Single-cell sequencing identifies a novel T-lymphocyte cluster expressing stemness genes in amyotrophic lateral sclerosis patients

Albert Ziye Ma[1#], Brian D. Kirk[2#], Deyu Yan[3], Qinghua Wang[4], Yan Han[5$], and Weihai Ying[6$]

[1]Shanghai American School, 1600 Lingbai Road, Shanghai, China 201201; [2]Multiscale Research Institute for Complex Systems, Fudan University, 653 Huatuo Road, Shanghai, China 201203; [3]Shanghai AI Laboratory, Shanghai, China; [4]Center for Biomolecular Innovation, Harcam Biomedicines, Shanghai, China; [5]Department of Neurology, Yueyang Hospital of Integrated Traditional Chinese and Western Medicine, Shanghai University of Traditional Chinese Medicine, Shanghai, China 200437; [6]School of Biomedical Engineering and Med-X Institute, Shanghai Jiao Tong University, Shanghai, China. [#]Co-first authors. [$]Senior authors.

[*]Corresponding author: Weihai Ying, School of Biomedical Engineering and Med-X Institute, Shanghai Jiao Tong University, Shanghai, China. Email: weihaiy@sjtu.edu.cn, Phone: (11)130-5221-8008.

The authors have declared that no conflict of interest exists.

June 24, 2022

Running Title: Single-cell sequencing of ALS PBMCs

Keywords: Amyotrophic lateral sclerosis, clonal expansion, novel T cell lymphocytes, single-cell sequencing

Abstract

Sporadic amyotrophic lateral sclerosis (sALS) is a fatal adult-onset neurodegenerative disease with unknown causes and few treatment options. This study employed single-cell RNA sequencing and immune profiling to analyze peripheral blood mononuclear cell samples from sALS patients and depicted the first cellular landscape at single-cell resolution. Compared to healthy controls, ALS patients exhibit markedly reduced naïve T cell and B cell populations, with significant clonal expansion of terminal effector T cells. However, no common recognition patterns were detected among CDR3 regions of expanded T cell receptor clonotypes, suggesting that each ALS patient may have been triggered by different antigens. We discovered a novel $CD8^+$ T cell type with heterogenous expression profiles of stemness genes and carrying signature genes specific for monocytes and terminal effector T cells, which may play a critical role in ALS pathogenesis. Using differentially expressed genes between the ALS and Control groups within the novel $CD8^+$ T cell type, we identified 80 genes that collectively distinguished ALS patients from healthy controls with an unprecedented accuracy of 0.929 (sensitivity: 0.928, specificity: 0.930). Our findings from this study have important implications for understanding the pathogenesis of sALS, and provide new strategies for accurate diagnosis and potential therapeutic development.

Summary

Single-cell RNA sequencing and immune profiling analysis of ALS patient PBMCs provided the first cellular landscape at single-cell resolution, unraveled possible causes and improved diagnosis.

Introduction

Amyotrophic lateral sclerosis (ALS) is an untreatable, adult-onset progressive neurodegenerative disease that causes motor neurons in the brain and spinal cord to die, leading to eventual respiratory muscle paralysis and death[1]. The average life span following diagnosis is only 2~5 years. More than 90% of all ALS cases are sporadic without known causes, while the remaining 10% are familial ALS cases with genetically dominant disorders. Genes linked to familial ALS occurrences include superoxide dismutase (SOD1), chromosome 9 open reading frame 72 (C9ORF72), TAR DNA binding protein (TDP-43), TATA-box binding protein associated factor 15 (TAF15), and FUS RNA binding protein (FUS)[2,3].

Recent studies have provided a link of neuroinflammation to ALS. The pathogenesis of ALS is believed to consist of an early neuroprotective stage by regulatory T lymphocytes that suppress responder T lymphocyte proliferation and a later neurotoxic stage from dysfunctional regulatory T lymphocytes[3-5]. Immune profiling studies of peripheral blood from ALS patients uncovered extensive immunophenotypic changes between ALS and healthy controls, including changes of overall lymphocytes, shifts in activated $CD4^+$ and $CD8^+$ T cells, increases of classical and non-classical monocytes in ALS patients[6-9]. Furthermore, decreased blood $CD4^+$ and $CD8^+$ T cells and B cells have been shown to predict cognitive impairment in ALS patients[10]. Clinical studies have found that genetic mutations observed in familial ALS can enhance neuroinflammation, thus underscoring the critical roles of dysfunctional immunity in ALS pathogenesis[1].

Reliable blood biomarkers of ALS could enable early and accurate diagnosis, improve patient management and survival, and probably provide novel therapeutic targets. Fourteen serum proteins with significant changes in ALS compared to controls were identified using

quantitative discovery proteomics[11]. Next-generation sequencing of coding and long non-coding RNA (lncRNA) in peripheral blood mononuclear cells (PBMCs) and spinal cord uncovered 293 differentially expressed (DE) lncRNA and 87 mRNA between controls and ALS patients[12]. An extensive microarray analysis on blood samples of ALS (n=396) and healthy controls (n=645) identified 850 genes that distinguished ALS patients from healthy controls with 87% accuracy (sensitivity of 86% and specificity of 87%)[13], which was the highest accuracy in the literature.

Single-cell sequencing technology is a revolutionary tool that, different from bulk sample-based RNA sequencing, can better account for the high heterogeneity that ALS samples likely display. A previous study using single-cell RNA-sequencing (scRNA-seq) of the brainstem of mutant SOD1 mice found perturbed cell types and pathways[14]. In the current study, by combining scRNA-seq together with single-cell T cell receptor (TCR) sequencing (scTCR-seq) and B cell receptor (BCR) sequencing (scBCR-seq), we unraveled for the first time the cellular landscape of PBMC samples from ALS patients at single-cell resolution, mapped out the detailed abundance changes of different PBMC cell types, discovered a novel $CD8^+$ T cell type expressing multiple stemness genes, and identified 80 genes selected from the novel $CD8^+$ T cell type that distinguished ALS patients from healthy controls at 0.929 accuracy. Our findings from this study have important implications for the pathogenesis of sporadic ALS and other immune-mediated diseases, as well as provide new strategies for accurate diagnosis and potential therapeutic development.

Results

Single-cell transcriptome and TCR/BCR analysis of ALS patients and healthy volunteers

In order to investigate the transcriptome and TCR/BCR landscapes of ALS patients, we collected PBMCs from clinically confirmed sporadic ALS patients (n=6) (**Supplementary Table 1**). The six patients in the ALS group had an age span of 38~70 years old, with an average of 61.3±11.7 years old. In order to provide adequate controls on age, sex and health conditions, we selected two sets of healthy volunteers as controls (n=6, and n=1, termed as Control 1 and Control 2, respectively). The volunteers in Control 1 had an average age of 26.5±3.3 years old, while the volunteer in Control 2 was in the range of 56~60 years old. The two sets of controls were collectively referred to as "Control". ScRNA-seq, scTCR-seq, and scBCR-seq were performed for the these clinical samples in parallel.

Comparison of single-cell transcriptome of ALS patients and healthy controls

We collected high-quality scRNA-seq data for a total of 16,091 PBMC cells. Clustering analysis using SingleR classified them into 22 clusters (Cluster 0~21) (**Fig.S1a~c**). The UMAPs for the PBMC cells from the two sets of healthy controls and the ALS patients have a similar distribution (**Fig.S1d**). These clusters were annotated as 17 different cell types (**Fig.1a~c**) based on their marker genes (**Supplementary Table 2, Fig.S1e**). The number of genes observed in each cell type was averaged at around 2,000 (**Fig.S1f**).

Clusters 0 and 13 were recognized as classical monocytes and non-classical monocytes, respectively, for the varying levels of expression for *S100A8, S100A9, LYZ, LST1, MS4A7,* and *SERPINA1* (**Fig.1d**). Clusters 1 and 17 high in *CD8A, CD8B* and *GZMH* were assigned as CD8+-terminal effector T cells. Clusters 2 and 16 were designated as $CD4^+$ central memory T cells. Clusters 3 and 6 (high in *CCR7, LEF1,* and *TCF7*) were $CD4^+$ and $CD8^+$ naïve T cells, respectively. Clusters 4, 5 and 20 (high in *GNLY, SPON2* and *GZMB*) were recognized as

natural killer cells. Cluster 7 was $CD8^+$ effector memory T cells. Clusters 8 and 18 were recognized as $CD4^+$ terminal effector T cells. Clusters 9 and 12 high in *MS4A1, Cd79A* and *HLA-DRA* expression were classified as memory and naïve B cells, respectively. Clusters 10 and 14 were Vd2-gd and non-Vd2-gd T cells, respectively. Cluster 11 was assigned as mucosal-associated invariant T (MAIT) cells. Cluster 15 displayed monocyte marker genes *S100A8, S100A9, LYZ*, T cell marker genes *CD3, CD8A,* and *CD8B*, as well as major histocompatibility complex (MHC) class II antigen *HLA-DRA* (**Fig.1d**). The level of CCL4L2 exhibited the highest fold change among all the differentially expressed (DE) genes (**Supplementary Table 2**). Therefore, Cluster 15 was labeled as $CD8^+$ T cell-$CCL4L2^{Hi}$ (or short for $^+$ T cell-$CCL4L2^{Hi}$) and further analyzed. Clusters 19 and 21 are myeloid dendritic cells and plasmacytoid dendritic cells, respectively.

The abundance of each major cell types was compared (**Fig.1e, 1f, Supplementary Table 3**). The distributions of major cell types in the two sets of control groups were very close to each other for most cell types (**Supplementary Table 3**). Relative to the two control groups, the distributions of major cell types in the ALS group with greater than $\log_2 Fc>0.7$ increase included T cell-$CCL4L2^{Hi}$, $CD8^+$-terminal effector T cells, non-Vd2-gd T cells, $CD4^+$-terminal effector T cells, and classical monocytes, while those with greater than $|\log_2 Fc|>0.7$ decrease in populations were $CD8^+$ naïve T cells, naïve B cells, $CD4^+$ naïve T cells, MAIT cells and memory B cells (**Fig.1e~g**). Of note, T cell-$CCL4L2^{Hi}$ exhibited the largest population increase in the ALS group comparing to the control group, increasing from 30 cells in the control group to 628 cells in the ALS group (**Supplementary Table 3**).

