Title

Cellular and molecular fine mapping pinpoints new immunopathology of lupus

Authors

Masahiro Nakano[1,21]*, Michihiro Kono[1,2,21], Kenichiro Asahara[1,3], Takayuki Katsuyama[4], Satoshi Kubo[5], Eri Katsuyama[4,6], Yuya Fujita[5], Hajime Inokuchi[1,2,7], Takahiro Nishino[1,2], Takahiro Arakawa[1], Tsugumi Kawashima[1], Shohei Noma[1], Reza Bagherzadeh[1,2], Yukiro Matsumoto[1,2,8], Jun Inamo[1,2], Haruka Takahashi[1,2], Bunki Natsumoto[1], Xuejun Zhang[9,10,11], Sang-Cheol Bae[12,13,14], Akari Suzuki[15], Hiroaki Hatano[1,2], Chikashi Terao[16,17,18], Yoshiya Tanaka[5,19], Yoshinori Matsumoto[4], Kazuhiko Yamamoto[15], Kazuyoshi Ishigaki[1,2,20]*

Affiliations

[1] Laboratory for Human Immunogenetics, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan

[2] Department of Microbiology and Immunology, Keio University School of Medicine, Tokyo, Japan

[3] Department of Ophthalmology, Graduate School of Medical Sciences, Kyushu University, Fukuoka, Japan

18 [4] Department of Nephrology, Rheumatology, Endocrinology and Metabolism, Okayama University
19 Faculty of Medicine, Dentistry and Pharmaceutical Sciences, Okayama, Japan
20 [5] The First Department of Internal Medicine, University of Occupational and Environmental Health,
21 Kitakyushu, Japan
22 [6] Faculty of Health Science, Okayama University Medical School, Graduate School of Health Sciences,
23 Okayama University, Okayama, Japan
24 [7] Division of Rheumatology, Department of Internal Medicine, Keio University School of Medicine, Tokyo,
25 Japan
26 [8] Department of Dermatology, Keio University School of Medicine, Tokyo, Japan
27 [9] Department of Dermatology, the First Affiliated Hospital of Anhui Medical University, Hefei, Anhui,
28 China
29 [10] Institute of Dermatology, Anhui Medical University, Hefei, Anhui, China
30 [11] Key Laboratory of Dermatology (Anhui Medical University), Ministry of Education, Hefei, Anhui, China
31 [12] Department of Rheumatology, Hanyang University Hospital for Rheumatic Diseases, Seoul, Korea
32 [13] Hanyang Institute of Bioscience and Biotechnology, Seoul, Korea
33 [14] Hanyang University Institute for Rheumatology Research, Seoul, Korea
34 [15] Laboratory for Autoimmune Diseases, RIKEN Center for Integrative Medical Sciences, Yokohama,
35 Japan

[16] Laboratory for Statistical and Translational Genetics, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan

[17] Clinical Research Center, Shizuoka General Hospital, Shizuoka, Japan

[18] The Department of Applied Genetics, The School of Pharmaceutical Sciences, University of Shizuoka, Shizuoka, Japan

[19] Department of Molecular Targeted Therapeutics, University of Occupational and Environmental Health, Kitakyushu, Japan

[20] Keio University Human Biology-Microbiome-Quantum Research Center (WPI-Bio2Q)

[21] These authors contributed equally.

*Correspondence:

Masahiro Nakano, M.D., Ph.D.

masahiro.nakano.fc@riken.jp

Kazuyoshi Ishigaki, M.D., Ph.D.

kazuyoshi.ishigaki@keio.jp

Abstract

Systemic lupus erythematosus (SLE) is a complex autoimmune disease with unknown etiology. To pinpoint new disease-relevant cell states in SLE and their detailed molecular profiles, we applied the deepest investigation strategy for single-cell data, cellular and molecular fine mapping, to the largest-scale multimodal SLE datasets, comprising ~2.1 million peripheral blood mononuclear cells from 232 SLE cases and 114 healthy controls. Among 123 fine-grained cell states in 27 cell types, we discovered that previously uncharacterized cell states, such as *GZMK*$^{+}$*GZMH*$^{+}$*HLA-DR*$^{+}$ effector memory CD8^{+} T cells (double-positive [DP] EMCD8), *FAM13A*$^{+}$*CDK6*$^{+}$ naive CD4^{+} T cells, and *ARHGAP15*$^{+}$*FOXO1*$^{+}$ T cells, expanded especially in severe SLE. Based on our new statistical model, which efficiently dissects gene signatures reflecting cell state abundance (quantitative change) and single-cell-level gene dysregulation (qualitative change) within each conventional cell type, we found that key biological processes and genetic risks of SLE were linked to these expanded cell states. Finally, by combining multimodal in-silico analyses, CRISPR perturbations, and cellular function assays, we identified cell-state-specific transcriptional regulators (e.g., *FOXO1* in *ARHGAP15*$^{+}$ *FOXO1*$^{+}$ T cells), key surface proteins (e.g., HLA-DR, CD27, CD38, and CD28 in DP EMCD8), T cell receptor features, and cis-regulatory elements. Notably, we newly revealed that DP EMCD8 cells exhibit chronic antigen exposure signatures and serve as a major source of key cytokines, such as IFN-γ, underscoring this cell state as a

73 central driver of SLE immunopathology. These findings provide new insights into the drug target

74 discovery in SLE.

75

Introduction

Systemic lupus erythematosus (SLE) is a prototypic autoimmune disease that affects multiple organs, triggered by genetic and environmental factors[1]. Since the pathophysiology of SLE involves various immune cell types, many bulk and single-cell transcriptome studies on peripheral blood and affected organs have been conducted[2–6]. While these efforts have begun to reveal key gene signatures and core cell populations of SLE[7,8], the fine-resolution pathophysiology remains unresolved, and the treatment options are still limited; to date, only two molecular-targeted drugs have been approved for this disease[9,10].

To understand the complex biology of SLE and discover new therapeutic targets, it is crucial to identify new disease-relevant cell states and profile their key molecules at fine cellular resolution. Recently, we reported the largest-scale bulk study of SLE with 27 immune cell types and identified cell-type-specific signatures linked to disease development and exacerbation[6]. However, as an inherent limitation of bulk studies, we failed to identify granular disease-relevant cell states that may exist within each cell type. While single-cell RNA sequencing (scRNA-seq) can address this problem, most studies have not fully leveraged the advantage of single-cell data. First, although scRNA-seq can, in theory, define hundreds of distinct and heterogeneous cell states, previous studies ended up roughly profiling only ~20 major cell types[5] or ~40 cell clusters[3,4] and did not pursue more granular and detailed biology. Second, although scRNA-seq can evaluate quantitative and qualitative changes in disease status at

single-cell resolution, most studies rely on the cell-type-level pseudobulk approach, which aggregates single-cell expression across each cell type and treats them as if bulk expression data. As a result, this approach obscures the distinction between altered cell state composition (quantitative changes) and dysregulated gene expression within individual cells (qualitative changes)[11], both of which may require different treatment strategies. These critical limitations have hindered us from the comprehensive identification of disease-relevant cell states and their functional characterization.

To tackle these issues, we took the deepest investigation strategy, cellular and molecular fine mapping, to pinpoint granular disease-relevant cell states and their key molecules at single-cell resolution and applied it to the largest-scale multimodal single-cell data of SLE with a total of ~2.1 million cells from 346 donors (Lupus-derived Peripheral immune cell Integrated panels [LuPIN]; **Fig. 1, left**). First, by employing a fine-grained clustering approach, we identified 123 distinct cell states in 27 immune cell types. Among them, we found 11 previously uncharacterized cell states expanded especially in clinically active SLE ("cellular fine mapping", **Fig. 1, middle**). Next, our new statistical model, which dissected gene signatures reflecting quantitative and qualitative changes, demonstrated that key immunological pathways and SLE risk variants were closely linked to cell-state compositional changes within each cell type. Finally, through multimodal in-silico analyses, CRISPR perturbations, and cellular function assays of disease-relevant cell state candidates, we successfully revealed their key molecular profiles, such as transcriptional regulators, cytokine profiles, surface marker proteins, T cell receptor

112 (TCR) repertoires, and epigenetic status ("molecular fine mapping", **Fig. 1, right**). Overall, our intensive
113 and comprehensive cellular and molecular fine mapping shed light on the new immunopathology of SLE
114 and provided guidance to design new treatment strategies.
115

116 **Results**

117 **Overview of this study**

118 LuPIN consists of three datasets, each with a different purpose. To robustly identify disease-relevant cell
119 states, we first used the largest, publicly available 3'-end scRNA-seq dataset from SLE[5] (dataset 1; **Fig.**
120 **1, left** and **Supplementary Table 1**). Our stringent multi-step quality control (QC) process yielded the
121 final dataset (1,456,041 peripheral blood mononuclear cells [PBMCs] from 97 healthy and 172 SLE
122 samples [159 unique SLE donors: 11 donors have 2~3 timepoints]; **Extended Data Fig. 1** and
123 **Methods**). Among 172 SLE samples, 21 had high disease activity (HDA), and 40 were clinically inactive
124 (**Methods**). Following our previous study scheme[6], we prepared two comparison axes: (i) disease-state
125 comparison between healthy controls and inactive SLE patients, reflecting the biology of disease
126 development, and (ii) disease-activity comparison between inactive vs. HDA SLE, reflecting the biology
127 of disease exacerbation. For replication analysis, we used another scRNA-seq dataset[4] (dataset 2,
128 344,235 PBMCs from 17 healthy and 41 SLE samples).

To overlay in-depth molecular information onto the fine-grained cell state map, we additionally prepared a new multimodal 5'-end dataset with cellular indexing of transcriptomes and epitopes by sequencing (CITE-seq), a technology profiling gene expression and surface marker proteins from single cells[12], as well as TCR-seq and enhancer RNA profiling[13,14] (dataset 3, 268,783 PBMCs from 32 SLE samples; **Methods**). Through the efficient integration with dataset 1, we searched for key surface proteins, TCR repertoires, and cis-regulatory elements specific to disease-relevant cell states.

Identification of 123 distinct cell states using fine-grained clustering

To identify granular cell states of SLE, we employed a three-step intensive clustering approach in dataset 1: lineage (layer 1), cell type (layer 2), and cell state (layer 3) (**Fig. 2**). First, by applying Louvain clustering, we separated 1,456,041 PBMCs into seven cell lineages (layer 1: $CD4^+$ T cells [CD4 T], $CD8^+$ and unconventional T cells [CD8 and other T], natural killer cells [NK], B cells, monocytes [Mono], dendritic cells [DC], and hematopoietic stem and progenitor cells [HSPC]; **Extended Data Fig. 1d-f**). We then separated these seven lineages into 27 coarse cell types (layer 2: **Fig. 2a-f**, **Extended Data Fig. 2a-f**, and **Supplementary Table 2**). Our elaborated analytical pipelines (e.g., curated clustering algorithms; **Supplementary Table 3** and **Methods**) achieved higher cell-type annotation accuracy than the original study[5] (**Extended Data Fig. 2g**).

146 While the original study defined only 20 coarse cell types, we further performed clustering on
147 each cell type (cell-type-level clustering) and identified 123 cell states with unique gene expression
148 profiles (layer 3: **Fig. 2a-f**, **Extended Data Fig. 3**, and **Supplementary Table 4**). Compared with
149 conventional lineage-level clustering, which primarily captures global cell-type structures (e.g., naive-
150 ness vs. memory-ness), cell-type-level clustering can efficiently capture subtle gene-program
151 heterogeneities within each cell type. As these cell-state definitions form the basis for all downstream
152 analyses, we performed rigorous approaches to ensure their robustness. First, we set the conservative
153 clustering resolution parameter for each cell type so that the distinct identity of each cluster was
154 supported by a wide range of literatures[7,8,15–38] (**Supplementary Table 5**). Indeed, upregulated marker
155 genes for each state showed minimal overlap within the same cell type (**Fig. 2g**), indicating that these
156 clusters were not redundant. Second, pathway analysis using these marker genes showed enrichment of
157 multiple cell-state-specific pathways (**Supplementary Table 6**), supporting their biological relevance.
158 Finally, our multimodal investigations using the independent dataset 3 identified cell-state-specific
159 surface proteins and TCR features (**Fig. 5-6**), providing orthogonal validation at the immune phenotypic
160 level.

161 Our fine-grained annotations successfully captured multiple disease-relevant cell state
162 candidates (**Fig. 2**), most of which were not reported in the original study due to insufficient resolution.
163 We first recapitulated well-known disease-relevant cell states such as (1) follicular and peripheral helper

T cells (Tfh and Tph) in central and effector memory CD4 T cells (CMCD4-3: *ICOS*$^{+}$*PASK*$^{+}$ and EMCD4-7: *PDCD1*$^{+}$*TIGIT*$^{+}$) and (2) autoimmune-associated B cells (ABC) in memory B cells (MemB-6: *ITGAX*$^{+}$*FCRL5*$^{+}$), which were important in the autoantibody production[7,8,23]. We also confirmed the subpopulations highly activated by IFN signaling (*ISG15*$^{+}$*MX1*$^{+}$ cell states), the hallmark signature of SLE[1,22], in most cell types. The shared gene program across cell types is not unique to IFN signaling; *JUN*$^{+}$*CD69*$^{+}$ cell states with upregulated AP-1 family TFs (e.g., *JUN, FOS*) and early activation markers (e.g., *CD69* and *NFKBIA*)[17] were identified in most lymphoid cell types (**Extended Data Fig. 3**). Importantly, we identified multiple previously uncharacterized cell states with distinct gene programs: some of them were later shown to expand especially in severe SLE (e.g., EMCD8-5: *GZMK*$^{+}$*GZMH*$^{+}$*HLA-DRB1*$^{+}$, **Fig. 3**). Moreover, our approach captured very rare subpopulations such as *ANXA2*$^{+}$*CTLA4*$^{+}$ activated double negative T cells (dnT-1)[31], *RORA*$^{+}$*GATA3*$^{+}$ group 2 innate lymphoid cells (ILC-1)[32], *MZB1*$^{+}$*PCNA*$^{+}$ plasmablasts (PlasmaB-1)[34], and *LAMP3*$^{+}$*CCR7*$^{+}$ DCs (cDC-5)[38]. Collectively, our three-step intensive clustering approach yielded the most granular cell state information for SLE, serving as the foundation for subsequent analyses.

Cellular fine mapping to pinpoint disease-relevant cell states of SLE

To evaluate which cell states expanded in SLE (cellular fine mapping), we next performed co-varying neighborhood analysis (CNA), an approach to identify single-cell-resolution transcriptional

182 neighborhoods associated with clinical phenotypes[39]. The extent of expansion (or depletion) of each cell
183 state was expressed as continuous values at the single-cell level ("neighborhood correlation", **Fig. 3a-b**).
184 As mentioned above, we performed CNA with two comparison axes: disease-state (inactive SLE vs.
185 healthy, reflecting disease development) and disease-activity CNA (HDA vs. inactive SLE, reflecting
186 disease exacerbation)[6].

187 We first examined the abundance of 26 coarse cell types across six cell lineages (excluding
188 HSPC), accounting for age, sex, and ancestry (lineage-level CNA; **Fig. 3a** and **Methods**). Our analysis
189 recapitulated the depletion of NaiveCD4 and mucosal-associated invariant T cells (MAIT) in disease-
190 state comparison[5] and newly showed that these depletion patterns were remarkable in disease-activity
191 comparison (**Fig. 3a** and **Extended Data Fig. 4a**). Similarly, regulatory T cells (Treg) and Proliferating
192 (Prolif) CD4/CD8/NK cells expanded both in disease-state and -activity CNA. In Mono-lineage,
193 monocytes with high C1q genes[36] ($C1Q^+$Mono) expanded in disease-activity CNA. Although
194 neighborhoods in these cell types exhibited universal expansion (or depletion) patterns, many other cell
195 types showed heterogeneous patterns in lineage-level CNA (**Fig. 3a**), underscoring the need for a more
196 fine-grained investigation.

197 To assess the cell state compositional changes within each cell type, we next applied CNA to 18
198 major cell types (>5,000 cells; cell-type-level CNA). Strikingly, we observed skewed neighborhood
199 abundance in 34 out of 36 tests (18 cell types × 2 comparisons, global $P < 0.05/36$), and 29 cell states

200 showed evidence of remarkable expansion (**Fig. 3b**, **Extended Data Fig. 4b**, and **Supplementary Table**
201 **7**). These expanded cell states tended to be rare within each cell type, supporting the importance of our
202 fine-grained strategy (**Fig. 3c** and **Extended Data Fig. 4c**).

203 Among the 29 cell states with remarkable expansion,18 recapitulated known signals. All IFN-
204 activated (*ISG15*$^{+}$*MX1*$^{+}$) cell states within each cell type expanded in disease-state and -activity CNA
205 (**Fig. 3d**), supporting our previous report that IFN signal contributes to both disease development and
206 exacerbation[6]. While Tfh (CMCD4-3) expanded primarily in disease-state, Tph (EMCD4-7) and ABC
207 (MemB-6) expanded especially in disease-activity CNA. The remaining 11 cell states represented
208 previously unreported expansions in SLE[3–5], nominating new disease-relevant cell state candidates. An
209 example is *GZMK*$^{+}$*GZMH*$^{+}$*HLA-DRB1*$^{+}$ cells (EMCD8-5), showing evidence of expansion both in disease-
210 state and -activity CNA (**Fig. 3b**). While other EMCD8 cell states express either *GZMK* or *GZMH*, this
211 population expresses both, as well as *GZMB* to a lesser extent (**Extended Data Fig. 3b**; hereafter, we
212 refer to this cell state as “double-positive [DP] EMCD8”). Notably, this population showed high
213 expression of multiple T cell activation markers, including *HLA-DR* genes, *PDCD1*, and *TIGIT* (**Extended**
214 **Data Fig. 3b**). Another example is *FAM13A*$^{+}$*CDK6*$^{+}$ cells (NaiveCD4-4), which further highlights the
215 significance of our fine-grained analysis: although NaiveCD4 was on average depleted within the CD4 T
216 lineage (**Fig. 3a**), this depletion pattern is not uniform across NaiveCD4 cell states. Lastly,

ARHGAP15^{+}*FOXO1*$^{+}$ cell states also expanded across most T-cell cell types in disease-state and -activity CNA (**Fig. 3b, d**).

Some cell states exhibited different patterns of abundance changes between disease-state and activity CNA. While *JUN*$^{+}$*CD69*$^{+}$ cell states expanded in disease-state CNA, they consistently depleted in disease-activity CNA in most cell types (**Fig. 3d**). These results suggest distinct roles of these cell states in SLE depending on the disease phase[6]. Our extensive analysis also identified ancestry-specific signals: some *ISG15*$^{+}$*MX1*$^{+}$ cell states and ABC (MemB-6) expanded in Asian SLE compared to European (EUR) SLE[8,40] (**Extended Data Fig. 4d**).

Our CNA results showed high reproducibility in multiple approaches. First, a cluster-based abundance testing (MASC)[26] showed concordant results with CNA (**Extended Data Fig. 4e**). Second, most CNA signals were globally concordant between Asian and EUR ancestries (**Extended Data Fig. 4f**). Third, by applying CNA independently to low disease activity (LDA) and moderate disease activity (MDA) cases that were excluded from the discovery set, we observed the concordant signals with a gradual increase in disease-activity effects, reflecting the continuous nature of SLE activity (**Fig. 3e** and **Methods**). Lastly, disease-state and activity CNA using the independent dataset 2 showed good replicability of CNA signals detected in dataset 1 (**Fig. 3f** and **Methods**). Overall, our intensive cellular fine mapping has robustly pinpointed novel disease-relevant cell-state candidates.

235 **A new single-cell model to dissect quantitative and qualitative changes within cell types**

236 Following the successful cellular fine mapping, we proceed to the molecular fine mapping. As a
237 stepping stone toward pinpointing key molecules in SLE, we introduce a new statistical model to
238 overcome limitations of the pseudobulk analyses, a standard approach in many studies. Since a typical
239 pseudobulk analysis aggregates single-cell-level expression within a cell type, the resulting differentially
240 expressed genes (DEGs) could be mixed up by the following two scenarios **(Fig. 4a)**: (i) a gene is
241 specifically expressed in a particular cell state whose abundance shifts with phenotype (quantitative
242 change) or (ii) a gene's per-cell expression is directly dysregulated with phenotype regardless of cell
243 state abundance (qualitative change). The pseudobulk approach cannot determine which scenario has a
244 more pathogenic effect on disease biology.

