Recurrent composite markers of cell types and states

Xubin Li* and Anil Korkut*

Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA

***Corresponding Authors**

Anil Korkut

Email: akorkut@mdanderson.org

Xubin Li

Email: xli27@mdanderson.org

Abstract

Determining concise sets of genomic markers that identify cell types and states within tissue ecosystems remains challenging. To address this challenge, we developed Recurrent Composite Markers for Biological Identities with Neighborhood Enrichment (RECOMBINE). Validations of RECOMBINE with simulation and transcriptomics data in bulk, single-cell and spatial resolutions demonstrated the method's ability for unbiased selection of composite markers that characterize biological subpopulations. RECOMBINE captured markers of mouse visual cortex from single-cell RNA sequencing data and provided a gene panel for targeted spatial transcriptomics profiling. RECOMBINE identified composite markers of CD8 T cell states including $GZMK^{+}HAVCR2^{-}$ effector memory cells associated with anti-PD1 therapy response. The method outperformed differential gene expression analysis in characterizing a rare cell subpopulation within mouse intestine. Using RECOMBINE, we uncovered hierarchical gene programs of inter- and intra-tumoral heterogeneity in breast and skin tumors. In conclusion, RECOMBINE offers a data-driven approach for unbiased selection of composite markers, resulting in improved interpretation, discovery, and validation of cell types and states.

Introduction

Tissue ecosystems carry diverse cell types and states that manifest biological structure and function[1,2]. The characterization of cell types and states within tissues has been a central focus in biology. It is particularly important to identify a minimal set of marker genes that distinguish not only distinct cell types, but also continuous cell states that are characterized by an overlapping spectrum of molecular features[3-5]. Identification of such composite markers that drive tissue heterogeneity can enable prediction of healthy or aberrant phenotypes such as immune states, developmental stages, and responses to therapy. Single cell omics technologies such as single cell transcriptomics provide opportunities to address the complexity of tissue ecosystems in both healthy and diseased states[6-10]. There is, however, a major need for analytical methods to decode the composite markers that drive tissue heterogeneity and function based on single cell omics data. In parallel, emerging technologies of targeted spatial transcriptomics, such as MERFISH[11], STARmap[12], and CosMx[13], require pre-selected marker panels to distinguish heterogeneous cell types and states in tissues. Marker panels, which can accurately represent biological processes underlying disease mechanisms, are also required for development of cost-effective clinical assays[14]. Despite the imperative need, identification of composite markers that optimally discriminate cell subpopulations with high granularity remains challenging.

Current computational methods for finding markers of cell subpopulations first perform partitional clustering to group cells into discrete clusters, followed by selecting markers to discriminate each cluster from others in a supervised way[15,16]. For example, the most popular method of finding markers involves statistical tests to generate a list of differentially expressed genes (DEGs) between a particular cell cluster and other clusters[15,16]. The list of DEGs is often very long, and its interpretation is tedious. More importantly, the selected markers are biased by pre-defined cluster labels and do not discriminate continuous cell states within clusters. One of the grand challenges, however, is flexibly choosing varying levels of resolution for mapping discrete cell types and continuous states[17]. Recently developed partition-free methods have addressed this challenge for different purposes. For example, MetaCell[18] and SEACells[19] were introduced to construct metacells representing highly granular and distinct cell states. By leveraging K nearest neighbor graph representation, Milo[20] identifies perturbed cell states that are obscured by partitional clustering. Yet, a partition-free and unbiased method that selects composite markers for characterization of continuous cell states in addition to discrete cell types is lacking.

We developed a novel computational framework, recurrent composite markers for biological identities with neighborhood enrichment (RECOMBINE). First, RECOMBINE selects an optimized set of markers that discriminate hierarchical cell subpopulations. Next, the method identifies recurrent composite markers (RCMs) for not only discrete cell types but also continuous cell states with high granularity. We tested and benchmarked our method to identify biological subpopulations and composite markers using simulation and actual biological data from breast cancer tumors. The tests with simulation data demonstrated the high recall and precision performance of RECOMBINE. In tests with protein and RNA expression data, RECOMBINE successfully recapitulated the well-known subtypes of breast cancers as well as the representative markers of the subtypes. In applications with single cell omics data, RECOMBINE identified composite markers that define cell subpopulations associated with (i) tissue heterogeneity that is spatially distributed to layers in mouse visual cortex, (ii) CD8 T cell states and responses to immunotherapy, (iii) rare and dominant biological identities in intestinal tissue, and (iv) intra- and inter-tumor heterogeneity in tumor tissues. From these studies, we demonstrated the excellent performance of RECOMBINE and its capability of unbiased marker selection for heterogeneous cell contexts with high granularity. RECOMBINE is a data-driven

approach for unbiased selection of composite markers that characterize discrete cell types and continuous cell states in tissue ecosystems.

Results

RECOMBINE selects recurrent composite markers of biological subpopulations

RECOMBINE is a computational framework for unbiased selection of RCMs that distinguish hierarchical cell subpopulations in two steps (Fig. 1). The first step is detection of features that discriminate hierarchical cell subpopulations based on high-dimensional data, such as scRNA-seq data, through clustering coupled with feature selection (Fig. 1a). For this purpose, we developed two novel algorithms based on the conventional sparse hierarchical clustering (SHC) algorithm[21]: SHC with spike-and-slab lasso (SHC-SSL) and SHC with fused lasso (SHC-FL) (Fig. 1b-c). The second step is extraction of RCMs from discriminant markers through construction of K nearest neighbor graphs of cells (Fig. 1a). The statistically significant RCMs are identified based on p-values of neighborhood recurrence tests followed by multiple hypothesis testing correction. The outcome is sets of RCMs characterizing individual biological subpopulations.

We evaluated the accuracy of SHC, SHC-SSL, and SHC-FL with a systematic benchmarking using simulation data (Fig. 1d, Supplementary Fig. 1). All three algorithms reached median precisions of 100%, indicating that selected features were true discriminant markers (Fig. 1d). SHC-SSL had the highest median values of recall and F1 score, and their variations were low, indicating it achieved the best feature selection accuracy with the lowest false negative rate (Fig. 1d). The prior information of feature order improved the performance of SHC-FL toward significantly better feature selection. The recall variation, however, was high for SHC-FL with a range of 100% to 20%, suggesting a high dependence on the concordance of priors with the data. In contrast, SHC consistently demonstrated low recall and F1 score compared to other algorithms. Next, we evaluated the clustering performances using four metrics: silhouette, concordance, adjusted Rand index, and purity (see Methods: Benchmarking of sparse hierarchical clustering algorithms). For all the metrics, the performance scores monotonically increased from SHC to SHC-FL to SHC-SSL, with SHC-SSL showing the best median performance with lowest variation (Fig. 1d). In conclusion, SHC-SSL is more accurate and robust for both feature selection and clustering compared to other SHC-based algorithms.

Next, we tested RECOMBINE with biological data involving bulk-level mRNA and protein expression measurements from tumor samples of breast cancer patients (source: TCGA, N=529)[22]. The goal was to test whether RECOMBINE could recapitulate known biomarkers of breast cancer subtypes as well as potential therapeutic targets. RECOMBINE identified discriminant mRNA and protein markers (N_{mRNA}=157, $N_{protein}$=25) that accurately distinguished breast cancer subtypes (Supplementary Fig. 1 and Supplementary Table 1). The top ranked discriminant markers included classical biomarkers of breast cancer subtypes: ESR1, AR, GATA3, ERBB3, ERBB4, and phopho-HER2 (for protein only). Interestingly, ERBB2 was not selected within the mRNA analysis. A linear regression analysis, however, showed that ERBB2 could be predicted well using the selected genes (Pearson correlation between predicted and actual ERBB2 expressions: 0.86), suggesting the discriminant effect of ERBB2 is largely explained by the selected markers. Based on the discriminant markers, Leiden clustering identified five major patient groups for both protein and mRNA analysis separating the luminal A and B (mRNA: clusters (c)1-3, protein: c2-5), $HER2^+$ (mRNA: c4, protein: c1), and basal-like (mRNA: c5, protein: c1) subtypes. Besides classical biomarkers of breast cancer subtypes, the RCM modules revealed other markers that had been less characterized previously. For example, basal-like tumors (mRNA: c5) were associated with high mRNA expression of BCL11A, a driver of TNBC

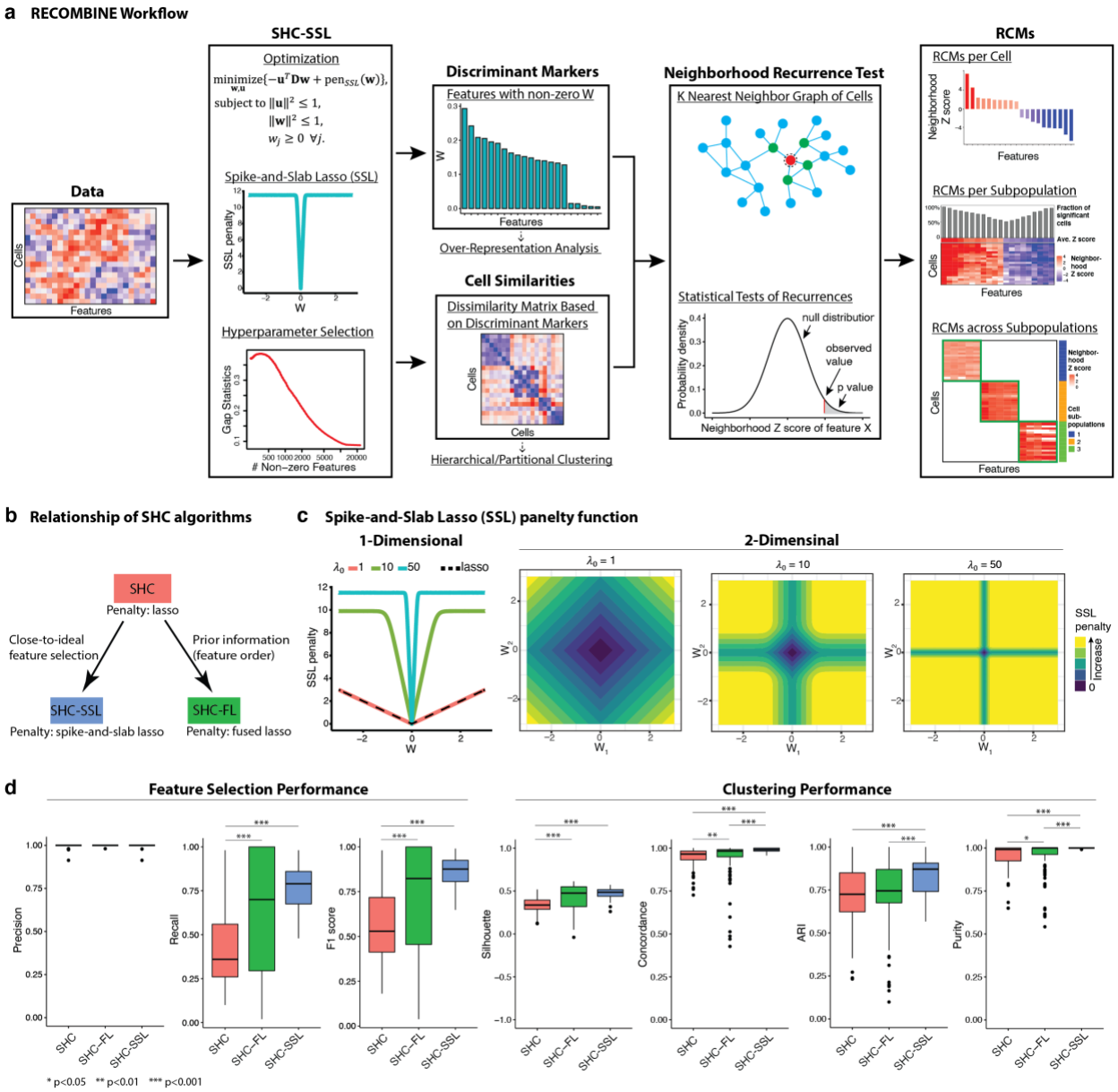

Fig. 1 | RECOMBINE: a computational framework enabling unbiased selection of recurrent composite markers for cell subpopulations with high granularity. a, *RECOMBINE workflow. From high-dimensional data such as single cell transcriptomics, RECOMBINE extracts markers that discriminate the population into hierarchical subpopulations using the sparse hierarchical clustering with spike-and-slab lasso (SHC-SSL) algorithm. SHC-SSL results in two outputs: (1) discriminant markers as features with non-zero weights and (2) cell similarities based on the discriminant markers. Next, based on the discriminant features, the recurrent composite markers (RCMs) are selected at single cell level through construction of K nearest neighbor graphs of cells. Next, a neighborhood Z score is calculated for each discriminant marker based on the distribution of the marker expression across all neighborhoods. Statistical significance of the RCMs is evaluated based on the p-values derived from the neighborhood Z scores followed by multiple hypothesis testing correction. To determine if a marker is recurrent for a cell subpopulation, fraction of significant cells and average neighborhood Z score are considered.* ***b,*** *The SSL penalty in SHC-SSL provides a continuum between lasso and L0-norm to enable more accurate selection*

of discriminant features compared to the lasso in SHC. The fused lasso (FL) in SHC-FL incorporates prior information of feature order. ***c,*** *The SSL penalty function. The SSL penalty as a function of* $\boldsymbol{w}$ *with* $\lambda_1 = 0.001$ *and* $\lambda_0 \in \{1,10,50\}$*, where* $\boldsymbol{w}$ *is a one-dimensional (left panel) or two-dimensional variable (other panels). The SSL penalty converges to the lasso penalty with decreasing* λ_0 *and to the L0-norm penalty with increasing* λ_0*. Therefore, the SSL penalty forms a data-adjustable bridge between an L0-norm and a lasso penalty, which empowers SHC-SSL without any* a priori *information.* ***d,*** *Benchmarking of the sparse hierarchical clustering algorithms for feature selection precision and recall, and clustering performance, using 100 randomly generated simulation datasets. Silhouette measures consistency of dissimilarity matrix based on discriminant features and ground truth labels of clusters, concordance measures clustering performance after hierarchical clustering, and adjusted Rand index (ARI) and purity measure clustering performance after Leiden clustering. For each metric, a higher value is better.*

progression[23], as well as low expression of FOXA1 and FOXP1, potential diagnostic and prognostic TNBC biomarkers[24]. Consistent with the promising clinical tests for targeting BCL2 in luminal cancers, luminal groups (mRNA: c3 and protein: c3-4) carried high expressions of BCL2 mRNA and protein. The cyclin B1 protein was highly expressed in patients with basal-like tumors within the protein cluster c1, indicating a potential therapeutic vulnerability for basal-like breast cancers, which is consistent with the reports on suppression of cell proliferation after cyclin B1 knockdown in basal-like subtype cell lines[25]. In conclusion, RECOMBINE accurately recapitulates the breast cancer subtypes and associated subtype-specific oncogenic, therapeutic, and prognostic markers.