A new T lymphocyte cluster in ALS patients

The T cell-CCL4L2^{Hi} cluster shared gene enrichment with many cell types (**Fig.1d**). We therefore further analyzed it at higher resolution. A total of 9 subclusters (subclusters 0~8) were discovered (**Fig.2a**). All the subclusters except subcluster 8 exhibited high expression in monocyte marker genes (such as *FCGR3A* (*CD16*), *S100A8, S100A9, LYZ*) and T cell marker genes (*CD3D, CD3E, CD3G, CD8A, CD8B*) in the majority of cell populations for most subclusters (**Fig.2b, Supplementary Table 4**). Out of the 658 cells in the T cell-CCL4L2^{Hi} cluster, 91.6% (603 cells) had at least one *CD3* genes expressed, 76.3% of which (460 cells) had at least one *CD8* gene, 22.2% (134 cells) had *CD4* gene, and 19.9% (120 cells) had both *CD8* and *CD4* genes (**Fig.2c**). In addition, 93.0% (612 cells) of the T cell-CCL4L2^{Hi} cluster had at least one S100 (S100A8 or S100A9) expressed, 63.7% (390 cells) of which expressed *FCGR3A* (*CD16*), 32.5% (199 cells) expressed *CD14*, and 23.2% (142 cells) expressed both *CD16* and *CD14* (**Fig.2d**). Among all the cells in the T cell-CCL4L2^{Hi} cluster, 88.4% (582 cells) had at least one *CD3* and one *S100* gene (**Fig.2e**). The co-existence of both T cell and monocyte marker genes in the majority of the T cell-CCL4L2^{Hi} cluster marked the novelty of this cell type. Furthermore, the DE genes with the highest fold changes among different subclusters were *TRBV, IgkV* and *IgLV* genes (**Fig.2b, Fig.S2b**), suggesting that the reason of the T cell-CCL4L2^{Hi} cell cluster to be heterogenous and separated into 9 subclusters was at least partially due to their differences in TCR and antibody expression.

A previous study suggested the existence of T cell-monocyte complex doublets in blood during immune perturbations such as infection[15]. In those T cell-monocyte complex doublets, proteins CD3 and CD14 were co-present, while increased levels of LFA1 (ITGB2) on T cells and ICAM1 on monocytes were needed to allow T cell-monocyte complex interaction[15]. In order to evaluate whether the T cell-CCL4L2^{Hi} cluster that we observed in this study was T cell-

monocyte complex doublets, we additionally analyzed the expression of *CD14, ITGB2*, and *ICAM1* genes in this cell type (**Fig.2b**). Among the 582 cells that co-expressed *CD3* and *S100* genes, only 4.98% (29 cells) co-expressed *ITGB2* and *ICAM1* genes (**Fig.2f**). Moreover, the average number of genes for the T cell-CCL4L2^{Hi} cell cluster was about 1500, similar to that of other cell types, while doublets tend to have a much higher average number of genes (in the quality control step of data processing, we performed DoubletFinder to detect and remove doublets). These multiple lines of evidence collectively provided strong support that the T cell-CCL4L2^{Hi} cluster is not T cell-monocyte complex doublet as described previously, instead, it represents a novel cell type with unique features, making it both monocyte- and CD8^{+} T cell-alike.

Next, we performed pseudotime analysis of the T cell-CCL4L2^{Hi} cluster together with monocytes (classical and non-classical) and CD8^{+} T cells (**Fig.2g**). The pseudo-temporal trajectory mapped CD8^{+} naïve T cells (Branch state 1) at the beginning of the trajectory, and 75% of CD8^{+} effector memory T cells (632 cells, termed as B1 EM T cells) along the increasing pseudotime scale on Branch 1, then the T cell-CCL4L2^{Hi} cluster at a critical three-way junction (**Fig.2g**). Following this junction, ~25% of CD8^{+} effector memory T cells (208 cells, termed as B2 EM T cells) were mapped along the increasing pseudotime scale of Branch 2, with CD8^{+} terminal effector T cells located at the end of this branch (Branch state 2), the monocytes (Branch state 3) were mapped to the end of Branch 3 along the pseudotime axis (**Fig.2g**). The novel T cell-CCL4L2^{Hi} cluster appeared to be intermediates among B1 and B2 EM T cells and monocytes. The location of the novel T cell-CCL4L2^{Hi} cluster on the pseudotime trajectory agreed very well with its enriched marker genes similar to monocytes and T cells. Furthermore, the majority of the novel T cell-CCL4L2Hi cluster was contributed by the ALS group (**Fig.2h**).

We analyzed the gene expression profiles along the pseudotime trajectory between $CD8^+$ naïve T cells (Branch state 1) and $CD8^+$ terminal effector T cells (Branch state 2) (**Fig.S2c**) and between $CD8^+$ naïve T cells (Branch state 1) and monocyte cells (Branch state 3) (**Fig.S2d**). The hallmark genes for $CD8^+$ naïve T cells, for instance the self-renewal and stemness genes such as *TCF7, MYC, CCR7, LEF1*, became low or extremely low in expression in $CD8^+$ terminal effector T cells or monocytes, as expected. Comparing to $CD8^+$ terminal effector T cells or monocytes, the novel T cell-CCL4L2^{hi} cluster (at regions surrounding the three-way junction indicated by ④ in **Fig.S2c, S2d, Fig.2g, 2h**) had considerably high expression levels of these genes (**Fig.S2c~e**). T-cell activation factors CD27 and IL7R, critical players in V(D)J recombination during T-lymphocyte development, also exhibited a similar trend in expression (**Fig.S2c~e**).

The ALS sample used for scRNA-seq was isolated from equally pooled blood cells from six ALS patients. To investigate whether the novel T cell-CCL4L2^{Hi} cluster came from only one or a few but not all the ALS patients, we further performed bulk RNA-seq analysis on each sample of the Control 1 group and ALS group (pooled blood samples of six individuals in each group). The rationale was that, given the fact that the novel T cell-CCL4L2^{Hi} cluster from the ALS group (628 cells) accounts for more than 8% of the total annotated 7,715 cells from this group, if this cluster was contributed by only one or a few but not all the ALS patients, the abundance of this cluster could vary by a much larger degree, for instance, zero percent in some samples, but 20% or higher in other samples. If this were the case, the signature genes for the novel T cell-CCL4L2^{Hi} cluster such as those mentioned above may display large sample-to-sample variations among the six ALS patients' samples, and this phenomenon should become even more pronounced for genes that only exhibited differential expressions in the novel T cell-

CCL4L2^{Hi} cluster but not any other clusters, for instance, *GSR* and *SPCS2* (**Supplementary Table 2**). Box plots of expression levels for selected stemness genes (*KLF2, CCR7, MYC, TCF7*), T-cell activation factor *CD27* and genes uniquely overexpressed in the novel T cell-CCL4L2^{Hi} cluster (*GSR* and *SPCS2*) showed rather small variations among different samples of the same group, and exhibited statistically significant differences between the Control and ALS groups (**Fig.2i, Supplementary Table 5**). These data suggested that the novel T cell-CCL4L2^{Hi} cluster was most likely contributed by all the ALS patients.

Clonal expansion of the TCR repertoire in ALS patients

The scRNA-seq and scTCR-seq data on the same samples allowed the identification of TCR information for the same cells. Analysis of scTCR-seq data yielded a total of 7,989 clonotypes with paired TCR α- and β-chains (**Fig.3a, Fig.S3a~c, Supplementary Table 6**). Among them, 96.4% (7,699 cells) were mapped to T cells identified by scRNA-seq in this study (**Supplementary Table 7**). The detected TCR for non-T cells was only 3.6%, suggesting the high specificity of scTCR-seq and confirming the accuracy of scRNA-seq-based T-cell assignment. The majority of T cell subtypes had a high percentage (79~92%) of cells with paired TCR detected, the exceptions were non-Vd2 gd T cells at 14~16%, Vd2 gd T cells and T cell-CCL4L2^{Hi} at ~5% (**Supplementary Table 7**).

As expected, TCR clonotypes in the less-developed T cells including CD4^{+} and CD8^{+} naïve T cells, and CD4^{+} central memory T cells were rarely expanded (**Fig.3b, Supplementary Table 7**). On the other hand, the more-differentiated T cells such as non-Vd2 gd T cells, CD4^{+} and CD8^{+} terminal effector T cells were moderately expanded in the Control group, at an average ratio of total TCR/unique TCR of 1.36, 2.02, 2.71, respectively (**Fig.3b, Supplementary**

Table 7). The overall ratio of total TCR/unique TCR for all cell types were 1.18 for the Control group. However, for the ALS group, the average ratios of total TCR/unique TCR were almost doubled, at 3.09, 4.16, 6.04, respectively, for non-Vd2 gd T cells, $CD4^+$ and $CD8^+$ terminal effector T cells, with an average of 1.73 for all cell types (**Fig.3b, Supplementary Table 7**). Since the ALS group had an average age of 61.3±11.7 years, to exclude the age effects in TCR clonal expansion, we separately calculated the total TCR/unique TCR ratios for Control 1 (age: 26.5±3.3 years) and Control 2 (age: 56~60 years) (**Supplementary Table 7.1**). The average total TCR/unique TCR ratios were 1.20 for Control 1 and 1.06 for Control 2, suggesting that the elevated clonal expansion as seen for the ALS group was unlikely due to age effects. In total, the ALS group had 2,062 unique TCR clonotypes, while the Control group had 3,517 unique TCR clonotypes (**Fig.3c, left panel**). Among them, 11.2% of all unique TCR clonotypes in the ALS group were observed in more than one T cell type, while this value was 3.1% for the Control group (**Fig.3c, right panel**).