245 To address this issue, we developed ScQuaD (Single-cell-level Quantitative and Qualitative
246 change Dissection), a generalized linear mixed effect model to evaluate the association of gene
247 expression levels with both quantitative and qualitative changes (**Fig. 4b**). In contrast to pseudobulk
248 analysis, ScQuaD treats each single-cell-level expression as an independent observation based on
249 negative binomial distribution and regresses this jointly by (i) a phenotype-driven cell-state abundance
250 term (quantitative change) and (ii) a direct phenotype term (e.g., inactive SLE vs. healthy) capturing per-
251 cell gene dysregulation (qualitative change), adjusted for donor-/batch-random effect and other fixed
252 effect covariates (**Methods**). Specifically, we incorporated single-cell-level neighborhood correlation

253 values in cell-type-level CNA (**Fig. 3**) into ScQuaD as metrics for cell-state abundance changes. We
254 applied this model to each gene across 18 major cell types. We termed (i) the genes associated with
255 quantitative changes as “abundance signature genes” and (ii) the genes directly associated with
256 qualitative changes, even after adjusting for quantitative changes, as “dysregulated signature genes”
257 (**Supplementary Table 8**). Permutation tests showed that this model is well calibrated (**Extended Data**
258 **Fig. 5a**). When we quantified the contribution of qualitative and quantitative changes to the top 1,000
259 pseudobulk DEGs for each cell type, around 50% of their variation was driven by quantitative change
260 (**Extended Data Fig. 5b**). For simplicity, we showed disease-state comparison results in **Fig. 4c-f** and
261 disease-activity results in **Extended Data Fig. 5c-d**.

262 *IFIH1* and *OAS1* were well-known interferon response genes (IRGs), and both are linked to the
263 significant genetic loci in genome-wide association studies of SLE[41] (SLE-GWAS; **Supplementary Table**
264 **9** and **Methods**). While pseudobulk analysis identified both as DEGs in NaiveCD4 (false discovery rate
265 [FDR] < 0.05), ScQuaD revealed different behaviors of qualitative and quantitative changes for these two
266 genes (**Fig. 4c**). While *OAS1* expression localized to NaiveCD4-7 (*ISG15*$^{+}$*MX1*$^{+}$), which expanded in
267 SLE (abundance signature gene), *IFIH1* did not show such localization and was globally overexpressed
268 in inactive SLE compared with healthy cells (dysregulated signature gene). Hierarchical clustering of 100
269 IRGs (**Supplementary Table 10**) based on ScQuaD model statistics across 18 cell types identified a
270 module representing abundance signature genes and another representing dysregulated signature

genes, suggesting the heterogeneous roles of IRGs on SLE pathophysiology (**Fig. 4d** and **Extended Data Fig. 5c**).

Next, we performed pathway enrichment analyses on these signatures for each cell type (**Fig. 4e**). We observed enrichment of IFN signaling in both dysregulated and abundance signatures across almost all cell types (FDR < 0.05; one-sided Fisher's exact test), demonstrating that IFN signaling exerts its pathogenic effects through both quantitative and qualitative changes. However, other classical immunological pathways were primarily enriched in abundance signature genes of relevant cell types (e.g., TCR signaling in T cells and complement signaling in myeloid cells; **Fig. 4e**, **Extended Data Fig. 5d**, and **Supplementary Table 11**), suggesting that key biological processes in SLE were more linked to abundance signatures rather than dysregulated signatures.

To estimate the causal roles of these two signatures in SLE, we integrated ScQuaD statistics with the largest-scale SLE-GWASs in East Asian (EAS)[42] and EUR[43]. We tested the polygenic enrichment of SLE risk variants around the signature genes using stratified linkage disequilibrium score regression (S-LDSC)[44] for EAS- and EUR-GWAS and meta-analyzed them **(Extended Data Fig. 5e**). While SLE risk variants were enriched around both dysregulated and abundance signature genes in multiple cell types (FDR < 0.05), the extent of enrichment was more remarkable in abundance signature genes ($P = 5.3 \times 10^{-4}$; Paired Wilcoxon test; **Fig. 4f**). Notably, this enrichment was observed in the abundance signature genes with the positive statistics, but not in those with the negative statistics

289 (**Extended Data Fig. 5f**). Therefore, our model demonstrated that cell state compositional changes
290 within a cell type (especially those of expanded cell states) can exert causal effects on SLE
291 pathogenesis, rather than reflecting the secondary effects during the chronic disease phase.
292

293 **Fine mapping of multimodal molecular profiles across cell states**

294 Having delineated cell state abundance and gene programs linked to SLE pathophysiology, we
295 next sought to profile these cell states in greater detail. To this end, we prepared a new multimodal
296 dataset with CITE-seq (130 surface marker proteins with gene expression), TCR-seq, and enhancer
297 RNA profiling (dataset 3; 32 SLE donors). Reference mapping[45] successfully integrated 268,783 PBMCs
298 from dataset 3 into the latent transcriptional space in dataset 1, enabling us to extend our fine-grained
299 strategy to the multimodal layers (**Fig. 5a** and **Methods**). First, we comprehensively investigated cell-
300 state-specific surface proteins and identified, on average, nine upregulated proteins per cell state (FDR <
301 0.05; **Extended Data Fig. 6a** and **Supplementary Table 12**). We recapitulated the well-known surface
302 proteins in some cell states: (1) CD278 (ICOS), CD279 (PD1), and TIGIT in Tph (EMCD4-7)[7], (2) CD11c
303 and CD19 in ABC (MemB-6)[8], and (3) CD38 and CD31 in *SOX4*$^{+}$ naive T cell states (NaiveCD4-3 and
304 NaiveCD8-3)[18]. Furthermore, we successfully identified several key proteins of new disease-relevant cell
305 states, which were utilized in the subsequent investigation (**Fig. 5-6**).

306 We next investigated TCR repertoire profiles. While we and others have shown that V gene
307 usage differs across cell types[46,47] and could be a new therapeutic target for autoimmunity[48], a
308 systematic cell-state-level analysis has been lacking. Principal component analysis (PCA) of V gene
309 frequency for each cell state confirmed that embeddings globally separated cell types (**Extended Data**
310 **Fig. 6b**). We next compared V gene usage across cell states. Notably, out of 94 V genes (45 α- and 49
311 β-chain), ~3.8 V genes per cell state showed skewed usage (FDR < 0.05; two-sided Fisher's exact test;
312 **Extended Data Fig. 6c** and **Supplementary Table 13**), highlighting distinct TCR signatures at the cell-
313 state level. While most of them were observed in EMCD8 cell states, we successfully identified several
314 cell-state-specific V genes in other cell types (e.g., NaiveCD4-5; **Fig. 5**).
315 Leveraging 5' transcript sequencing, we next mapped transcription start sites (TSSs) at single-
316 cell resolution. Recent studies have shown that bidirectional transcription around TSSs marks enhancer
317 activity at the locus[13,14]. Using this strategy, we called ~1,564 transcribed enhancers per cell type
318 (**Methods**), which were enriched in PBMC enhancer marks than others[49] (**Extended Data Fig. 6d**).
319 Among them, we identified ~68 cell-state-specific enhancers per cell type (**Supplementary Table 14**).
320 Thus, the efficient integration of our multimodal data with large-scale scRNA-seq data
321 successfully provided comprehensive molecular information on fine-grained cell states.
322

323 ***FOXO1* is a candidate driver of the pathogenic CD4$^+$ T cell state gene program**

We next advanced molecular fine-mapping to investigate disease-relevant cell states. Motivated by our findings that key signal pathways and SLE causal variants were more enriched in abundance signatures than dysregulated signatures (**Fig. 4e-f**), we focused subsequent analyses on several new cell states that expanded in both disease-state and activity CNA (**Fig. 3b**).

Although some expanded CD4$^+$ T cell states such as Tph (EMCD4-7) and *ISG15*$^+$*MX1*$^+$ cell states have been well-characterized, the molecular features of our newly identified cell states remain unclear: (1) *FAM13A*$^+$*CDK6*$^+$ cells (NaiveCD4-4) in NaiveCD4, and (2) *ARHGAP15*$^+$*FOXO1*$^+$ cell states in multiple T cell types (e.g., NaiveCD4-5 and CMCD4-4, **Fig. 3b**). Pathway analyses of upregulated marker genes showed the enrichment of JAK-STAT signaling for *FAM13A*$^+$*CDK6*$^+$ cells and TCR signaling for *ARHGAP15*$^+$*FOXO1*$^+$ cells, demonstrating the distinct immunological roles for these cell states (**Fig. 5b** and **Supplementary Table 6**). We observed a unique TCR V gene usage pattern in *ARHGAP15*$^+$*FOXO1*$^+$ cells (NaiveCD4-5; **Fig. 5c** and **Supplementary Table 13**). We also identified cell-state-specific enhancers near the marker genes, suggesting unique epigenetic regulations (**Extended Data Fig. 7a** and **Supplementary Table 14**).

We next examined the localization of GWAS-prioritized genes for each cell state. In *FAM13A*$^+$*CDK6*$^+$ cells, *ARID5B*, a transcription coactivator, was linked to one SLE-GWAS locus[41] (**Fig. 5d** and **Supplementary Table 9**). In *ARHGAP15*$^+$*FOXO1*$^+$ cells, *ARHGAP15*, encoding a GTPase-activating protein, was linked to another SLE-GWAS locus. Furthermore, we identified several key

surface proteins associated with these cell states. CD127 (IL-7 receptor α) was upregulated in *FAM13A*$^+$*CDK6*$^+$ cells (**Fig. 5e**), reflecting the promotion of their differentiation[50]. On the other hand, a signaling lymphocyte activation molecule [SLAM] family protein CD244 was consistently upregulated in multiple *ARHGAP15*$^+$*FOXO1*$^+$ cell states (NaiveCD4-5, CMCD4-4, and NaiveCD8-4; **Supplementary Table 12**).

Finally, to identify transcriptional regulators driving the gene program of *FAM13A*$^+$*CDK6*$^+$ and *ARHGAP15*$^+$*FOXO1*$^+$ cells, we took an experimental approach. We selected 22 TFs upregulated in these cells and knocked out (KO) each TF, using a CRISPR array screen in primary CD4 T cells from two healthy donors (**Fig. 5f**, **Supplementary Table 15**). Genes downregulated by *FOXO1*-KO were closest to *ARHGAP15*$^+$*FOXO1*$^+$ upregulated signature genes (Bonferroni corrected $P < 0.05$; one-sided Fisher's exact test; **Fig. 5g** and **Supplementary Table 16**). A validation experiment from the 3rd donor showed concordant results (**Extended Data Fig. 7b**). Indeed, *ARHGAP15*$^+$*FOXO1*$^+$ signature genes showed enrichment for FOXO signaling pathway (**Extended Data Fig. 7c** and **Supplementary Table 6**). When focused on EMCD4 cells, *FOXO1* showed the highest expression in Tph (EMCD4-7; **Fig. 5h** and **Extended Data Fig. 3a**). Consistently, trajectory analysis[51] on whole CD4$^+$ T cells showed the differentiation of *ARHGAP15*$^+$*FOXO1*$^+$ cell states toward Tph (and ProlifCD4) rather than other effector cells. Given that *FOXO1* has recently been reported to be responsible for T cell memory-ness promotion and clinical efficacy in cancer immunotherapy[20,21], the inhibition of *FOXO1* could be a new treatment

360 option for autoimmunity by modulating the aberrant gene program in CD4 T cells. On the other hand, we
361 observed that the genes downregulated by *ARID5B*-KO did not account for the *FAM13A*$^+$*CDK6*$^+$ cell
362 state, indicating that *ARID5B* does not play a significant role in defining the gene program of this
363 population (**Extended Data Fig. 7d**). Together, our molecular fine mapping pinpointed key regulator
364 candidates in disease-relevant CD4 T cell states of SLE.
365

366 **Multimodal molecular fine mapping on *GZMK*$^+$*GZMH*$^+$*HLA-DRB1*$^+$ EMCD8 cells**

367 Among CD8$^+$ T cell states exhibiting remarkable expansions both in disease-state and activity
368 CNA, the most uncharacterized one is *GZMK*$^+$*GZMH*$^+$*HLA-DRB1*$^+$ EMCD8 cells (EMCD8-5, DP EMCD8).
369 Among upregulated signature genes in DP EMCD8, *RGS1*, known as a chemokine signaling regulator
370 and a cell exhaustion marker[52], was linked to one SLE-GWAS locus[41] (**Fig. 6a**). Indeed, we observed the
371 highest classical exhaustion signature scores (e.g., *PDCD1* and *TIGIT*, **Supplementary Table 17**) on
372 this cell state in EMCD8 (**Fig. 6b**). Notably, this pattern was comparable to Tph cells in EMCD4,
373 suggesting chronic antigen exposure to these cell states[53]. Consistently, DP EMCD8 showed evidence of
374 TCR clonal expansion in multiple SLE donors (**Fig. 6c**). Some of the expanded clonotypes overlapped
375 with those that reacted to epitopes of cytomegalovirus or Epstein-Barr virus (**Supplementary Tables**
376 **18**). While many of these clones were shared with other EMCD8 cell states, we observed 9 clonotypes
377 and 9 V genes specific to DP EMCD8 (FDR < 0.05; two-sided Fisher's exact test; **Extended Data Fig.**

8a-b, Supplementary Tables 13 and **18**). Pathway analyses showed the enrichment of oxidative phosphorylation signaling in DP EMCD8 and Tph (**Extended Data Fig. 8c** and **Supplementary Table 6**), suggesting the shared metabolomic reprogramming between these two cell states. We also examined the cytokine profile of DP EMCD8. Intriguingly, we found that this population served as a major source of proinflammatory cytokines, including *IFNG*, *IL16*, *IL32*, *CCL5*, and *CD70* (**Fig. 6d**). Accumulating evidence has suggested that *IFNG* (type II IFN) is mainly produced by $CD8^+$ T cells and plays a pivotal role in SLE as well as type I IFN[54–56]. Thus, our fine mapping strategy successfully pinpointed the cell-state origin of this key cytokine.

We next sought to identify key surface proteins of DP EMCD8. Integration of the multimodal dataset 3 into dataset 1 through reference mapping showed that HLA-DR, CD27, CD38, and CD28 as the most upregulated proteins in DP EMCD8 (**Fig. 6e** and **Supplementary Table 12**). Notably, CD45RA was downregulated in DP EMCD8, suggesting this population does not overlap with terminally differentiated effector memory (TEMRA) $CD8^+$ T cells[57]. An inhibitory marker, GPR56, was also downregulated in DP EMCD8. In-silico gating on HLA-$DR^+CD27^+CD38^+CD28^+$ successfully achieved a purity of >90% for DP EMCD8 (**Extended Data Fig. 8d**), enabling us to investigate this population using flow cytometry (FCM) and fluorescence-activated cell sorting (FACS).

We first FACS-sorted target EMCD8 cells gated on $CCR7^-$HLA-$DR^+CD27^+CD38^+CD28^+$ from purified $CD8^+$ T cells and performed bulk RNAseq and Assay for Transposase-Accessible Chromatin

396 with sequencing (ATAC-seq) (**Fig. 6f**, **Extended Data Fig. 8e**, and **Supplementary Table 19**). After
397 confirming that these sorted cells showed a highly comparable gene expression profile to the DP EMCD8
398 (**Extended Data Fig. 8f**), we identified 113 differentially accessible regions (DARs) in this population
399 compared with whole EMCD8 (FDR < 0.1; **Extended Data Fig. 8g** and **Supplementary Table 20**). Motif
400 analysis for these DARs revealed the enrichment of Fli1 and IRF motifs (**Fig. 6g**). Since there were no
401 DARs around the *IFNG* region in standard ATAC-seq, we enriched open chromatin peaks around this
402 region using UNIChro-seq[58] (**Methods** and **Supplementary Table 21**). Around 20-fold enrichment of
403 peak counts detected one DAR, suggesting the epigenetic dysregulation of *IFNG* in this population. (**Fig.**
404 **6h**).

405 In intracellular FCM, we confirmed that a larger proportion of HLA-DR$^+$CD27$^+$CD38$^+$CD28$^+$
406 EMCD8 cells co-express GZMK and GZMB at the single-cell level, compared with whole EMCD8
407 (**Extended Data Fig. 8h**). Importantly, these cells showed high PD1 and TIGIT expression even at the
408 protein level (**Fig. 6i** and **Extended Data Fig. 8h**). Upon stimulation with Phorbol myristate acetate
409 (PMA)/Ionomycin, this population exhibited higher IFN-γ expression than whole EMCD8.

410 Identification of cell-state-specific surface proteins could have significant implications for drug
411 target discovery. The current therapeutic agents in autoimmunity target ubiquitous surface markers,
412 which can deplete normal cells as well as autoreactive cells (e.g., CD19 and CD20 in B cells[59,60]).
413 Narrowing down the target to disease-relevant cell states based on their specific markers could provide

more efficient treatment options. Collectively, our extensive approaches revealed proinflammatory molecular phenotypes of DP EMCD8, underscoring this population as a potential key player in SLE.

Discussion

In this study, we developed an analytical framework, cellular and molecular fine mapping, to pinpoint disease-relevant cell states and molecules at single-cell resolution, applying it to ~2.1 million cells from 346 donors. Fine-grained clustering resolved 123 cell states across 27 cell types, including 11 previously uncharacterized ones that expanded especially in severe SLE. Subsequent in-silico multimodal analyses, CRISPR perturbations, and functional experiments successfully defined their key transcriptional regulators, cytokines, TCR features, and surface proteins, guiding the efficient design of therapeutics for clinical settings.

In cellular fine mapping, we noted an important caveat: to pinpoint pathogenic cell states, careful fine-grained clustering is essential. The most disease-relevant cell states are often low-frequency populations and are easily obscured by coarse clustering, consistent with other studies that successfully identified rare pathogenic populations[7,8,24]. Based on the extensive literature review[7,8,15–38] and rigorous validation, we reconstructed 123 cell states with distinct immunological features from the largest, publicly available single-cell data[5] (~1.5 million PBMCs). While the recent single-cell studies on inflamed tissues in autoimmune diseases have identified several critical parenchymal subpopulations[24,61], the resolution of

immune cells is limited (~60 immune cell states from 0.2~0.4 million cells). To benefit the research community, we released the most fine-grained immune cell reference panels with multimodal features, called LuPIN (**Methods**). Researchers can map their datasets onto our reference panels for cell-state annotations, providing an opportunity to uncover new pathophysiology of multiple diseases.

In this study, we proposed a new statistical model to overcome the limitation of the conventional cell-type-level pseudobulk approach. To pinpoint molecular therapeutic targets from single-cell data, distinguishing quantitative and qualitative changes linked to disease is a critical step, since each could play different roles in disease pathophysiology. Some researchers are tackling this issue with different statistical concepts from ours[11]. While such approaches are expected to be the mainstream of next-generation single-cell research, it must be noted that our model relies on the latent space defined by current sparse single-cell gene expression. The distinction between abundance and dysregulated signatures may be refined in the future as single-cell technology advances.

Despite rapid growth in single-cell datasets, biological insights have not kept pace with the rising number of reported association signals. This trajectory mirrors early era of GWAS, when efforts were focused primarily on variant discovery rather than mechanistic understanding[62]. GWAS ultimately matured by shifting from simple association to fine-mapping of causal variants followed by functional dissection[63] (the so-called *post-GWAS era*); similarly, single-cell studies should progress from lists of coarse associations toward fine-mapping of pathogenic cell states and their key driver molecules (a *post-*

450 *cell atlas era*). Given the abundance of public datasets now available, we advocate reallocating
451 resources from generating ever more data to performing targeted functional characterizations.

452 This study has some limitations. First, the current study focused on the functional investigation
453 of three cell states in T cells. Further characterization of other disease-relevant population candidates will
454 deepen understanding of complex pathophysiology in SLE. Second, because this study focused on
455 PBMCs, we may have missed key tissue-resident populations. Our fine-mapping approach should next
456 be extended to inflamed tissues in multiple diseases. Third, our study design lacked in in-vivo functional
457 validation of disease-relevant cell states.