RECOBMINE selects unbiased marker panel for targeted spatial transcriptomics

Targeted spatial transcriptomics technologies (e.g., MERFISH[11], STARmap[12], and CosMx[13]) have enabled spatially resolved molecular profiling at single cell resolution. The applications of targeted spatial transcriptomics, however, require *a priori* selection of a gene panel that can discriminate diverse cell types and states. We addressed this problem using RECOMBINE, which selects markers from paired scRNA-seq data without any bias or prior knowledge of cell type-specificity or number of markers. We applied RECOMBINE to an scRNA-seq dataset of mouse visual cortex[26] and validated the selected markers using a STARmap dataset of the same tissue[12]. Using the scRNA-seq data covering 14,249 cells from 23 cell types across three major compartments (excitatory neurons, inhibitory neurons, and non-neuronal cells), RECOMBINE selected 366 genes as discriminant markers (Fig. 2a, Supplementary Fig. 2, and Supplementary Table 2). The top ranked discriminant markers included genes identifying inhibitory neurons (Gad1, Gad2, Vip, Npy, Synpr, Cck, Sst, and Rab3b) and excitatory neurons (Slc17a7, Pcp4, and Nrgn). The UMAP projection of all cells based on the discriminant markers clearly separated all 23 cell types (Fig. 2b). Furthermore, we investigated whether these discriminant markers could be captured in DEGs between cell types. Comparison of RECOMBINE markers and DEGs showed that all the RECOMBINE markers were captured in DEGs, but the number of RECOMBINE markers (N=366) was only 12% of the number of DEGs (N=3096), suggesting that RECOMBINE reached an optimized discrimination power with substantially fewer markers (Fig. 2c). Moreover, the RECOMBINE markers were enriched in more significant DEGs but also included markers in the region of less significant DEGs, suggesting that the less significant DEGs could be important discriminants of hierarchical cell subpopulations (Fig. 2d). We then compared the 366 discriminant markers and the 1,020 genes that were spatially profiled in the STARmap dataset, resulting in 248 overlapped markers (Fig. 2e). Based on the overlapped markers, Leiden clustering of cells from the STARmap dataset identified 11 cell clusters of

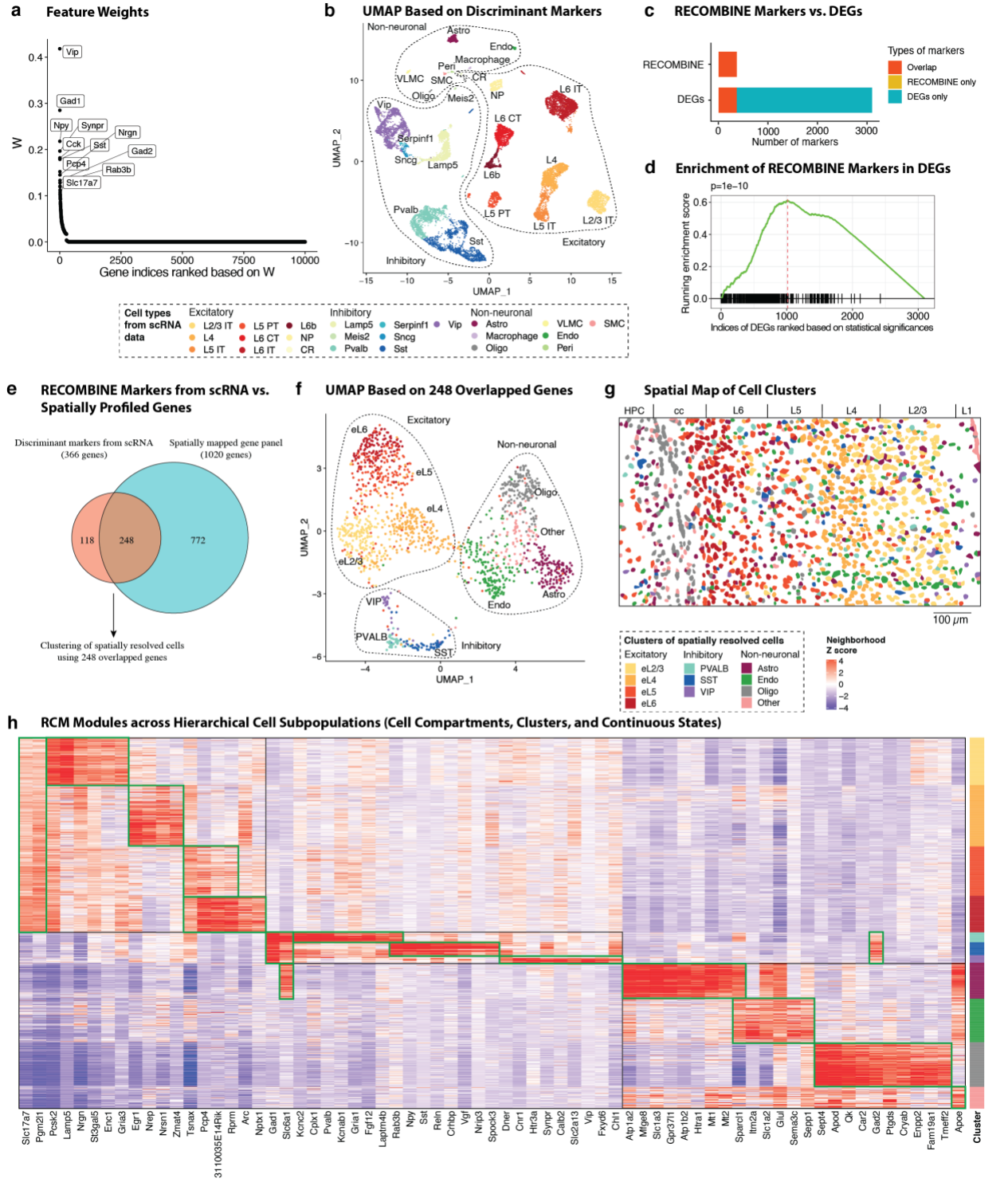

Fig. 2 | RECOMBINE selects composite markers from scRNA data that can be used as a targeted panel for spatial profiling of mouse visual cortex. a, *The contribution of each gene to discrimination of cell subpopulations are quantified as feature weights (w). The top discriminant markers are labeled.* ***b,*** *UMAP of cells based on expression of the discriminant markers, where colors indicate cell type labels from Tasic et al. (Nature, 2018). Cell type abbreviations: IT, intratelencephalic; PT, pyramidal tract; CT, corticothalamic; NP, near-projecting; CR, Cajal–*

Retzius cell; Astro, astrocyte; Oligo, oligodendrocyte; VLMC, vascular leptomeningeal cell; Endo, endothelial cell; Peri, pericyte; and SMC, smooth muscle cell. ***c,*** *Comparison of RECOMBINE discriminant markers and DEGs. DEGs were obtained by Wilcoxon tests of each cell type with respect to the rest of cell types and filtered based on absolute $\log_2$ fold-change > 0.25 and false discovery rate < 0.05.* ***d,*** *Enrichment of RECOMBINE discriminant markers within the DEGs ranked by their significance levels (Wilcoxon test).* ***e,*** *Venn diagram of discriminant markers in the scRNA data and the genes profiled in the spatial data.* ***f,*** *UMAP of spatially resolved cells based on the selected markers (N=248), where colors indicate clusters identified by Leiden clustering based on the discriminant markers.* ***g,*** *Spatial map of cells with the same color code of cell clusters in (e). L1-6, the six neocortical layers; cc, corpus callosum; HPC, hippocampus.* ***h,*** *Heatmap of neighborhood Z scores showing recurrent composite marker (RCM) modules across cell clusters of the spatial data. The RCMs shown here are the top 10 significantly enriched markers in more than 50% of cells in each cluster. Cells are hierarchically clustered within each Leiden cluster. Black boxes indicate RCM modules across the three cell compartments: excitatory neurons, inhibitory neurons, and non-neuronal cells. Green boxes indicate RCM modules across cell clusters.*

the excitatory neuron, inhibitory neuron, and non-neuronal compartments. (Fig. 2f). Mapping the cluster annotation of cells to their physical locations recapitulated the layered structure of excitatory neurons and sparsely distributed inhibitory neurons[12] (Fig. 2g). The RCMs formed distinct modules across cell clusters and defined cell subpopulations (Fig. 2h). Individual markers, however, could be shared across clusters or compartments, indicating the importance of modular RCMs for cell typing and continuous nature of expression events across cell types. For example, although both Slc6a1 and Gad2 were enriched in inhibitory neurons, Slc6a1 was also enriched in astrocytes, and Gad2 was also enriched in oligodendrocytes. Finally, we hypothesized that the RCMs with highest neighborhood enrichment may provide a marker panel with reduced size and yet minimal loss of discriminant power for cell typing. To test this hypothesis, we extracted the top 10 enriched markers in each cluster and aggregated them, forming a panel of 69 markers (Fig. 2h). UMAP projection of all cells based on the 69 most enriched markers separated all the cell subpopulations and revealed a similar pattern as the one based on all the discriminant markers, except that cluster eL4 was far away from cluster eL5 (Supplementary Fig. 2). Collectively, these findings show that RECOMBINE enables unbiased and data-driven selection of a marker panel based on paired scRNA-seq data for targeted spatial transcriptomics.

RECOMBINE identifies CD8 T cell states and markers associated with response to cancer immunotherapy

Improvement of cancer immunotherapies such as immune checkpoint blockade and adoptive cell therapy requires a deep understanding of T cell states and associated markers[27-30]. We applied RECOMBINE to identify RCMs of CD8 T cell states using a pan-cancer scRNA-seq dataset of CD8 T cells (N=234,550)[31]. RECOMBINE selected 320 discriminant markers that discriminated sub-groups of CD8 T cells (Fig. 3a, Supplementary Fig. 3, and Supplementary Table 3). The top 10 ranked markers included cytotoxic molecules (FGFBP2, FCGR3A, NKG7, GZMA, GZMB, and GZMH), and cytokine and growth factor-related molecules (CCL5, CCR6, CCR7, and FGFBP2), suggesting effector and memory functions. Leiden clustering based on the discriminant markers identified 15 clusters (Fig. 3b). Each cluster was characterized by its RCMs, which consisted of a small set of markers either unique to the cluster or shared by clusters with similar CD8 T cell states (Fig. 3c and d). We further looked at expressions of checkpoint genes

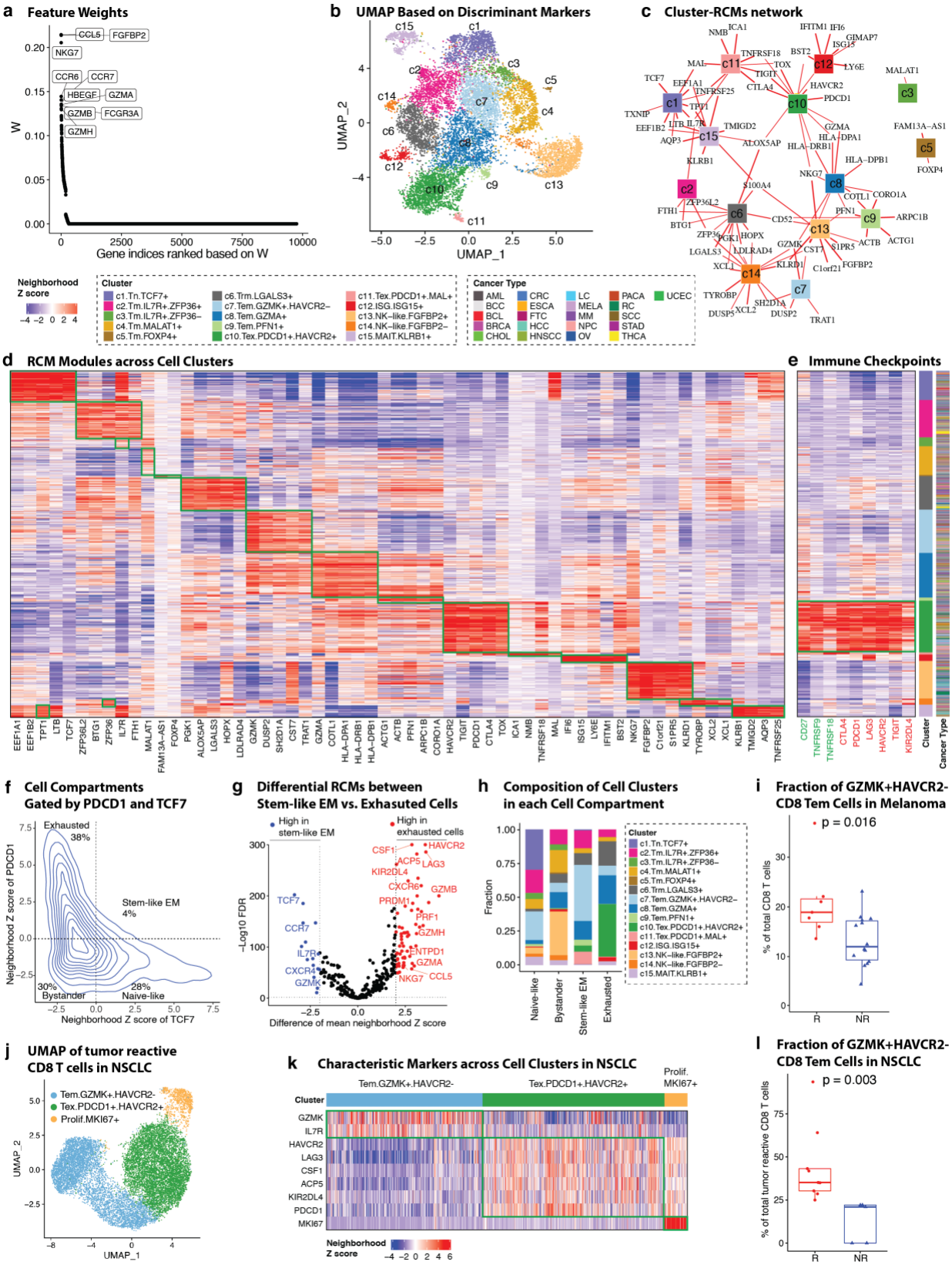
a Feature Weights
Gene indices ranked based on W
b UMAP Based on Discriminant Markers
UMAP_1
UMAP_2
c Cluster-RCMs network
Neighborhood Z score
Cluster
c1.Tn.TCF7+
c2.Tm.IL7R+.ZFP36+
c3.Tm.IL7R+.ZFP36−
c4.Tm.MALAT1+
c5.Tm.FOXP4+
c6.Trm.LGALS3+
c7.Tem.GZMK+.HAVCR2−
c8.Tem.GZMA+
c9.Tem.PFN1+
c10.Tex.PDCD1+.HAVCR2+
c11.Tex.PDCD1+.MAL+
c12.ISG.ISG15+
c13.NK−like.FGFBP2+
c14.NK−like.FGFBP2−
c15.MAIT.KLRB1+
Cancer Type
AML
BCC
BCL
BRCA
CHOL
CRC
ESCA
FTC
HCC
HNSCC
LC
MELA
MM
NPC
OV
PACA
RC
SCC
STAD
THCA
UCEC
d RCM Modules across Cell Clusters
e Immune Checkpoints
f Cell Compartments Gated by PDCD1 and TCF7
Exhausted 38%
Stem-like EM 4%
30% Bystander
28% Naive-like
Neighborhood Z score of PDCD1
Neighborhood Z score of TCF7
g Differential RCMs between Stem-like EM vs. Exhausted Cells
High in stem-like EM
High in exhausted cells
−Log10 FDR
Difference of mean neighborhood Z score
h Composition of Cell Clusters in each Cell Compartment
Fraction
Naive-like
Bystander
Stem-like EM
Exhausted
i Fraction of GZMK+HAVCR2- CD8 Tem Cells in Melanoma
p = 0.016
% of total CD8 T cells
R
NR
j UMAP of tumor reactive CD8 T cells in NSCLC
Tem.GZMK+.HAVCR2-
Tex.PDCD1+.HAVCR2+
Prolif.MKI67+
k Characteristic Markers across Cell Clusters in NSCLC
Cluster
GZMK
IL7R
HAVCR2
LAG3
CSF1
ACP5
KIR2DL4
PDCD1
MKI67
Neighborhood Z score
l Fraction of GZMK+HAVCR2- CD8 Tem Cells in NSCLC
p = 0.003
% of total tumor reactive CD8 T cells