To investigate TCR specificity, we performed CDR3 pattern analysis using GLIPH2[16] (**Supplementary Table 8**). GLIPH2 is an algorithm to cluster TCRs that recognize the same epitope according to the shared similar patterns in the CDR3 region of TCR β-chains, particularly across multiple individuals[16]. In the published study, GLIPH2 successfully identified 354 specificity groups harboring at least 3 unique TCRs from 3 or more individuals infected with *Mycobacterium tuberculosis*[16]. By using GLIPH2 on the total of 7,699 paired TCR clonotypes detected on T cells in this study, we identified 710 GLIPH specificity groups with at least 2 members in each group (**Supplementary Table 8**), the top 10 most abundant specificity groups were shown in **Fig.3d**, which had copy numbers ranging from 48~137 copies. For the top 10 specificity groups (**Fig.3d**), only Group 1 and 10 had two composite sequences that

differed by one residue, Serine (S) *vs.* Threonine (T) for Group 1 and Alanine (A) *vs.* Serine (S) for Group 10, in the CDR3 sequences of TCR β-chains (TCRb), while the TRBJ usage and TCR α-chains (TCRa) sequence were the same. All other 8 GLIPH groups were composed of one single paired TCRb and TCRa sequences (**Fig.3d**). These highly homogenous sequences indicated that each of these sequences likely existed and expanded in one donor. Among these top 10 GLIPH groups, only Group 2 was from the Control 1 group, whilst all other groups were found in the ALS group, agreeing with the higher expansion as seen for the TCR repertoire of the ALS group (**Fig.3b, Supplementary Table 7**). The lack of shared CDR3 patterns among different ALS donors suggests that, although each ALS patient was immune-elevated, there lacked a common antigenic trigger for the immune-elevation. The TRBV and TRBJ gene usage did not differ much between the ALS and Control groups, with the only noticeable exception being TRBV6-5 which became the 6th most frequently used in ALS group (**Fig.3f**), comparing to the 9th place in Control group (**Fig.3e**).

The scBCR-seq data on the same samples were also mapped to the cells identified by scRNA-seq (**Supplementary Table 9**). Analysis of scBCR-seq data yielded a total of 1,182 cells with paired BCR chains (**Fig.S3d~g, Supplementary Table 9**). Among them, 97.8% (1,124 cells) were mapped to B cells identified by scRNA-seq in this study (**Supplementary Table 10**). The detected BCR for non-B cells was only 4.9%, suggesting high specificity of scBCR-seq and confirming the accuracy of scRNA-seq-based B-cell assignment. In both the Control and ALS groups, BCR clonotypes in naïve B cells were rarely expanded (**Fig.S3h, Supplementary Table 10**). However, a slight expansion was seen for memory B cells in the ALS group, with the top two clonotypes having 13 and 4 copies, respectively (**Fig.S3i**).

Furthermore, the ALS group had 358 unique BCR clonotypes, while the Control group had 746 unique BCR clonotypes (**Fig.S3h, Supplementary Table 10**).

Transcriptome signatures of the ALS patients

We analyzed the DE genes between the ALS and the Control groups for each cell type (**Supplementary Table 11**). Overall, between the two groups, there were a much larger number of genes downregulated in the ALS group (1,039 DE genes) than those upregulated in the ALS group (300 DE genes) (**Fig.4a**), with a total of 1,339 DE genes. The novel T cell-CCL4L2^{Hi} cluster had the most DE genes (127 upregulated and 415 downregulated genes) between the two groups (**Fig.4a**). The plasmacytoid dendritic cells did not have any DE gene (**Supplementary Table 11**). Gene ontology (GO) and pathway analysis revealed that the upregulated DE genes were mostly involved in GO terms and pathways directly related to immune response (**Fig.S4a, Supplementary Table 12**), while the downregulated DE genes were largely related to GO terms and pathways in immune response and basic cellular functions such as cytoplasmic ribosome and hemopoiesis (**Fig.S4b, Supplementary Table 12**).

Among all cell types, 43 genes were observed in at least 5 cell types (**Fig.4b**). The heatmap displaying the relative expression level of these 43 genes was shown in **Fig.4c**. Major histocompatibility complex (MHC) class II antigen genes *HLA-DQA2, HLA-DRB5* and class I antigen gene *HLA-B* were over-expressed in the majority of cell types in the ALS group (**Fig.4c, 4d, Fig.S4c**), suggesting an upregulated antigen-presenting activity in ALS donors. Genes that were also upregulated in at least five cell types included *DUSP1* (Dual Specificity Phosphatase 1) that is critical for dealing with environmental stress, *TSC22D3* (TSC22 domain family member 3) with anti-inflammatory and anti-apoptotic activities, ribosomal gene *RPS4Y1*

(Ribosomal Protein S4 Y-Linked 1) and lncRNA *SNHG5* (Small Nucleolar RNA Host Gene 5), the S100 family proteins *S100A9* and *S100A4* (**Fig.4c, 4d, Fig.S4c**). Interestingly, *TXNIP* (Thioredoxin Interacting Protein) that inhibits thioredoxin and leads to the accumulation of reactive oxygen species and cellular stress and MHC class I antigen gene *HLA-C* were only upregulated in the novel T cell-CCL4L2^{Hi} cluster of the ALS donors.

Out of the 43 genes observed in at least 5 cell types, the majority genes were downregulated in ALS donors. These included *PCBP1* (Poly(RC) Binding Protein 1), *RPS2* that codes for a ribosomal protein, *CD81* that promotes antigen presentation, *H1FX* (H1.10 Linker Histone) that is involved in chromatin compaction, and *LYN* (LYN Proto-oncogene) in regulation of immune responses and hematopoiesis (**Fig.4c, 4d, Fig.S4c**). The self-renewal and stemness gene *KLF2* was strongly downregulated in the majority (12 out of 16) of cell types, underscoring the reduced stemness of PBMCs in ALS patients and agreeing with the overall reduced number of naïve cells in ALS (**Fig.1g**). Similarly, *TCF7* was downregulated in five cell types (**Fig.4c, 4d**).

Transcriptome signatures of the novel T cell-CCL4L2^{Hi} cluster

Along the pseudo-temporal trajectory of CD8^{+} T cells and monocytes (**Fig.2g**), the novel cell cluster T cell-CCL4L2^{Hi} adjoined two subgroups of CD8^{+} effector memory T cells, B1 and B2 CD8^{+} EM, respectively, and monocytes. B1 CD8^{+} EM T cells were between naïve CD8^{+} T cells and T cell-CCL4L2^{Hi} cluster, while B2 CD8^{+} EM T cells were between T cell-CCL4L2^{Hi} cluster and CD8^{+} terminal effector T cells. We analyzed the DE genes for the following four pairs: cells in the T cell-CCL4L2^{Hi} cluster between the ALS and Control groups; all cells in the T cell-CCL4L2^{Hi} cluster with B1 or B2 CD8^{+} EM T cells, or with monocytes (**Supplementary**

Table 13). This analysis identified a total of 3,649 DE genes, with the comparison between T cell-CCL4L2^{Hi} cluster and monocytes yielded the largest number of DE genes (1,818 genes in total) (**Fig.5a**). GO and pathway analysis revealed that the significantly enriched GO terms and pathways by upregulated and downregulated DE genes were largely overlapping, including immune response, cell killing, cell activation and response to pathogens, and hemopoiesis (**Fig.S5a, S5b, Supplementary Table 14**).

Agreeing with their identity as CD8$^+$ effector memory T cells, the comparisons of B1 and B2 CD8$^+$ EM T cells with T cell-CCL4L2^{Hi} cluster were very similar (**Fig.5b, 5c**). However, noticeable differences were found between these two comparisons. For example, the stemness factors *CCR7*, *SELL, LEF1* and T cell activation factor *IL7R* were significantly highly expressed in B1 than in B2 CD8$^+$ EM T cells (**Fig.5d**). On the other hand, cytotoxicity genes such as *S100A8, S100A9* and activated T cell genes such as *HLA-DRA*, *CD8A* and *CCL5* were significantly expressed at a higher level in B2 than in B1 CD8$^+$ EM T cells (**Fig.5d**). These observations were consistent with the respective positions of B1 and B2 CD8$^+$ EM T cells on the pseudo-temporal trajectory between naïve CD8$^+$ T cells and CD8$^+$ terminal effector T cells (**Fig.2g**). Indeed, these results from the pseudo-temporal analysis were in agreement with the initial separation of B1 CD8$^+$ EM T cells as Cluster 1 and B2 CD8$^+$ EM T cells as Cluster 17, despite their shared identity as CD8$^+$ EM T cells (**Fig.1a**).

The stemness factor *TCF7* and T cell activation factor *IL7R*, *CD27* exhibited higher expression levels in the T cell-CCL4L2^{Hi} cluster than in monocytes (**Fig.5b, 5c**), in agreement with the more stem-like status of the T cell-CCL4L2^{Hi} cluster as we discussed earlier (**Fig.2, Fig.S2**). For the signature genes in monocytes (such as *S100A8, S100A9, LYZ, CST3, IFI30, TYROBP*), although their expression levels in the T cell-CCL4L2^{Hi} cluster were significantly

lower than those in monocytes, as expected (**Fig.5b~d**), they were constantly overexpressed in T cell-CCL4L2^{Hi} cluster from the ALS donors than from the Control group, suggesting the more-developed status of this cluster along the pseudo-temporal trajectory towards the monocytes in ALS. Moreover, these genes were constantly overexpressed in the T cell-CCL4L2^{Hi} cluster than in B1 or B2 CD8^{+} EM T cells, highlighting the unique features of this cluster comparing to other T cells. Meantime, the lineage signature genes such as *CD3D, CD3E, CD8A*, *CD8B* and *CCL5* were of higher expression levels in the T cell-CCL4L2^{Hi} cluster compared to monocytes (**Fig.5b~d**), confirming its identity as T cells. Furthermore, the cells in the T cell-CCL4L2^{Hi} cluster from the ALS donors had higher expression levels in these lineage signature genes than in the Control group, suggesting the more-developed status of this cluster in ALS than in Control group.