458 Together, our intensive cellular and molecular fine mapping can contribute to a better
459 understanding of SLE pathophysiology and the discovery of new treatment targets.
460

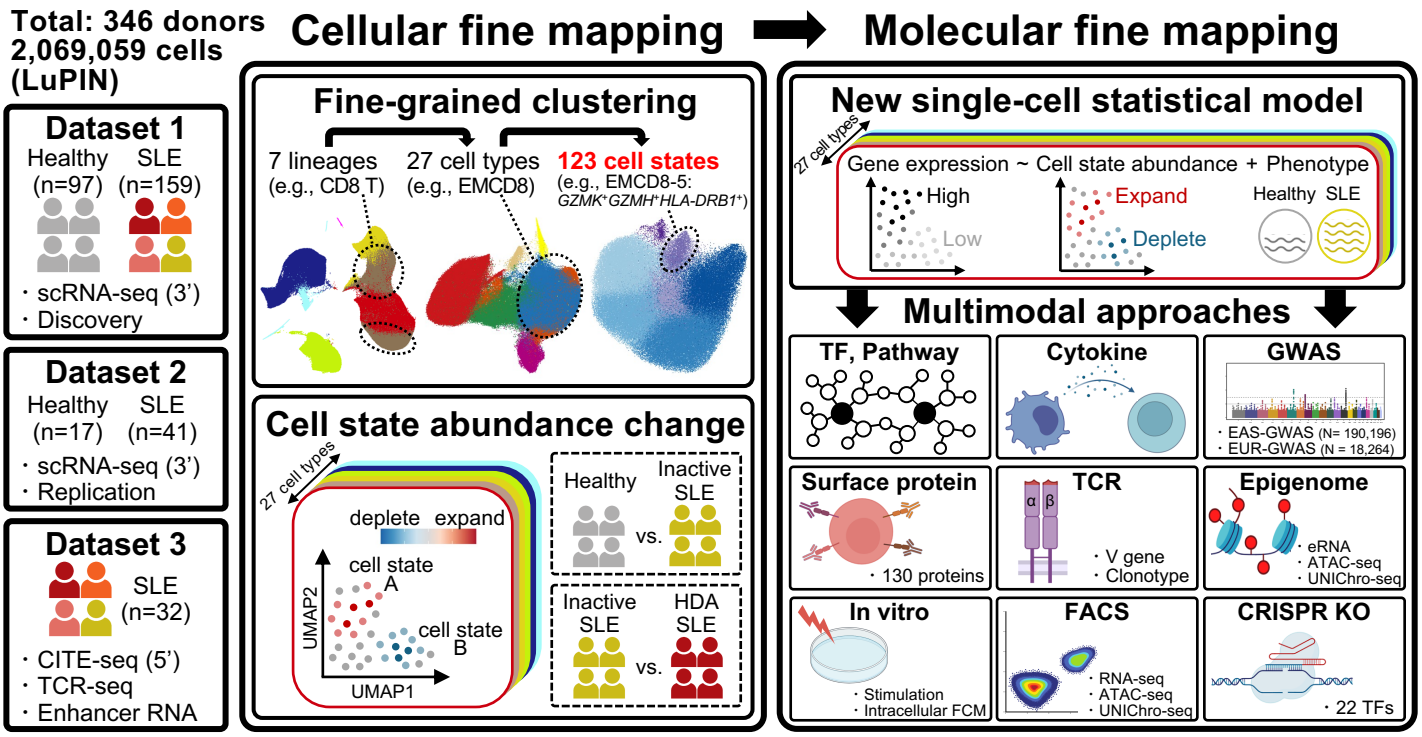


Figure 1. Overview of this study. (left) We applied cellular and molecular fine mapping to the largest-scale multimodal single-cell datasets of SLE, comprising ~2.1 million PBMCs from 346 donors. **(middle)** Through fine-grained clustering, we first identified 123 distinct cell states in 27 immune cell types. Among them, we found previously uncharacterized cell states expanded especially in clinically active SLE. **(right)** Next, our new statistical model, which dissected gene signatures reflecting quantitative and qualitative changes linked to phenotypes, demonstrated that key pathways and SLE risk variants were linked to cell-state compositional changes within each cell type. Finally, through multimodal in-silico and experimental approaches, we successfully revealed key molecular profiles of these disease-relevant cell states. CITE-seq, cellular indexing of transcriptomes and epitopes by sequencing; TCR, T cell receptor; EMCD8, effector memory $CD8^+$ T cells; HDA, high disease activity; TF, transcription factor; GWAS, genome-wide association studies; eRNA, enhancer RNA; UNIChro-seq, unique molecular identifier counting of regional chromatin accessibility with sequencing; FCM, flowcytometry; FACS, fluorescence-activated cell sorting.

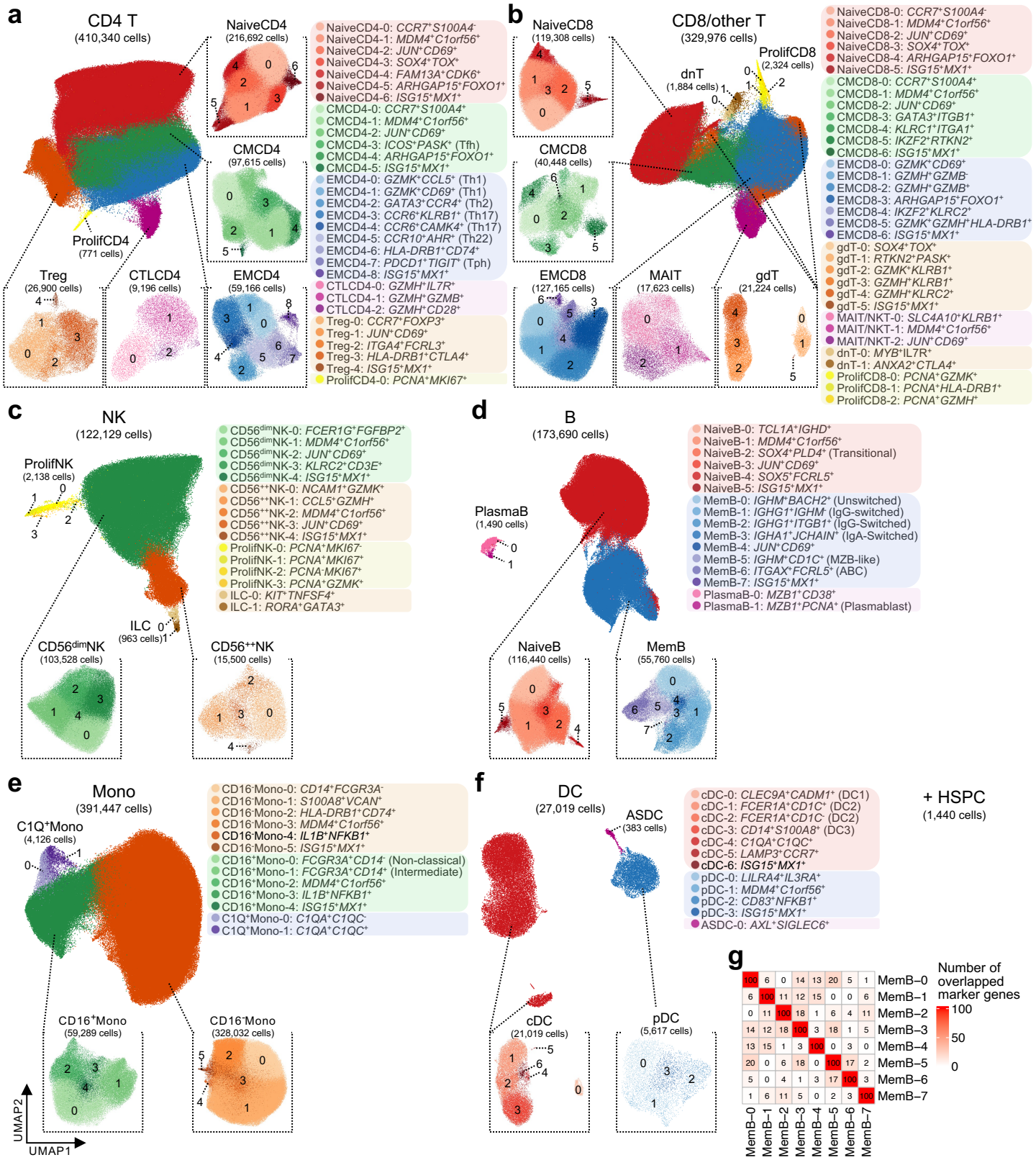


Figure 2. Identification of 123 distinct cell states using fine-grained clustering. (a-f) UMAP plots showing 27 cell types and 123 cell states identified by fine-grained clustering in **(a)** CD4+ T cell (CD4 T), **(b)** CD8+ and unconventional T cell (CD8/other T), **(c)** natural killer cell (NK), **(d)** B cell, **(e)** monocyte (Mono), and **(f)** dendritic cell (DC) lineages (dataset 1; N=256). Hematopoietic stem and progenitor cell (HSPC) lineage is not shown since it has only one cell type and one cell state. In **(a-f)**, large UMAPs represent lineage-level UMAPs, colored by cell types. Small UMAPs represent cell-type-level UMAPs, colored and numbered by cell states. For each panel, cell state annotations are listed on the right. **(g)** Heatmap showing the number of overlaps across 100 upregulated marker genes for each cell state in memory B cells (MemB). See **Supplementary Table 3** for the description of cell type abbreviations.

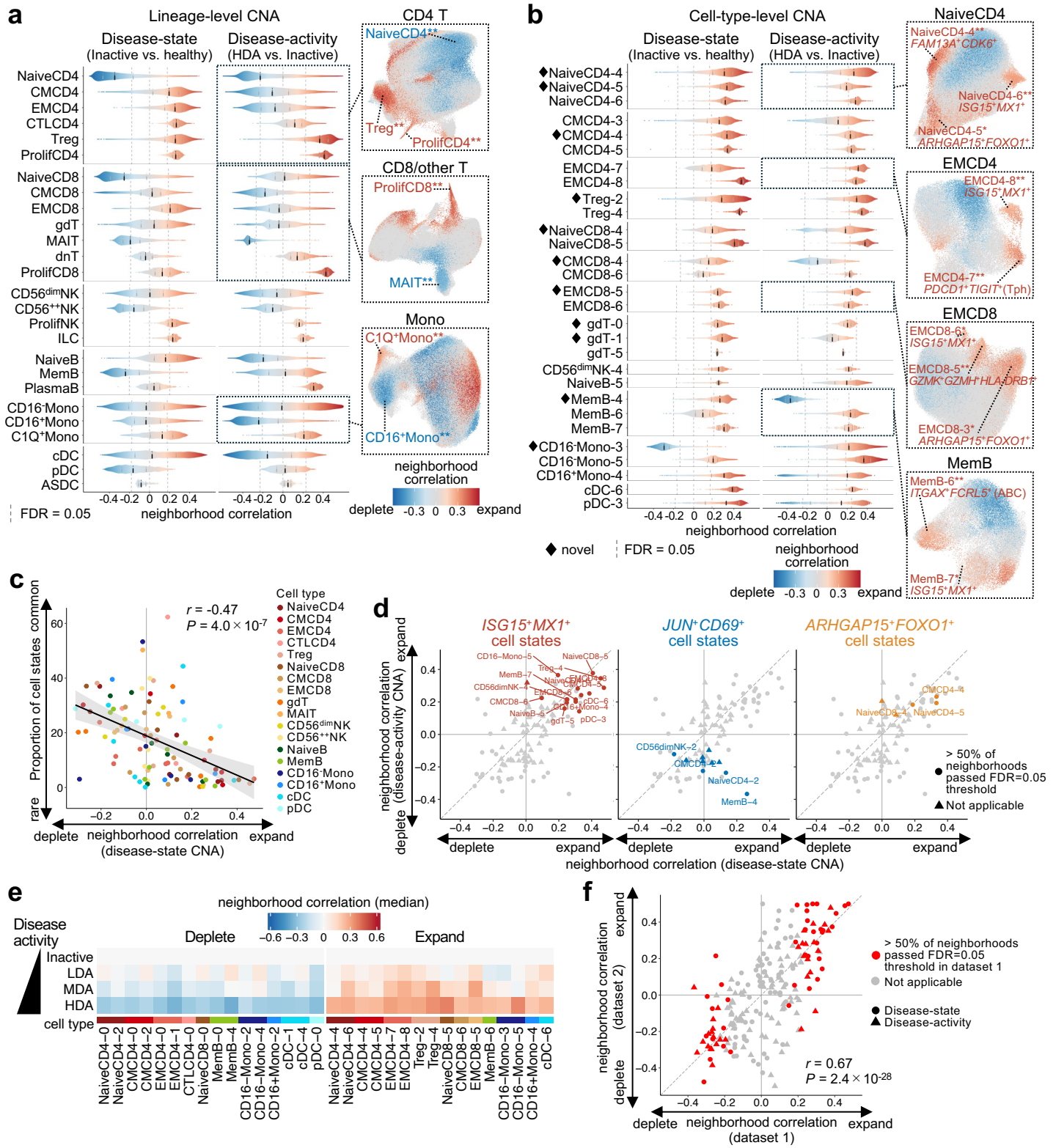


Figure 3. Cellular fine mapping to pinpoint disease-relevant cell states of SLE. (a-b) (left) Distributions of neighborhood correlations in disease-state or disease-activity comparisons **(a)** for each cell type in lineage-level-CNA and **(b)** for each cell state in cell-type-level-CNA (dataset 1; N=256). In **(b)**, 29 cell states with > 50% of neighborhoods positively associated with phenotypes at false discovery rate (FDR) < 0.05 are shown (See **Extended Data Fig. 4b** for details). **(right)** UMAP plots in representative lineages (or cell types). Cell types (or cell states) are annotated if > 50% (**) or > 25% (*) of neighborhoods associated with phenotypes. **(c)** A scatter plot showing the association between median neighborhood correlation for each cell state in cell-type-level disease-state CNA and cell-state proportion over cell type. Shaded regions represent 95% confidence intervals. **(d)** Scatter plots comparing median neighborhood correlation for each cell state between cell-type-level disease-state (x-axis) and disease-activity CNA (y-axis). Cell states related to each panel title are colored. **(e)** A heatmap showing the median neighborhood correlation for each cell state for low disease activity (LDA) vs. Inactive SLE, moderate disease activity (MDA) vs. Inactive SLE, and high disease activity (HDA) vs. Inactive SLE comparisons. 32 cell states with > 50% of neighborhoods associated with disease-activity comparison are shown. **(f)** Scatter plots comparing median neighborhood correlation for each cell state between dataset 1 (x-axis) and dataset 2 (y-axis; N=58). In **(c)** and **(f)**, Pearson correlation coefficients (*r*) between x- and y-axis and their *P* values are shown.

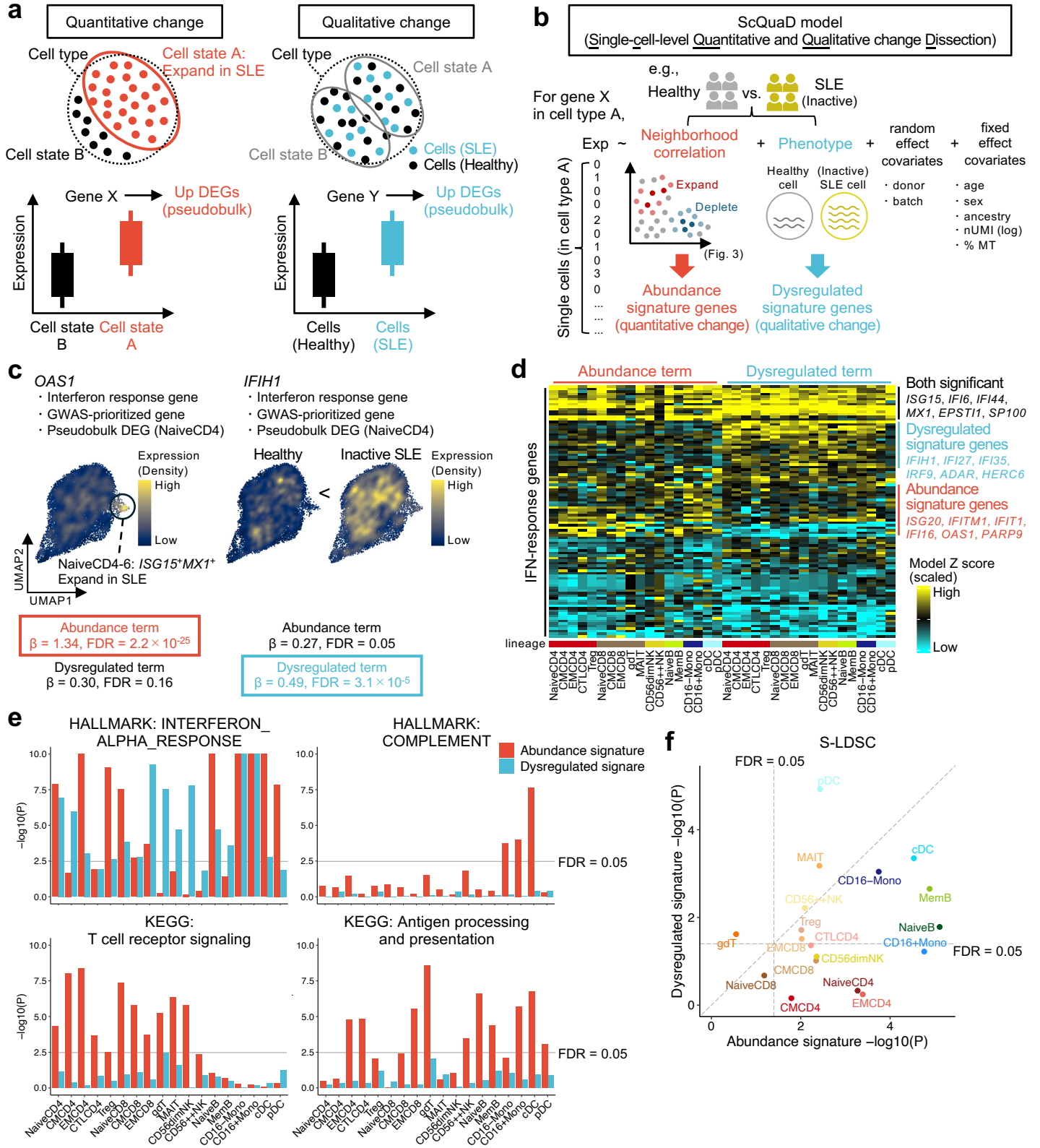


Figure 4. A new single-cell model to dissect quantitative and qualitative changes within cell types. (a) Shema of two different scenarios (quantitative and qualitative changes) to drive cell-type-level pseudobulk DEGs (differentially expressed genes) in SLE. **(b)** Shema of a new statistical model (ScQuaD) to dissect quantitative and qualitative changes (dataset 1; N=256). **(c)** Representative examples of abundance (*OAS1*) and dysregulated (*IFIH1*) signature genes. In UMAPs, cells are colored in expression density. Model statistics for both genes are shown in the bottom. **(d)** Heatmap showing the ScQuaD model Z scores of abundance and dysregulated terms across 18 major cell types for 100 IFN-response genes. Gene orders are based on hierarchical clustering using Z scores across cell types. **(e)** Bar plots showing the enrichment of representative pathways for each signature for each cell type. *P*, *P* values in one-sided Fisher's exact test. **(f)** Scatter plots comparing the enrichment of SLE risk variants around abundance (x-axis) and dysregulated signature genes (y-axis) for each cell type in S-LDSC (stratified LD score regression).