Fig. 3 | RECOMBINE identifies CD8 T cell states and RCMs associated with cancer immunotherapy response. a, *Feature weights of all genes in which the top 10 discriminant markers are labeled.* ***b,*** *UMAP showing clusters identified by Leiden method based on discriminant markers.* ***c,*** *Bipartite graph showing inter-connection of clusters and their RCMs.* ***d,*** *Heatmap of neighborhood Z scores showing RCM modules across clusters. The RCMs shown in (c and d) are top 5 significantly enriched markers in more than 50% of cells in each cluster. Cells are hierarchically clustered within each Leiden cluster. Green boxes indicate RCM modules.* ***e,*** *Heatmap of neighborhood Z scores showing stimulatory and inhibitory immune checkpoint genes identified by RECOMBINE.* ***f,*** *CD8 T cell compartments gated by neighborhood Z scores of TCF7 and PDCD1. EM: effector memory.* ***g,*** *Differentially expressed markers between stem-like EM and exhausted T cell compartments defined in (f).* ***h,*** *Compositions of cell clusters across the cell compartments defined in (f).* ***i,*** *Differential enrichment of GZMK$^+$HAVCR2$^-$ T_{em} (c7 in RECOMBINE analysis) CD8 T cell fractions in immunotherapy responders (R, N=7) vs. non-responders (NR, N=14) of melanoma patients treated with anti–PD-1 therapy (Sade-Feldman et al. 2018).* ***j,*** *UMAP embedding of 13,403 tumor-reactive CD8 T cells of anti-PD-1–treated and treatment-naïve patients with non-small cell lung cancer (N=34) (Liu et al. 2022).* ***k,*** *Heatmap of neighborhood Z scores showing characteristic markers across all the clusters in (j).* ***l,*** *Differential enrichment of GZMK$^+$HAVCR2$^-$ T_{em} cell fractions within tumor-reactive CD8 T cells in anti–PD-1 therapy responder (R, N=9) vs. non-responder (NR, N=5) patients with non-small cell lung cancer (Liu et al. 2022).*

identified by RECOMBINE. Notably, cluster c10 was marked with high levels of both stimulatory (CD27, TNFRSF9, and TNFRSF18) and inhibitory (CTLA4, PDCD1, LAG3, HAVCR2, and TIGIT) checkpoint molecules, indicating a terminally exhausted state (Fig. 3e). Expression of KIR2DL4, an inhibitory checkpoint molecule expressed on NK cells, was also enriched in c10, consistent with a recent report that an NK cell–like transition was a signature of CAR T cell dysfunction[32]. In summary, RECOMBINE successfully identified CD8 T cell states characterized by RCMs.

Next, we asked whether characteristic markers and T cell subpopulations were associated with responses to immunotherapies such as PD-1 blockade. Previous studies suggested that exhausted CD8 T cells (TCF7$^-$PDCD1$^+$) acquire a distinct epigenetic profile that is stable and minimally remodeled after anti-PD-1 therapy[33,34], whereas stem-like effector memory (EM) CD8 T cells (TCF7$^+$PDCD1$^+$) provide tumor control and proliferative burst in response to anti–PD-1 therapy[35-37]. To decipher characteristic markers and specific cell subpopulations responded to anti–PD-1 therapy, we divided all CD8 T cells by gating the neighborhood Z scores of TCF7, a stemness marker, and PDCD1, an activation and exhaustion marker, into 4 compartments: naïve-like, bystander, stem-like EM, and exhausted cells (Fig. 3f). To find markers distinguishing the stem-like EM and exhausted cells, we performed differential analysis of the neighborhood Z scores of discriminant markers between these two populations (Fig. 3g and Supplementary Table 1). As expected, many naïve and memory markers (TCF7, CCR7, IL7R, and CXCR4) were significantly enriched in stem-like EM cells, whereas many cytotoxic and terminal differentiation markers (GZMA, GZMB, GZMH, PRF1, NKG7, CCL5, CXCR6, ENTPD1, and PRDM1) were significantly enriched in exhausted cells. Notably, GZMK expression was significantly high in stem-like EM cells while inhibitory checkpoint molecules HAVCR2, LAG3, and KIR2DL4 (but not PDCD1, CTLA4, and TIGIT) were significantly enriched in exhausted cells, suggesting crucial roles of GZMK in the early-stage development of EM cells and HAVCR2, LAG3, and KIR2DL4 in the late-stage development of terminally exhausted cells. Accordingly, c7 (GZMK$^+$HAVCR2$^-$ T_{em}) and c10 (PDCD1$^+$HAVCR2$^+$ T_{ex}) were the largest subpopulations of stem-like EM and exhausted cells, respectively (Fig. 3h).

As stem-like EM CD8 T cells are likely the bona fide target of anti–PD-1 therapy[28,35-37], we hypothesized that GZMK$^+$HAVCR2$^-$ T_{em} cells may play an essential role in anti–PD-1 responses. To test this hypothesis, we explored the association of GZMK$^+$HAVCR2$^-$ T_{em} cells with clinical responses in a subset of anti-PD-1–treated melanoma patients (N=21) whose clinical outcomes were described by Sade-Feldman et al.[38]. Consistent with our hypothesis, the enrichment of c7 (GZMK$^+$HAVCR2$^-$ T_{em}) cells was significantly higher in melanoma patients who responded to anti-PD-1 therapy (Fig. 3i). To validate this finding, we applied RECOMBINE to an independent scRNA-seq dataset that was not included in the pan-cancer data. The dataset had 13,403 tumor-reactive CD8 T cells from both anti-PD-1–treated and treatment-naïve patients with non-small cell lung cancer (N=34)[39]. Three major cell states with characteristic markers were identified: GZMK$^+$HAVCR2$^-$ T_{em}, PDCD1$^+$HAVCR2$^+$ T_{ex}, and MKI67$^+$ proliferative cells (Fig. 3j and k). Using the samples from this cohort obtained after anti-PD-1–therapy (N=14), we analyzed the association of GZMK$^+$HAVCR2- T_{em} cells with clinical responses. Consistent with the result from melanoma patients, the GZMK$^+$HAVCR2$^-$ T_{em} cell population after anti–PD-1 therapy was significantly enriched in responders (Fig. 3l). Finally, the top RCMs of exhausted cells (HAVCR2, LAG3, KIR2DL4, CSF1, and ACP5) in the pan-cancer dataset (Fig. 3g) were recapitulated in the lung cancer dataset (Fig. 3k), confirming their role in T cell exhaustion. It is important to acknowledge that KIR2DL4, CSF1, and ACP5 have received less characterization in relation to T cell exhaustion. Therefore, further studies are necessary to delve into these factors, which may unveil valuable insights. In conclusion, through RECOMBINE analyses, we identified characteristic markers of stem-like EM and exhausted CD8 T cells, demonstrated enrichment of GZMK$^+$HAVCR2$^-$ T_{em} cell population in responders to anti–PD-1 therapy, and nominated novel markers related to CD8 T cell exhaustion that may be targeted to improve responses to cancer immunotherapy.

RECOMBINE identifies RCMs of common and rare intestinal cell subpopulations

To demonstrate the ability of RECOMBINE to identify common and rare cell types, a major goal in single-cell omics analysis[17], we analyzed an scRNA-seq dataset of mouse intestinal organoids[40]. Leiden clustering of all cells based on the 183 RECOMBINE markers identified four clusters corresponding to enterocytes, enterocyte precursor cells, transit amplifying cells, and secretory cells, with substantial intra-cluster heterogeneity (Fig. 4a-c, Supplementary Fig. 4, and Supplementary Table 4). The top 10 RECOMBINE markers discriminated the common subpopulation, enterocytes, from other cell subpopulations. The enterocyte precursor cell cluster had two subpopulations: late-stage cells, which were more differentiated toward enterocytes, and early-stage cells, which were closer to transit amplifying cells. Analysis of RCMs revealed three gene modules corresponding to enterocytes, transit amplifying cells, and secretory cells (Fig. 4c). Over-representation analysis of the RECOMBINE markers showed that transit amplifying cells were enriched with genes encoding ribosomal proteins and mRNAs while enterocytes were enriched with metabolic pathway genes (hypergeometric test, false discovery rate [FDR] < 0.001) (Fig. 4c-d). Taken together, these analyses showed that RECOMBINE identified four cell subpopulations characterized by three modules of RCMs in intestinal organoids.

Next, we tested whether RECOMBINE can detect rare cell subpopulations and associated markers. As Reg4$^+$ enteroendocrine cells were previously discovered as a rare subpopulation of secretory intestinal cells, we investigated whether RECOMBINE could identify the Reg4$^+$ enteroendocrine subpopulation[40]. Hierarchical clustering based on the RECOMBINE markers of secretory cells revealed the rare population of Reg4$^+$ enteroendocrine cells. The rare Reg4$^+$ population was characterized by relatively high expression of not only Reg4 but also other genes, including Cd24a, Actg1, Hopx, Jun, Tff3, Chgb, Egr1, Tac1, Chga, and Fos (Fig 4e). Importantly, Reg4, Chga, and Chgb, which were previously used to characterize this population in Grün et al.[40], were also expressed in a fraction of enterocytes and enterocyte precursor cells (Fig. 4f).

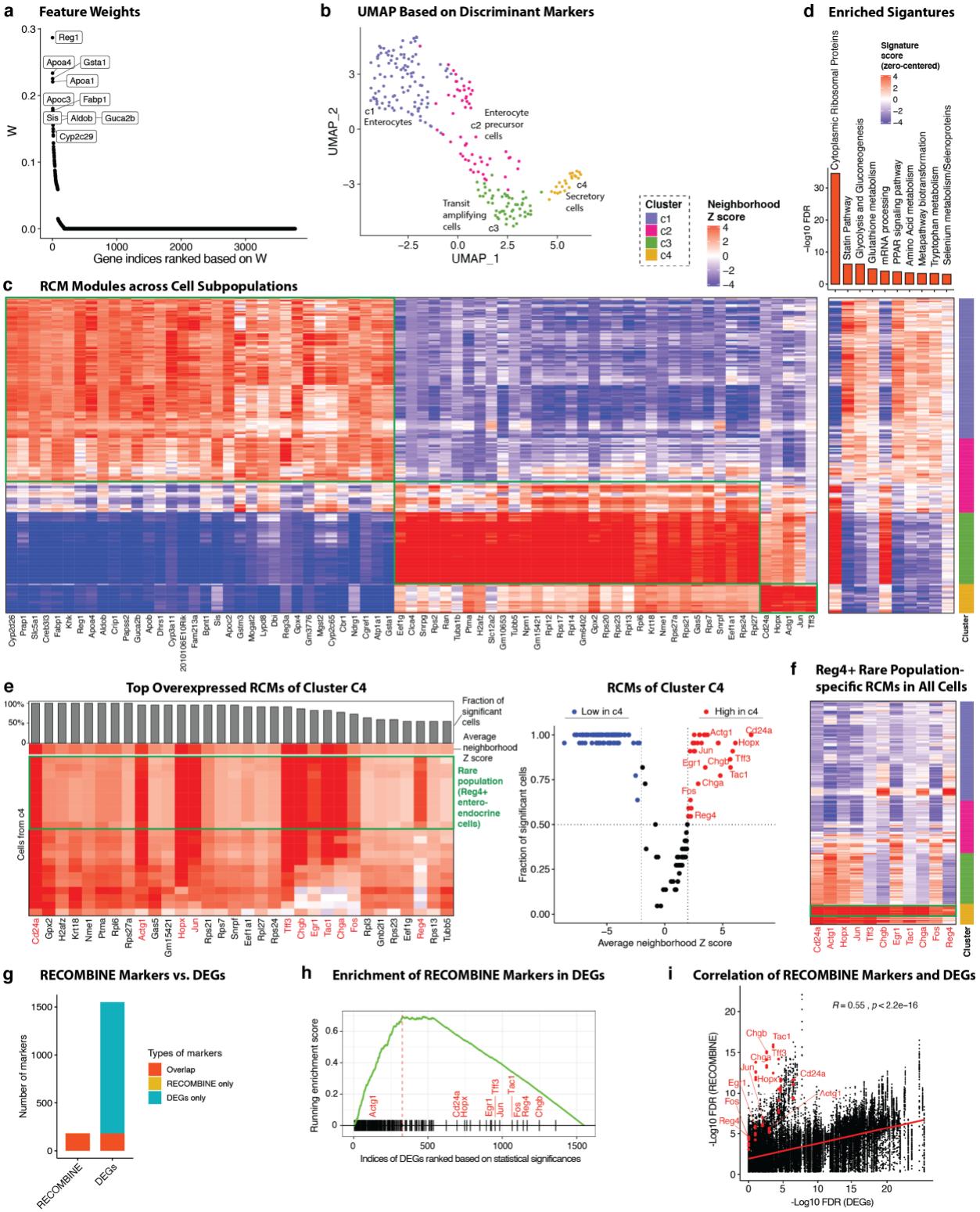

Fig. 4 | RECOMBINE identifies RCMs of both major and rare cell subpopulations of intestine. ***a,*** *Feature weights of all genes where the top 10 discriminant markers are labeled.* ***b,*** *UMAP showing clusters identified by Leiden method based on discriminant markers.* ***c,*** *Heatmap of neighborhood Z scores showing RCM modules across clusters. The RCMs shown here are top*

20 significantly enriched markers in more than 50% of cells in each cluster. Cells are hierarchically clustered within each Leiden cluster based on discriminant marker expressions. Green boxes indicate RCM modules. ***d,*** *Enrichment of mouse pathway signatures in discriminant markers (top) and heatmap of signature scores across cells (bottom). Top 10 significantly enriched signatures are shown. In the heatmap, cells are ordered according to the hierarchical clustering in Fig. 1c.* ***e,*** *RCMs of a rare subpopulation within cluster c4. Left panel: heatmap showing cluster c4-specific RCMs. The top bar plot shows the fraction of cells having significantly high expression of each marker. The green box indicates a rare subpopulation of cells characterized by RCMs including Cd24a, Actg1, Hopx, Jun, Tff3, Chgb, Egr1, Tac1, Chga, Fos, and Reg4 (highlighted in red). Right panel: volcano plot of cluster c4-specific RCMs.* ***f,*** *Heatmap of neighborhood Z scores showing the rare Reg4$^+$ subpopulation's RCMs across all cells.* ***g,*** *Comparison of RECOMBINE markers and DEGs. DEGs were obtained by Leiden clustering of all cells based on all genes followed by Wilcoxon test for each cluster with respect to the rest of the clusters. RECOMBINE markers were statistically significant RCMs with false discovery rate (FDR) < 0.05, and DEGs were filtered based on absolute log_2 fold-change > 0.25 and FDR < 0.05.* ***h,*** *Enrichment plot of RECOMBINE markers in the list of DEGs ranked by their significance from Wilcoxon tests.* ***i,*** *Comparison of FDRs between RECOMBINE and DEGs for all cells. As FDRs of DEGs were obtained at the cluster level, we used the same FDRs for cells within the same cluster. The rare Reg4$^+$ subpopulation's RCMs identified in (e) are highlighted in red in (h and i).*

Therefore, these three genes could not selectively mark the Reg4$^+$ enteroendocrine cells. To fully characterize Reg4$^+$ enteroendocrine cells and discriminate from other subpopulations, additional markers, particularlyTff3 or Tac1, should be included, as detected by RECOMBINE (Fig. 4e-f).