Interestingly, the T cell-CCL4L2^{Hi} cells also expressed a high level of selected immunoglobulin genes including *JCHAIN, IGKC, IGHM* relative to B1 and B2 CD8^{+} EM T cells and monocytes (**Fig.5b~d**). For this group of genes, the T cell-CCL4L2^{Hi} cluster from the Control group had an overall higher expression level than those from the ALS donors. In addition, compared the T cell-CCL4L2^{Hi} cluster from the Control group, antigen presenting genes including *HLA-DRA, CD2* and *IFI30* had much higher expression levels for those from the ALS donors (**Fig.5b~d, Fig.S5c**).

Collectively, the T cell-CCL4L2^{Hi} cluster exhibited a novel combinatorial feature in gene expression profile, this includes the signature genes in monocytes (such as *S100A8, S100A9, LYZ*), stemness genes and T cell activation genes including *TCF7, LEF1, SELL, CD27, IL7R* and *HLA-DRA*, lineage signature genes such as *CD3D, CD3E, CD8A* and *CD8B,* and immunoglobulin genes including *JCHAIN, IGKC,* and *IGHM*. The expression profiles of the T

cell-CCL4L2^{Hi} cluster confirmed it as a critical inter-junction connecting the pseudotime trajectory among CD8^{+} naïve T cells, CD8^{+} effector memory T cells, CD8^{+} terminal effector T cells and monocytes, and established it as a novel cell type that was highly enriched in ALS patients.

Prediction of ALS diagnosis using gene lists derived from this study

The DE gene lists identified from comparing the T cell-CCL4L2^{Hi} cells between the ALS and Control groups (1,753 genes, Genelist 1), or the T cell-CCL4L2^{Hi} with CD8^{+} T cells (531 genes, Genelist 2) or the T cell-CCL4L2^{Hi} with monocytes (1,505 genes, Genelist 3) (**Supplementary Table 15**) were used for initial feature selection towards more accurate prediction of ALS diagnosis.

Blood gene expression profiling brings a new way to ALS diagnosis. William *et al.*[13] evaluated the diagnostic performance of 850 genes (468 Principal Components (PCs)) selected from microarray data, and achieved an average accuracy of 0.87 using Support vector machine (SVM). The small sample size and high dimensions of the dataset mean that feature selection is critical to improving the generalization ability of the classification model[17]. Therefore, the three DE genelists from our analysis were used for initial feature selection in our study. Starting from the 1,753 DE genes in Genelist 1, after ranking and combination of ReliefF and RFE, the average accuracy across the 10,000 simulation trials was 0.929 (sensitivity: 0.928, specificity: 0.930) using 80 input genes for ALS diagnosis (**Fig.6a, Supplementary Table 16**). In addition, the 531 DE genes in Genelist 2 between all the T cell-CCL4L2^{Hi} cells and CD8^{+} T cells for feature selection resulted in an average accuracy of 0.916 (sensitivity: 0.919, specificity: 0.912) when 60 input genes were used as predictors (**Fig.6a, Fig. S6a, S6b, Supplementary Table 16**). With

1,505 DE genes in Genelist 3 between T cell-CCL4L2^{Hi} cells and monocytes, the average accuracy was 0.920 (sensitivity: 0.913, specificity: 0.926, 90 input genes) (**Fig.6a, Fig. S6c, S6d, Supplementary Table 16**).

Discussion

This study aims to shed new light on understanding the causes and improving the diagnosis of ALS. Towards this end, scRNA-seq was combined with single-cell immune profiling on PBMC cells from ALS patients and two sets of heathy donors. From a total of 16,091 cells with high-quality data, we identified 17 different cell types, of them 10 cell types belonging to T cells, 2 cell types as B cells, 2 cell types as monocytes, along with NK cells, myeloid dendritic cells and plasmacytoid dendritic cells. Compared to the Control group, the ALS group has significantly increased abundance in five cell types (T cell-CCL4L2^{Hi}, CD8^{+}-terminal effector T cells, non-Vd2-gd T cells, CD4^{+}-terminal effector T cells, and classical monocytes), and decreased abundance in five cell types (CD8^{+} naïve T cells, naïve B cells, CD4^{+} naïve T cells, MAIT cells and memory B cells). Among them, the most significant change in cell abundance was found for a novel T cell-CCL4L2^{Hi} cluster (increased by 22-fold in the ALS group relative to the Control group). The significantly reduced cell abundance in all three types of naïve cells (B cells, CD4^{+} T cells and CD8^{+} T cells) was not previously reported. The decrease in cell abundance of MAIT and monocytes as well as increases in terminal effector T cells were previous reported[6-9]. However, the scRNA-seq approach used in the current study provided a more-complete and -detailed picture of the changes in cell abundance between the ALS and Control groups. These changes in cell abundance collectively suggests an overall elevated immune cell activation in ALS donors.

Detailed analysis of scTCR-seq data of the same samples allowed an understanding of T cell activation at single-cell level. Compared to the Control group, the T cells from the ALS patients exhibited much higher expansion in $CD8^+$ terminal effector T cells, $CD4^+$ terminal effector T cells and non-Vd2 gd T cells (in decreasing order) (**Fig.3b, Supplementary Table 7**). However, out of the total of 7,699 paired TCR clonotypes detected on T cells in this study, we found no CDR3 patterns that were likely contributed by more than one ALS individuals. These results suggest that although ALS donors exhibited a high degree of T cell activation, there exists no common antigen responsible for T cell activation in these ALS patients. One intriguing scenario is that each ALS patient was triggered by one or more different antigens prior to the onset of ALS development, however, all ALS patients shared dysfunctional regulatory T lymphocytes, thus immune cell activation was initiated, leading to drastic reduction of naïve T and B cell populations and high clonal expansion of terminal effector immune cells, resulting in neuroinflammation and rapid progression of ALS pathogenesis. This scenario not only explains why the causes of sporadic ALS remain unknown to date, but also unifies the pathogenesis of sporadic ALS and familial ALS, because no matter whether it is an antigenic trigger in sporadic ALS or genetic mutations in familial ALS, the same chain reaction (dysfunctional regulatory T lymphocytes – T cell activation – clonal expansion - inflammation) eventually leads to ALS.

The T cell-$CCL4L2^{Hi}$ cluster that was increased in the ALS group by more than 20 folds exhibited a novel combinatorial feature in gene expression profile, including signature genes in monocytes (such as *S100A8, S100A9, LYZ*), stemness genes and T cell activation genes including *TCF7, LEF1, SELL, CD27, IL7R* and *HLA-DRA*, lineage signature genes such as *CD3D, CD3E, CD8A* and *CD8B* and immunoglobulin genes including *JCHAIN, IGKC,* and *IGHM*. Interestingly, the T cell-$CCL4L2^{Hi}$ cluster expressed multiple stemness genes with a small

degree of overlap (**Fig.S2f**), which may be fundamental to its high heterogenicity and functional adaptability. In a recently published study, an autoimmune stem-like $CD8^+$ T cell population was identified as a root cause of type 1 diabetes[18]. These cells behave as autoimmune progenitor cells and constantly generate new autoimmune mediators to attack pancreas β-cells[19]. In another study, stem-like $CD8^+$ T cells act as progenitors that mediate response of adoptive cell immunotherapy against human cancer[18]. Intriguingly, the T cell-$CCL4L2^{Hi}$ cluster as discovered in this current study shares certain stemness gene markers such as *CCR7, LEF1, SELL, TCF7* with the stem-like progenitor $CD8^+$ T cells in type 1 diabetes[19]. Additionally, both type 1 diabetes and ALS are linked with autoimmune disorders, although type 1 diabetes is linked with far greater certainty. These similarities and the analyses conducted in this study suggest that it is highly plausible that the T cell- $CCL4L2^{Hi}$ cluster in ALS patients may perform a similar autoimmune progenitor role, one where they constantly give rise to autoimmune terminal effector T cells to attack ALS target tissues such as the blood-CSF (central spinal fluid) barrier[20], brain or spinal cord causing motor neuron neurodegeneration.

Interestingly, the stemness genes highlighted above are in general less expressed in the ALS group than the corresponding cells in the Control group for many cell types. The expression of the stemness genes is extremely high in naïve T cells and lower or none in effector cell types. Furthermore, the abundance of cells in the naïve state for the $CD8^+$ T cell, $CD4^+$ T cell, and B cell species were all decreased in the ALS group, while on the other hand, the number of $CD8^+$ and $CD4^+$ terminal effector T cells were all increased in the ALS patients. One possible interpretation of these observations is that ALS patients develop a higher rate of differentiation in these adaptive immune cell types, probably directed by the novel T cell-$CCL4L2^{Hi}$ cluster, which is further supported by the fact that propagation of cytotoxic T cells could lead to the observed

enrichment of pathways including adaptive immune response, cytokine signaling in immune system, response to oxidative stress and response to external stimuli in GO terms and pathway enrichment analysis.

One goal of this study was to improve accuracy of ALS diagnosis. There have been a number of attempts in the literatures. William *et al.*[13] achieved the best accuracy of 0.87 using 850 genes. By using DE genes identified in comparing the ALS and Control cells within the novel T cell- $CCL4L2^{Hi}$ cluster, here we acquired the highest prediction score of ALS diagnosis to date, with an average accuracy of 0.929 using 80 input genes.

Methods

Blood collection and PBMC isolation

All human blood samples were collected at Yueyang Hospital of Integrated Traditional Chinese and Western Medicine, Shanghai University of Traditional Chinese Medicine (Shanghai, China), with the personal information de-identified. None of the participants had a history of prior conditions other than the clinical diagnosis of ALS for ALS donors. A total of 5 mL fresh whole blood sample was collected from each participant into a tube containing EDTA. PBMCs were isolated from whole blood within 2 hours of sample collection using Lympholyte - H (Cedarlane Labs, Cat. # C5015) following the manufacturer's instructions. For pooled samples (ALS group and Control 1 group), equal volume of fresh whole blood sample from each participant was mixed together prior to PBMC isolation.