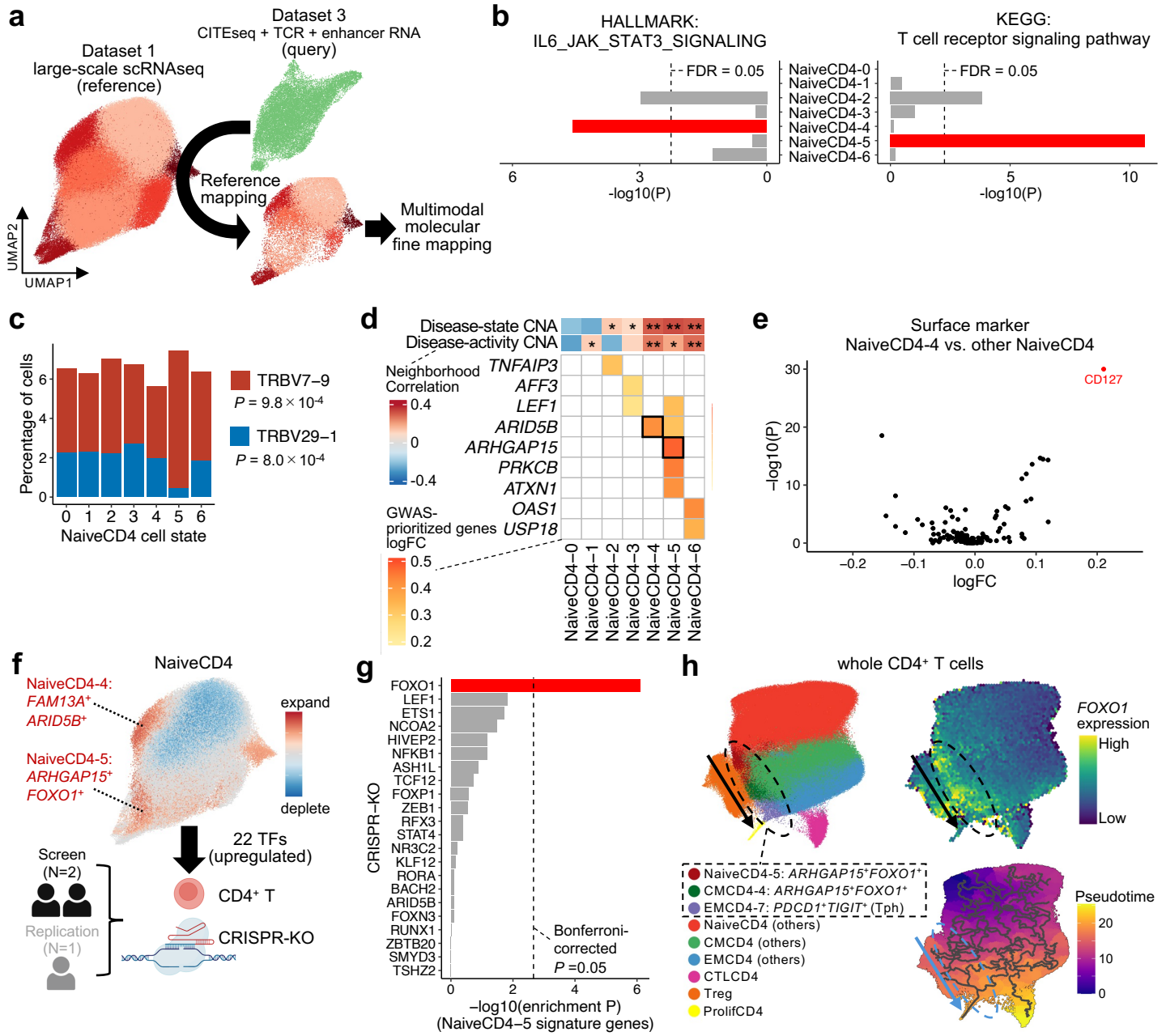


Figure 5. *FOXO1* is a candidate driver of the pathogenic CD4⁺ T cell state gene program. (a) Shema of the integration of multimodal dataset 3 (query; N=32) into larger scRNA-seq dataset 1 (reference; N=256) through reference mapping. **(b)** Bar plots showing the enrichment of representative pathways for upregulated signature genes for each NaiveCD4 cell state. *P*, *P* values in one-sided Fisher's exact test. **(c)** A bar plot showing the percentage of two TCR V gene usages that are skewed in NaiveCD4-5 compared with other NaiveCD4 cells. *P*, *P* values in two-sided Fisher's exact test. **(d)** Heatmap showing the logFC of GWAS-prioritized genes between one vs. other NaiveCD4 cell states. Up to three genes with the highest logFC per cell state are shown. Top annotation color represents the median neighborhood correlation of each cell state in disease-state and activity CNA (**Fig. 3**). **(e)** A volcano plot showing the differential expression of surface proteins between NaiveCD4-4 vs. other NaiveCD4 cell states. *P*, *P* values in differential expression with FindMarker function (Seurat; **Methods**). **(f)** Experiment design of CRISPR-KO array screening to primary CD4⁺ T cells. 22 TFs upregulated in NaiveCD4-4 or NaiveCD4-5 are selected for array screening (**Methods**). **(g)** A bar plot showing the enrichment of downregulated genes by each KO condition for NaiveCD4-5 upregulated signature genes. **(h)** Whole CD4⁺ T cell UMAP plots, colored by **(left)** the focused cell states and other cell types, **(upper right)** *FOXO1* expression, and **(lower right)** pseudotime in trajectory analysis. Edges represent trajectory graph.

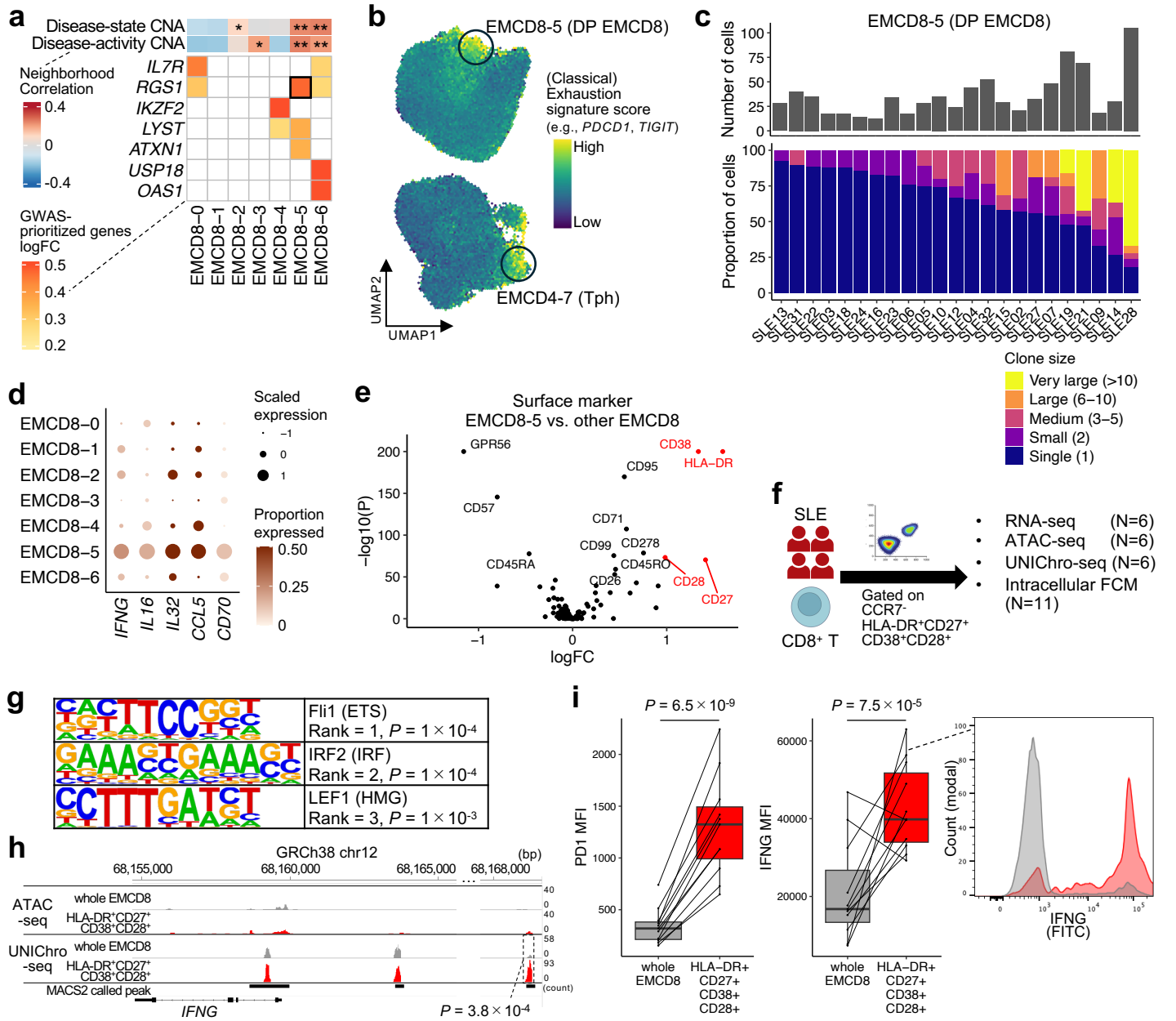
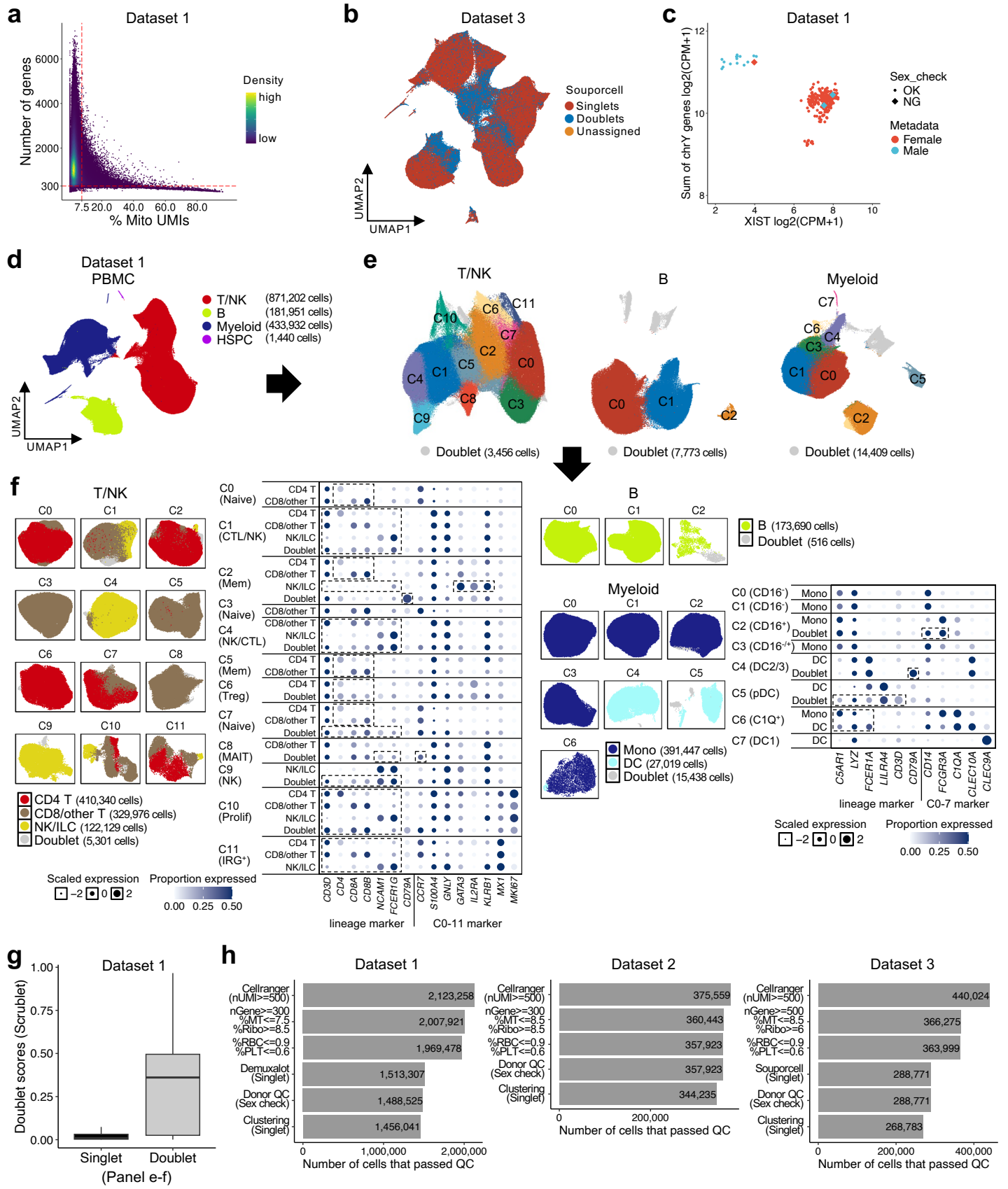
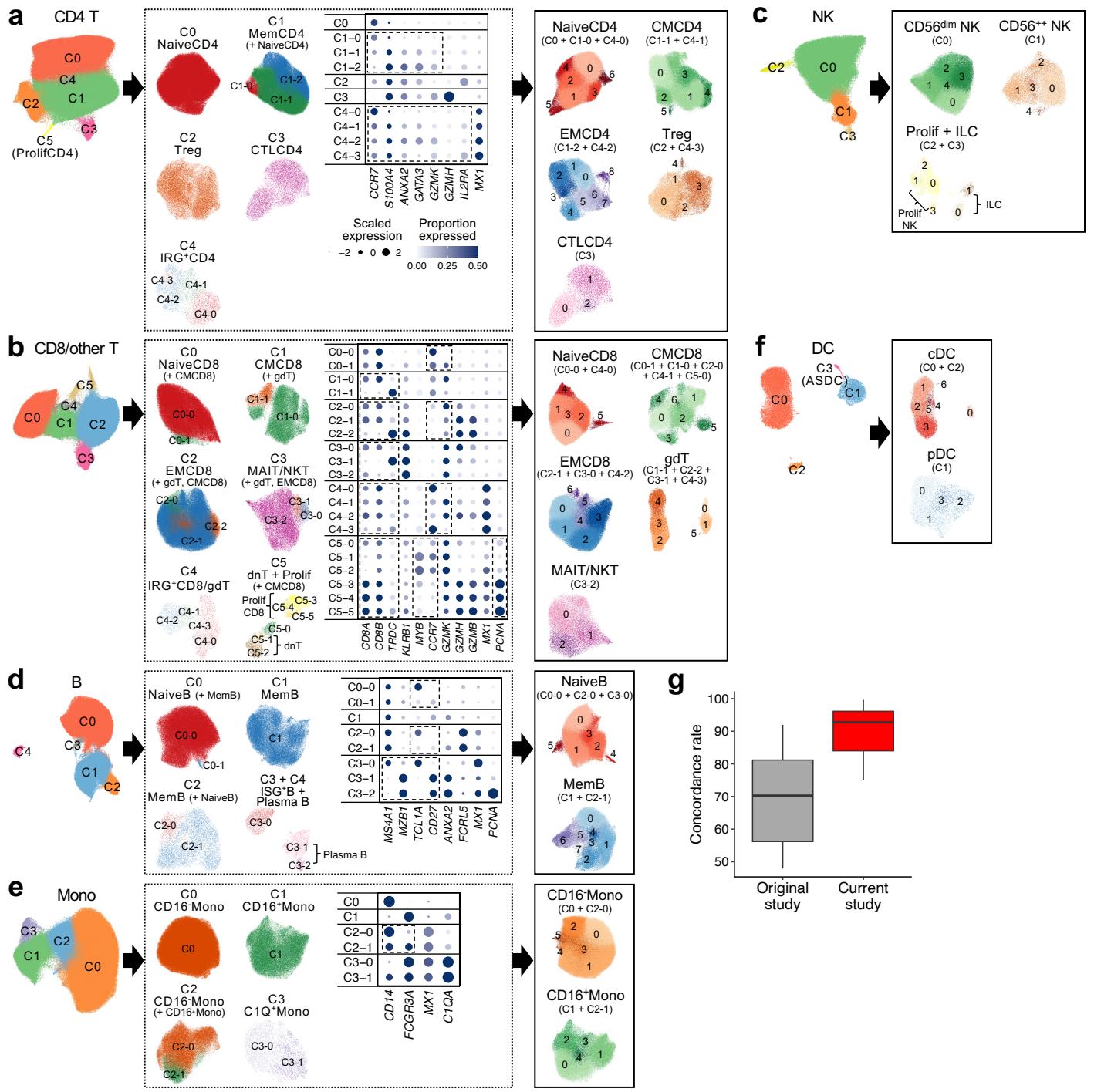


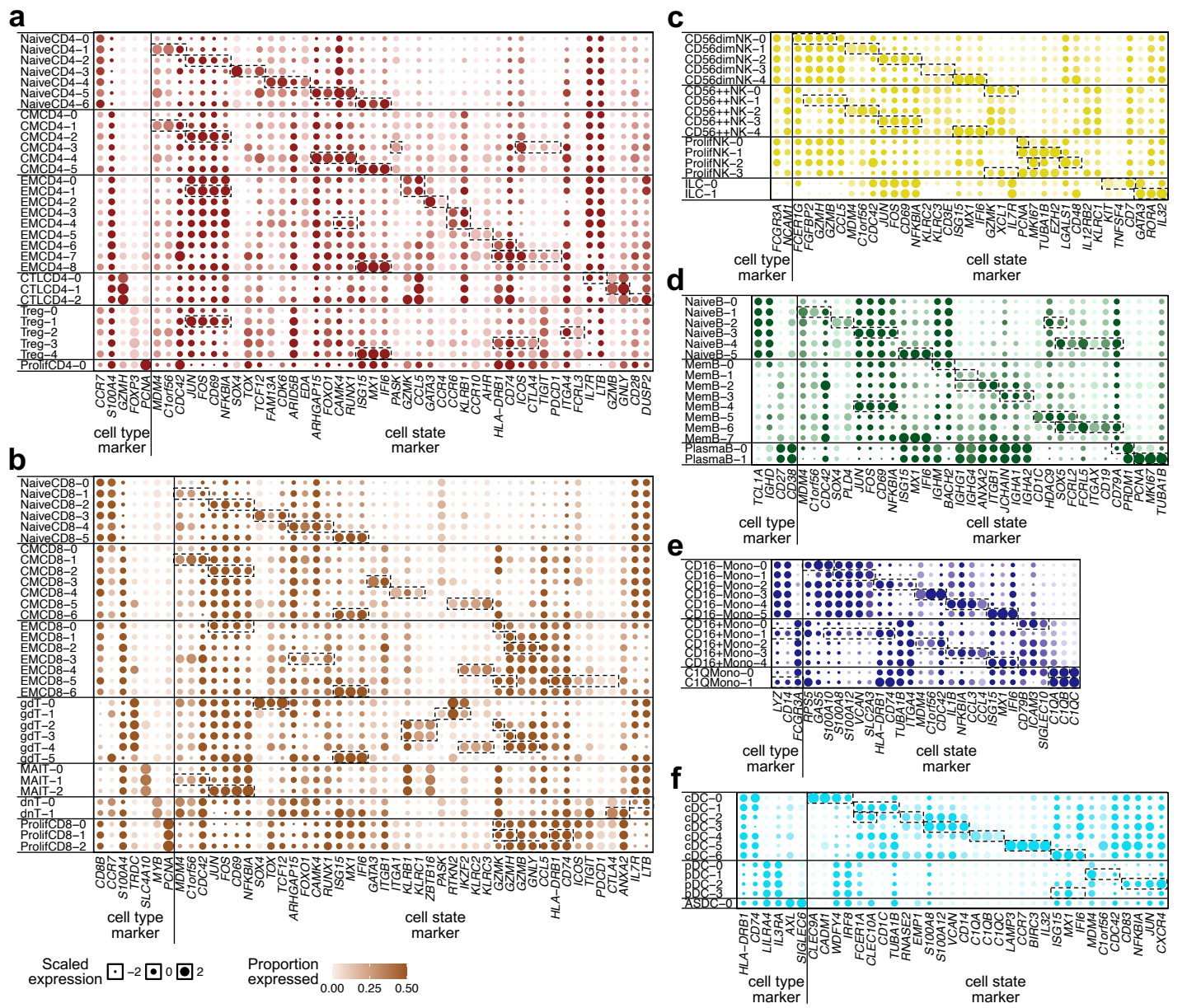
Figure 6. Multimodal molecular fine mapping on *GZMK*⁺*GZMH*⁺*HLA-DRB1*⁺ EMCD8 cells. (a) Heatmap showing the logFC of GWAS-prioritized genes between one vs. other EMCD8 cell states (dataset 1; N=256). Top annotation color represents the median neighborhood correlation in CNA. **(b) (top)** EMCD8 and **(bottom)** EMCD4 UMAP plots, colored by exhaustion signature scores. **(c)** Bar plots showing **(top)** the number of EMCD8-5 cells for each donor and **(bottom)** the proportion of TCR clonotypes that are categorized into five clone-size groups in EMCD8-5 (dataset3; N=32; 23 donors with >10 cells are shown). **(d)** A dot plot showing the expressions of upregulated cytokines in EMCD8-5. Color represents the proportion of cells expressing each gene. **(e)** A volcano plot showing the differential expression of surface proteins between EMCD8-5 vs. other EMCD8 cells. *P*, *P* values in differential expression (**Methods**). **(f)** Experiment design of flowcytometry (FCM) and FACS on HLA-DR⁺CD27⁺CD38⁺CD28⁺ EMCD8 cells. **(g)** TF motif logos enriched for differentially accessible peaks in HLA-DR⁺CD27⁺CD38⁺CD28⁺ cells compared with whole EMCD8 cells. *P*, enrichment *P* values in HOMER. **(h)** ATAC-seq and UNIChro-seq peaks around *IFNG* region in whole EMCD8 cells and HLA-DR⁺CD27⁺CD38⁺CD28⁺ cells from a SLE donor. *P*, logistic regression *P* values (N=6; **Methods**). **(i) (left)** Boxplots comparing the mean fluorescence intensities (MFI) of PD1 and IFNG between HLA-DR⁺CD27⁺CD38⁺CD28⁺ and whole EMCD8 cells. Samples from the same donor are connected with lines. *P*, *P* values in linear mixed models (N=11). **(right)** A histogram comparing the IFNG FI between HLA-DR⁺CD27⁺CD38⁺CD28⁺ and whole EMCD8 cells from a SLE donor.



