145996
ICB validation 1 - lung	GEO	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE111414
ICB validation 2 - melanoma	GEO	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE168204
ICB validation 3 - liver	GEO	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE140901
ICB validation 4 - melanoma scRNA	GEO	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE120575
Breast cancer ICB scRNA-seq	GEO	http://tisch.comp-genomics.org/gallery/?cancer=BRCA&species=Human
Breast cancer spatial transcriptomic	10x	https://www.10xgenomics.com/resources/datasets/human-breast-cancer-ductal-carcinoma-in-situ-invasive-carcinoma-ffpe-1-standard-1-3-0
Ovarian cancer ICB scRNA-seq	GEO	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE154600
Ovarian cancer spatial transcriptomic	10x	https://www.10xgenomics.com/datasets/human-ovarian-cancer-1-standard
Intestine cancer ICB scRNA-seq	GEO	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE139555
Intestine cancer spatial transcriptomic	10x	https://www.10xgenomics.com/datasets/human-intestine-cancer-1-standard
Deconvolution Signature Matrix	GEO	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE156728

RESOURCE AVAILABILITY

Lead contact

- Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Ken Chen (kchen3@mdanderson.org).

Materials availability

- This study did not generate new unique reagents.

Data and code availability

- All RNA-seq data had been deposited at GEO and are publicly available. DOIs are listed in the key resources table.
- All original code to generate the figures and tables presented in the paper can be accessed through https://github.com/chloehe1129/immune-hallmark.

METHOD DETAILS

Bulk RNA-seq data curation

We referenced the 389 immunology relevant studies identified in the ImmuneSigDB publication[12] and curated the corresponding available bulk RNA-seq datasets deposited in NCBI Gene Expression Omnibus (GEO).[96] Bulk RNA-seq datasets sourced from human samples with more than 4 individuals were kept for downstream analysis. Since we focused on immune pathways represented in these datasets, mitochondria genes and ribosomal genes were removed. In addition, we grouped the datasets based on broad immune lineages.[97] As a result, 47 datasets sourced from lymphoid lineages and 36 datasets sourced from myeloid lineages were kept for further analysis.

"Data-driven" statistical analysis

We performed the following analysis for each of the 47 lymphoid dataset and 36 myeloid datasets. Non-negative matrix factorization (NMF)[15] is a common approach to establish genes that have coordinated expression or consistently opposite expression pattern that underpin the data. We performed NMF on each of the qualifying dataset with number of latent factors K = 4, 5, 6, and each NMF program is composed of genes with top 50 loadings from one latent factor. We column-combined the loadings of 529 NMF programs and normalized them. To ensure that the programs are not only non-redundant but also generalizable, we curated robust NMF programs,[16] defined as programs that (1) would reoccur across datasets (have at least 20% overlapping genes with at least one NMF program derived from outside of the current data) and (2) is non-redundant within the same dataset (have less than 20% overlapping genes with all NMF programs derived from the current dataset). To further reduce redundancy, we iteratively merged and updated the robust NMF programs into meta programs (MP) based on algorithm proposed by Tirosh et al.[16]: we initiated the process by selecting two robust NMF programs with the greatest gene overlap. Combining these two robust NMF programs, we generated a fresh gene set consisting of 50 genes, comprising of both the common genes present in both programs and selected genes unique to each program. The selected genes were chosen from all the unique genes

arranged in descending order of their loadings to ensuring that the gene set was populated to its full capacity of 50 genes. We then selected another robust NMF programs with the highest overlap with this fresh gene set and repeated the same process until the overlap is < 5, in which we restarted the process by selecting two robust NMF programs with the greatest gene overlap.

Functional annotation

We performed over-representation enrichment analysis (ORA)[98] for each MP using KEGG[10] pathways, Hallmarks,[9] and biological process terms from GO.[11] ORA was performed with "clusterProfiler" package.[99] Annotation for each MP was determined by highly enriched terms based on core enrichment count and False Discovery Rate (FDR) adjusted p-value < 0.05.[100] The annotations were then manually reviewed and refined by immunologists from MD Anderson Cancer Center.

Protein-protein interaction network for each MP

We overlaid the genes from each gene set onto protein-protein interaction (PPI) network from string database[23] using *STRINGdb* package in R.

Benchmarking

We calculated the Jaccard distance between our gene sets with Hallmark, KEGG and GO to benchmark our gene sets.

$$\text{Jaccard}(A, B) = \frac{|A \text{ Ç } B|}{|A \text{ È } B|}$$

Significance of each pairwise Jaccard distance was computed by randomly choosing gene set of the same length 1000 times and p-value was calculated as the number of times a permutated jaccard distance is larger than the actual jaccard distance divided by 1000. Multiple testing was adjusted using FDR adjusted p-values.

Cell type annotation for CITE-seq

We performed cell type annotation using *SingleR*[37] package in R with CHETAH[101] as reference.

Cell dichotomization in CITE-seq

We dichotomized cells into two groups based on their RNA activity levels of each irMP of interest using Gaussian Mixture Models[102] with G=2 using *mclust* package in R and compared the relevant protein abundance between the irMP-high and the irMP-low groups with two sample t-test.

Gene set scoring

We performed ssGSEA using *ssgsea* function in the *corto* package in R with min.size=0. We performed GSDensity when scoring cell cycle and metabolism activities using *gsdensity* package in R.