Single-cell RNA sequencing

Droplet-based encapsulation was performed using the Chromium Controller (10x Genomics). Libraries were generated using Chromium Single Cell 5' Feature Barcode Library Kit, Chromium Next GEM Single Cell 5' Kit, and Chromium Single Cell Human TCR Amplification Kit or Chromium Single Cell Human BCR Amplification Kit (10x Genomics). The libraries were multiplexed and sequenced on Illumina Novaseq 6000.

Single-cell RNA-seq data analysis

We used the 10x Genomics Cellranger 7.0.0 pipeline[21] to process the raw sequencing data, which included demultiplexing, read alignment (with human genome GRCh38 as the reference), barcode counting, and UMI counting. The resulting feature-barcode matrix was then imported into Seurat (v4.0.3)[22] for downstream analysis, including quality control, normalization, dimensionality reduction, and clustering. Quality control involved first removing genes that were expressed in less than three cells across all conditions. Subsequently, cells that were identified as singlets (doublets identified by DoubletFinder (v2.0.3)[23], had mitochondrial gene percentages <10%, ribosomal gene percentages >5%, and gene counts between 200 and 4500 were kept for downstream analysis. The expression data for genes detected in each sample were then normalized with the SCTransform function. The normalized data for each experimental condition was then consolidated and integrated into a master Seurat object (IntegrateData). This master Seurat object contained 19,100 genes across 16,091 cells. The top 30 principal components (RunPCA) were used to perform dimensionality reduction on the integrated matrix via UMAP projection (RunUMAP). Various clustering resolutions were then interrogated with the clustree (v0.4.4) R package[24], and subsequently, clustering was performed

with FindClusters at a resolution of 0.6, yielding 22 clusters. Finally, visualization was done with the DimPlot feature.

Cell type annotations were determined with the R package SingleR (v1.4.1)[25], which compared the transcriptome of each individual cell to three reference databases: Database of Immune Cell Expression, Novershtern hemopoietic data, and Monaco immune data. A score heatmap and delta distribution plot were produced to validate SingleR's conclusions. SingleR annotations were then refined, returned to Seurat, and transferred to the cell clusters**.** The expression of selected characteristic genes was also plotted (VlnPlot) to further validate the final annotations.

Differential expression analysis

We utilized Seurat to generate two distinct sets of differentially expressed (DE) genes. The first set of DE genes was generated with the FindAllMarkers ($\log_2$FC threshold of 0.25) function by extracting each cell type from all samples (for example, Monocyte Classical) and comparing its expression profile against the expression profile of all other cell types combined. DE genes with a Bonferroni-adjusted p-value <0.05 and an absolute $\log_2$FC threshold greater than 0.25 were considered significant. The top 50 significantly upregulated genes for each cell type were visualized with a heatmap and a dot plot.

The second DE gene set was generated with FindMarkers by comparing the expression levels of each cell type in the ALS group to the expression levels of the same cell type in the Control Group. This was repeated for all 17 cell types. DE genes with absolute $\log_2$ fold change values greater than 0.25 and Bonferroni-adjusted p-values less than 0.05 were considered significant and kept for further investigation. Preserved DE genes were then passed on for GO

analysis. The expression patterns of 43 DE genes which were expressed in five or more cell types were also visualized with the pheatmap (v1.0.12) R package (RRID:SCR_016418). Selected DE genes underwent further visualization with VlnPlots.

Reclustering of the novel T cell-CCL4L2^{Hi} cluster

The cells in cluster labelled as T cell-CCL4L2^{Hi} by SingleR were extracted and underwent a second, separate round of clustering. Altogether, 658 cells were extracted from the master Seurat object. Normalization, scaling, dimensionality reduction, and clustering were performed again on the raw count data of this subset, producing 9 clusters (FindClusters; resolution = 0.3). Differential expression analysis (FindAllMarkers; $\log_2$FC threshold = 0.25, min.pct = 0.25) was then reperformed between these 9 clusters, identifying the top upregulated marker genes for each individual cluster. The top 10 upregulated marker genes of each cluster were visualized in a heatmap.

Pseudotemporal analysis

We used the R package Monocle (v2.18.0)[26] to perform pseudotime and branched chain analysis of the CD8$^+$ T cells, T cell-CCL4L2^{Hi}, and monocytes. The top 1000 most differentially expressed genes (identified by differentialGeneTest) between the selected cell types were used to order genes for DDRTree analysis (reduceDimension). The cells were then ordered along pseudotime with the function "OrderCells" and the cell fate trajectories were plotted. We utilized Monocle's branched expression analysis modelling (BEAM) in order to identify the most branch-dependent genes between the naïve CD8$^+$ T cell (Branch state 1) and terminal effector CD8$^+$ T cell (Branch state 2) branches and between the naïve CD8$^+$ T cell (Branch state 1) and

monocyte (Branch state 3) branches. The top 150 most branch-dependent genes for each comparison (as determined by *q*-values) were then smoothed and visualized with heatmaps. Pseudotime data was also extracted and appended to the master Seurat object by unique cell barcodes. Expression patterns for selected genes were then projected onto the cell trajectory UMAP.

Differential expression analysis of the novel T cell-CCL4L2^{Hi} cluster

DE genes were identified with FindMarkers ($\log_2$FC threshold of 0.25) by testing the following expression profiles: 1) T cell-CCL4L2^{Hi} in the ALS group against T cell-CCL4L2^{Hi} in the Control group (labeled as "T cell-CCL4L2^{Hi} ALS vs Ctrl"); 2) all T cell-CCL4L2^{Hi} against all CD8^{+} Effector Memory T cells present on branch 1 of the pseudotime plot (labeled as "T cell-CCL4L2^{Hi} vs B1 CD8^{+} EM"); 3) all T cell-CCL4L2^{Hi} against all CD8^{+} Effector Memory T cells present in branch 2 of the pseudotime plot (labeled as "T cell-CCL4L2^{Hi} vs B2 CD8^{+} EM"); 4) all the T cell-CCL4L2^{Hi} against all monocytes (labeled as "T cell-CCL4L2^{Hi} vs monocytes"); and 5) all the T cell-CCL4L2^{Hi} in both ALS and Control groups against all CD8^{+} T cells (labeled as "T cell-CCL4L2^{Hi} vs CD8^{+} T cells"). A blanket Bonferroni-adjusted *p*-value threshold of <0.05 was applied to these four DE gene lists. Genes which passed the $\log_2$FC and *p*-value thresholds were considered significant and passed on for subsequent GO enrichment analysis. Additionally, the top 20 upregulated genes and the top 15 downregulated genes of each of the DE gene lists 1~4 were extracted, merged into an upregulated and downregulated gene list, respectively, and filtered for overlaps. Their expression patterns across all these tests were then visualized with heatmaps.

TCR and BCR repertoire analysis

Raw scTCR-seq and scBCR-seq data for each sample was assembled with the 10x Genomics Cellranger V(D)J 7.0.0 pipeline, using human reference dataset GRCh38. TCR and BCR data was then aggregated respectively using Cellranger's aggr pipeline. Only TCRs with paired chains (TRA and TRB) and BCRs with paired chains (IGH and IGL/IGK) were considered productive and preserved. This preserved data were then passed to Seurat, thus removing TCR/BCR expressing cells that did not belong to the final 16,091 cells that survived the quality control in scRNA-seq data analysis. Immune expression was projected onto the Seurat UMAP with the DimPlot function. Subsequently, clonotype abundance data for TCRs among T cell subtypes and BCRs among B cell subtypes were exported for each experimental condition and visualized for both total and unique counts. The top 20 TCR clonotypes and the top 9 BCR clonotypes (defined by the number of cells expressing a specific clonotype) were also visualized separately.

CDR3β, TRBV, TRBJ, CDR3α, sample and frequency data were also passed to the GLIPH2 algorithm[16] to test for TCR antigen specificity. Finally, V-J junction frequency for all TCR expressing cells was analyzed with VDJtools (v1.2.1)[27]. The data for Control group and ALS group were analyzed separately.

Gene Ontology analysis

DE gene sets were passed to Metascape[28] for Gene Ontology enrichment analysis. The first set of DE genes consisted of the gene lists generated by comparing each cell type in the ALS group against the same cell type in the Control group. GO analyses for the upregulated and downregulated genes were run separately for each cell type, totaling 34 runs. The most

significantly enriched pathways generated from each gene list (based on absolute -$\log_{10}P$ values) were then consolidated and visualized with dot plots in the ggplot2 (v3.3.5) R package[29].

The second DE gene set passed to Metascape for GO enrichment analysis was composed of four gene lists: T cell-CCL4L2^{Hi} ALS vs Ctrl, T cell-CCL4L2^{Hi} vs B1 CD8^{+} EM, T cell-CCL4L2^{Hi} vs B2 CD8^{+} EM, and T cell-CCL4L2^{Hi} vs monocytes. GO analyses were run separately for each upregulated and downregulated gene list in each comparison, for a total of eight separate GO runs. The 10 enriched pathways with the highest absolute -$\log_{10}P$ values from each run were extracted, and subsequently, enriched pathways derived from the upregulated gene lists and enriched pathways generated from the downregulated gene lists were merged together respectively. Duplicate pathways were filtered out from both enrichment lists. The enriched pathways were then visualized with dot plots.

Feature selection and machine learning construction for accurate ALS diagnosis

We used three DE gene lists, T cell-CCL4L2^{Hi} ALS vs Ctrl (Genelist 1), T cell-CCL4L2^{Hi} vs CD8^{+} T cells (Genelist 2) and T cell-CCL4L2^{Hi} vs monocytes T cell (Genelist 3) to uncover sets of genes for ALS diagnosis. We utilized microarray data from blood samples of ALS patients, ALS mimic disease, and healthy controls (GSE112676 and GSE112680). Expression profiles were preprocessed using background correction, quantile normalization, and batch correction. For subsequent comparisons, the data processing steps were kept consistent with William *et al*[13]. A total of 792 samples (296 ALS patients vs. 296 CTL/MIM subjects) were used for model construction, of which 592 subjects were used for training (296 ALS patients vs. 296 CTL/MIM subjects) and 200 subjects were used for testing (100 ALS patients vs. 100 CTL/MIM subjects).