To compare RECOMBINE and DEG analysis, a commonly used approach to find cluster-specific markers, we interrogated the overlap and difference between RECOMBINE markers and DEGs. All but one of the RECOMBINE markers were captured in DEGs, but the number of RECOMBINE markers (N=183) was substantially smaller than that of DEGs (N=1,550) (Fig. 4g). While the RECOMBINE markers were enriched in more significant DEGs, some RECOMBINE markers, including nine of the 10 characteristic markers of Reg4$^+$ enteroendocrine cells, were statistically less significant DEGs (Fig. 4h-i). For example, Reg4 was ranked toward the less significant region of DEGs, but it had much greater statistical significance based on RECOMBINE analysis. In conclusion, compared with DEGs, RECOMBINE selects a smaller set of markers with greater detection sensitivity for the rare population of Reg4$^+$ enteroendocrine cells.

RECOMBINE identifies gene programs of inter- and intra-tumoral heterogeneity

To study gene programs of inter-tumoral heterogeneity across cancer patients, we applied RECOMBINE to an scRNA-seq dataset of 1,252 tumor cells from 14 melanoma patients[41]. RECOMBINE identified 269 discriminant markers (Fig. 5a, Supplementary Fig. 5, and Supplementary Table 5). The most discriminant markers were related to melanocyte lineage (PMEL, DCT, MIA, and TYR), cell migration (SERPINA3 and A2M), lipid metabolism (APOC2 and APOD), and immune activation (IFI27 and IFITM1). Through Leiden clustering based on expressions of discriminant markers, we identified 10 clusters that mainly separated individual patients (Fig. 5b-c). The discriminant markers were significantly enriched in pathways such as coagulation, epithelial-mesenchymal transition (EMT), Myc targets and immune signatures involving IFNα/IFNγ responses as well as complement activity (overrepresentation analysis with hypergeometric test, FDR < 0.001, Fig. 5d). Interestingly, we observed a mutually exclusive pattern between Myc vs. immune signatures (Fig. 5d; Pearson correlation: -0.67, p < 1.0e-5, Fig. 5e). A validation with the TCGA melanoma dataset[42] confirmed the negative association between

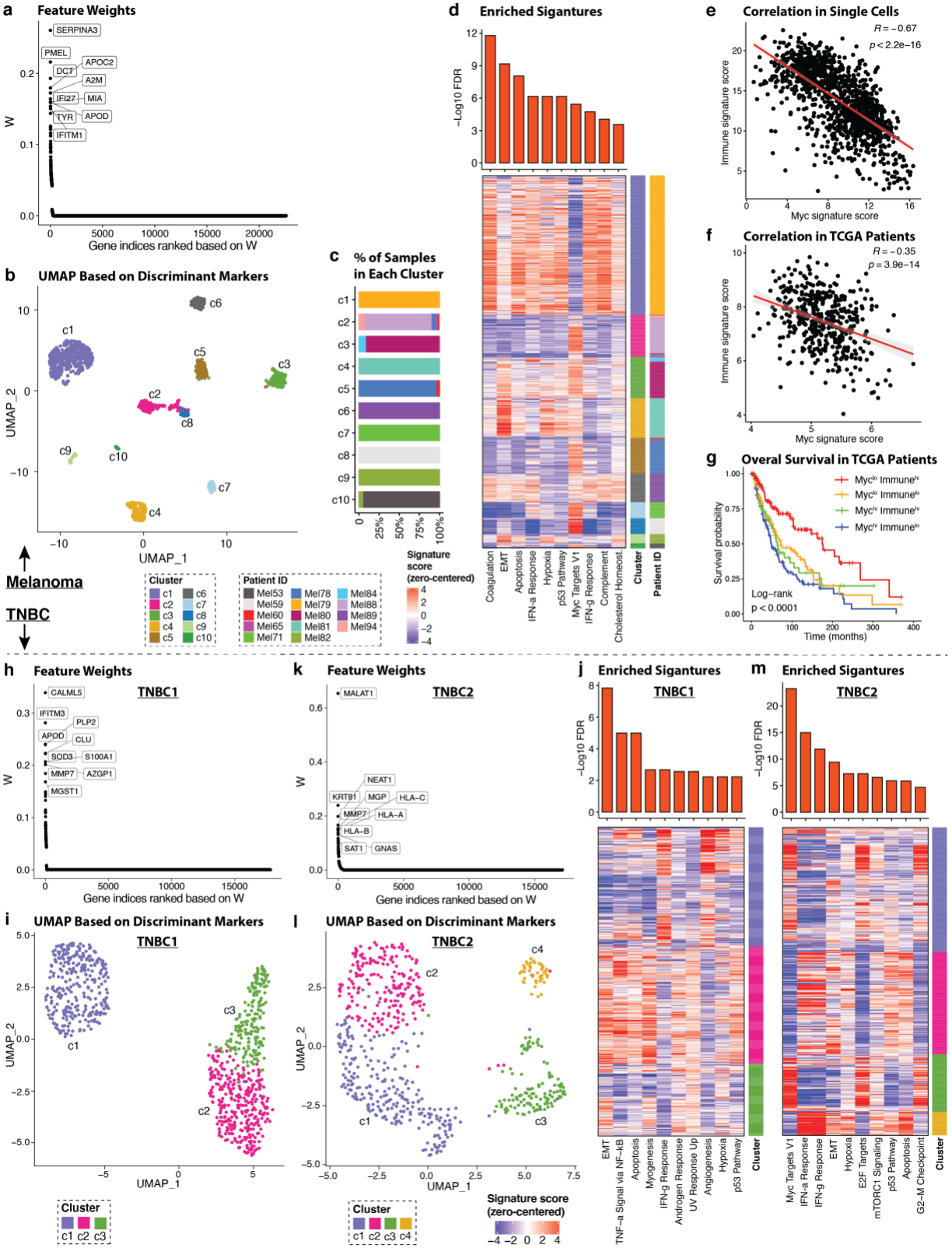

Fig. 5 | RECOMBINE reveals expression programs of inter- and intra-tumoral heterogeneity. *Analysis of malignant cells from 14 melanoma patients (a-g) and two triple-negative breast cancer (TNBC) patients (h-j: TNBC1; k-m: TNBC2).* ***a, h*** *and* ***k,*** *Feature weights of all genes where the*

top 10 discriminant markers are labeled. ***b, i*** *and* ***l,*** *UMAP showing clusters identified by Leiden method based on discriminant markers.* ***c,*** *Composition of cells from patients in each cluster.* ***d, j*** *and* ***m,*** *Enrichment of MSigDB hallmark signatures in discriminant markers (top) and heatmap of signature scores across cells (bottom). Top 10 significantly enriched signatures are shown. In the heatmap, cells are hierarchically clustered within each Leiden cluster based on discriminant marker expressions.* ***e,*** *Correlation plot of Myc and immune signature scores across cells. The red line shows linear regression.* ***f,*** *Correlation plot of Myc and immune signature scores calculated from bulk mRNA expressions across 443 melanoma patients (TCGA).* ***g,*** *Overall survival of TCGA melanoma patients stratified by Myc and immune signature scores.*

the two signatures (Pearson correlation: -0.37, p <1.0e-5, Fig. 5f). A survival analysis suggested that low Myc and high immune signatures were associated with prolonged survival in melanoma (p < 0.001). This finding is consistent with the previous findings that Myc upregulates immunosuppressive pathways and that targeting Myc enhances antitumor immune responses in checkpoint inhibition therapy[43,44]. Taken together, these results reveal that RECOMBINE identified genes and gene programs associated with inter-tumoral heterogeneity of melanoma patients.

Next, we applied RECOMBINE to characterize and compare intra-tumoral heterogeneity of gene expression programs in two TNBC patients based on scRNA-seq data from malignant cells (patient 1: 797 cells, patient 2: 620 cells)[45]. Through RECOMBINE analysis, we identified the discriminant markers for each patient (patient 1: 124 markers, patient 2: 414 markers) (Fig. 5h-i and Supplementary Fig. 5). Using Leiden clustering with the discriminant markers, we identified three clusters for patient 1 and four clusters for patient 2 (Fig. 5j-k). The discriminant markers from the two patients shared enrichment with EMT and interferon response (Fig. 5l-m). Each patient, however, carried unique subpopulations marked with apoptosis, androgen response, and TNFα signaling in patient 1 and a unique Myc signature–enriched subpopulation in patient 2 (hypergeometric test, FDR < 0.01 for both patients). The malignant cells displayed hierarchical gene programs across clusters and between cells within clusters. For example, for patient 1, c1 had low EMT scores and c2 had high EMT scores while c3 had at least two cell states that existed with either high or low EMT scores. Taken together, these studies show that RECOMBINE selects genes and gene programs underlying intra-tumoral heterogeneity of TNBC patients who carry both shared and unique transcriptomic signatures.

Discussion

Here we present the development and applications of RECOMBINE, a computational framework for unbiased selection of RCMs that discriminate cell subpopulations with high granularity. RECOMBINE is a semi-supervised statistical learning method consisting of two steps. In the first step, RECOMBINE employs our novel SHC-SSL algorithm to select a compact set of discriminant markers. As the marker selection is coupled with hierarchical clustering of cells, the selected markers can discriminate both cell types and states with high granularity and continuity. This is different from the partition-based methods, including DEGs, which find markers separating pre-specified clusters without considering the structure of tissue and cell hierarchy. The simultaneous optimization of marker selection and hierarchical clustering empowers RECOMBINE to detect markers at all levels of cell and tissue hierarchy, including distinct cell types, continuous cell states, and rare cell subpopulations. In the second step, RECOMBINE employs neighborhood recurrence tests to extract RCMs at both cell and subpopulation levels. The subpopulations may be defined in a supervised way, either by user's interests or by clustering of cells. For any cell subpopulation, RECOMBINE maps the RCMs using neighborhood Z scores and statistical significances for marker enrichment within the subpopulation. In summary,

RECOMBINE identifies RCMs that characterize cell subpopulations with high granularity and map the hierarchical structure of cell states within tissues.

We established RECOMBINE through systematic benchmarking and applications in diverse biological domains. Benchmarking using simulation data revealed that SHC-SSL outperforms other SHC algorithms for both optimal marker selection and accurate clustering. Benchmarking with bulk-level omics data demonstrated the method's ability to capture canonical breast cancer subtypes. Applied to scRNA data of mouse visual cortex, RECOMBINE selected discriminant markers of cell types and states. Importantly, these markers successfully discriminated spatially resolved cell subpopulations of the same tissue from a STARmap dataset, demonstrating RECOMBINE's excellent performance of marker selection for targeted spatial transcriptomics. When applied to scRNA data from pan-cancer CD8 T cells, RECOMBINE identified RCMs of CD8 T cell states, $GZMK^{+}HAVCR2^{-}$ EM cells associated with anti–PD-1 therapy response, and novel markers related to CD8 T cell exhaustion. In an scRNA dataset from mouse intestine tissue, RECOMBINE successfully identified RCMs of common cell subpopulations and outperformed DEGs for characterizing a rare subpopulation of $Reg4^{+}$ enteroendocrine cells with a minimal set of composite markers. Finally, in the applications to scRNA data from melanoma and triple-negative breast tumors, RECOMBINE uncovered gene programs of inter- and intra-tumoral heterogeneity. In conclusion, RECOMBINE is useful in diverse applications for data-driven extraction of RCMs to characterize cell subpopulations based on single-cell omics data.

RECOMBINE has several key algorithmic advances compared to other methods. First, RECOMBINE doesn't require a pre-specified marker panel size but instead finds an optimal value through a statistically principled and data-driven approach (i.e., gap statistic). In contrast, for example, SCMER[46] requires an input of the marker panel size, whose optimal value is often *a priori* unknown. This is important because an arbitrarily chosen marker panel size is likely to be suboptimal for discriminating and identifying cell subpopulations. Second, RECOMBINE enforces sparsity of discriminant markers through a spike-and-slab lasso penalty function, which has better performance of marker selection than the lasso penalty function that is used in SHC and other methods such as scGeneFit[47] and RANKCORR[48]. Third, RECOMBINE performs neighborhood enrichment tests to map RCMs for cell subpopulations based on discriminant markers. This approach allows for the generation of meaningful biological interpretations of cell types and states. Finally, an advantage of RECOMBINE is the selection of an optimized, compact set of discriminant markers. Compared to DEGs, RECOMBINE can select a substantially compact set of markers with high sensitivity. This compactness leads to easy and better interpretation of the marker list, but it comes with a cost that the list may not be comprehensive. As in our benchmark study, SHC-SSL had a median precision of 100% but a median recall of 79%. The imperfect recall can be explained by the fact that some markers may be missed if the selected markers reach the maximum possible discrimination of cell subpopulations.

RECOMBINE is a scalable algorithm that can be used to characterize hundreds of thousands of cells as exemplified in the CD8 T cell analysis (N_{cells}=234,550). To reach such scalability, we have overcome the challenge that RECOMBINE's marker selection algorithm, SHC-SSL, operates with a dissimilarity matrix of dimension n^2 by p, (n: cells, p: features), leading to quadratic increases with n in memory use. We addressed this challenge through merging cells with shared transcriptomic profiles into pseudo-cells and identifying the markers that discriminate the pseudo-cells. Provided that pseudo-cells are formed from fine-grained cell groups, the pseudo-cell strategy leads to an approximate solution for selecting discriminant markers of individual cells. An added advantage of the pseudo-cell strategy is its ability to reduce technical noise and stabilize variance, which can contribute to more robust marker selection. Notably, similar to the pseudo-cell strategy used here, the MetaCell[18] and SEACells[19] methods were

developed to construct highly granular and distinct cell groups to overcome sparsity of single-cell data while preserving the fine-grained heterogeneity of cell states and scaling up to large datasets. The outputs from MetaCell and SEACells methods, which involve the granular cell group representations, can also be integrated to RECOMBINE as an alternative to pseudo-cell strategy. Such representations can serve as an input to RECOMBINE. With these strategies as well as further refinements, we expect RECOMBINE to be highly scalable for applications in large-scale projects such as the Human Cell Atlas[49].

RECOMBINE is a generalized framework for unbiased selection of composite markers to characterize biological subpopulations. The applications of RECOMBINE, as demonstrated here in diverse examples, can lead to discoveries on how multicellular systems are organized into distinct and continuous states that mediate biological functions. The RECOMBINE results in composite markers of cell subpopulations and may detect the drivers of biological functions. One example is the detection of immune cell-state specific marker sets for cell subpopulations that may drive resistance to immunotherapy. In addition, the RECOMBINE-selected composite markers can guide new assays for detection of unique processes and cell types and states within tissues based on single cell omics experiments. Although our studies focused on single cell expression profiles, RECOMBINE can be applied to any type of omics data modality, either in single-cell or bulk resolution, including protein expressions, epigenomic profiles, and genomic alterations. Therefore, we expect RECOMBINE will have a broad impact on diverse biological analyses. Through the discovery of novel composite drivers of biological functions, the ability to characterize both common and rare cell subpopulations in complex tissues, and the guidance of strategies to manipulate biological processes (e.g., selection of therapeutic interventions), RECOMBINE can have a transformative impact on our understanding and ability to manipulate functions in biological systems.