Antigen-related pathways validation

Persistent tumor mutational burden (pTMB) was calculated for each sample in TCGA, as the total number of single copy mutations and multiple copy mutations.[51] We grouped samples based on pTMB, dichotomized at mean. We performed Differential gene expression (DGE) analysis between samples with high and low pTMB using function DESeq2.[103] Results from DGE analysis with the whole transcriptome were further analyzed in Gene set enrichment analysis (GSEA) with the 19 lymphoid MPs using function *GSEA*[104] from the *clusterProfiler* package.

Cytotoxicity pathway validation

We calculated the IFN.gamma-induced cytotoxicity (L_MP17) score and exhaustion score based on GS Density.[105] Exhaustion was calculated using the expression of the following genes: *TIGIT, TOX, PDCD1, CD160, CD244, CTLA4, BTLA, HAVCR2, and LAG3.*

TCGA immune subtypes refinement

We calculated the single sample gene set enrichment score (ssGSEA score)[27] for each of the MP across all patient samples in TCGA pan-cancer data and projected the samples onto a two-dimensional space through UMAP using *umap* function. We colored each sample with corresponding TCGA immune subtype information. In addition, we fitted Kaplan Meier (KM) survival curves for each of the four TCGA immune subtypes. We also performed k-mean clustering with *kmean* function using the ssGSEA matrix with the value of K=6 selected by low within-cluster sum of square and fitted KM survival curves for each identified cluster.

ICB response model

We calculated ssGSEA score for each patient from the combined ICB studies[106–108] using (1) 50 MSigDB Hallmark pathways, (2) 29 immune-related KEGG pathways (3) lymphoid-derived irMPs, (4) myeloid-derived irMPs and (5) lymphoid-myeloid combined irMPs. We fitted five

generalized linear model with LASSO regularization using *glmnet* with binary response variable adjusting for each set of the above-mentioned gene sets using 70% of the full data as training data. In cases where ICB response is defined in more than two categories, we grouped complete and partial responder into responders and stable and progressive into non-responders. Using the best model selected from 10-fold cross-validation, we fitted the model on the test data and evaluated model performance by comparing classification accuracy. Classification accuracy is defined as the average of sensitivity and specificity. To deal with randomness in data splitting, we performed the procedure 1000 times and computed a range of classification accuracy from each of the five models.

Marker genes for immune signatures

Exhaustion: *PDCD1, TIGIT, HAVCR2, CTLA4, LAG3.*

Cytotoxicity: *NKG7, CCL4, CST7, PRF1, GZMA, GZMB, IFNG, CCL3.*

Tissue resident memory: *ITGAE, ZNF683, ITGA1, CD69, CXCR6, CXCL13, PDCD1.*

Immune signatures were scored based on ssGSEA.

Pseudo-bulk aggregation

scRNA-seq data was converted into pseudo-bulk gene expression by using the function *AggregateExpression* with group.by = "sample".

Cell type deconvolution

We used breast cancer infiltrated T cells subtypes annotated by Zheng et al.[73] to deconvolute the bulk RNA samples using CIBERSORTx[109] with the following configuration: B-mode batch correction and 50 permutations for significance calculation.

Cell chat analysis

We performed cell chat analysis using the *CellChat* package in R.

Doublet detection and annotation

We used *scrublet* to calculate a doublet score for each cell and identified doublets with top 5% doublet score. We used *FindAllMarkers* with adjusted p-value<0.05 and log fold change>0.25 to define differentially expressed genes for each cell type and calculate the cell type ssGSEA for each doublet. We double-annotated each doublet with the top two positive scoring cell types.

Single cell level annotation on spatial transcriptomic

We calculated the irMPs score on each spatial spot using *AddModuleScore* function in R. We performed *CellTrek* to achieve single-cell level spatial cell type mapping on breast cancer spatial transcriptomic (ST) data with an independent scRNAseq (SC) data as reference. Default arguments were used for *celltrek* function in R.

scRNAseq data processing

All single cell RNA sequencing data were processed using standard Seurat procedure implemented by *Seurat* package in R. We performed the following procedures for all scRNAseq data: *NormalizeData(), FindVariableGenes(), ScaleData(), RunPCA(), FindNeighbors(),FindClusters()* and *RunUMAP()* with 10 PCs.

Differential gene expression and pathway analysis

For single cell data, we used *FindAllMarkers()* to identify differentially expressed genes. For BulkRNAseq data, we used *Deseq2* to identify differentially expressed genes. Pathway enrichment analysis was performed using *clusterProfiler* package in R. For gene set enrichment, statistical significance was determined by a rank-based test, in which the enrichment score (ES) reflects the degree to which genes in a set are over-represented at the extremes of a ranked list, utilizing a Kolmogorov–Smirnov-like statistic. Then the statistical significance of the ES is determined through a permutation test based on phenotypic labels, generating a null distribution for comparison to ascertain the gene set's dependency on these labels.

QUANTIFICATION AND STATISTICAL ANALYSIS

All statistical analysis were performed in R version 4.2.1 and Seurat version 4.3.0. Differences between continuous distributions were tested using Standardized t test in case of two variables or ANOVA in cases of more than two variables. Multiple testing correction for false discovery was performed using False Discovery Rate (FDR) with function *p. adjust*, and statistical significance is defined at FDR adjusted p-value < 0.05.