The importance of each gene for a given gene list was scored using the Relief (Python package: skrebate; function: ReliefF), a feature weighting algorithm that assigns different weights to features based on the relevance of each feature and category. The Top 300 ranked genes were selected for further feature combination. A fixed number of genes from the 300 mentioned above were selected using the REF algorithm (Recursive Feature Elimination, Python package: sklearn; function: RFE) and used as features for model training.

Support vector machine (SVM) is a supervised machine learning model, and the principle is to find the hyperplane that correctly classifies the training data set and maximizes the margins. Average accuracy, specificity and sensitivity were used to assess the performance of the model.

Statistics

The R package was used for statistical analysis.

Study approval

This study was approved by the Ethics Committee of the Yueyang Hospital of Integrated Traditional Chinese and Western Medicine, Shanghai University of Traditional Chinese Medicine (Approval #2020-060-02). Written informed consent was received prior to participation.

Author contributions

A.Z.M. conceived the study, A.Z.M. and W.Y. designed the study, Y.H. provided critical clinical samples, A.Z.M. and B.D.K. performed data analysis, D.Y. performed the prediction,

A.Z.M. generated all the figures except Fig.6 and Fig.S6, A.Z.M., Q.W. and W.Y. wrote the manuscript.

Acknowledgement

This work was partially supported by Major Special Program Grant of Shanghai Municipality (Award Number 2017SHZDZX01), National Key Research and Development Program of China (Award Number 2021YFF1200400) and Shanghai Municipal Science and Technology Major Project (Award Number 2018SHZDZX01).

Figure legends

Figure 1. Transcriptome landscape of human PBMCs from ALS patients. a~c). UMAP visualization of cell types. Each dot represents a single cell, which is colored according to cell type. The identity of each cell type is indicated. **a).** Pooled data; **b).** Control group; **c).** ALS group. **d).** Dot plot of marker genes for distinct cell types. **e).** Cell type abundance in the Control group. **f).** Cell type abundance in the ALS group. **g).** Bubble plot of cell type abundance changes ($|\log_2 FC| \geq 0.7$).

Figure 2. Characterization of T cell-CCL4L2^{Hi} cluster. a). UMAP of T cell-CCL4L2^{Hi} subtypes at higher resolution. **b).** Dot plot of marker genes for distinct cell subtypes. **c).** Venn diagram showing the number of cells in T cell-CCL4L2^{Hi} cluster expressing *CD3 (CD3D, CD3E, CD3G), CD8* (*CD8A, CD8B*) or *CD4* genes. **d).** Venn diagram showing the number of cells in T cell-CCL4L2^{Hi} cluster expressing at least one *S100* (*S100A8* or *S100A9), CD16* or *CD14* genes. **e).** Venn diagram showing the number of cells in T cell-CCL4L2^{Hi} cluster expressing at least one *CD3* and/or *S100* genes. **f).** Venn diagram showing the number of cells

expressing *ICAM1* and/or *ITGB2* within the cells co-expressing *CD3* and/or *S100* genes. **g).** Pseudotime analysis of T cell-CCL4L2^{Hi} together with all monocytes and all CD8^{+} T cells colored by cell types. **h)** Pseudotime trajectory of T cell-CCL4L2^{Hi} together with all monocytes and all CD8^{+} T cells colored by the Control group (in cyan color) or the ALS group (in red color). In g) and h), the three branches of the pseudotime trajectory were labeled as Branch 1, 2 or 3, and the cells located at the apex of each branch were defined as "Branch state 1, 2 or 3", respectively (shown as 1, 2, or 3 in blue color). **i)** Box plots of transcription levels of selected genes in bulk RNA sequencing. Each group (Control 1 and ALS) contains six individual RNA sequencing samples. The "x" marks the mean, and the line marks the median. The asterisks indicate the *p* values: *: $p<0.05$; **: $p<0.005$; ***: $p<0.001$.

Figure 3. Characterization of TCR and BCR landscapes. a). Projection of detected TCR landscape on UMAP visualization. **b).** Pie charts of TCR clonotypes detected. Top row: Total clonotypes; bottom row: unique clonotypes. For both rows, left for the Control group and right for the ALS group. The values on the pie chart indicate the number of detected clonotypes in each cell type under the given condition. **c).** Bar charts showing the total number of unique TCR clonotypes in the Control or ALS group (Left) and the number of TCR clonotypes that were not shared among different T cell types (right). **d).** TCR beta-chain V and J-gene usage in the Control group. **e).** TCR beta-chain V and J-gene usage in the Control group. **f).** TCR beta-chain V and J-gene usage in the ALS group. In e) and f), V and J genes are shown as arcs, scaled according to their occurrence. V-J pairings are represented by ribbons. Plot generated using VDJtools.

Fig.4. Transcriptional profiles of ALS patients. a). Bar chart showing the number of DE genes comparing the Control group and the ALS group. The upregulated genes are shown in red

color while the downregulated genes are shown in green color. **b).** Bar chart showing the number of DE genes between the Control group and the ALS group and their occurrences. **c).** Heatmap of the expression levels of selected genes that are differentially expressed in at least five cell types in the ALS group. **d).** Violin plots of expression levels of selected marker genes.

Fig.5. Transcriptome signatures of the novel T cell-CCL4L2^{Hi} cluster. a). Bar chart showing the number of DE genes comparing the novel T cell-CCL4L2^{Hi} cluster with other cell types. The upregulated genes are shown in red color while the downregulated genes are shown in green color. **b).** Heatmap of the expression levels of selected upregulated genes. The top 20 most upregulated genes in each list were selected. **c).** Heatmap of the expression levels of the selected downregulated genes. The top 15 most downregulated genes in each list were selected. **d).** Box plots of expression levels of selected marker genes.

Fig.6. Classifiers for ALS diagnosis using Genelist 1. a). Performance comparison of models using different gene sets. **b).** SVM cost and input genes parameters using Genelist 1 derived from the DE genes for the novel T cell-CCL4L2^{Hi} cluster comparing ALS group and Control group. **c**). SVM prediction accuracy using Genelist 1. **d)**. Venn diagram of the relationship among the three gene lists used for prediction.

References

1. Beers DR, Appel SH. Immune dysregulation in amyotrophic lateral sclerosis: mechanisms and emerging therapies. Lancet Neurol 2019;18(2):211-220. DOI: 10.1016/S1474-4422(18)30394-6.
2. Ince PG, Highley JR, Kirby J, et al. Molecular pathology and genetic advances in amyotrophic lateral sclerosis: an emerging molecular pathway and the significance of glial pathology. Acta Neuropathol 2011;122(6):657-71. DOI: 10.1007/s00401-011-0913-0.
3. Kapeli K, Pratt GA, Vu AQ, et al. Distinct and shared functions of ALS-associated proteins TDP-43, FUS and TAF15 revealed by multisystem analyses. Nat Commun 2016;7:12143. DOI: 10.1038/ncomms12143.

4. Beers DR, Zhao W, Wang J, et al. ALS patients' regulatory T lymphocytes are dysfunctional, and correlate with disease progression rate and severity. JCI Insight 2017;2(5):e89530. DOI: 10.1172/jci.insight.89530.
5. Sheean RK, McKay FC, Cretney E, et al. Association of Regulatory T-Cell Expansion With Progression of Amyotrophic Lateral Sclerosis: A Study of Humans and a Transgenic Mouse Model. JAMA Neurol 2018;75(6):681-689. DOI: 10.1001/jamaneurol.2018.0035.
6. Jin M, Gunther R, Akgun K, Hermann A, Ziemssen T. Peripheral proinflammatory Th1/Th17 immune cell shift is linked to disease severity in amyotrophic lateral sclerosis. Sci Rep 2020;10(1):5941. DOI: 10.1038/s41598-020-62756-8.
7. Xie Y, Luo X, He H, Tang M. Novel Insight Into the Role of Immune Dysregulation in Amyotrophic Lateral Sclerosis Based on Bioinformatic Analysis. Front Neurosci 2021;15:657465. DOI: 10.3389/fnins.2021.657465.
8. Rolfes L, Schulte-Mecklenbeck A, Schreiber S, et al. Amyotrophic lateral sclerosis patients show increased peripheral and intrathecal T-cell activation. Brain Commun 2021;3(3):fcab157. DOI: 10.1093/braincomms/fcab157.
9. Gustafson MP, Staff NP, Bornschlegl S, et al. Comprehensive immune profiling reveals substantial immune system alterations in a subset of patients with amyotrophic lateral sclerosis. PLoS One 2017;12(7):e0182002. DOI: 10.1371/journal.pone.0182002.
10. Yang Y, Pan D, Gong Z, et al. Decreased blood CD4+ T lymphocyte helps predict cognitive impairment in patients with amyotrophic lateral sclerosis. BMC Neurol 2021;21(1):157. DOI: 10.1186/s12883-021-02185-w.
11. Katzeff JS, Bright F, Lo K, et al. Altered serum protein levels in frontotemporal dementia and amyotrophic lateral sclerosis indicate calcium and immunity dysregulation. Sci Rep 2020;10(1):13741. DOI: 10.1038/s41598-020-70687-7.
12. Gagliardi S, Zucca S, Pandini C, et al. Long non-coding and coding RNAs characterization in Peripheral Blood Mononuclear Cells and Spinal Cord from Amyotrophic Lateral Sclerosis patients. Sci Rep 2018;8(1):2378. DOI: 10.1038/s41598-018-20679-5.
13. Swindell WR, Kruse CPS, List EO, Berryman DE, Kopchick JJ. ALS blood expression profiling identifies new biomarkers, patient subgroups, and evidence for neutrophilia and hypoxia. J Transl Med 2019;17(1):170. DOI: 10.1186/s12967-019-1909-0.
14. Liu W, Venugopal S, Majid S, et al. Single-cell RNA-seq analysis of the brainstem of mutant SOD1 mice reveals perturbed cell types and pathways of amyotrophic lateral sclerosis. Neurobiol Dis 2020;141:104877. DOI: 10.1016/j.nbd.2020.104877.
15. Burel JG, Pomaznoy M, Lindestam Arlehamn CS, et al. Circulating T cell-monocyte complexes are markers of immune perturbations. Elife 2019;8. DOI: 10.7554/eLife.46045.
16. Huang H, Wang C, Rubelt F, Scriba TJ, Davis MM. Analyzing the Mycobacterium tuberculosis immune response by T-cell receptor clustering with GLIPH2 and genome-wide antigen screening. Nat Biotechnol 2020;38(10):1194-1202. DOI: 10.1038/s41587-020-0505-4.
17. Li Z, Xie W, Liu T. Efficient feature selection and classification for microarray data. PLoS One 2018;13(8):e0202167. DOI: 10.1371/journal.pone.0202167.