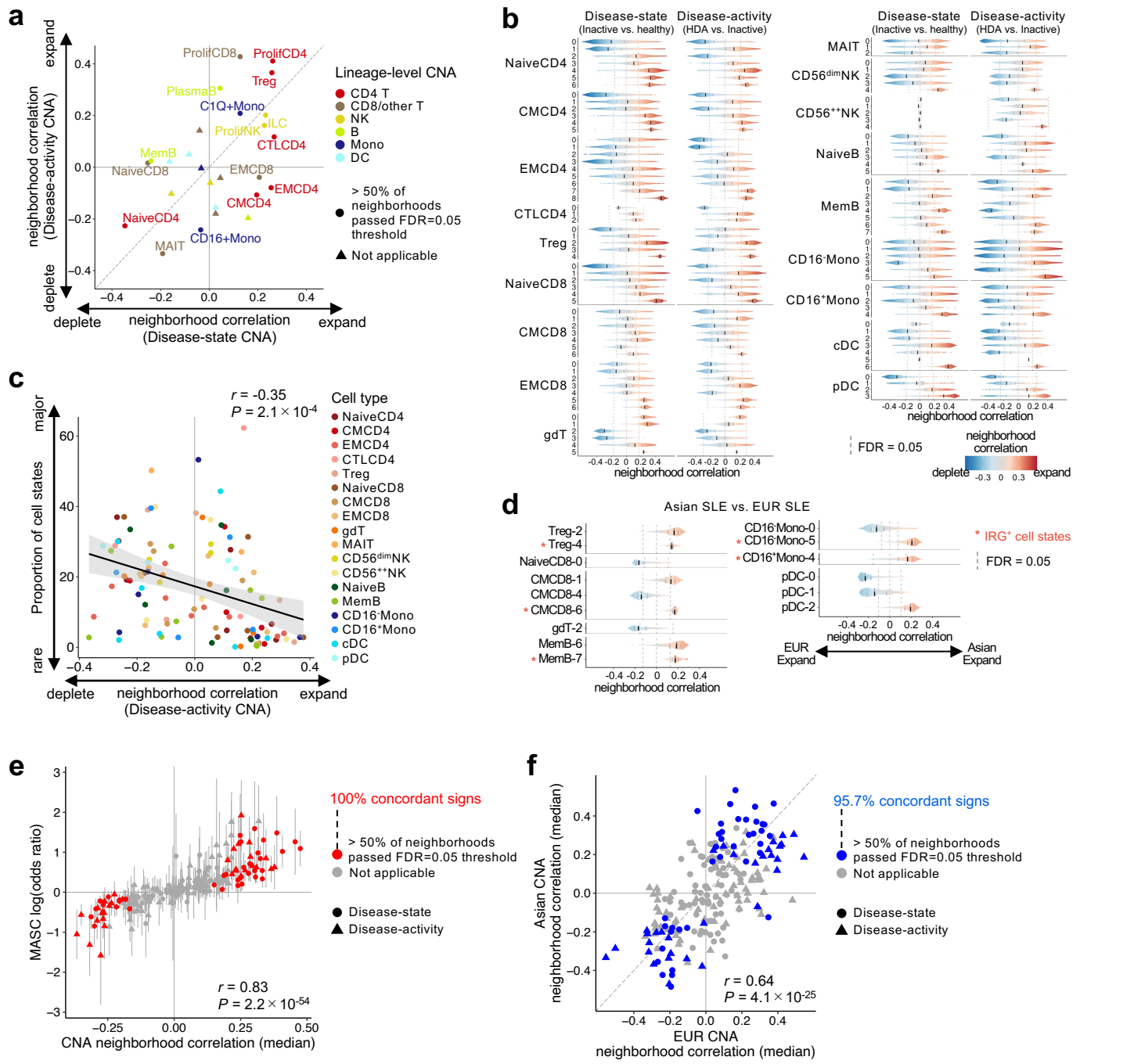
Extended Data Figure 1. Quality control of datasets. (a) A density plot showing the cell density based on the number of genes detected and the percentage of UMIs from mitochondrial genes (%Mito) in dataset 1 (N=256). **(b)** A UMAP plot showing PBMCs in dataset 3 (N=32), colored by the diagnosis of Souporcell. Due to the computational cost, dataset 3 cells are visualized instead of dataset 1. **(c)** A scatter plot showing donor-level pseudobulk gene expression of *XIST* and Y chromosome-related genes in dataset 1. Color represents the sex information from the metadata. **(d)-(f)** Flow chart to separate PBMCs into seven lineages in dataset 1 (**Methods**). In **(f)**, dot plots show the scaled expression levels of lineage or subcluster marker genes. Color represents the proportion of cells expressing each gene. **(g)** A boxplot comparing doublet scores in Scrublet between singlets and doublets detected in **(e)-(f)**. **(h)** Barplots showing the number of cells that passed each QC step in datasets 1, 2, and 3.



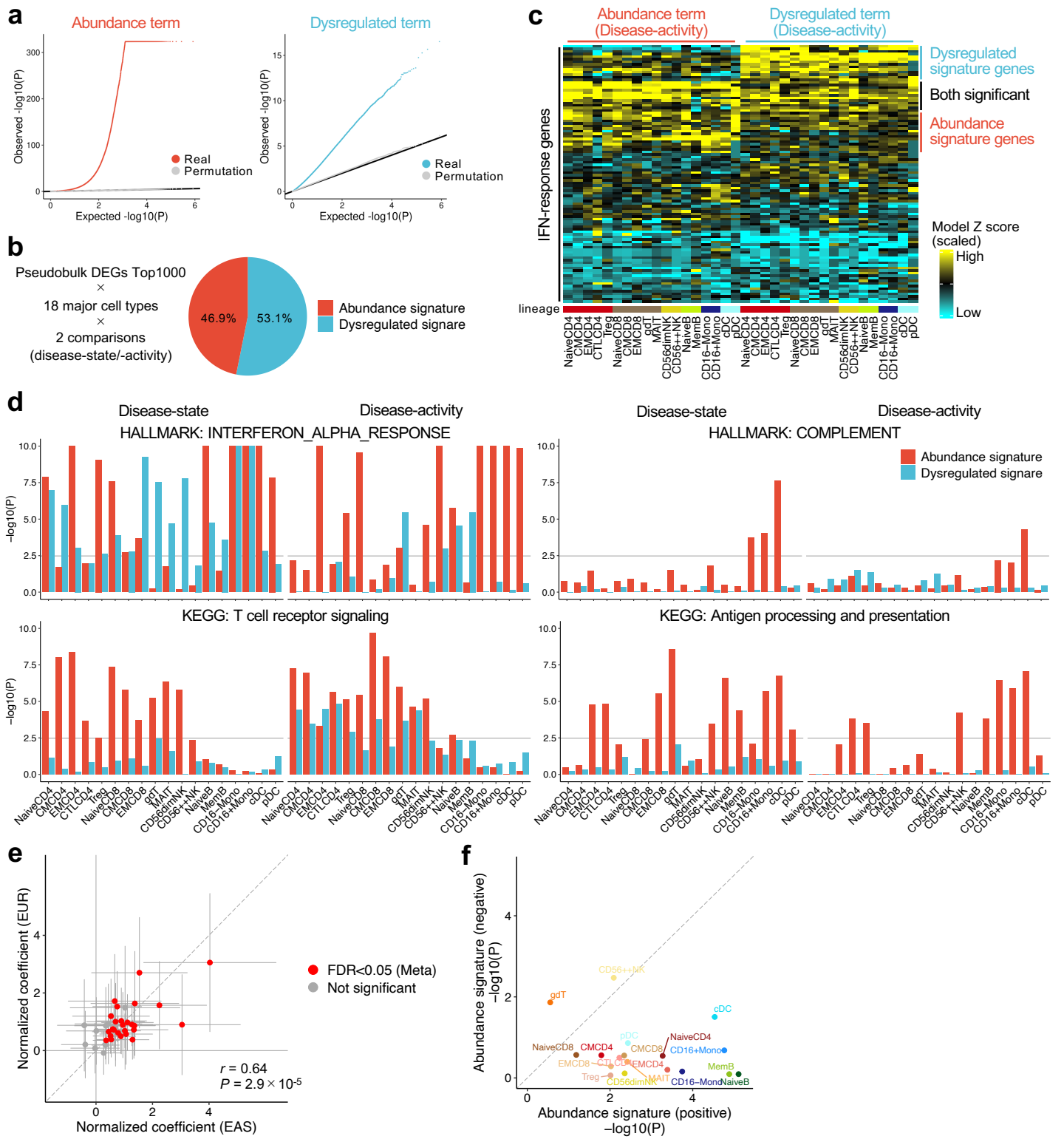
Extended Data Figure 2. Identification of 123 distinct cell states using fine-grained clustering. (a-f) Flow chart to separate lineage cells into cell types and cell states in **(a)** CD4+ T cell (CD4 T), **(b)** CD8+ and unconventional T cell (CD8/other T), **(c)** natural killer cell (NK), **(d)** B cell, **(e)** monocyte (Mono), and **(f)** dendritic cell (DC) lineages (dataset 1; N=256; **Methods**). Hematopoietic stem and progenitor cell (HSPC) lineage is not shown since it has only one cell type and one cell state. Dot plots show the scaled expression levels of cell type marker genes. Color represents the proportion of cells expressing each gene. **(g)** A box plot showing the concordance rates of cell-type-level annotation in the original and the current study with the predicted cell-type labels in the reference mapping onto the independent PBMC dataset (**Methods**).



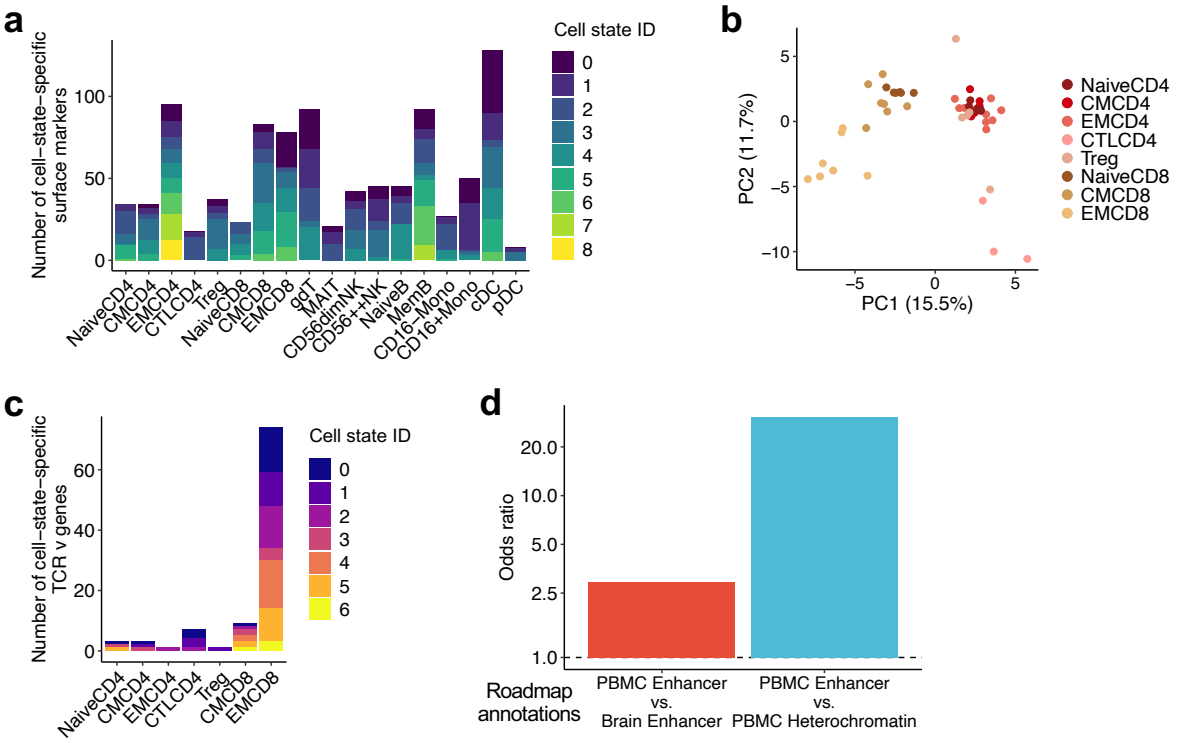
Extended Data Figure 3. Identification of 123 distinct cell states using fine-grained clustering. Dot plots showing the scaled expression levels of cell type and cell state marker genes in **(a)** CD4⁺ T cell (CD4 T), **(b)** CD8⁺ and unconventional T cell (CD8/other T), **(c)** natural killer cell (NK), **(d)** B cell, **(e)** monocyte (Mono), and **(f)** dendritic cell (DC) lineages (dataset 1; N=256). Color represents the proportion of cells expressing each gene.



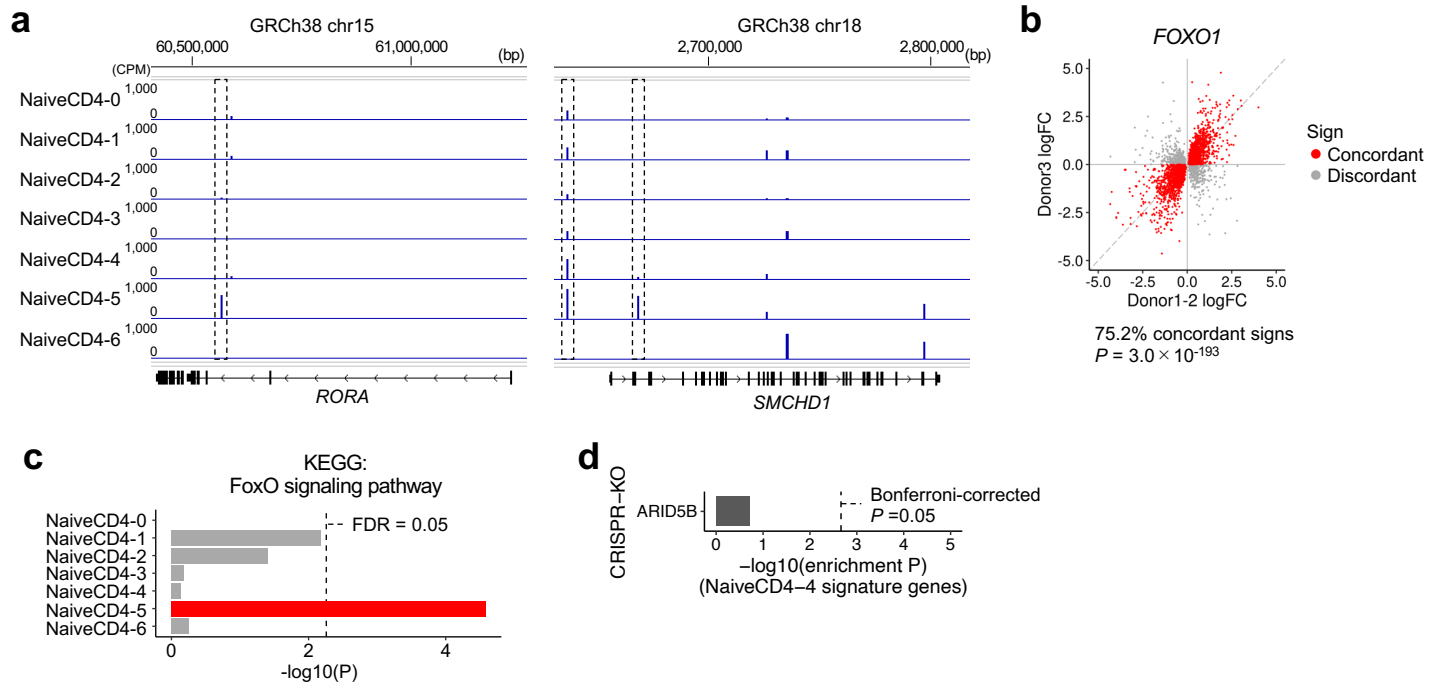
Extended Data Figure 4. Cellular fine mapping to pinpoint disease-relevant cell states of SLE. (a) A scatter plot comparing the median neighborhood correlation of each cell type between disease-state (x-axis) and activity (y-axis) comparisons in lineage-level CNA (dataset 1; N=256). **(b)** Distributions of neighborhood correlations for each cell state with disease-state and activity comparisons in cell-type-level-CNA. **(c)** A scatter plot showing the association between median neighborhood correlation of each cell state in cell-type-level disease-activity CNA and cell-state proportion over cell type. Shaded regions represent 95% confidence intervals. **(d)** Distributions of neighborhood correlations for each cell state with Asian vs. EUR SLE comparison in cell-type-level-CNA. 15 cell states with > 50% of neighborhoods associated with phenotype at false discovery rate (FDR) < 0.05 are shown. **(e)** A scatter plot comparing median neighborhood correlation in cell-type-level CNA and effect size in MASC for each cell state with disease-state and activity comparisons. **(f)** A scatter plot comparing median neighborhood correlation in cell-type-level CNA in European and cell-type-level CNA in Asian with disease-state and activity comparisons. In **(c)**, **(e)** and **(f)**, Pearson correlation coefficients (*r*) between x- and y-axis and their *P* values are shown.