Methods

Overview of RECOMBINE

Given single-cell omics data, RECOMBINE uncovers RCMs for cell subpopulations with high granularity in two steps, where RCMs are composed of an optimized set of markers that delineate hierarchical cell subpopulations. In the first step, RECOMBINE employs the SHC-SSL algorithm to select the markers that best discriminate cells while performing hierarchical clustering. The discriminant markers consist of a minimal set that characterize the data and can be subjected to over-representation analysis for biological interpretation. Besides hierarchical clustering, the cell dissimilarity matrix based on the discriminant markers can be used for partitional clustering (e.g., Leiden clustering) to group cells into discrete subpopulations. In the second step, RECOMBINE uses neighborhood recurrence tests to extract RCMs from the discriminant markers. Specifically, the cell dissimilarity matrix is used to build a K nearest neighbor graph of cells. For any individual cell, its nearest neighbors compose its local neighborhood context. Then, a neighborhood Z score is calculated to quantify the recurrence of each of the discriminant markers for each cell's local neighborhood. Finally, based on statistical significances of neighborhood Z scores, RCMs are identified at the single cell level; based on fractions of cells having significant neighborhood Z scores, RCMs are identified at the subpopulation level.

In below sections, we first describe SHC-based algorithms by reviewing the previously developed SHC algorithm and introducing SHC-SSL and another related algorithm that we developed, SHC-FL. As these algorithms involve hyperparameters, we next describe a method for hyperparameter selection. Then, we give details of simulation data and metrics for benchmarking the three algorithms. Next, as SHC-SSL is computationally expensive when the sample size is large, we describe a pseudo-cell strategy for reducing its computational cost. Then, we introduce a neighborhood recurrence test for extracting RCMs from discriminant markers. Finally, we give details of applying RECOMBINE to various biological datasets.

Here, we describe some notations. For any vector $\mathbf{v} \in \mathbb{R}^n$, $\|\mathbf{v}\|$ denotes the L2 norm (Euclidian norm) of $\mathbf{v}$, and $\|\mathbf{v}\|_1$ denotes the L1 norm of $\mathbf{v}$.

Sparse hierarchical clustering

First, we summarize the previously developed SHC algorithm[21], which we applied to identify recurrent oncogenic co-alterations[50]. Given an $n \times p$ data matrix $\mathbf{X}$ with n samples by p features, we denote x_i as a sample that is a p-dimensional vector indexed by i, where $1 \leq i \leq n$ and $d(x_i, x_{i'})$ as a measure of the dissimilarity between samples x_i and $x_{i'}$. Assuming the dissimilarity measure is additive in features, that is, $d(x_i, x_{i'}) = \sum_j d_{i,i',j}$, where $d_{i,i',j}$ is the dissimilarity measure between samples x_i and $x_{i'}$ along feature j, the dissimilarity matrix becomes the solution of a constrained optimization problem:

$$\mathbf{U}^* = \underset{\mathbf{U}}{\operatorname{argmax}} \left\{ \sum_j \sum_{i,i'} d_{i,i',j} U_{i,i'} \right\} \text{ subject to } \sum_{i,i'} U_{i,i'}^2 \leq 1.$$

It can be shown as the optimal solution $U_{i,i'}^* \propto \sum_j d_{i,i',j}$. Thus, with this reformulation, hierarchical clustering performed on $\mathbf{U}^*$ is equivalent to that performed on $\{d(x_i, x_{i'})\}_{n \times n}$. To calculate $d_{i,i',j}$, one could use the squared Euclidean distance $d_{i,i',j} = \left(X_{ij} - X_{i'j}\right)^2$ or the absolute difference $d_{i,i',j} = \left|X_{ij} - X_{i'j}\right|$.

The weighted dissimilarity matrix was introduced in the SHC method. SHC provides a means to select a small number of discriminant features that contribute to a weighted dissimilarity matrix. Specifically, SHC solves the optimization problem

$$\underset{\mathbf{w},\mathbf{U}}{\text{maximize}}\left\{\sum_j w_j \sum_{i,i'} d_{i,i',j} U_{i,i'}\right\},$$

$$\text{subject to} \sum_{i,i'} U_{i,i'}^2 \leq 1, \qquad \|\mathbf{w}\|^2 \leq 1, \qquad \|\mathbf{w}\|_1 \leq s, \qquad w_j \geq 0 \ \ \forall j,$$

where $\mathbf{w}$ is a weight vector for features and s is a tuning parameter that controls sparsity of $\mathbf{w}$. When s is small, the lasso, or L1, penalty of $\mathbf{w}$ encourages many elements of $\mathbf{w}$ to be zero. The L2 penalty of $\mathbf{w}$ is necessary because without it, only one element of $\mathbf{w}$ would be nonzero. Let $\mathbf{U}^{**}$ optimize the above criterion, from which we have $U_{i,i'}^{**} \propto \sum_j w_j d_{i,i',j}$, where the dissimilarity along each feature is weighted according to w_j. Due to these constraints, a small value of s makes w sparse, so $\mathbf{U}^{**}$ depends only on a small subset of features. Therefore, clustering based on $\mathbf{U}^{**}$ depends only on the selected features.

To simplify representation, let's rearrange $\{d_{i,i',j}\}_{n\times n\times p}$ to get a matrix $\mathbf{D} \in \mathbb{R}^{n^2\times p}$ in which column j consists of ordered elements $\{d_{i,i',j}\}_{n\times n}$ stacked into a vector. Accordingly, let $\mathbf{u} \in \mathbb{R}^{n^2}$ be the vector by stacking $\{U_{i,i'}\}_{n\times n}$. Then, the SHC optimization criterion is in the form

$$\underset{\mathbf{w},\mathbf{u}}{\text{minimize}}\{-\mathbf{u}^T\mathbf{D}\mathbf{w}\},$$

$$\text{subject to } \|\mathbf{u}\|^2 \leq 1, \qquad \|\mathbf{w}\|^2 \leq 1, \qquad \|\mathbf{w}\|_1 \leq s, \qquad w_j \geq 0 \ \ \forall j.$$

We can iteratively optimize $\mathbf{u}$ and $\mathbf{w}$. With $\mathbf{w}$ fixed, this criterion is equivalent to

$$\underset{\mathbf{u}}{\text{minimize}}\{-\mathbf{u}^T\mathbf{D}\mathbf{w}\} \ \text{ subject to } \|\mathbf{u}\|^2 \leq 1.$$

The optimal $\mathbf{u} = \frac{\mathbf{D}\mathbf{w}}{\|\mathbf{D}\mathbf{w}\|}$, and $u_i \geq 0$ for $1 \leq i \leq n^2$ since all the elements of $\mathbf{D}$ and $\mathbf{w}$ are nonnegative. With $\mathbf{u}$ fixed, the criterion takes the form

$$\underset{\mathbf{w}}{\text{minimize}}\{-\mathbf{u}^T\mathbf{D}\mathbf{w}\} \ \text{ subject to } \|\mathbf{w}\|^2 \leq 1, \qquad \|\mathbf{w}\|_1 \leq s,$$

where we drop the non-negativity constraint because it is redundant when all the elements of $\mathbf{u}$ and $\mathbf{D}$ are non-negative. The optimal $\mathbf{w} = \frac{S(\mathbf{D}^T\mathbf{u},\ \Delta)}{\|S(\mathbf{D}^T\mathbf{u},\ \Delta)\|}$, where S is the soft-thresholding operator $S(x,c) = \text{sign}(x)(|x| - c)_+$, and $\Delta = 0$ if this results in $\|\mathbf{w}\|_1 \leq s$; otherwise, $\Delta > 0$ is chosen such that $\|\mathbf{w}\|_1 = s$.

Sparse hierarchical clustering with spike-and-slab lasso

Here, we introduce the SHC-SSL algorithm, which uses a SSL penalty. We start with a spike-and-slab Laplace prior distribution of $\mathbf{w}$, from which an SSL penalty function follows[51,52]. Suppose the prior distribution of $\mathbf{w}$ is in the form

$$\pi(\mathbf{w} \mid \gamma) = \prod_{j=1}^{p}\left[(1-\gamma_j)\psi_0(w_j) + \gamma_j\psi_1(w_j)\right],$$

$$\psi_0(w) = \frac{\lambda_0}{2}e^{-\lambda_0|w|},$$

$$\psi_1(w) = \frac{\lambda_1}{2}e^{-\lambda_1|w|},$$

where $\boldsymbol{\gamma} := (\gamma_j)_{1\le j\le n}$, $\gamma_j \in \{0,1\}$. Each w_j is a mixture of two Laplace distributions, ψ_0 and ψ_1, and γ_j is the latent variable indicating the mixture components to which w_j belongs. With λ_0 large, $\psi_0(w)$ serves a restrictive, point-mass-like "spike distribution" for modeling a small w; in contrast, with λ_1 small, $\psi_1(w)$ serves a diffusive, heavy-tailed "slab distribution" for modeling a large w. Suppose the latent variable $\boldsymbol{\gamma}$ has a hierarchical prior distribution as

$$\pi(\boldsymbol{\gamma} \mid \theta) = \prod_{j=1}^{p} \theta^{\gamma_j}(1-\theta)^{1-\gamma_j},$$

$$\pi(\theta) = \mathrm{Beta}(a,b),$$

where θ is shared across γ_j's and $\mathrm{Beta}(a,b)$ is the beta distribution with shape parameters $a, b > 0$. As we can see, $\pi(\gamma_j = 1 \mid \theta) = \theta$; thus, θ acts as the prior probability of w_j being large. Since θ is shared across w_j's, intuitively θ reflects the expected fraction of large w_j's. By marginalizing out $\boldsymbol{\gamma}$, we obtain

$$\pi(\mathbf{w} \mid \theta) = \prod_{j=1}^{p} \left[(1-\theta)\psi_0(w_j) + \theta\psi_1(w_j)\right].$$

Furthermore, by marginalizing out θ, we have

$$\pi(\mathbf{w}) = \int_0^1 \prod_{j=1}^{p} \left[(1-\theta)\psi_0(w_j) + \theta\psi_1(w_j)\right] d\pi(\theta).$$

With this prior marginal distribution of $\mathbf{w}$, we define the SSL penalty as

$$\mathrm{pen}_{SSL}(\mathbf{w}) = -\log \frac{\pi(\mathbf{w})}{\pi(\mathbf{0}_p)},$$

which is centered so that $\mathrm{pen}_{SSL}(\mathbf{0}) = 0$. Note that our definition of the SSL penalty is the negative of the one described by Ročková and George[51]. We choose this definition because it is consistent with the definition of the lasso penalty that is the negative logarithm of the Laplace distribution. When $\lambda_0 \to \infty$ and $\lambda_1 \to 0$, the SSL penalty is equivalent to an L0-norm penalty; when $\lambda_0 = \lambda_1$, it is equivalent to a lasso penalty. Therefore, the SSL penalty forms a continuous bridge between an L0-norm and a lasso penalty. We keep λ_1 to a small constant and tune λ_0 during hyperparameter selection.

Now, we replace the lasso with the SSL penalty in the objective of SHC, yielding the optimization criterion as

$$\underset{\mathbf{w},\mathbf{u}}{\mathrm{minimize}}\{-\mathbf{u}^T\mathbf{D}\mathbf{w} + \mathrm{pen}_{SSL}(\mathbf{w})\},$$

$$\text{subject to } \|\mathbf{u}\|^2 \le 1, \qquad \|\mathbf{w}\|^2 \le 1, \qquad w_j \ge 0 \ \ \forall j.$$

We iteratively optimize $\mathbf{u}$ and $\mathbf{w}$. With $\mathbf{w}$ fixed, the optimal $\mathbf{u}$ is calculated in the same way as the one in SHC. With $\mathbf{u}$ fixed, the SHC-SSL objective takes the form

$$\underset{\mathbf{w}}{\mathrm{mimimize}}\{-\mathbf{u}^T\mathbf{D}\mathbf{w} + \mathrm{pen}_{SSL}(\mathbf{w})\} \ \text{ subject to } \|\mathbf{w}\|^2 \le 1.$$

To find stationary points of the above objective, we rewrite it as a Lagrangian function

$$\mathcal{L}(\mathbf{w},\alpha) = -\mathbf{u}^T\mathbf{D}\mathbf{w} + \frac{1}{2}\alpha\|\mathbf{w}\|^2 + \mathrm{pen}_{SSL}(\mathbf{w}),$$

where $\alpha \geq 0$ is the Lagrange multiplier and a factor of $1/2$ is included for later convenience. Following the sub-differential calculus, we obtain

$$\frac{\partial \mathcal{L}(\mathbf{w}, \alpha)}{\partial w_j} = -y_i + \alpha w_j + \frac{\partial \mathrm{pen}_{SSL}(\mathbf{w})}{\partial |w_j|} h(w_j),$$

where $\mathbf{y} = \mathbf{D}^T \mathbf{u}$ and $h(w)$ is a subgradient of $|w|$, meaning $h(w) = \mathrm{sign}(w)$ when $w \neq 0$, and $h(w) \in [-1, 1]$ when $w = 0$.

Let's present the derivative of the SSL penalty. Using Lemma 2 and 3 described by Ročková and George[51], the derivative of the SSL penalty with respect to $|w_j|$ is in the form

$$\frac{\partial \mathrm{pen}_{SSL}(\mathbf{w})}{\partial |w_j|} = \lambda^*(w_j;\ \mathbf{w}_{\backslash j}),$$

where

$$\lambda^*(w_j;\ \mathbf{w}_{\backslash j}) = \left(1 - p^*(w_j;\ \mathbf{w}_{\backslash j})\right) \lambda_0 + p^*(w_j;\ \mathbf{w}_{\backslash j}) \lambda_1,$$

$$p^*(w_j;\ \mathbf{w}_{\backslash j}) = \frac{\theta_j \psi_1(w_j)}{(1 - \theta_j)\psi_0(w_j) + \theta_j \psi_1(w_j)},$$

$$\theta_j = \mathrm{E}[\theta \mid \mathbf{w}_{\backslash j}],$$

where $\mathrm{E}[\theta \mid \mathbf{w}_{\backslash j}]$ is the posterior expectation of θ given $\mathbf{w}_{\backslash j}$. Here, $\lambda^*(w_j;\ \mathbf{w}_{\backslash j})$ can be interpreted as a weighted average of λ_0 and λ_1, and the weight $p^*(w_j;\ \mathbf{w}_{\backslash j})$ is the conditional probability of w_j being from $\psi_1(w_j)$ rather than from $\psi_0(w_j)$. Thus, a larger w_j tends to have a smaller $\lambda^*(w_j;\ \mathbf{w}_{\backslash j})$. Both $\lambda^*(w_j;\ \mathbf{w}_{\backslash j})$ and $p^*(w_j;\ \mathbf{w}_{\backslash j})$ depend on θ_j. To calculate $\theta_j = \mathrm{E}[\theta \mid \mathbf{w}_{\backslash j}]$, note that when p is large, $\mathrm{E}[\theta \mid \mathbf{w}_{\backslash j}]$ should be similar to $\mathrm{E}[\theta \mid \mathbf{w}]$. From Lemma 4 in Ročková and George[51], θ_j can be approximated by

$$\theta_j \approx \mathrm{E}[\theta \mid \mathbf{w}] \approx \frac{a + \|\mathbf{w}\|_0}{a + b + p},$$

where $\|\mathbf{w}\|_0$ is the L0 norm of $\mathbf{w}$.