18. Krishna S, Lowery FJ, Copeland AR, et al. Stem-like CD8 T cells mediate response of adoptive cell immunotherapy against human cancer. Science 2020;370(6522):1328-1334. DOI: 10.1126/science.abb9847.
19. Gearty SV, Dundar F, Zumbo P, et al. An autoimmune stem-like CD8 T cell population drives type 1 diabetes. Nature 2022;602(7895):156-161. DOI: 10.1038/s41586-021-04248-x.
20. Winkler EA, Sengillo JD, Sullivan JS, Henkel JS, Appel SH, Zlokovic BV. Blood-spinal cord barrier breakdown and pericyte reductions in amyotrophic lateral sclerosis. Acta Neuropathol 2013;125(1):111-20. DOI: 10.1007/s00401-012-1039-8.
21. Zheng GX, Terry JM, Belgrader P, et al. Massively parallel digital transcriptional profiling of single cells. Nat Commun 2017;8:14049. DOI: 10.1038/ncomms14049.
22. Hao Y, Hao S, Andersen-Nissen E, et al. Integrated analysis of multimodal single-cell data. Cell 2021;184(13):3573-3587 e29. DOI: 10.1016/j.cell.2021.04.048.
23. McGinnis CS, Murrow LM, Gartner ZJ. DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors. Cell Syst 2019;8(4):329-337 e4. DOI: 10.1016/j.cels.2019.03.003.
24. Zappia L, Oshlack A. Clustering trees: a visualization for evaluating clusterings at multiple resolutions. Gigascience 2018;7(7). DOI: 10.1093/gigascience/giy083.
25. Aran D, Looney AP, Liu L, et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. Nat Immunol 2019;20(2):163-172. DOI: 10.1038/s41590-018-0276-y.
26. Qiu X, Hill A, Packer J, Lin D, Ma YA, Trapnell C. Single-cell mRNA quantification and differential analysis with Census. Nat Methods 2017;14(3):309-315. DOI: 10.1038/nmeth.4150.
27. Shugay M, Bagaev DV, Turchaninova MA, et al. VDJtools: Unifying Post-analysis of T Cell Receptor Repertoires. PLoS Comput Biol 2015;11(11):e1004503. DOI: 10.1371/journal.pcbi.1004503.
28. Zhou Y, Zhou B, Pache L, et al. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. Nat Commun 2019;10(1):1523. DOI: 10.1038/s41467-019-09234-6.
29. Wickham H. ggplot2: Elegant Graphics for Data Analysis: Springer-Verlag New York, 2016.

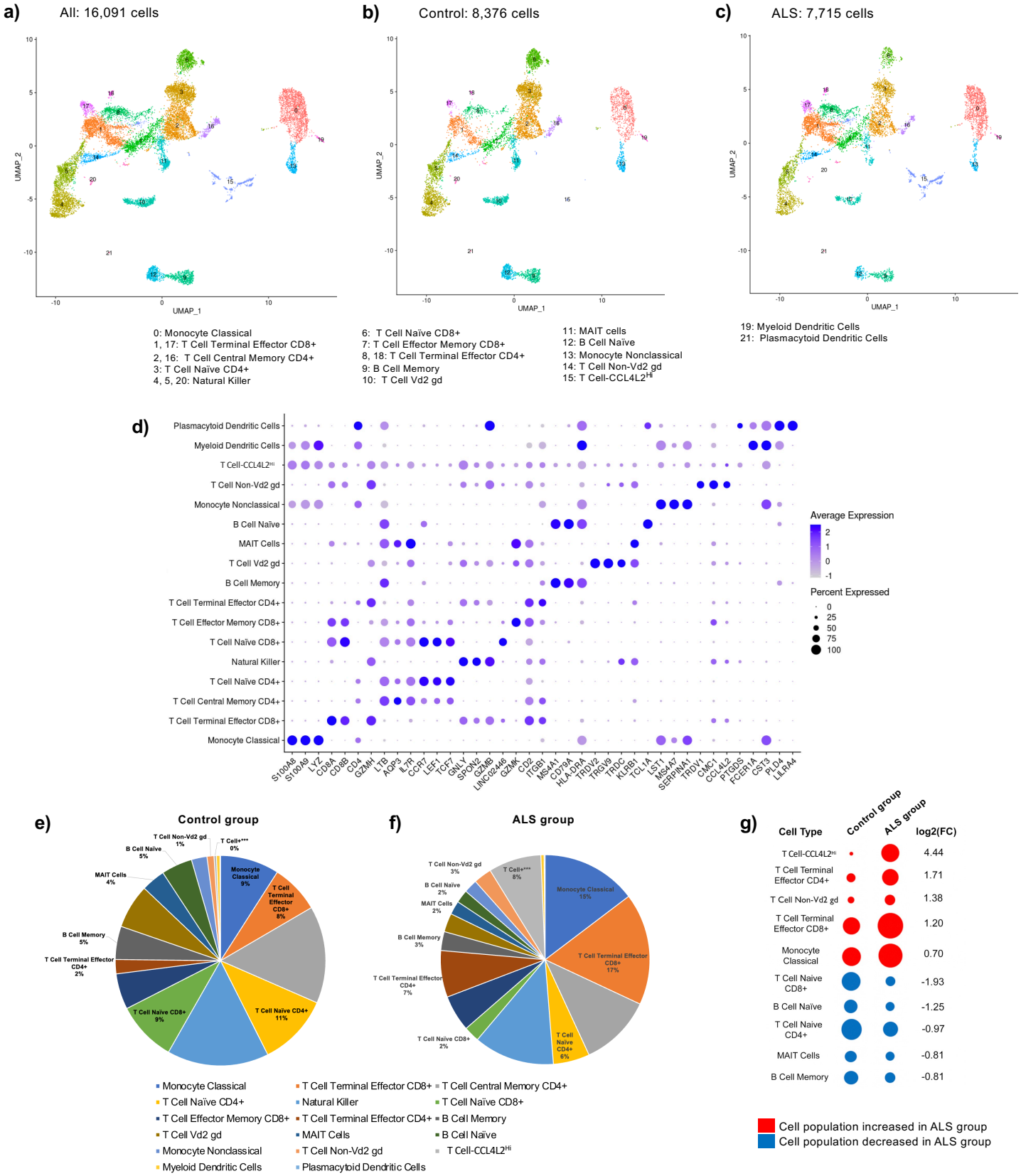

a)
All: 16,091 cells
b)
Control: 8,376 cells
c)
ALS: 7,715 cells
UMAP_1
UMAP_2
0: Monocyte Classical
1, 17: T Cell Terminal Effector CD8+
2, 16: T Cell Central Memory CD4+
3: T Cell Naïve CD4+
4, 5, 20: Natural Killer
6: T Cell Naïve CD8+
7: T Cell Effector Memory CD8+
8, 18: T Cell Terminal Effector CD4+
9: B Cell Memory
10: T Cell Vd2 gd
11: MAIT cells
12: B Cell Naïve
13: Monocyte Nonclassical
14: T Cell Non-Vd2 gd
15: T Cell-CCL4L2Hi
19: Myeloid Dendritic Cells
21: Plasmacytoid Dendritic Cells
d)
Plasmacytoid Dendritic Cells
Myeloid Dendritic Cells
T Cell-CCL4L2Hi
T Cell Non-Vd2 gd
Monocyte Nonclassical
B Cell Naïve
MAIT Cells
T Cell Vd2 gd
B Cell Memory
T Cell Terminal Effector CD4+
T Cell Effector Memory CD8+
T Cell Naïve CD8+
Natural Killer
T Cell Naïve CD4+
T Cell Central Memory CD4+
T Cell Terminal Effector CD8+
Monocyte Classical
S100A8 S100A9 LYZ CD8A CD8B CD4 GZMH LTB AQP3 IL7R CCR7 LEF1 TCF7 GNLY SPON2 GZMB LINC02446 GZMK CD2 ITGB1 MS4A1 CD79A HLA-DRA TRDV2 TRGV9 TRDC KLRB1 TCL1A LST1 MS4A7 SERPINA1 TRDV1 CMC1 CCL4L2 PTGDS FCER1A CST3 PLD4 LILRA4
Average Expression
2
1
0
-1
Percent Expressed
0
25
50
75
100
e)
Control group
Monocyte Classical 9%
T Cell Terminal Effector CD8+ 8%
T Cell Naïve CD4+ 11%
T Cell Naïve CD8+ 9%
T Cell Terminal Effector CD4+ 2%
B Cell Memory 5%
MAIT Cells 4%
B Cell Naïve 5%
T Cell Non-Vd2 gd 1%
T Cell+*** 0%
f)
ALS group
Monocyte Classical 15%
T Cell Terminal Effector CD8+ 17%
T Cell Naïve CD4+ 6%
T Cell Naïve CD8+ 2%
T Cell Terminal Effector CD4+ 7%
B Cell Memory 3%
MAIT Cells 2%
B Cell Naïve 2%
T Cell Non-Vd2 gd 3%
T Cell+*** 8%
Monocyte Classical
T Cell Naïve CD4+
T Cell Effector Memory CD8+
T Cell Vd2 gd
Monocyte Nonclassical
Myeloid Dendritic Cells
T Cell Central Memory CD4+
Natural Killer
T Cell Naïve CD8+
T Cell Terminal Effector CD4+
MAIT Cells
T Cell Non-Vd2 gd
Plasmacytoid Dendritic Cells
B Cell Memory
B Cell Naïve
T Cell-CCL4L2Hi
g)
Cell Type
Control group
ALS group
log2(FC)
T Cell-CCL4L2Hi 4.44
T Cell Terminal Effector CD4+ 1.71
T Cell Non-Vd2 gd 1.38
T Cell Terminal Effector CD8+ 1.20
Monocyte Classical 0.70
T Cell Naive CD8+ -1.93
B Cell Naïve -1.25
T Cell Naive CD4+ -0.97
MAIT Cells -0.81
B Cell Memory -0.81
Cell population increased in ALS group
Cell population decreased in ALS group