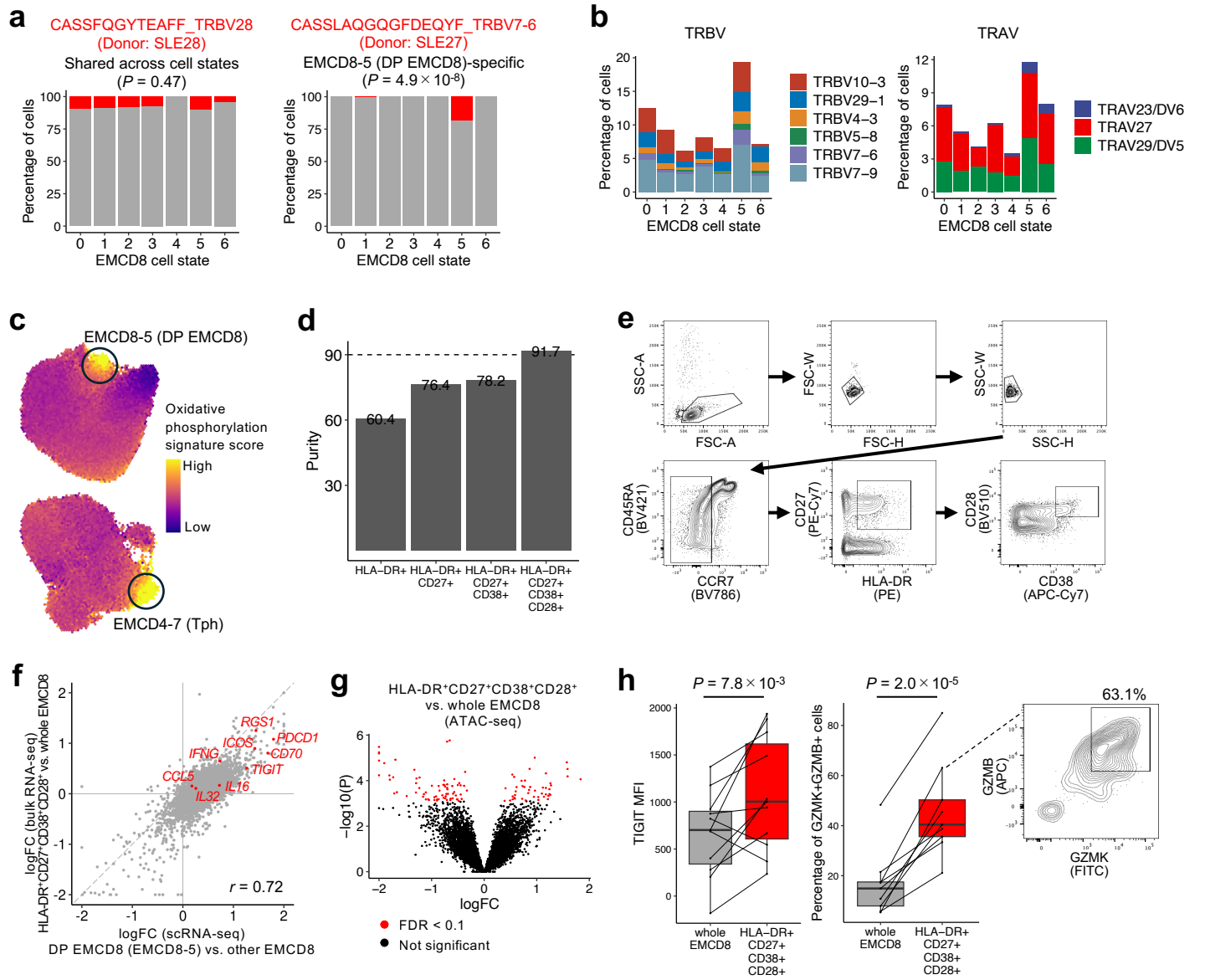
Extended Data Figure 5. A new single-cell model to dissect quantitative and qualitative changes within cell types. (a) Quantile-quantile plots comparing the observed *P* values in abundance (red) and dysregulated (blue) terms in the ScQuaD model across all cell types and comparisons (dataset 1; N=256). **(b)** Relative proportion of the variance explained by abundance (red) and dysregulated (blue) terms for the union of top 1,000 P-value pseudobulk DEGs in each cell type and each comparison. **(c)** Heatmap showing the ScQuaD model Z scores of abundance and dysregulated terms across 18 major cell types (row) for 100 IFN-response genes (column) in disease-activity comparisons. Gene orders are based on hierarchical clustering using Z scores across cell types. **(d)** Bar plots showing the enrichment of representative pathways for each signature for each cell type with disease-state and activity comparisons. *P*, *P* values in one-sided Fisher's exact test. **(e)** Scatter plots comparing the normalized coefficients in S-LDSC (stratified LD score regression) for EAS and EUR SLE-GWAS. Each dot represents each signature across 18 major cell types. Error bars represent 2 × standard errors. Pearson correlation coefficient (*r*) between them and *P* values are shown. **(f)** Scatter plots comparing the enrichment of SLE risk variants around abundance signature genes with positive (x-axis) and negative (y-axis) model statistics for each cell type in S-LDSC.



Extended Data Figure 6. Fine mapping of multimodal molecular profiles across cell states. (a) A bar plot showing the number of cell-state-specific upregulated surface marker proteins (dataset 3; N=32). Color represents cell state IDs for each cell state (e.g., NaiveCD4-0). **(b)** A PCA plot of TCR V gene usage frequency for each cell state, colored by cell types. **(c)** A bar plot showing the number of cell-state-specific TCR V genes. For each cell state, more frequently used V genes are counted. **(d)** A bar plot showing the enrichment of called transcribed enhancers across all cell types for Roadmap project PBMC enhancer annotation compared with (left) Brain enhancer and (right) PBMC heterochromatin.



Extended Data Figure 7. *FOXO1* is a candidate driver of the pathogenic CD4+ T cell state gene program. **(a)** Representative examples of cell-state-specific transcribed enhancers near the upregulated marker genes in NaiveCD4-5 (dataset 3; N=32). **(b)** Scatter plots comparing the logFC by *FOXO1*-KO in donors 1-2 and donor 3 (replication) for each gene. Genes with concordant signs between donors 1-2 and donor 3 are colored in red. *P*, *P* values in one-sided binomial sign test. **(c)** A Bar plot showing the enrichment of Kyoto Encyclopedia of Genes and Genomes (KEGG) FoxO signaling pathway for upregulated signature genes for each NaiveCD4 cell state. *P*, *P* values in one-sided Fisher's exact test. **(d)** A bar plot showing the enrichment of downregulated genes by *ARID5B*-KO for NaiveCD4-4 upregulated signature genes.



Extended Data Figure 8. Multimodal molecular fine mapping on *GZMK*⁺*GZMH*⁺*HLA-DRB1*⁺ EMCD8 cells. (a) Bar plots showing the percentage of **(left)** a representative clonotype that is shared across EMCD8 cell states and **(right)** one that is specific to EMCD8-5 (dataset 3; N=32). *P*, *P* values in two-sided Fisher's exact test. **(b)** Bar plots showing the percentage of **(left)** six TRBV and **(right)** three TRAV gene usages that are positively skewed in EMCD8-5 compared with other EMCD8 cells. *P*, *P* values in two-sided Fisher's exact test. **(c) (top)** EMCD8 and **(bottom)** EMCD4 UMAP plots, colored by oxidative phosphorylation signature scores. **(d)** A bar plot showing the percentage of EMCD8-5 cells on each in-silico gate using dataset 3. **(e)** Gating strategy to define HLA-DR⁺CD27⁺CD38⁺CD28⁺ EMCD8 cells in flowcytometry (FCM) and FACS. **(f)** Scatter plots comparing the scRNA-seq dataset 1 logFC (EMCD8-5 vs. other EMCD8 cells; N=256) and bulk RNA-seq logFC (HLA-DR⁺CD27⁺CD38⁺CD28⁺ vs. whole EMCD8 cells; N=6). **(g)** A volcano plot showing the differential peaks between HLA-DR⁺CD27⁺CD38⁺CD28⁺ vs. whole EMCD8 cells in bulk ATAC-seq (N=6). **(h) (left)** Boxplots comparing the mean fluorescence intensities (MFI) of TIGIT and the percentage of GZMK⁺GZMB⁺ (double-positive) cells between HLA-DR⁺CD27⁺CD38⁺CD28⁺ and whole EMCD8 cells. Samples from the same donor are connected with lines. *P*, *P* values in linear mixed models (N=11). **(right)** A FCM plot to show the percentage of double-positive cells in HLA-DR⁺CD27⁺CD38⁺CD28⁺ EMCD8 cells from a SLE donor.

Methods

Data availability

Single-cell count matrices of gene expression, surface protein, and TCR features in the CITE-seq dataset will be available at the National Bioscience Database Center (NBDC) Human Database (https://humandbs.dbcls.jp/en/) as of the date of publication. Public data used in this study are available at https://cellxgene.cziscience.com/collections. Reference panels for each lineage and each cell type are available on GitHub (https://github.com/MasahiroNakano-hub/LuPIN).

Code availability

Original codes used in this study are available on GitHub (https://github.com/MasahiroNakano-hub/LuPIN).

Subjects

In this study, we utilized two publicly available 3'-end scRNA-seq datasets (datasets 1 and 2). For dataset 1, we downloaded FASQ files from Chan Zuckerberg CELL by GENE (https://cellxgene.cziscience.com/collections) and raw genotype data from dbGaP (phs002812.v1.p1) with authorized access[5]. SLE samples were recruited from the California Lupus Epidemiology Study (CLUES) cohort, and healthy samples were obtained from the CLUES and the ImmVar Consortium. For

479 dataset 2, we downloaded the count data from dbGaP (phs002048.v1.p1)[4]. For our new 5'-end dataset
480 with CITE-seq, TCR-seq, and enhancer RNA profiling (dataset 3) and additional FCM and FACS
481 experiments (**Fig. 6**), patients were recruited at Okayama University Hospital and Hospital of the
482 University of the Occupational and Environmental Health based on the approval by the Ethics
483 Committees of RIKEN Center for Integrative Medical Sciences (RIKEN−Y-2025-006, 008). Written
484 informed consent was obtained from all participants.

485 Across all datasets, we defined four disease-activity categories in SLE[64]: i) inactive as SLE
486 disease activity index (SLEDAI) = 0, ii) low disease activity (LDA) as 1 ≤ SLEDAI ≤ 4, iii) moderate
487 disease activity (MDA) as 5 ≤ SLEDAI ≤ 9, and iv) high disease activity (HDA) as SLEDAI ≥ 10
488 (**Supplementary Table 1**).

489

490 **Sample processing and quality control**

491 To generate dataset 3, we processed single-cell suspensions of PBMCs from 32 SLE donors through the
492 10x Genomics Chromium. We prepared gene expression (GEX), antibody-derived tag (ADT), and TCR
493 (VDJ) libraries using the Chromium Next GEM Single Cell 5' Kit v2, Feature Barcode Kit, Human TCR
494 Amplification Kit (10x Genomics), and TotalSeq™-C Human Universal Cocktail v1.0 (Biolegend,
495 399905), following the manufacturer's protocol. This dataset comprises two experimental batches and
496 eight cell libraries. For each batch, we pooled cells from 16 samples into each of four independent

497 libraries. These libraries were sequenced on the Illumina NovaSeq 6000 to generate 150-base paired-
498 end reads.

499 In datasets 1 and 3, FASTQ files were processed using Cell Ranger v7.0.1 (10x Genomics) with
500 the developer's reference (refdata-gex-GRCh38-2020-A). We first filtered out the droplets with < 500
501 total unique molecular identifiers (UMIs) or those annotated as antibody aggregates. We next excluded
502 low-quality cells with a low number of genes (nGene), a high percentage of UMIs from mitochondrial
503 genes (% MT), or a low percentage of UMIs from ribosomal protein genes (% Ribo) based on their
504 distributions for each dataset (dataset 1: nGene < 300, %MT > 7.5%, %Ribo < 8.5%; dataset 2: nGene <
505 300, %MT > 8.5%, %Ribo < 8.5%; dataset 3: nGene < 500, %MT > 8.5%, %Ribo < 6%; **Extended Data**
506 **Fig. 1a**). We also excluded cells with a high percentage of UMIs from hemoglobin genes (> 0.9%, *HBB*,
507 *HBA1*, *HBA2*, *HBD*) or platelet genes (> 0.6%, *PF4*, *PPBP*, *CAVIN2*, *GNG11*).

508 In datasets 1 and 3, multiple samples were pooled in each library (dataset 1: 8-19 samples per
509 library, dataset 3: 16 samples per library). For dataset 1, we used Demuxalot, a reference-genotype-
510 based demultiplexing software[65], to demultiplex these samples. We first performed genotype imputation
511 for autosomal variants using SHAPEIT2[66] and Minimac3[67] after quality control (QC). Next, we ran the
512 Demultiplexer.learn_genotypes function for each library, using ~5 million variants with Minimac3 R2 >
513 0.95. We additionally demultiplexed three samples with missing genotype using Souporcell, a reference-
514 genotype-free demultiplexing software[68]. For dataset 3, we ran Souporcell using common variants (minor

515 allele frequency > 1%) from the 1000 Genomes Project (GRCh38) and identified 16 genotype clusters
516 per library. In this step, we identified and excluded putative doublets from different genotypes for each
517 dataset (**Extended Data Fig. 1b**). For dataset 1, we excluded two samples with < 200 cells recovered
518 and three samples that showed discordance between the sex estimated from pseudobulk expressions of
519 *XIST* and Y chromosome-related genes and the metadata sex information (**Extended Data Fig. 1c**). For
520 dataset 3, to link the Souporcell genotype clusters to distinct individuals, we performed multiplex PCR,
521 amplifying genomic regions that cover 20 common variants using genomic DNA from each sample, and
522 sequenced the libraries on the Illumina MiSeq. For each sample, we identified the corresponding
523 genotype cluster whose genotypes of these 20 variants matched those obtained from multiplex PCR.
524 To separate PBMCs into seven cell lineages (CD4 T, CD8/other T, NK, B, Mono, DC, and
525 HSPC), we performed Louvain clustering with multiple steps (**Extended Data Fig. 1d-f**). First, we
526 normalized and scaled gene expression of PBMCs for each batch using SCTransform[69] in R package
527 Seurat (v5.1.0)[70] with parameters vst.flavor = “v1”, method = "glmGamPoi", and min_cells = 30 (low-
528 expression gene filter). We then integrated data from each batch using 3,000 shared highly variable
529 genes (HVGs) in the SelectIntegrationFeatures function, followed by PCA and batch correction using R
530 package Harmony (v1.2.3)[71] with parameter group.by.vars = c(“library”, “sample”). Mitochondrial genes,
531 Y chromosome-related genes, hemoglobin genes, platelet genes, and cell-cycle genes were excluded
532 from HVGs. We used 30 harmony-corrected PCs for UMAP visualization and shared nearest neighbor

533 (SNN) graph construction. Next, we performed Louvain clustering based on the SNN graph and
534 separated PBMCs into four large cell lineages (T/NK, B, Myeloid, and HSPC, **Extended Data Fig. 1d**).
535 We then repeated the above procedures for each of T/NK, B, and Myeloid cells and identified 12, 7, and
536 3 clusters, respectively (**Extended Data Fig. 1e**). Since most of these clusters were mixed by multiple
537 cell lineages, we further performed sub-clustering for each cluster (**Extended Data Fig. 1f**). We
538 annotated each sub-cluster as relevant cell lineage based on their lineage marker gene expression. For
539 some clusters, we iterated this step to maximize the purity of each cell lineage. Detailed
540 hyperparameters for each step were summarized in **Supplementary Table 3**. During this step, we
541 identified several sub-clusters representing remaining doublets (even after the exclusion of putative
542 doublets in Demuxalot or Souporcell), which showed high expression of marker genes from multiple
543 lineages (**Extended Data Fig. 1f**) and higher doublet scores in Python package Scrublet[72] (**Extended**
544 **Data Fig. 1g**), and excluded them from the final dataset.
545 Through these multi-step QCs, we finally kept 1,456,041 PBMCs (159 SLE and 97 healthy
546 donors) for dataset 1, 344,235 PBMCs (41 SLE and 17 healthy) for dataset 2, and 268,783 PBMCs (32
547 SLE) for dataset 3 (**Extended Data Fig. 1h**).
548
549 **Three-step clustering to define fine-grained 123 cell states**

550 To identify fine-grained cell states within each cell type, we employed a three-step intensive clustering
551 approach in dataset 1: lineage (layer 1), cell type (layer 2), and cell state (layer 3). After separating the
552 PBMCs into seven cell lineages in the QC step, we again performed SCTransform, HVGs selection,
553 PCA, Harmony batch correction, and Louvain clustering for each lineage with optimal parameters
554 (**Extended Data Fig. 2a-f, left** and **Supplementary Table 3**). Since some of the initial coarse clusters
555 were mixed by multiple cell types, we further performed sub-clustering for each cluster. We then
556 annotated each sub-cluster as a relevant cell type based on their cell type marker gene expression
557 (**Extended Data Fig. 2a-f, middle**). Finally, we repeated the above procedures for each cell type using
558 optimal parameters (**Extended Data Fig. 2a-f, right** and **Supplementary Table 3**) and identified a total
559 of 123 cell states. To evaluate the cell-type annotation accuracy in this study and the original study, we
560 performed reference mapping of dataset 1 onto the independent PBMC dataset (Azimuth)[15]. We then
561 calculated the concordance rates of our annotation and the original annotation with the predicted cell-
562 type labels in the reference mapping for each cell type (**Extended Data Fig. 2g**).

563 To identify cell-state-specific marker genes, we ran the FindMarkers function in Seurat (v4.3.0)[15]
564 with parameter min.pct = 0.05, comparing one cell state vs. other cell states within the same cell type
565 (**Extended Data Fig. 3** and **Supplementary Table 4**). We used cell-state-specific upregulated marker
566 genes (FDR < 0.05 and logFC > 0.1) for pathway enrichment analyses (**Pathway enrichment analysis**
567 **and gene signature score**). Since the default logFC values in FindMarkers (Seurat) are small due to the

568 parameter pseudocount.use=1, we also provided logFC with pseudocount.use=0, more intuitive values
569 for differential expression (**Supplementary Table 4**).
570

571 **Co-varying neighborhood analysis**

572 To examine cell state abundance within each cell type (or cell type abundance within each lineage), we
573 applied co-varying neighborhood analysis (Python package CNA v0.1.6), an approach to identify single-
574 cell-resolution transcriptional "neighborhoods" associated with clinical phenotypes[39], to the 18 major cell
575 types with a total of > 5,000 cells (cell-type-level CNA) and six cell lineages (lineage-level CNA) in
576 dataset 1. Following our previous study scheme[6], we prepared two comparison axes: (i) disease-state
577 comparison between healthy controls and inactive SLE, and (ii) disease-activity comparison between
578 inactive and HDA SLE.

579 We used harmony-corrected PCs as inputs and inferred the neighborhood abundance matrix
580 (NAM) using all CLUES samples and cells, followed by PCA (NAM-PC). ImmVar healthy samples were
581 excluded from all disease-state comparisons. We then tested the association between individual-level
582 phenotype labels and NAM-PCs with a linear model, adjusting for age, sex, ancestry, and the number of
583 cells per cell type. Given the substantial correlation between disease activity and age (r = -0.60) in
584 dataset 1, we excluded age from covariates in disease-activity comparisons. For CNAs across 18 cell
585 types in disease-state and activity comparisons, we defined a significant association as a global

586 permutation $P < 0.05/36$ (Bonferroni correction). CNA calculates "neighborhood correlations" from the
587 above linear model statistics, which represent the correlation between each neighborhood and
588 phenotypes (positive values as expansion and negative values as depletion). We aggregated these
589 single-cell-level neighborhood correlation values into each cell state (**Fig. 3** and **Supplementary Table**
590 **6**) and defined the cell states with more than 50% of neighborhoods correlated with phenotypes at
591 empirical FDR < 0.05 as "remarkable expansion (depletion)" and those with more than 25% of
592 neighborhoods as "moderate expansion (depletion)".

593 We additionally applied CNA to dataset 1 with (1) inactive vs. LDA SLE, (2) inactive vs. MDA
594 SLE, (3) Asian vs. EUR SLE with the same pipelines (**Fig. 3e** and **Extended Data Fig. 4d**). For
595 replication analysis, we applied CNA to dataset 2 with disease-state and disease-activity comparisons
596 (**Fig. 3f**). We projected the dataset 2 cells into the latent transcriptional space in dataset 1 using
597 Symphony[45] and utilized symphony-corrected PCs as inputs of CNA. Since CNA requires ~30 samples
598 for stable performance[39], we compared 14 SLE patients with 0 <= SLEDAI <= 2 vs. 17 healthy for
599 disease-state comparison, and 15 patients with SLEDAI >= 6 vs. 14 patients with 0 <= SLEDAI <= 2 for
600 disease-activity comparison in dataset 2. We also applied MASC[26] to dataset 1 to test the association of
601 cell-state abundance with disease-state and disease-activity, adjusting for age, sex, and ancestry
602 (**Extended Data Fig. 4e**).
603

604 **ScQuaD model to dissect quantitative and qualitative changes**

605 To decompose cell-type-level pseudobulk gene expression into quantitative and qualitative changes, we
606 developed ScQuaD (Single-cell-level Quantitative and Qualitative change Dissection), a single-cell-
607 level generalized linear mixed effect model (GLMM, **Fig. 4a-b** and **Supplementary Table 8**). This model
608 fitted each single-cell-level expression to the following GLMM with a negative binomial distribution:

609 $$\log(E_i) = \beta_{abundance}X_{i,abundance} + \beta_{dysregulated}X_{d,dysregulated} + \beta_{age}X_{d,age} + \beta_{sex}X_{d,sex}$$

610 $$+\beta_{ancestry}X_{d,ancestry} + \beta_{nUMI}X_{i,nUMI} + \beta_{MT}X_{i,MT} + (\phi_d|d) + (\kappa_b|b) + \theta$$

611 where *E* is the expression of the gene in cell *i*, *θ* is an intercept, and all other *β*s represent fixed effects
612 as indicated (nUMI = number of UMI and MT = % MT) in cell *i*, donor *d*, or batch *b*. Donor and batch are
613 modeled as random effect intercepts. The first exploratory variable is the neighborhood correlation
614 values for cell *i* in cell-type-level CNA with a specific comparison (e.g., disease-state CNA), representing
615 cell-state abundance changes in that comparison (abundance signature term). The second one is binary
616 phenotype information (e.g., healthy = 0 and inactive SLE = 1), representing single-cell-level gene
617 dysregulation (dysregulated signature term). We applied this model to all genes that passed the low-
618 expression filter (>=3 counts in >= 10% of samples) across 18 major cell types (> 5,000 cells) for
619 disease-state and activity comparisons in dataset 1, using R package lme4[73]. To determine significance,
620 we used a likelihood ratio test (LRT) comparing the models with and without the abundance term or the
621 dysregulated term.