Now, we solve for the optimal $\mathbf{w}$. Substituting the derivative of $\mathrm{pen}_{SSL}(\mathbf{w})$, the derivative of $\mathcal{L}(\mathbf{w}, \alpha)$ can be expressed as

$$\frac{\partial \mathcal{L}(\mathbf{w}, \alpha)}{\partial w_j} = -y_i + \alpha w_j + \lambda^*(w_j;\ \mathbf{w}_{\backslash j}) h(w_j).$$

Let $\widehat{\mathbf{w}}$ denote an estimate of the global optimum of $\mathbf{w}$. At stationary points, with the Karush-Kuhn-Tucker conditions, $\widehat{\mathbf{w}}$ needs to satisfy

$$y_i - \alpha \widehat{w}_j = \lambda^*(\widehat{w}_j;\ \widehat{\mathbf{w}}_{\backslash j}) \mathrm{sign}(\widehat{w}_j) \quad \text{for } w_j \neq 0,$$

$$|y_i - \alpha \widehat{w}_j| \leq \lambda^*(\widehat{w}_j;\ \widehat{\mathbf{w}}_{\backslash j}) \quad \text{for } w_j = 0.$$

This can be written equivalently as

$$\widehat{w}_j = \frac{1}{\alpha} S\left(y_i, \lambda^*(\widehat{w}_j;\ \widehat{\mathbf{w}}_{\backslash j})\right).$$

where the soft thresholding operator $S(x, c) = \mathrm{sign}(x)(|x| - c)_+$. At optimum, α is chosen such that $\|\widehat{\mathbf{w}}\|^2 = 1$.

It should be noted that the above necessary conditions are not sufficient for the global mode under the SSL penalty[51]. Using Theorem 3.1 from Ročková[52], the global mode $\widehat{\mathbf{w}}$ under the SSL penalty can be expressed as

$$\widehat{w}_j = \frac{1}{\alpha}\tilde{S}\big(y_i, \lambda^*(\widehat{w}_j;\ \widehat{\mathbf{w}}_{\backslash j}), \Delta\ \big),$$

$$\Delta \approx \sqrt{-2\log\big[p^*\big(0;\ \widehat{\mathbf{w}}_{\backslash j}\big)\big]} + \lambda_1,$$

where $\tilde{S}$ is a generalized thresholding operator defined as $\tilde{S}(x, c, \Delta) = \mathrm{sign}(x)(|x| - c)_+ \mathbb{I}(|x| > \Delta)$, and α is chosen such that $\|\widehat{\mathbf{w}}\|^2 = 1$. We can think of $\tilde{S}$ as a composition operator consisting of an L0-type hard thresholding operator $S_0(x, \Delta) = x\mathbb{I}(|x| > \Delta)$ followed by a lasso-type soft thresholding operator $S_1(x, c) = \mathrm{sign}(x)(|x| - c)_+$. Both operators adaptively depend on the value of $\widehat{\mathbf{w}}$. Each coordinate of $\widehat{\mathbf{w}}$ has its own self-adaptive shrinkage, which borrows information across other coordinates. The hard thresholding operator S_0 depends on Δ, which in turn depends on $p^*\big(0;\ \widehat{\mathbf{w}}_{\backslash j}\big)$, the conditional probability of $\widehat{w}_j$ being zero is from the slab distribution. A larger $p^*\big(0;\ \widehat{\mathbf{w}}_{\backslash j}\big)$, corresponding to a smaller Δ, induces a smaller hard thresholding effect on $\widehat{w}_j$ (i.e., S_0 tends to leave $\widehat{w}_j$ as is). In contrast, a larger hard thresholding effect is induced if $\widehat{w}_j$ being zero is likely from the spike distribution (i.e., S_0 tends to set $\widehat{w}_j$ as 0). On the other hand, the soft thresholding operator S_1 adaptively depends on $\lambda^*\big(\widehat{w}_j;\ \widehat{\mathbf{w}}_{\backslash j}\big)$. As we discussed before, a larger $\widehat{w}_j$ has a smaller $\lambda^*\big(\widehat{w}_j;\ \widehat{\mathbf{w}}_{\backslash j}\big)$. When $\widehat{w}_j$ is large, it is shrunk by a small amount (close to λ_1); when $\widehat{w}_j$ is small, it is shrunk by a large amount (close to λ_0). Therefore, in contrast to the constant shrinkage of the lasso penalty, the SSL penalty induces a self-adaptive shrinkage that depends on itself and borrows information from other coordinates. In other words, while lasso constantly biases $\widehat{w}_j$ toward zero[53], SSL shrinks strongly small $\widehat{w}_j$'s toward zeros but only introduces very slight biases for large $\widehat{w}_j$'s. For example, as shown in Fig. 1c, by fixing $\lambda_1 = 0.001$ and increasing λ_0 from 1 to 50, the SSL penalty forms a continuum between a lasso and an L0-norm penalty. When $\lambda_0 = 1$, the SSL penalty function is hardly distinguishable from lasso. When $\lambda_0 = 50$, the SSL penalty function has a sharp transition from zero to non-zero weights of features, approaching the L0-norm penalty that is known to be ideal for feature selection.

SHC-SSL algorithm

Input: stacked dissimilarity matrix $\mathbf{D}$, shape parameters a and b, and hyperparameter λ_0 and λ_1, where λ_1 is a small number and $\lambda_0 > \lambda_1$.

Output: feature weights $\mathbf{w}$, dissimilarity matrix $\mathbf{U}$.

(1) Initialize $\mathbf{w}$ as $w_1 = \cdots = w_p = \frac{1}{\sqrt{p}}$.

(2) Iterate until convergence of $\mathbf{w}$:

(a) Update $\mathbf{u} = \frac{\mathbf{D}\mathbf{w}}{\|\mathbf{D}\mathbf{w}\|}$.

(b) Update $\mathbf{w}$:

(i) Calculate $\theta = \frac{a + \|\mathbf{w}\|_0}{a + b + p}$.

(ii) Calculate $\mathbf{p}^* = \frac{\theta\psi_1(\mathbf{w})}{(1-\theta)\psi_0(\mathbf{w}) + \theta\psi_1(\mathbf{w})}$, where $\psi_0\big(w_j\big) = \frac{\lambda_0}{2}e^{-\lambda_0|w_j|}$ and $\psi_1\big(w_j\big) = \frac{\lambda_1}{2}e^{-\lambda_1|w_j|}$ for $1 \leq j \leq p$.

(iii) Calculate $\boldsymbol{\lambda}^* = (\mathbf{1} - \mathbf{p}^*)\lambda_0 + \mathbf{p}^*\lambda_1$.

(iv) Calculate $\Delta = \sqrt{-2\log[\mathbf{p}^*]} + \lambda_1$.

(v) Update $\mathbf{w} = \frac{\tilde{S}\left(\mathbf{D}'^T\mathbf{u}, \boldsymbol{\lambda}^*, \Delta\right)}{\left\|\tilde{S}\left(\mathbf{D}'^T\mathbf{u}, \boldsymbol{\lambda}^*, \Delta\right)\right\|}$ where $\tilde{S}(x, c, \Delta) = \mathrm{sign}(x)(|x| - c)_+ \mathbb{I}(|x| > \Delta)$.

(3) Calculate dissimilarity matrix:

(a) Update $\mathbf{u} = \frac{\mathbf{Dw}}{\|\mathbf{Dw}\|}$.

(b) Rewrite $\mathbf{u}$ as a $n \times n$ matrix $\mathbf{U}$.

(4) Perform clustering or dimension reduction using $\mathbf{U}$.

Sparse hierarchical clustering with fused lasso

Here, we introduce the SHC-FL algorithm. An FL penalty contains L1 norms of both features and their successive differences[54], which is defined as

$$\mathrm{pen}_{FL}(\mathbf{w}) = \lambda_1 \sum_{j=1}^{p} |w_j| + \lambda_2 \sum_{j=2}^{p} |w_j - w_{j-1}|,$$

where $\lambda_1, \lambda_2 \geq 0$ are hyperparameters. While λ_1 encourages sparsity of features, λ_2 encourages neighboring features to be similar and some to be identical.

We replace the lasso with the FL penalty in the objective of SHC, yielding the optimization criterion as

$$\underset{\mathbf{w},\mathbf{u}}{\mathrm{minimize}}\{-\mathbf{u}^T\mathbf{Dw} + \mathrm{pen}_{FL}(\mathbf{w})\},$$

$$\text{subject to } \|\mathbf{u}\|^2 \leq 1, \qquad \|\mathbf{w}\|^2 \leq 1, \qquad w_j \geq 0 \ \ \forall j.$$

We iteratively optimize $\mathbf{u}$ and $\mathbf{w}$. With $\mathbf{w}$ fixed, the optimal $\mathbf{u}$ is calculated in the same way as the one in SHC. With $\mathbf{u}$ fixed, the SHC-FL objective takes the form

$$\underset{\mathbf{w}}{\mathrm{minimize}}\{-\mathbf{u}^T\mathbf{Dw} + \mathrm{pen}_{FL}(\mathbf{w})\} \ \ \text{subject to } \|\mathbf{w}\|^2 \leq 1,$$

which is convex, but not smooth and not separable. To find stationary points of the above objective, we rewrite it as a Lagrangian function

$$\mathcal{L}(\mathbf{w}, \alpha) = -\mathbf{u}^T\mathbf{Dw} + \frac{1}{2}\alpha\|\mathbf{w}\|^2 + \mathrm{pen}_{FL}(\mathbf{w}),$$

where $\alpha \geq 0$ is the Lagrange multiplier. The derivative of $\mathcal{L}(\mathbf{w}, \alpha)$ can be expressed as

$$\frac{\partial \mathcal{L}(\mathbf{w}, \alpha)}{\partial w_j} = -y_j + \alpha w_j + \frac{\partial \mathrm{pen}_{FL}(\mathbf{w})}{\partial w_j},$$

where $\mathbf{y} = \mathbf{D}^T\mathbf{u}$. At stationary points, the optimal w_j's needs to satisfy

$$-y_j + \alpha w_j + \frac{\partial \mathrm{pen}_{FL}(\mathbf{w})}{\partial w_j} = 0.$$

Let's present the derivative of the FL penalty. Following the sub-differential calculus, the derivative of $\mathrm{pen}_{FL}(\mathbf{w})$ is in the form

$$\frac{\partial \mathrm{pen}_{FL}(\mathbf{w})}{\partial w_j} = \begin{cases} \lambda_1 h(w_1) - \lambda_2 h(w_2 - w_1), & j = 1 \\ \lambda_1 h(w_j) + \lambda_2 \left[h(w_j - w_{j-1}) - h(w_{j+1} - w_j) \right], & 1 < j \leq p, \end{cases}$$

where $h(x)$ is a subgradient of $|x|$, meaning $h(x) = \mathrm{sign}(x)$ when $x \neq 0$ and $h(x) \in [-1, 1]$ when $x = 0$. It is easy to see $\partial \mathrm{pen}_{FL}(\alpha \mathbf{w}) / \partial(\alpha w_j) = \partial \mathrm{pen}_{FL}(\mathbf{w}) / \partial w_j$, where α is a positive constant. Thus, a scaling of $\mathbf{w}$ does not affect the derivative of $\mathrm{pen}_{FL}(\mathbf{w})$.

Now, we solve for the optimal $\mathbf{w}$. Let $\mathbf{w}' = \alpha \mathbf{w}$. Because $\partial \mathrm{pen}_{FL}(\mathbf{w}') / \partial w_j' = \partial \mathrm{pen}_{FL}(\mathbf{w}) / \partial w_j$, the necessary condition of the stationary points of $\mathbf{w}$ becomes

$$-y_j + w_j' + \frac{\partial \mathrm{pen}_{FL}(\mathbf{w}')}{\partial w_j'} = 0.$$

To solve $\mathbf{w}'$ in the above equation, we construct an axillary convex optimization problem whose solution coincides with $\mathbf{w}'$:

$$\underset{\mathbf{a}}{\mathrm{minimize}} \left\{ \frac{1}{2} \|\mathbf{a} - \boldsymbol{y}\|^2 + \mathrm{pen}_{FL}(\mathbf{a}) \right\},$$

where $\mathbf{a} \in \mathbb{R}^p$. This is a signal approximation problem under the fused lasso penalty, with the solution called the fused lasso signal approximator (FLSA). We employ an FLSA solver proposed by Liu et al.[55], because of its high computational efficiency given a pair of λ_1 and λ_2, and strong scalability in high dimension[56]. Once $\mathbf{w}'$ is obtained, we get $\mathbf{w} = \mathbf{w}' / \|\mathbf{w}'\|$ such that $\|\mathbf{w}\|^2 = 1$.

SHC-FL algorithm

The algorithm is same as the algorithm of SHC-SSL, with the following differences:

Input: stacked dissimilarity matrix $\mathbf{D}$ in which features are ordered, outlier indicator $\mathbf{z}$, and hyperparameter $\lambda_1, \lambda_2 \geq 0$.

During the iteration, step 2b is replaced with:

(2b) Update $\mathbf{w}$:

(i) Find $\mathbf{w}' = \underset{\mathbf{a}}{\mathrm{argmin}} \left\{ \frac{1}{2} \left\| \mathbf{a} - \mathbf{D}'^T \mathbf{u} \right\|^2 + \lambda_1 \sum_{j=1}^{p} |a_j| + \lambda_2 \sum_{j=2}^{p} |a_j - a_{j-1}| \right\}$ using an FLSA solver.

(ii) Update $\mathbf{w} = \frac{\mathbf{w}'}{\|\mathbf{w}'\|}$.

Hyperparameter selection

We now turn to the selection of hyperparameters. In the above algorithms, the hyperparameters include: (1) s in SHC; (2) (λ_1, λ_2) in SHC-FL; and (3) λ_0 in SHC-SSL, where we keep λ_1 to a small constant. Let $\boldsymbol{\lambda}$ denote a single or a pair of hyperparameters in any one of the above algorithms. As in SHC, we employ a permutation-based approach, the gap statistic, to select the optimal value of $\boldsymbol{\lambda}$[21,57]. Given a hyperparameter $\boldsymbol{\lambda}$, the gap statistic measures the strength of the clustering based on real data with respect to the one based on randomly permuted data that are supposed to have no cluster. We quantify the strength of the clustering as

$$O(\boldsymbol{\lambda}, \mathbf{X}) = \sum_j w_j \sum_{i,i'} d_{i,i',j} U_{i,i'},$$

where $\mathbf{X}$ denotes the dataset used. Let $\mathbf{X}$ be the original real dataset, and $\mathbf{X}_1, \dots, \mathbf{X}_B$ denote B randomly permuted datasets by permuting samples within each feature. The gap statistic is calculated by

$$\mathrm{Gap}(\boldsymbol{\lambda}) = \log O(\boldsymbol{\lambda}, \mathbf{X}) - \frac{1}{B}\sum_{b=1}^{B} \log O(\boldsymbol{\lambda}, \mathbf{X}_b).$$

The optimal value of $\boldsymbol{\lambda}$ is obtained at the largest gap statistic.

In addition, to accelerate convergence of the presented algorithms for a series of N hyperparameter candidates $\{\boldsymbol{\lambda}_1, \dots, \boldsymbol{\lambda}_N\}$, we employ a "warm start" strategy: since solutions for similar hyperparameters are close, $\mathbf{w}$ is initialized as a solution for a nearby hyperparameter so that its solution can be found relatively quickly.

Simulation data

We generated simulation datasets by adapting a simulation procedure described by Brodinová et al.[58]. Each dataset has n samples characterized by p features, with $p = p_d + p_n$, where p_d is the number of discriminant features and p_n is the number of uninformative features. Each sample comes from one of C clusters. Let n_t be the size of the cluster t, and we have $n = \sum_{t=1}^{C} n_t$. In this study, we choose $p_d = 50$, $p_n = 950$, $C = 3$, and $n_t = 40$ for $t = 1, 2, 3$.

Samples of each cluster are fully characterized by the discriminant features, which follow a multivariate normal distribution with a mean vector $\mu_t \in \mathbb{R}^{p_d}$ and a covariance matrix $\Sigma_t \in \mathbb{R}^{p_d \times p_d}$, for $t = 1, \dots, C$. The elements of the mean vector μ_t are constructed as

$$\mu_{tj} = \begin{cases} \mu, & j = mC + t \\ 0, & \text{otherwise} \end{cases}$$

where $1 \le j \le p_d$, μ is a constant, and m is any nonnegative integer. The larger μ, the farther away the clusters are located geometrically. The covariance matrix Σ_t is constructed as

$$\Sigma_t = \mathbf{Q}\begin{pmatrix} 1 & \rho_t & \dots & \rho_t \\ \rho_t & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \rho_t \\ \rho_t & \dots & \rho_t & 1 \end{pmatrix}\mathbf{Q}^T,$$

where $\mathbf{Q}$ is a random orthogonal matrix and ρ_t is a random number sampled from a uniform distribution $U(a_\rho, b_\rho)$ with $0 < a_\rho, b_\rho < 1$. Here, we choose $\mu = 2$, $a_\rho = 0.1$, and $b_\rho = 0.9$ to generate clusters that are close to each other, rendering a nontrivial problem for sparse clustering algorithms. For the uninformative features of each sample, we generate p_n random numbers that are independently and identically distributed from the standard normal distribution.

Benchmarking of sparse hierarchical clustering algorithms

Using the above procedure, we randomly generated 100 simulation datasets for benchmarking performances of the presented algorithms. In all the analyses, the dissimilarity metric between features was the squared distance, and the hyperparameters were chosen based on gap statistic profiles. In the SHC-SSL analyses, we set $\lambda_1 = 0.001$ and chose the optimal λ_0 based on gap statistic profiles.

To evaluate the performances of the presented algorithms, we used the following performance metrics.

Recovery of discriminant features. To evaluate feature selection, given $\mathbf{w}$ from the result of the algorithms, we regard features $\{j | w_j > 0\}$ as predicted discriminant features, and all the rest of

features are predicted uninformative features. By comparing them with the true discriminant and uninformative features, we obtain precision, recall, and F1 score of the discriminant features.

Silhouette. To measure the clustering effect, we calculate the silhouette value using the dissimilarity matrix $\mathbf{U}$ and true cluster labels of samples. We use the silhouette value to measure the consistency of $\mathbf{U}$ from the true clusters rather than the consistency of a predicted partition from a fixed dissimilarity matrix.

Concordance of clusters along dendrogram after hierarchical clustering. The resulting dissimilarity matrix $\mathbf{U}$ is subjected to hierarchical clustering. To evaluate consistency of the resulting hierarchy of samples, represented as a dendrogram, with their true cluster labels, we calculate the concordance of cluster along the dendrogram as

$$\text{Concordance} = \frac{\sum_{i=2}^{n} \mathbb{I}(L_i = L_{i-1})}{n - C},$$

where L_i is the label of the cluster to which the sample i belongs. The denominator $n - C$ normalizes the range of the concordance to be within 0 and 1. The larger the concordance, the better the clustering. In the case of the worst clustering, when successive samples always belong to different clusters, the concordance becomes 0; in the case of an ideal clustering, when all within-cluster samples are tightly connected, the concordance is 1.

This metric measures a combined effect of hierarchical clustering and the chosen linkage criterion. Given $\mathbf{U}$, different linkage criteria generally produce different hierarchies. To compare different methods, we used the average linkage throughout this study.

Adjusted Rand index and purity of partitioned clusters after Leiden clustering. The resulting dissimilarity matrix $\mathbf{U}$ is used to find a shared k-nearest neighbor graph, where $k = 20$. The neighbor graph is then subjected to Leiden clustering, in which the resolution parameter is set as 1, to partition samples into clusters. To evaluate consistency of the partitioned clusters with the true cluster labels, we calculate adjusted Rand index and purity. The larger these metrics, the better the clustering.

Pseudo-cell strategy for reducing computational cost of SHC-SSL

All SHC-based algorithms involve a dissimilarity matrix $\mathbf{D} \in \mathbb{R}^{n^2 \times p}$, making the requirements of memory and computation scale quadratically with the sample size n. Thus, it is expensive to naïvely apply SHC-SSL when n is large. To reduce the computational cost, one can group cells with similar transcriptomes to form pseudo-cells, which are then subjected to SHC-SSL to select discriminant features. The selected features are used in the final clustering of all cells. An additional advantage of the pseudo-cell strategy is that it reduces technical noise and stabilizes variance, which may lead to more robust feature selection.

To construct pseudo-cells, specifically, one can use any efficient clustering algorithm, such as the Leiden algorithm, to group cells into many small clusters. Then, within each small cluster, pseudo-cells can be formed by calculating the mean of expressions of its constituent cells. This strategy is similar to the one used in Vision[59], where pseudo-cells were termed “micro-clusters”. We employed this strategy for datasets with a large number of cells to reduce computational cost and select discriminant markers that best discriminate pseudo-cells.

Neighborhood recurrence test

To test if a marker of a cell is recurrently up- or down-regulated in a biological subgroup rather than randomly perturbed, we devised a statistical test based on the idea that a recurrent marker should be enriched in the local neighborhood where the cell is located. The neighborhood recurrence test consists of two steps. First, we construct a K nearest neighbor graph of cells using

their similarities based on discriminant markers. We define $S_k(i)$ to be the local neighborhood of cell i, for $1 \leq i \leq n$, where $S_k(i)$ includes the k-nearest neighbors of i. Second, we perform a statistical test to determine if marker j is significantly up- or down-regulated in the local neighborhood $S_k(i)$. Let μ_j and σ_j^2 denote the mean and variance of the expression of marker j, which can be estimated as $\mu_j = \sum_i X_{ij}/n$ and $\sigma_j^2 = \sum_i \left(X_{ij} - \mu_j\right)^2/(n-1)$, respectively. The mean expression of marker j for cells in the local neighborhood $S_k(i)$ can estimated as

$$\bar{X}_{S_k(i),j} = \frac{\sum_{i' \in S_k(i)} X_{i'j}}{k}.$$

By the central limit theorem, when k is sufficiently large, $\bar{X}_{S_k(i),j}$ will be approximately normal distributed with its mean as μ_j and variance as σ_j^2/k. In our studies, we chose $k = 20$. Next, we calculate the neighborhood Z score of marker j for cell i as

$$Z_{ij} = \frac{\bar{X}_{S_k(i),j} - \mu_j}{\sigma_j/\sqrt{k}}.$$

Thus, the distribution of the neighborhood Z score is approximately the standard normal distribution, i.e., $Pr(Z) = N(0,1)$. Accordingly, the p-value of marker j that is recurrently up- or down-regulated in the local neighborhood of a cell i is $Pr\left(Z > Z_{ij}\right)$ or $Pr\left(Z < Z_{ij}\right)$, respectively. Finally, we adjust the p-values of all tested markers for each cell using the Benjamini-Hochberg procedure.

Applying RECOMBINE to biological datasets

We applied RECOMBINE to bulk and single-cell omics datasets. In each analysis, we set $\lambda_1 = 0.0001$, and chose the optimal λ_0 based on gap statistic profile. The squared distance was used as the dissimilarity metric in SHC-SSL. The dissimilarity matrix based on discriminant markers was subjected to building a K nearest neighbor graph, where $\mathrm{K} = 20$, followed by Leiden clustering, where the choice of resolution parameter depends on the datasets. The cells were visualized in heatmaps showing Leiden clusters, where cells within each Leiden cluster were hierarchically clustered using the average linkage.

Discovering biomarkers for subtypes of breast cancer. We downloaded mRNA expression (microarray, $\log_2$-transformed mRNA Z scores relative to all samples) and protein expression (reverse-phase protein arrays), median-normalized $\log_2$ intensities) data from the TCGA breast cancer[22] dataset from cBioPortal[60]. To focus on the cancer-driving genes, we only kept genes from the COSMIC Cancer Gene Census[61] for the mRNA expression data. After removing missing values, the mRNA expression data had 526 samples by 634 COSMIC Cancer Gene Census genes. For the protein expression data, all the proteins were kept, and there were 408 samples by 165 (phospho-)proteins. We ran RECOMBINE on these two datasets to identify RCMs as biomarkers that discriminate subtypes of breast cancer. In Leiden clustering, the resolution parameter was set to 0.8.

Selecting targeted panel from scRNA data for spatial molecular profiling of mouse visual cortex. For scRNA data, we downloaded an scRNA-seq dataset of mouse visual cortex from Allen Brain Atlas[26] (http://celltypes.brain-map.org/api/v2/well_known_file_download/694413985) and filtered out cells with low quality. In total, we obtained 14,249 cells of 23 annotated cell types belonging to glutamatergic neurons, GABAergic neurons, and non-neuronal cells. We employed Seurat to normalize the data using the “LogNormalize” method and selected 10,000 highly variable genes using the “vst” method. Before applying RECOMBINE to the scRNA data, to reduce computational cost of SHC-SSL imposed by the large numbers of cells, we generated 287 pseudo-cells corresponding to clusters identified by Leiden clustering with a resolution of 50. Then, we used

RECOMINE to select discriminant markers from 10,000 highly variable genes of these pseudo-cells. As a result, 366 genes were selected as discriminant markers of the mouse visual cortex scRNA data.

For spatial data, we downloaded a STARmap dataset of 1020 genes mapped in mouse primary visual cortex[12] (https://www.dropbox.com/sh/f7ebheru1lbz91s/AADm6D54GSEFXB1feRy6OSASa/visual_1020/20180505_BY3_1kgenes). The spatial location coordinate of each cell was extracted from "labels.npz" according to the method provided by the original paper describing STARmap[12]. In total, there were 1,549 cells. We employed Seurat to normalize the data using the "LogNormalize" method with a scale factor of 100. After clustering, we annotated cell clusters into cell types based on expression of marker genes. Specifically, Slc17a7 and Gad1 were used to annotate excitatory and inhibitory neurons, respectively; Nov, Rorb, Sulf2, and Pcp4 were used to annotate eL2/3, eL4, eL5, and eL6 neurons, respectively; Vip, Sst, and Pvalb were used to annotate VIP, SST, and PVALB neurons, respectively; and Pdgfra, Aqp4, Enpp2, and Bsg were used to annotate microglia, astrocyte, oligodendrocyte, and endothelial cells, respectively. As the paired scRNA data of mouse visual cortex did not include microglia cells, we filtered out microglia cells from the spatial data, resulting in 1,486 spatially resolved cells remaining in our analysis. Of 366 discriminant markers identified from the scRNA data, 248 genes were mapped in the spatial data of mouse visual cortex. Next, we performed Leiden clustering of 1,486 cells from the spatial data using the 248 discriminant markers, where a resolution of 0.5 was used. Cells were annotated based on their markers and mapped to their physical locations. Finally, we performed RECOMBINE neighborhood recurrence tests to identify RCMs that characterize cell subpopulations in the spatial data.

Identifying RCMs that characterize cell states of CD8 tumor-infiltrating T cells across cancer types. We downloaded SingleCellExperiment object of scRNA-seq dataset of pan-cancer CD8 T cells (https://zenodo.org/record/5461803), which were obtained by integrating 234,550 tumor-infiltrating CD8 T cells of 316 patients across 21 cancer types from multiple sources[31]. To reduce noise, cells within each source had been partitioned into small groups ("miniclusters"). In total, the expression matrix had a dimension of 11,972 miniclusters by 11,772 genes. The downloaded data had been $\log_2$-transformed, and 1,500 highly variable genes were provided. Before applying RECOMBINE to the dataset, we generated pseudo-cells using Seurat to reduce computational cost of SHC-SSL. Specifically, the 1,500 highly variable genes were scaled and subjected to principal component analysis, from which the top 50 principal components were used for embedding. The embeddings were corrected using Harmony to mitigate batch effects across data sources. Then, shared nearest neighbor graphs were constructed, and Leiden clustering with a resolution of 30 was used to generate 205 pseudo-cells. We applied RECOMBINE to the pseudo-cell data to select discriminant markers and extract RCMs from the discriminant markers for all cells. In Leiden clustering of all cells based on the selected discriminant markers, the resolution parameter was set to 2, which identified 15 clusters, including naïve cells expressing TCF7 (c1); memory cells expressing IL7R, MALAT1, or FOXP4 (c2-5); resident memory cells expressing LGALS3 (c6); effector memory cells expressing GZMK, GZMA, or PFN1, (c7-9); exhausted cells simultaneously expressing high levels of PDCD1, CTLA4, TIGIT, and TOX (c10-11); interferon-stimulated genes (ISG)-positive cells (c12); natural killer (NK)-like cells expressing both NKG7 and KLRD1 (c13-14); and mucosal-associated invariant T cells (MAIT) expressing both KLRB1 and TMIGD2 (c15).

To study compositions of cell clusters associated with anti–PD-1 therapy response, we downloaded the clinical response data from the report by Sade-Feldman et al.[38] (Supplementary Table 1, sheet patient-scRNA data) and used data of patients treated with anti–PD-1 therapy, where responders (N=7) included patients with complete and partial responses, and non-

responders (N=14) included those with stable or progressive disease based on RECIST. The pan-cancer dataset included 3,853 CD8 T cells from these 21 patients. The statistical comparison of cell compositions between responders and non-responders was based on Wilcoxon rank sum test.

To validate the association between the composition of $GZMK^{+}HAVCR2^{-}$ EM T cells and anti-PD-1 response, we used a dataset of CD8 T cells from non-small-cell lung cancer patients (N=34) treated with anti–PD-1 therapy in the report by Liu et al.[39], which was not included in the pan-cancer CD8 T cell dataset. We downloaded scRNA-seq, scTCR-seq, and metadata from the Gene Expression Omnibus (GEO; GSE179994), and response data from Liu et al.[39] (Supplementary Table 1). Of 34 patients, 11 patients had post-treatment samples (N=14) with annotated clinical responses (9 responders vs. 5 non-responders). Based on the scTCR-seq and metadata, we extracted all CD8 T cells having TCR clonotypes shared by exhausted CD8 T cells, resulting in 13,403 cells defined as tumor-reactive CD8 T cells. After constructing 286 pseudo-cells, we applied RECOMBINE to this dataset and selected 236 discriminant genes. Leiden clustering of all tumor-reactive CD8 T cells with the discriminant genes identified 3 clusters: $GZMK^{+}HAVCR2^{-}$ EM, exhausted, and proliferative T cells. Finally, we compared compositions of $GZMK^{+}HAVCR2^{-}$ EM T cells from post-treatment samples between responders and non-responders after anti–PD-1 therapy.

Identifying RCMs that characterize both common and rare subpopulations of mouse intestine. We downloaded an scRNA-seq dataset of intestinal organoid cells from the GEO (GSE62270). The data were preprocessed as described previously[40]. Mitochondrial genes, ERCC spike-ins, and genes associated with clustering artifacts (Rn45s, Malat1, Kcnq1ot1, A630089N07Rik, and Gm17821) were excluded as described previously[40,62]. The resulting dataset had an expression matrix of 238 cells by 3,773 genes. Finally, $\log_2$ transformation was applied. Using this preprocessed dataset, we ran RECOMBINE to identify RCMs that discriminate both common and rare cell subpopulations. In Leiden clustering, the resolution parameter was set to 0.6.