To assess the type I error calibration, we performed permutation tests across all cell types for each comparison (**Extended Data Fig. 5a**). For the abundance term, we randomly shuffled neighborhood correlation values across all single cells. For the dysregulated term, we shuffled phenotype labels across all target samples.

To quantify the contribution of quantitative and qualitative changes to cell-type-level pseudobulk gene expression, we first selected the top 1,000 P-value pseudobulk DEGs from R package edgeR[74] for each cell type. For each gene, we calculated the explained variances by abundance and dysregulated terms as Nagelkerke r2 from the above GLMM. We then calculated the mean across all genes for each term and their proportions (**Extended Data Fig. 5b**).

We also investigated the abundance and dysregulated term statistics of IRGs for disease-state and activity comparisons (**Fig. 4c-d** and **Extended Data Fig. 5c**). We selected 100 IRGs based on a previous study[4] (**Supplementary Table 10**) and performed hierarchical clustering of them with Ward's method, based on the Euclidean distances of the abundance and dysregulated term Z scores for each cell type.

Pathway enrichment analysis and gene signature score

To test the enrichment of pathway annotations for abundance and dysregulated signature genes, we performed over-representation analyses (ORA) using the top 1,000 LRT P-value genes for each

640 signature term in each cell type, with one-sided Fisher's exact test in R package clusterProfiler[75] (**Fig. 4e**
641 and **Extended Data Fig. 5d**). For pathway annotations, we used the MsigDB hallmark gene set
642 collection and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways. To capture the cell-type-
643 specific biology for both signature genes, we used the union of the top 1,000 LRT P-value genes across
644 all cell types as the background gene set.

645 For pathway analyses of cell-state-specific upregulated marker genes (FDR < 0.05 in
646 FindMarkers [Seurat]), we performed ORA with one-sided Fisher's exact test (**Fig. 5b**, **Extended Data**
647 **Fig. 7c**, and **Supplementary Table 6**). We set all genes that passed the low-expression filter (>=3
648 counts in >= 10% of samples) for the corresponding cell type as the background gene sets.

649 To calculate classical exhaustion signature scores in EMCD8 and EMCD4 cells, we prepared a
650 classical exhaustion signature gene set based on previous studies[3,5] (**Supplementary Table 17**). We
651 first log-normalized and scaled each gene expression in the gene set. We then calculated the signature
652 score for each cell as the sum of these scaled expressions across all genes (**Fig. 6b**).

653

654 **Stratified LD score regression analysis**

655 To evaluate the enrichment of the genome-wide distribution of all SLE risk variants irrespective of their
656 effect sizes (heritability) around abundance and dysregulated signature genes, we performed stratified
657 linkage disequilibrium score regression (S-LDSC)[44] with the sc-linker pipeline[76] (**Fig. 4f** and **Extended**

658 **Data Fig. 5f**). We first inferred min-max-scaled gene scores based on abundance and dysregulated term
659 Z scores across all tested genes for each cell type. We then calculated LD scores for EAS and EUR,
660 using the above gene scores as gene weights. In S-LDSC, we examined enrichment of SLE heritability
661 for common variants within 100-kb windows on either side of the transcription start sites (TSS) of all
662 genes with positive Z scores for both signatures across all cell types, adjusting for baseline model v1[44].
663 For this analysis, we used two large-scale SLE GWAS summary statistics from EAS and EUR
664 ancestries[42,43]. Since the regression coefficients of S-LDSC are influenced by GWAS heritability, we
665 normalized the coefficients by dividing them by the mean per-SNP heritability as reported in a previous
666 report[44]; we then reported the normalized coefficients. Finally, we performed fixed-effect meta-analysis of
667 EAS and EUR results using the inverse-variance weighting method with normalized coefficients and their
668 standard errors (**Extended Data Fig. 5e**). We reported *P* values to test whether the regression
669 coefficient is significantly positive.
670

671 **Localization of GWAS prioritized genes for each cell state**

672 To assign the most likely causal genes to SLE-GWAS significant variants ($P < 5 \times 10^{-8}$), we used the
673 L2G score from Open Targets Platform[41]. We defined 124 genes that showed L2G score > 0.5 for SLE-
674 GWAS variants as SLE-GWAS prioritized genes (**Supplementary Table 9**). For each cell state, we

identified overlapping genes between the SLE-GWAS prioritized genes and cell-state-specific upregulated marker genes with FDR < 0.05 and logFC > 0.2 (FindMarkers [Seurat]; **Fig. 5d** and **Fig. 6a**).

Reference mapping

To integrate three single-cell datasets in LuPIN, we performed reference mapping using R package Symphony[45], setting the larger dataset 1 as reference and datasets 2 and 3 as query (**Fig. 5a**). For dataset 1, we built lineage-level and cell-type-level reference objects based on the corresponding harmony objects used in fine-grained clustering. For datasets 2 and 3, we first separated PBMCs into seven lineages with the same pipeline as in dataset 1 (**Sample processing and quality control**). After normalizing and scaling gene expression with the same parameters in dataset 1, we next projected the cells from each lineage into the latent transcriptional space of the same lineage in dataset 1 (lineage-level reference mapping). For each cell in datasets 2 and 3, we annotated the best-predicted cell type labels using k-nearest neighbor classifiers (k = 5) based on the cell type labels of dataset 1. We then repeated the above procedure at the cell-type level (cell-type-level reference mapping) and annotated the best-predicted cell state labels for each cell in datasets 2 and 3.

Surface marker protein analysis

692 After obtaining fine-grained cell state information for multimodal dataset 3 through the reference
693 mapping, we investigated the cell-state-specific surface marker profiles. For dataset 3, we profiled 130
694 surface marker proteins and 7 isotype controls using TotalSeq™-C Human Universal Cocktail v1.0
695 (Biolegend, 399905). For each cell, we normalized each protein expression with centered-log ratio (CLR)
696 across 130 proteins.

697 To identify cell-state-specific surface markers, we ran the FindMarkers function (Seurat),
698 comparing one cell state vs. other cell states within the same cell type. We reported the proteins with
699 FDR <0.05 and logFC >0.2 as cell-state-specific upregulated surface markers (**Fig. 5e**, **6e**, **Extended**
700 **Data Fig. 6a**, and **Supplementary Table 12**).

701

702 **TCR analysis**

703 In dataset 3, we used Cell Ranger-called TCR UMI counts of the cells that passed the following criteria:
704 (1) cells that passed in our GEX-based QC steps (**Sample processing and quality control**), (2) cells
705 annotated as NaiveCD4, CMCD4, EMCD4, CTLCD4, Treg, NaiveCD8, CMCD8, or EMCD8 (**Reference**
706 **mapping**), and (3) cells that passed Cell Ranger TCR QC criteria.

707 To identify cell-state-specific V gene usage, we compared the V gene frequencies between one
708 cell state vs. other cell states within the same cell type. To detect both positive and negative
709 enrichments, we used two-sided Fisher's exact tests (**Fig. 5c**, **Extended Data Fig. 6c**, **8b**, and

710 **Supplementary Table 11**). Among all 45 TRAV and 49 TRBV genes, we tested V genes that passed the
711 low expression filter (>= 10 counts in total) for each cell type.

712 To assess the TCR clonal expansion in EMCD8-5 (DP EMCD8), we counted the number of
713 clonotypes in DP EMCD8 across SLE donors (**Fig. 6c**). Clonotypes were defined as the combination of
714 CDR3 amino acid sequence and V gene on the β-chain. To identify DP EMCD8-specific clonotypes, we
715 compared the clonotype frequencies between one cell state vs. other cell states within the same cell type
716 for each donor. To detect both positive and negative enrichments, we used two-sided Fisher's exact
717 tests (**Extended Data Fig. 8a** and **Supplementary Table 18**). To identify the known cognate epitopes
718 for each clonotype, we used VDJdb browser (**Supplementary Table 18**).

719

720 **Enhancer RNA analysis**

721 To detect transcribed enhancers in dataset 3, we followed the ReapTEC (read-level prefiltering and
722 transcribed enhancer call) pipeline[14]. We sequenced 10X libraries on the Illumina NovaSeq 6000 to
723 generate 150 base-paired-end reads for Read1 as well as Read2, and mapped them using STARsolo[77]
724 to obtain reads with unencoded G. Reads were deduplicated, and those with cell barcodes for each cell
725 type were extracted. TSS peaks were generated by merging TSSs located within 10 bp of each other. To
726 identify bidirectionally transcribed candidate enhancers for each cell type, TSS peak pairs were detected
727 with the developers' scripts, masking promoter regions. To test their enrichment for Roadmap

epigenomes core 15 reference annotations[49], we merged called enhancers across all cell types and compared their overlap with (1) PBMC enhancers (6_EnhG and 7_Enh in E062), (2) PBMC heterochromatin (9_Het in E062), and (3) Brain hippocampus middle enhancers (6_EnhG and 7_Enh in E062) using two-sided Fisher's exact tests (**Extended Data Fig. 6d**).

To identify cell-state-specific transcribed enhancers, we calculated cell-state-specific scores using entropy as previously described[14]. We first excluded enhancers with <5 mapped reads per each cell type and aggregated reads across donors for each cell state. We then calculated the entropy and cell-state-specific scores for each enhancer with the following equations:

$$\mathrm{Entropy}(x) = -\sum_{i=1}^{n} p_i * \ log_2(p_i);\ p_i = \frac{x_i}{\sum_{i=1}^{n} x_i}$$

$$\mathrm{Specificity\ score}(x) = 1 - \frac{\mathrm{entropy}(x)}{log_2(n)}$$

where x_i is the transcript level (count-per-million, CPM) for enhancer x in cell state i, and n represents the number of cell states in each cell type. We defined cell-state-specific enhancers as those with a cell-state-specific score >0.7 (**Extended Data Fig. 7a** and **Supplementary Table 14**).

Trajectory analysis

To infer the trajectory graph in the whole CD4$^+$ T cell lineage, we used Monocle3[51] with default parameters (**Fig. 5h**). We set a cell in NaiveCD4-3 (SOX4$^+$TOX$^+$) as the root cell for pseudotime

745 definition based on the recent report that these cells (recent thymic emigrants) are the most immature

746 naive T cell population[18].

747

748 **Arrayed CRISPR knock-out screening**

749 To identify transcriptional regulators of disease-relevant cell state candidates in CD4 T cells, we

750 performed arrayed CRISPR knock-out (KO) screening. We first selected upregulated TFs (FDR < 0.05

751 and logFC > 0.3 in FindMarkers [Seurat]) in NaiveCD4-5 (22 TFs) and NaiveCD4-4 (1 TF [ARID5B];

752 overlapped with NaiveCD4-5), which expanded both in disease-state and activity CNA (**Fig. 3b**). Each

753 TF was targeted with two different guide RNAs (gRNAs) selected from a previous report[78]

754 (**Supplementary Table 15**; for each donor, two control samples were prepared).

755 We first isolated human primary $CD4^+$ T cells from freshly thawed PBMCs of three healthy

756 donors using $CD4^+$ T Cell Isolation Kit (Miltenyi, 130-096-533). For expansion, we activated $CD4^+$ T cells

757 for 2 days with Dynabeads Human T-Activator CD3/CD28 (Thermo, DB11132) at a 1:1 bead-to-cell ratio

758 and cultured them in T cell medium in the presence of 30 IU/mL IL-2 (Peprotech, 200-02), 5 ng/mL IL-7

759 (Peprotech, 200-07), and 5 ng/mL IL-15 (Peprotech, 200-15) at 37 °C and 5% CO_2. We replaced the

760 medium every three days to maintain CD4+ T cells at ~1×10^6/mL, and we collected and froze them

761 seven days after activation.

762 Prior to electroporation, expanded CD4 T cells from three donors were reactivated with
763 Dynabeads Human T-Activator CD3/CD28 at a 1:1 bead-to-cell ratio in the presence of IL-2, IL-7, and IL-
764 15, with beads removed after 24 h. We assembled CRISPR ribonucleoprotein (RNP) complexes by
765 combining recombinant Cas9 protein (IDT, 1081058) with two synthetic sgRNAs (IDT) at a 1:1:2 molar
766 ratio, and then incubated them at room temperature for 15 min. We performed electroporation using the
767 P3 Primary Cell 96-well Nucleofector kit and 4D-Nucleofector (Lonza, program EH-140). Each reaction
768 contained 1×10^6 expanded CD4 T cells, 3 μL Cas9 RNP, and 0.75 μL 100 μM electroporation enhancer
769 (IDT). After electroporation, cells were divided into three well replicates and cultured in T cell medium
770 supplemented with 30 IU/mL IL-2. Control cells were electroporated with equal amounts of non-targeting
771 Cas9 RNP and electroporation enhancer. For two donors, we knocked out all 22 TFs (discovery data,
772 **Fig. 5f-g**). For the 3rd donor, we knocked out *FOXO1* as a replication analysis (**Extended Data Fig. 7b**).
773 For bulk RNA-seq, we isolated RNA 5 days after transfection using MagMAX-96 Total RNA
774 Isolation Kit (Thermo Fisher Scientific, AM1830). Libraries were prepared using the RamDA-seq method
775 as previously described[79] with 100 pg of total RNA as input and sequenced on the Illumina NovaSeq
776 6000 to generate 150 base paired-end reads.
777

778 **FCM and FACS experiments on EMCD8 cells**

779 For in-depth characterization for EMCD8-5 cells (DP EMCD8), we performed functional assays of this
780 population with FCM and FACS using FACSMelody (BD Biosciences) based on four key surface
781 markers (HLA-DR, CD27, CD38, and CD28) with highest logFCs (FindMarkers [Seurat]; **Fig. 6e-f** and
782 **Supplementary Table 12**). We prepared three FCM panels (**Supplementary Table 19**): Panel 1 for
783 FACS and surface marker FCM (PD1 and TIGIT), Panel 2 for intracellular FCM (GZMK and GZMB), and
784 Panel 3 for intracellular FCM (IFNG).

785 For the FACS experiment, we first isolated $CD8^+$ T cells from six SLE donors using the $CD8^+$ T
786 Cell Isolation Kit (Miltenyi, 130-096-495). After blocking non-specific binding with anti-human Fc-gamma
787 receptor antibodies (Thermo Fisher Scientific) and surface marker staining (Panel 1) with the
788 manufacturer's protocol, we FACS-sorted target EMCD8 cells gated on $CCR7^-$HLA-
789 $DR^+CD27^+CD38^+CD28^+$ from each donor (**Extended Data Fig. 8e**) for bulk RNA-seq, ATAC-seq, and
790 UNIChro-seq (unique molecular identifier counting of regional chromatin accessibility with sequencing;
791 5,000 cells per sample). We also FACS-sorted whole EMCD8 cells gated on $CCR7^-$ from each donor as
792 a control (5,000 cells per sample). Total RNA from each sorted sample was extracted using MagMAX-96
793 Total RNA Isolation Kits. RNA-seq libraries were prepared using SMART-seq mRNA (Takara Bio
794 634772). For ATAC-seq, we washed sorted cells twice with cold PBS and performed tagmentation using
795 1.25 μL TDE1, 12.5 μL 2× TD buffer (Illumina 20034198), 0.25 μL 1% digitonin (Promega G944A), and 6
796 μL ddH_2O. We purified the tagmented DNA (Zymo D4014), amplified it for 15−18 cycles with KAPA HiFi

797 ReadyMix (Nippon Genetics KK2602) and Nextera XT Index Kit v2 (Illumina), and size-selected
798 fragments (180–1000 bp) using AMPure XP beads. UNIChro-seq libraries were prepared with custom
799 primers as previously described[58] to quantify chromatin accessibility at the five loci near the *IFNG* gene,
800 using an accessible region near the *UBE2D2* gene as a reference control (**Supplementary Table 21**).
801 RNA-seq and ATAC-seq libraries were sequenced on the Illumina NovaSeq 6000 to generate 150 base
802 paired-end reads. UNIChro-seq libraries were sequenced on the illumina Miseq platform using a custom
803 R1 sequencing primer (5'- ACACTCTTTCCCTACACGACGCTCTTCAGGTCT).

804 For FCM experiments, we isolated $CD8^+$ T cells from 11 SLE donors using $CD8^+$ T Cell Isolation
805 Kit (Miltenyi, 130-096-495). For surface marker staining (Panels 1-3), we took the same procedures as
806 the above FACS experiment. Dead cells were stained and removed using Zombie Violet Fixable Viability
807 Kit (Biolegend, 423113). For intracellular staining of GZMK and GZMB (Panel 2), we used Pharmingen™
808 Transcription Factor Buffer Set (BD Biosciences, 562574), following the manufacturer's protocol. For
809 intracellular staining of IFNG (Panel 3), we first stimulated $CD8^+$ T cells in culture medium with 50 ng/ml
810 PMA (MP Biomedicals, 16561-29-8) and 1 μg/ml ionomycin (1 μg/ml). One hour after stimulation, we
811 added GolgiStop™ Protein Transport Inhibitor (BD Biosciences, 554724) and cultured them for an
812 additional three hours. We then performed intracellular staining using Cytofix/Cytoperm™
813 Fixation/Permeabilization Kit (BD Biosciences, 554714), following the manufacturer's protocol. In all
814 panels, we compared the mean fluorescence intensities (MFI) of target proteins between $CCR7^-$HLA-

815 DR$^+$CD27$^+$CD38$^+$CD28$^+$ cells and whole EMCD8 cells with a linear mixed model, adjusting for age and
816 sex as fixed effects and donor as a random effect (**Fig. 6i** and **Extended Data Fig. 8h**). All analyses
817 were performed with FlowJo™ Software (BD Biosciences).
818

819 **Bulk RNA-seq, ATAC-seq, and UNIChro-seq analyses**

820 For both RNA-seq and ATAC-seq samples, adaptor sequences in FASTQ files were trimmed using
821 Cutadapt and short reads with < 50 base pairs were removed.

822 For RNA-seq samples, reads were mapped to the GRCh38 genome with STAR[80] using the
823 same reference files with Cell Ranger (refdata-gex-GRCh38-2020-A) for single-cell data. Expression was
824 counted with HTSeq[81], setting the parameter "stranded=no" for the unstranded libraries.

825 Among a total of 153 CRISPR KO RNA-seq samples (discovery: [22 TFs + 2 controls] × 3
826 replicates × 2 donors, replication: [1 TFs + 2 controls] × 3 replicates × 1 donors; **Arrayed CRISPR**
827 **knock-out screening**), we first excluded three samples with unique read counts < 4 × 10^6 as low-quality
828 samples. We kept the genes that passed the low-expression filter (>=10 counts in >= 15% of samples),
829 followed by a trimmed mean of M values (TMM) normalization[74]. For differential expression analysis
830 between TF-KO conditions and controls, we fitted each count to the following GLMM with a negative
831 binomial distribution:

832 $$\log(E_s) = \beta_{KO} X_{s,KO} + (\phi_d | d) + \text{offset}(\log[libsize]) + \theta$$

where *E* is the expression of the gene in sample *s*, *θ* is an intercept, and *β* represent fixed effect for KO condition (one KO condition vs. control). Donor *d* is modeled as a random effect intercept. TMM-normalized library size is modeled as an offset term. To determine the significance of this model, we used LRT comparing the models with or without the KO condition term in lme4[73]. We tested the enrichment of the downregulated genes by each KO condition (LRT FDR < 0.05; **Supplementary Table 16**) for NaiveCD4-4 and NaiveCD4-5 upregulated signature genes (FDR < 0.05 in Seurat FindMarkers) using one-sided Fisher's exact test (**Fig. 5g** and **Extended Data Fig. 7d**).