Selecting RCMs that characterize inter-tumoral heterogeneity in malignant cells from melanoma patients. From the GEO (GSE72056)[41], we downloaded an scRNA-seq dataset that had been integrated from 14 melanoma patients and $\log_2$-transformed. Following the method described by Tirosh et al.[41], we processed the data by filtering out cells with either fewer than 1,700 detected genes or an mean housekeeping expression below 3. We kept only malignant cells in our analysis. Furthermore, we removed ERCC spike-ins and excluded genes with a pooled expression below 4. The resulting data had 1,252 cells by 22,568 genes. Before applying RECOMBINE, we generated pseudo-cells using Seurat. Next, 2,000 highly variable genes were selected, scaled, and subjected to principal component analysis, from which the top 10 principal components were used for embedding and clustering. Then, Leiden clustering with a resolution of 20 was used to generate 106 pseudo-cells. We used RECOMBINE to select discriminant markers that characterize the heterogeneity of these pseudo-cells. In Leiden clustering of all cells based on the selected discriminant markers, the resolution parameter was set to 0.3.

To calculate Myc and immune signature scores of patients from TCGA, we downloaded the TCGA melanoma RNA-seq V2 expression dataset including 443 patients[42] and $\log_2$-transformed the RSEM-based gene quantifications to calculate signature scores. We also downloaded the corresponding clinical data for survival analysis. According to Myc signature scores, patients were classified as either Myc^{lo} (less than the median) or Myc^{hi} (greater than the median); and according to Immune signatures, patients were classified as either $Immune^{lo}$ (less than the median) or $Immune^{hi}$ (greater than the median). Together, we stratified patients into 4 groups (Myc^{lo} $Immune^{lo}$, Myc^{lo} $Immune^{hi}$, Myc^{hi} $Immune^{lo}$, and Myc^{hi} $Immune^{hi}$) and compared their survival times.

Selecting RCMs that characterize transcriptional intra-tumoral heterogeneity of TNBC. We downloaded scRNA-seq datasets of two TNBC patients, denoted as TNBC1 and TNBC2, from the GEO (GSE148673) reported by Gao et al.[45]. In these datasets, cells had been annotated as either tumor or normal cells. To study the heterogeneity of tumor cells, we removed the normal cells, resulting in 797 cells in TNBC1 and 620 cells in TNBC2. Next, 2,000 highly variable genes were selected, scaled, and subjected to principal component analysis, from which the top 10 principal components were used for embedding and clustering. Leiden clustering with a resolution of 10 was used to generate 48 and 56 pseudo-cells for TNBC1 and TNBC2, respectively. Then, we ran RECOMBINE to select discriminant markers that characterize transcriptional intra-tumoral heterogeneity in each TNBC individual. In Leiden clustering of all cells based on the selected discriminant markers, the resolution parameters were set to 0.3 for both TNBC1 and TNBC2.

Functional enrichment of discriminant markers and signature score calculation

To relate RECOMBINE-selected discriminant markers with biological functions, we performed functional enrichment using Enrichr[63] with the MSigDB hallmark signatures[64] and WikiPathways[65] provided by the Enrichr web server. Let's denote e_{ij} as the normalized and log-transformed expression of gene j in cell i, and G as the gene set of a signature. Following the signature score defined in Vision[59], the signature score can be calculated as

$$s_i = \frac{\sum_{j \in G} e_{ij}}{|G|},$$

where $|G|$ denotes the size of the gene set G. Since discriminant markers can optimally distinguish cells, we calculate the signature score using only discriminant markers as

$$s_i = \frac{\sum_{j \in G} e_{ij} \mathbb{I}(w_j > 0)}{\sum_{j \in G} \mathbb{I}(w_j > 0)},$$

where $\mathbb{I}(w_j > 0)$ is an indicator function of gene j selected in discriminant markers. As reported by DeTomaso et al.[59], even if the expression values are normalized, the signature scores calculated as described above may still depend on technical variability across cells. To remove this variability, the signature score can be normalized using the expected mean and variance of the score of a random gene set having the same number of genes. For a random gene set, the expected mean of the signature score in cell i is

$$\mathrm{E}[s_i] = \overline{e}_i,$$

where $\overline{e}_i$ is the mean of expressions in cell i, and the expected variance of the signature score is

$$\mathrm{Var}[s_i] = \frac{\sigma_i^2}{m},$$

where σ_i^2 is the variance of the expressions in cell i, and $m = \sum_{j \in G} \mathbb{I}(w_j > 0)$. Finally, the signature score is normalized as

$$s_i \leftarrow \frac{s_i - \mathrm{E}[s_i]}{\sqrt{\mathrm{Var}[s_i]}}.$$

Data availability

The simulation data are available and can be reproduced using the RECOMBINE R package. All the biological datasets are public and were found in previous papers as described in the Methods.

Code availability

The RECOMBINE R package is available at https://github.com/korkutlab/recombine.

Acknowledgements

We thank Lele Zhu for his comments on the pan-cancer CD8 T cell analysis. Editorial support was provided by Bryan Tutt, Scientific Editor, Research Medical Library at MD Anderson Cancer Center. This work is supported with grants from MDACC Support Grant P30 CA016672 (the Bioinformatics Shared Resource) (AK), U01CA253472 (AK).

Author contributions

X.L. and A.K designed the study, developed the algorithms, performed data analyses, and wrote the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper.

References

1. Osumi-Sutherland, D. *et al.* Cell type ontologies of the Human Cell Atlas. *Nature Cell Biology* **23**, 1129–1135 (2021).
2. Zhang, Y., Gao, S., Xia, J. & Liu, F. Hematopoietic hierarchy – an updated roadmap. *Trends Cell Biol* **28**, 976–986 (2018).
3. Becker, W. R. *et al.* Single-cell analyses define a continuum of cell state and composition changes in the malignant transformation of polyps to colorectal cancer. *Nature Genetics* **54**, 985–995 (2022).
4. Lange, M. *et al.* CellRank for directed single-cell fate mapping. *Nature Methods* **19**, 159–170 (2022).
5. Kong, W. *et al.* Capybara: A computational tool to measure cell identity and fate transitions. *Cell Stem Cell* **29**, 635-649.e11 (2022).
6. Aldridge, S. & Teichmann, S. A. Single cell transcriptomics comes of age. *Nature Communications* **11**, 1–4 (2020).
7. Liu, Z. & Zhang, Z. Mapping cell types across human tissues. *Science* **376**, 695–696 (2022).
8. Elmentaite, R., Domínguez Conde, C., Yang, L. & Teichmann, S. A. Single-cell atlases: shared and tissue-specific cell types across human organs. *Nature Reviews Genetics* **23**, 395–410 (2022).
9. Rao, A., Barkley, D., França, G. S. & Yanai, I. Exploring tissue architecture using spatial transcriptomics. *Nature* **596**, 211–220 (2021).
10. Longo, S. K., Guo, M. G., Ji, A. L. & Khavari, P. A. Integrating single-cell and spatial transcriptomics to elucidate intercellular tissue dynamics. *Nature Reviews Genetics* **22**, 627–644 (2021).
11. Moffitt, J. R. *et al.* Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. *Science* **362**, (2018).
12. Wang, X. *et al.* Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science* **361**, (2018).
13. He, S. *et al.* High-plex imaging of RNA and proteins at subcellular resolution in fixed tissue by spatial molecular imaging. *Nature Biotechnology* **40**, 1794–1806 (2022).
14. Gohil, S. H., Iorgulescu, J. B., Braun, D. A., Keskin, D. B. & Livak, K. J. Applying high-dimensional single-cell technologies to the analysis of cancer immunotherapy. *Nature Reviews Clinical Oncology* **18**, 244–256 (2020).
15. Luecken, M. D. & Theis, F. J. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol Syst Biol* **15**, e8746 (2019).
16. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587.e29 (2021).
17. Lähnemann, D. *et al.* Eleven grand challenges in single-cell data science. *Genome Biology* **21**, 1–35 (2020).
18. Baran, Y. *et al.* MetaCell: Analysis of single-cell RNA-seq data using K-nn graph partitions. *Genome Biol* **20**, 1–19 (2019).
19. Persad, S. *et al.* SEACells infers transcriptional and epigenomic cellular states from single-cell genomics data. *Nature Biotechnology* 1–12 (2023).
20. Dann, E., Henderson, N. C., Teichmann, S. A., Morgan, M. D. & Marioni, J. C. Differential abundance testing on single-cell data using k-nearest neighbor graphs. *Nature Biotechnology* **40**, 245–253 (2021).
21. Witten, D. M. & Tibshirani, R. A framework for feature selection in clustering. *J Am Stat Assoc* **105**, 713–726 (2010).
22. Koboldt, D. C. *et al.* Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70 (2012).

23. Khaled, W. T. *et al.* BCL11A is a Triple-negative breast cancer gene with critical functions in stem and progenitor cells. *Nat Commun* **6**, 1–10 (2015).
24. Dai, X. *et al.* FOXA1 is Prognostic of Triple Negative Breast Cancers by Transcriptionally Suppressing SOD2 and IL6. *Int J Biol Sci* **15**, 1030–1041 (2019).
25. Rajput, S. *et al.* Inhibition of cyclin dependent kinase 9 by dinaciclib suppresses cyclin B1 expression and tumor growth in triple negative breast cancer. *Oncotarget* **7**, 56864–56875 (2016).
26. Tasic, B. *et al.* Shared and distinct transcriptomic cell types across neocortical areas. *Nature* **563**, 72–78 (2018).
27. Waldman, A. D., Fritz, J. M. & Lenardo, M. J. A guide to cancer immunotherapy: from T cell basic science to clinical practice. *Nature Reviews Immunology* **20**, 651–668 (2020).
28. van der Leun, A. M., Thommen, D. S. & Schumacher, T. N. CD8+ T cell states in human cancer: insights from single-cell analysis. *Nature Reviews Cancer* **20**, 218–232 (2020).
29. Chow, A., Perica, K., Klebanoff, C. A. & Wolchok, J. D. Clinical implications of T cell exhaustion for cancer immunotherapy. *Nature Reviews Clinical Oncology* **19**, 775–790 (2022).
30. Lowery, F. J. *et al.* Molecular signatures of antitumor neoantigen-reactive T cells from metastatic human cancers. *Science* **375**, 877–884 (2022).
31. Zheng, L. *et al.* Pan-cancer single-cell landscape of tumor-infiltrating T cells. *Science* **374**, (2021).
32. Good, C. R. *et al.* An NK-like CAR T cell transition in CAR T cell dysfunction. *Cell* **184**, 6081-6100.e26 (2021).
33. Sen, D. R. *et al.* The epigenetic landscape of T cell exhaustion. *Science* **354**, 1165–1169 (2016).
34. Pauken, K. E. *et al.* Epigenetic stability of exhausted T cells limits durability of reinvigoration by PD-1 blockade. *Science* **354**, 1160–1165 (2016).
35. Im, S. J. *et al.* Defining CD8+ T cells that provide the proliferative burst after PD-1 therapy. *Nature* **537**, 417–421 (2016).
36. Siddiqui, I. *et al.* Intratumoral Tcf1+ PD-1+ CD8+ T Cells with Stem-like Properties Promote Tumor Control in Response to Vaccination and Checkpoint Blockade Immunotherapy. *Immunity* **50**, 195-211.e10 (2019).
37. Huang, Q. *et al.* The primordial differentiation of tumor-specific memory CD8+ T cells as bona fide responders to PD-1/PD-L1 blockade in draining lymph nodes. *Cell* **185**, 4049-4066.e25 (2022).
38. Sade-Feldman, M. *et al.* Defining T cell states associated with response to checkpoint immunotherapy in melanoma. *Cell* **175**, 998-1013.e20 (2018).
39. Liu, B. *et al.* Temporal single-cell tracing reveals clonal revival and expansion of precursor exhausted T cells during anti-PD-1 therapy in lung cancer. *Nature Cancer* **3**, 108–121 (2021).
40. Grün, D. *et al.* Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature* **525**, 251–255 (2015).
41. Tirosh, I. *et al.* Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352**, 189–196 (2016).
42. Cancer Genome Atlas Network. Genomic classification of cutaneous melanoma. *Cell* **161**, 1681–1696 (2015).
43. Casey, S. C. *et al.* MYC regulates the antitumor immune response through CD47 and PD-L1. *Science* **352**, 227–231 (2016).
44. Wu, X. *et al.* MYC oncogene is associated with suppression of tumor immunity and targeting Myc induces tumor cell immunogenicity for therapeutic whole cell vaccination. *J Immunother Cancer* **9**, e001388 (2021).

45. Gao, R. *et al.* Delineating copy number and clonal substructure in human tumors from single-cell transcriptomes. *Nat Biotechnol* 1–10 (2021) doi:10.1038/s41587-020-00795-2.
46. Liang, S. *et al.* Single-cell manifold-preserving feature selection for detecting rare cell populations. *Nat Comput Sci* **1**, 374–384 (2021).
47. Dumitrascu, B., Villar, S., Mixon, D. G. & Engelhardt, B. E. Optimal marker gene selection for cell type discrimination in single cell analyses. *Nature Communications* **12**, 1–8 (2021).
48. Vargo, A. H. S. & Gilbert, A. C. A rank-based marker selection method for high throughput scRNA-seq data. *BMC Bioinformatics* **21**, 1–51 (2020).
49. Rood, J. E., Maartens, A., Hupalowska, A., Teichmann, S. A., & Regev, A. Impact of the Human Cell Atlas on medicine. *Nature Medicine* **28**(12), 2486–2496. (2022).
50. Li, X. et al. (2022). Precision combination therapies based on recurrent oncogenic coalterations. *Cancer Discovery*, 12(6), 1542–1559. (2022).
51. Ročková, V. & George, E. I. The Spike-and-Slab LASSO. *J Am Stat Assoc* **113**, 431–444 (2018).
52. Ročková, V. Bayesian estimation of sparse signals with a continuous spike-and-slab prior1. *Ann Stat* **46**, 401–437 (2018).
53. Hastie, T., Tibshirani, R. & Wainwright, M. *Statistical learning with sparsity: The lasso and generalizations*. *Statistical Learning with Sparsity: The Lasso and Generalizations* (CRC Press, 2015).
54. Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. & Knight, K. Sparsity and smoothness via the fused lasso. *J R Stat Soc Series B Stat Methodol* **67**, 91–108 (2005).
55. Liu, J., Yuan, L. & Ye, J. An efficient algorithm for a class of fused Lasso problems. in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 323–332 (2010).
56. Xin, B., Kawahara, Y., Wang, Y. & Gao, W. Efficient generalized fused lasso and its application to the diagnosis of Alzheimer's disease. in *Association for the Advancement of Artificial Intelligence* (2014).
57. Tibshirani, R., Walther, G. & Hastie, T. Estimating the number of clusters in a dataset via the gap statistic. *J R Stat Soc Series B Stat Methodol* **63**, 411–423 (2001).
58. Brodinová, Š., Filzmoser, P., Ortner, T., Breiteneder, C. & Rohm, M. Robust and sparse k-means clustering for high-dimensional data. *Adv Data Anal Classif* **13**, 905–932 (2019).
59. DeTomaso, D. *et al.* Functional interpretation of single cell similarity maps. *Nat Commun* **10**, 1–11 (2019).
60. Cerami, E. *et al.* The cBio Cancer Genomics Portal: An open platform for exploring multidimensional cancer genomics data. *Cancer Discov* **2**, 401–404 (2012).
61. Sondka, Z. *et al.* The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nature Reviews Cancer* ***18***, 696–705 (2018).
62. Gehart, H. *et al.* Identification of enteroendocrine regulators by real-time single-cell differentiation mapping. *Cell* **176**, 1158-1173.e16 (2019).
63. Chen, E. Y. *et al.* Enrichr: Interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* **14**, (2013).
64. Liberzon, A. *et al.* The molecular signatures database hallmark gene set collection. *Cell Syst* **1**, 417–425 (2015).
65. Kelder, T., *et al*. WikiPathways: building research communities on biological pathways. *Nucleic Acids Research* **40**, D1301–D1307 (2012).