For EMCD8 bulk RNA-seq samples, all 12 samples (6 donors × 2 cell subsets [target and controls]; **FCM and FACS on EMCD8 cells**) passed QCs (unique read counts > 2×10^7). We kept the genes that passed the low-expression filter (>=10 counts in >= 3 samples), followed by TMM normalization. For differential expression analysis between target (HLA-DR$^+$CD27$^+$CD38$^+$CD28$^+$ EMCD8) and control (whole EMCD8) cell populations, we fitted each count to the following GLMM with a negative binomial distribution in lme4[73] (**Extended Data Fig. 8f**):

$$\log(E_s) = \beta_{subset} X_{s,subset} + \beta_{age} X_{d,age} + \beta_{sex} X_{d,sex} + (\phi_d | d) + \mathrm{offset}(\log[libsize]) + \theta$$

where *E* is the expression of the gene in sample *s*, *θ* is an intercept, and *β* represent fixed effect for cell subsets (target vs. control). Donor *d* is modeled as a random effect intercept. TMM-normalized library size is modeled as an offset term.

For EMCD8 bulk ATAC-seq and UNIChro-seq samples, reads were mapped to the GRCh38 genome with Bowtie2[82]. We filtered out low-quality reads by samtools[83] with parameters "-f 2 -q 30", duplicated reads by Picard, with and blacklisted regions in ENCODE blacklist for hg38. We then called ATAC narrow peaks using MACS2[84]. We kept the peaks that passed the low-expression filter (>=5 counts in >= 3 samples), followed by TMM normalization. To detect differentially accessible regions (DARs) between target and control populations in ATAC-seq, we fitted each count to the same GLMM equation with RNA-seq (**Extended Data Fig. 8g**). For motif enrichment analysis with HOMER, we used the DARs with LRT FDR < 0.1 and logFC >0 (**Fig. 6g**). For UNIChro-seq samples, we extracted UMI counts as previously described[58] at the target five loci near the *IFNG* gene and a reference control region. To detect DARs between target and control cell populations, we used a logistic mixed effect model, regressing a binary phenotype vector (control populations: 0, target populations: 1) for each UMI by region information (reference region: 0, target region: 1), adjusting for donor random effect. We used Integrative Genomics Viewer for the visualization of focused genomic regions (**Fig. 6h**).

Acknowledgements

This study was supported by the Japan Agency for Medical Research and Development (AMED) (JP22tm0424223, JP22ek0410099, JP23tm0524005, JP25ek0410139, JP22ama121015 and JP223fa627010, JP22gm1810002, JP22gm1810005), JSPS Grants-in-Aid for Scientific Research (JP21K20647, JP22H03114, JP23KJ2185, JP23K15361, JP25K19621), the Uehara Memorial Foundation, the Nakajima Foundation, Japan Intractable Diseases Research Foundation, GSK Japan Research Grant, Daiichi Sankyo Foundation of Life Science, Mochida Memorial Foundation for Medical and Pharmaceutical Research, Takeda COCKPI-T Funding, Kanzawa Medical Research Foundation, SENSHIN Medical Research Foundation, Japan Research Foundation for Clinical Pharmacology, the Japan Rheumatism Foundation, the Ichiro Kanehara Foundation, Keio University Academic Development Funds, and Keio University Program for the Advancement of Next Generation Research Projects. The super-computing resource was provided by Human Genome Center, Institute of Medical Sciences, The University of Tokyo. We appreciate the generosity of the donors for a RIKEN solicited donation project, Support for SLE research and development. We appreciate the RIKEN-IMS Genome Platform for its help in the sequencing experiments. We thank AIDA consortium and RIKEN-IMS Laboratory for Gene Structure and Regulation for helpful feedback.

Author contributions

M.N. and K.I. designed the study. M.N., K.I., and M.K. wrote the manuscript with critical inputs from T.N., Y.T., and K.Y. M.N. conducted analyses with support from K.I., M.K., K.A., H.I., T.N., J.I., H.T., and H.H. M.K. M.N., and K.A. conducted experiments with support from T. A., T. K., S. N., R. B., Y.M., B.N., A.S., and K.Y. T.K., S.K., E.K., Y.F., Y.T., and Y.M. contributed to sample collection. X.Z., S.B. and C.T. managed EAS SLE-GWAS data. All authors contributed to the final manuscript.

Competing interests

RIKEN has filed a patent application related to UNIChro-seq (WO/2024/242037). K.I, M.K, and T.A are inventors on this patent. S.K. has received speaking fees from Eli Lilly, GlaxoSmithKline, Bristol-Myers, Abbvie, Eisai, Pfizer, Astra-Zeneca, and research grants from Daiichi-Sankyo, Abbvie, Boehringer Ingelheim, and Astellas.

Materials and Correspondence

Correspondence and requests for materials should be addressed to Masahiro Nakano and Kazuyoshi Ishigaki.

899 **Reference**

900 1. Tsokos, G. C., Lo, M. S., Reis, P. C. & Sullivan, K. E. New insights into the immunopathogenesis of systemic lupus
901 erythematosus. *Nat. Rev. Rheumatol.* **12**, 716–730 (2016).

902 2. Banchereau, R. *et al.* Personalized Immunomonitoring Uncovers Molecular Networks that Stratify Lupus Patients. *Cell* **165**,
903 551–565 (2016).

904 3. the Accelerating Medicines Partnership in SLE network *et al.* The immune cell landscape in kidneys of patients with lupus
905 nephritis. *Nat. Immunol.* **20**, 902–914 (2019).

906 4. Nehar-Belaid, D. *et al.* Mapping systemic lupus erythematosus heterogeneity at the single-cell level. *Nat. Immunol.* **21**,
907 1094–1106 (2020).

908 5. Perez, R. K. *et al.* Single-cell RNA-seq reveals cell type–specific molecular and genetic associations to lupus. *Science* **376**,
909 eabf1970 (2022).

910 6. Nakano, M. *et al.* Distinct transcriptome architectures underlying lupus establishment and exacerbation. *Cell* **185**, 3375-
911 3389.e21 (2022).

912 7. Rao, D. A. *et al.* Pathologically expanded peripheral T helper cell subset drives B cells in rheumatoid arthritis. *Nature* **542**,
913 110–114 (2017).

914 8. Jenks, S. A. *et al.* Distinct Effector B Cells Induced by Unregulated Toll-like Receptor 7 Contribute to Pathogenic
915 Responses in Systemic Lupus Erythematosus. *Immunity* **49**, 725-739.e6 (2018).

916 9. Navarra, S. V. *et al.* Efficacy and safety of belimumab in patients with active systemic lupus erythematosus: a randomised,
917 placebo-controlled, phase 3 trial. *The Lancet* **377**, 721–731 (2011).
918 10. Morand, E. F. *et al.* Trial of Anifrolumab in Active Systemic Lupus Erythematosus. *N. Engl. J. Med.* **382**, 211–221 (2020).
919 11. Missarova, A., Dann, E., Rosen, L., Satija, R. & Marioni, J. Leveraging neighborhood representations of single-cell data to
920 achieve sensitive DE testing with miloDE. *Genome Biol.* **25**, 189 (2024).
921 12. Stoeckius, M. *et al.* Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods* **14**, 865–868
922 (2017).
923 13. Moody, J. *et al.* A single-cell atlas of transcribed *cis* -regulatory elements in the human genome. Preprint at
924 https://doi.org/10.1101/2023.11.13.566791 (2023).
925 14. Oguchi, A. *et al.* An atlas of transcribed enhancers across helper T cell diversity for decoding human diseases. *Science*
926 **385**, eadd8394 (2024).
927 15. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587.e29 (2021).
928 16. Watanabe, M., Moon, K. D., Vacchio, M. S., Hathcock, K. S. & Hodes, R. J. Downmodulation of Tumor Suppressor p53 by
929 T Cell Receptor Signaling Is Critical for Antigen-Specific CD4+ T Cell Responses. *Immunity* **40**, 681–691 (2014).
930 17. Castellanos, M. C. *et al.* Expression of the leukocyte early activation antigen CD69 is regulated by the transcription factor
931 AP-1. *J. Immunol. Baltim. Md 1950* **159**, 5463–5473 (1997).
932 18. Bohacova, P. *et al.* Multidimensional profiling of human T cells reveals high CD38 expression, marking recent thymic
933 emigrants and age-related naive T cell remodeling. *Immunity* **57**, 2362-2379.e10 (2024).

934 19. Yasumizu, Y. *et al.* Single-cell transcriptome landscape of circulating CD4+ T cell populations in autoimmune diseases. *Cell*

935 *Genomics* **4**, 100473 (2024).

936 20. Chan, J. D. *et al.* FOXO1 enhances CAR T cell stemness, metabolic fitness and efficacy. *Nature* **629**, 201–210 (2024).

937 21. Doan, A. E. *et al.* FOXO1 is a master regulator of memory programming in CAR T cells. *Nature* **629**, 211–218 (2024).

938 22. Crow, M. K. Type I Interferon in the Pathogenesis of Lupus. *J. Immunol.* **192**, 5459–5468 (2014).

939 23. He, J. *et al.* Circulating Precursor CCR7loPD-1hi CXCR5+ CD4+ T Cells Indicate Tfh Cell Activity and Promote Antibody

940 Responses upon Antigen Reexposure. *Immunity* **39**, 770–781 (2013).

941 24. Zhang, F. *et al.* Deconstruction of rheumatoid arthritis synovium defines inflammatory subtypes. *Nature* **623**, 616–624

942 (2023).

943 25. Koga, T. *et al.* CaMK4-dependent activation of AKT/mTOR and CREM-α underlies autoimmunity-associated Th17

944 imbalance. *J. Clin. Invest.* **124**, 2234–2245 (2014).

945 26. Fonseka, C. Y. *et al.* Mixed-effects association of single cells identifies an expanded effector $CD4^+$ T cell subset in

946 rheumatoid arthritis. *Sci. Transl. Med.* **10**, eaaq0305 (2018).

947 27. Patil, V. S. *et al.* Precursors of human $CD4^+$ cytotoxic T lymphocytes identified by single-cell transcriptome analysis. *Sci.*

948 *Immunol.* **3**, eaan8664 (2018).

949 28. Luo, Y. *et al.* Single-cell transcriptomic analysis reveals disparate effector differentiation pathways in human Treg

950 compartment. *Nat. Commun.* **12**, 3913 (2021).

951 29. Li, J. *et al.* KIR$^+$ CD8$^+$ T cells suppress pathogenic T cells and are active in autoimmune diseases and COVID-19. *Science*

952 **376**, eabi9591 (2022).

953 30. Jonsson, A. H. *et al.* Granzyme K$^+$ CD8 T cells form a core population in inflamed human tissue. *Sci. Transl. Med.* **14**,

954 eabo0686 (2022).

955 31. Velikkakam, T., Gollob, K. J. & Dutra, W. O. Double-negative T cells: Setting the stage for disease control or progression.

956 *Immunology* **165**, 371–385 (2022).

957 32. Colonna, M. Innate Lymphoid Cells: Diversity, Plasticity, and Unique Functions in Immunity. *Immunity* **48**, 1104–1117

958 (2018).

959 33. Morgan, D. & Tergaonkar, V. Unraveling B cell trajectories at single cell resolution. *Trends Immunol.* **43**, 210–229 (2022).

960 34. Nutt, S. L., Hodgkin, P. D., Tarlinton, D. M. & Corcoran, L. M. The generation of antibody-secreting plasma cells. *Nat. Rev.*

961 *Immunol.* **15**, 160–171 (2015).

962 35. Mulder, K. *et al.* Cross-tissue single-cell landscape of human monocytes and macrophages in health and disease. *Immunity*

963 **54**, 1883-1900.e5 (2021).

964 36. Leffler, J., Bengtsson, A. A. & Blom, A. M. The complement system in systemic lupus erythematosus: an update. *Ann.*

965 *Rheum. Dis.* **73**, 1601–1606 (2014).

966 37. Villani, A.-C. *et al.* Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors.

967 *Science* **356**, eaah4573 (2017).

968 38. Zhang, Q. *et al.* Landscape and Dynamics of Single Immune Cells in Hepatocellular Carcinoma. *Cell* **179**, 829-845.e20
969 (2019).
970 39. Reshef, Y. A. *et al.* Co-varying neighborhood analysis identifies cell populations associated with phenotypes of interest from
971 single-cell transcriptomics. *Nat. Biotechnol.* **40**, 355–363 (2022).
972 40. Rector, I. *et al.* Differential regulation of the interferon response in systemic lupus erythematosus distinguishes patients of
973 Asian ancestry. *RMD Open* **9**, e003475 (2023).
974 41. Mountjoy, E. *et al.* An open approach to systematically prioritize causal variants and genes at all published human GWAS
975 trait-associated loci. *Nat. Genet.* **53**, 1527–1533 (2021).
976 42. Yin, X. *et al.* Meta-analysis of 208370 East Asians identifies 113 susceptibility loci for systemic lupus erythematosus. *Ann.*
977 *Rheum. Dis.* **80**, 632–640 (2021).
978 43. Langefeld, C. D. *et al.* Transancestral mapping and genetic load in systemic lupus erythematosus. *Nat. Commun.* **8**, 16021
979 (2017).
980 44. Finucane, H. K. *et al.* Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell
981 types. *Nat. Genet.* **50**, 621–629 (2018).
982 45. Kang, J. B. *et al.* Efficient and precise single-cell reference atlas mapping with Symphony. *Nat. Commun.* **12**, 5890 (2021).
983 46. Lagattuta, K. A. *et al.* Repertoire analyses reveal T cell antigen receptor sequence features that influence T cell fate. *Nat.*
984 *Immunol.* **23**, 446–457 (2022).

985 47. Nagafuchi, Y. *et al.* T cell plasticity in systemic lupus erythematosus revealed by large-scale T cell receptor repertoire and
986 transcriptome studies. Preprint at https://doi.org/10.1101/2025.01.06.24319648 (2025).
987 48. Britanova, O. V. *et al.* Targeted depletion of TRBV9+ T cells as immunotherapy in a patient with ankylosing spondylitis. *Nat.*
988 *Med.* **29**, 2731–2736 (2023).
989 49. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330
990 (2015).
991 50. Meyer, A., Parmar, P. J. & Shahrara, S. Significance of IL-7 and IL-7R in RA and autoimmunity. *Autoimmun. Rev.* **21**,
992 103120 (2022).
993 51. Cao, J. *et al.* The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566**, 496–502 (2019).
994 52. Bai, Y., Hu, M., Chen, Z., Wei, J. & Du, H. Single-Cell Transcriptome Analysis Reveals RGS1 as a New Marker and
995 Promoting Factor for T-Cell Exhaustion in Multiple Cancers. *Front. Immunol.* **12**, 767070 (2021).
996 53. Saggau, C. *et al.* Autoantigen-specific CD4+ T cells acquire an exhausted phenotype and persist in human antigen-specific
997 autoimmune diseases. *Immunity* **57**, 2416-2432.e8 (2024).
998 54. Chiche, L. *et al.* Modular Transcriptional Repertoire Analyses of Adults With Systemic Lupus Erythematosus Reveal Distinct
999 Type I and Type II Interferon Signatures. *Arthritis Rheumatol.* **66**, 1583–1595 (2014).
1000 55. Munroe, M. E. *et al.* Altered type II interferon precedes autoantibody accrual and elevated type I interferon activity prior to
1001 systemic lupus erythematosus classification. *Ann. Rheum. Dis.* **75**, 2014–2021 (2016).

1002 56. Fava, A. *et al.* Integrated urine proteomics and renal single-cell genomics identify an IFN-γ response gradient in lupus
1003 nephritis. *JCI Insight* **5**, e138345 (2020).
1004 57. Perl, A. & Morel, L. Expanding scope of TEMRA in autoimmunity. *eBioMedicine* **90**, 104520 (2023).
1005 58. Kono, M. *et al.* Accurate, sensitive, and efficient chromatin accessibility quantification at target loci using UNIChro-seq.
1006 *medRxiv* 2025.07.29.25332340 (2025) doi:10.1101/2025.07.29.25332340.
1007 59. Taylor, R. P. & Lindorfer, M. A. Drug Insight: the mechanism of action of rituximab in autoimmune disease—the immune
1008 complex decoy hypothesis. *Nat. Clin. Pract. Rheumatol.* **3**, 86–95 (2007).
1009 60. Müller, F. *et al.* CD19 CAR T-Cell Therapy in Autoimmune Disease — A Case Series with Follow-up. *N. Engl. J. Med.* **390**,
1010 687–700 (2024).
1011 61. Gurajala, S. *et al.* A population-scale atlas of blood and tissue in lupus nephritis. *bioRxiv* 2025.08.11.669754 (2025)
1012 doi:10.1101/2025.08.11.669754.
1013 62. Ishigaki, K. Beyond GWAS: from simple associations to functional insights. *Semin. Immunopathol.* **44**, 3–14 (2022).
1014 63. Lappalainen, T. & MacArthur, D. G. From variant to function in human disease genetics. *Science* **373**, 1464–1468 (2021).
1015 64. Bombardier, C. *et al.* Derivation of the sledai. A disease activity index for lupus patients. *Arthritis Rheum.* **35**, 630–640
1016 (1992).
1017 65. Rogozhnikov, A. *et al.* Demuxalot: scaled up genetic demultiplexing for single-cell sequencing. Preprint at
1018 https://doi.org/10.1101/2021.05.22.443646 (2021).

1019 66. Delaneau, O., Zagury, J.-F. & Marchini, J. Improved whole-chromosome phasing for disease and population genetic
1020 studies. *Nat. Methods* **10**, 5–6 (2013).
1021 67. Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat. Genet.* **48**, 1284–1287 (2016).
1022 68. Heaton, H. *et al.* Souporcell: robust clustering of single-cell RNA-seq data by genotype without reference genotypes. *Nat.*
1023 *Methods* **17**, 615–620 (2020).
1024 69. Hafemeister, C. & Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative
1025 binomial regression. *Genome Biol.* **20**, 296 (2019).
1026 70. Hao, Y. *et al.* Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nat. Biotechnol.* **42**, 293–304
1027 (2024).
1028 71. Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296
1029 (2019).
1030 72. Wolock, S. L., Lopez, R. & Klein, A. M. Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic
1031 Data. *Cell Syst.* **8**, 281-291.e9 (2019).
1032 73. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting Linear Mixed-Effects Models Using **lme4**. *J. Stat. Softw.* **67**, (2015).
1033 74. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR : a Bioconductor package for differential expression analysis of
1034 digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
1035 75. Xu, S. *et al.* Using clusterProfiler to characterize multiomics data. *Nat. Protoc.* **19**, 3292–3320 (2024).

1036 76. Jagadeesh, K. A. *et al.* Identifying disease-critical cell types and cellular processes by integrating single-cell RNA-
1037 sequencing and human genetics. *Nat. Genet.* **54**, 1479–1492 (2022).
1038 77. Kaminow, B., Yunusov, D. & Dobin, A. STARsolo: accurate, fast and versatile mapping/quantification of single-cell and
1039 single-nucleus RNA-seq data. Preprint at https://doi.org/10.1101/2021.05.05.442755 (2021).
1040 78. Freimer, J. W. *et al.* Systematic discovery and perturbation of regulatory genes in human T cells reveals the architecture of
1041 immune networks. *Nat. Genet.* **54**, 1133–1144 (2022).
1042 79. Hayashi, T. *et al.* Single-cell full-length total RNA sequencing uncovers dynamics of recursive splicing and enhancer RNAs.
1043 *Nat. Commun.* **9**, 619 (2018).
1044 80. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
1045 81. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data.
1046 *Bioinformatics* **31**, 166–169 (2015).
1047 82. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
1048 83. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
1049 84. Zhang, Y. *et al.* Model-based Analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
1050