Figure 1

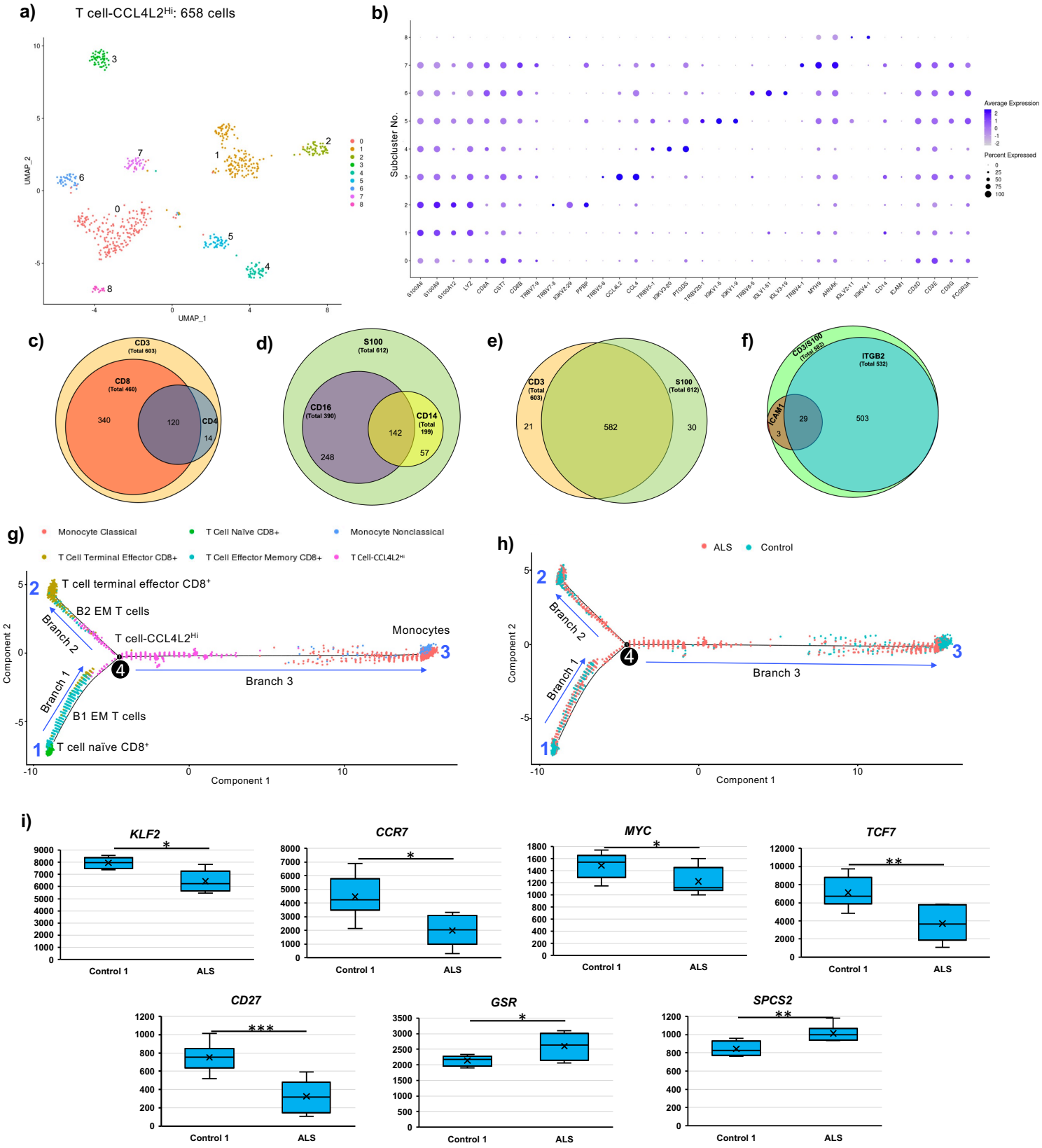
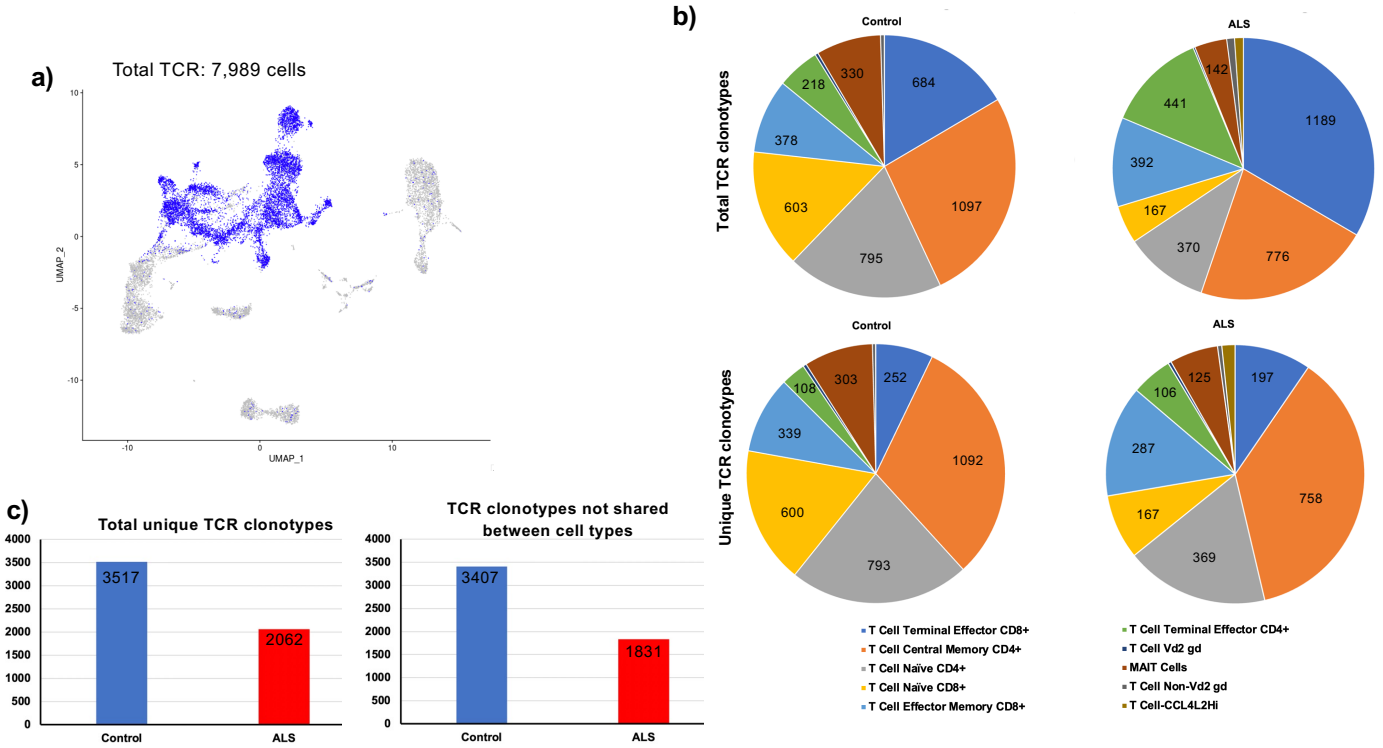

a)
T cell-CCL4L2Hi: 658 cells
UMAP_2
UMAP_1
b)
Subcluster No.
Average Expression
Percent Expressed
c)
CD3 (Total 603)
CD8 (Total 460)
CD4
340
120
14
d)
S100 (Total 612)
CD16 (Total 390)
CD14 (Total 199)
142
248
57
e)
CD3 (Total 603)
S100 (Total 612)
21
582
30
f)
CD3/S100 (Total 992)
ITGB2 (Total 532)
ICAM1
29
503
3
g)
Monocyte Classical
T Cell Naïve CD8+
Monocyte Nonclassical
T Cell Terminal Effector CD8+
T Cell Effector Memory CD8+
T Cell-CCL4L2Hi
T cell terminal effector CD8+
B2 EM T cells
T cell-CCL4L2Hi
Monocytes
Branch 2
Branch 1
Branch 3
B1 EM T cells
T cell naïve CD8+
Component 1
Component 2
h)
ALS
Control
Branch 2
Branch 1
Branch 3
Component 1
Component 2
i)
KLF2
CCR7
MYC
TCF7
CD27
GSR
SPCS2
Control 1
ALS

Figure 2



d)

	GLIPH ID	Donor group	TCRb sequence	Frequency	V	J	TCRa sequence
			CASSPPSGSYEQYF	37/137	TRBV7-9	TRBJ2-7	CAVDNFNKFYF
Group 1	229	ALS group	CASSPPTGSYEQYF	100/137	TRBV7-9	TRBJ2-7	CAVDNFNKFYF
Group 2	4503	Control group	CASSVGGLDYEQYF	112/112	TRBV9	TRBJ2-7	CAARGEKQAGTALIF
Group 3	3667	ALS group	CASSKLGNIQYF	89/89	TRBV27	TRBJ2-4	CAETYTNAGKSTF
Group 4	4391	ALS group	CASSLGPPMTEAFF	72/72	TRBV19	TRBJ1-1	CAANGRDQAGTALIF
Group 5	4790	ALS group	CASTQDRNYEQYF	64/64	TRBV11-2	TRBJ2-7	CAAAPNFGNEKLTF
Group 6	949	ALS group	CASSPVTGTGDYGYTF	62/62	TRBV6-5	TRBJ1-2	CARNTGNQFYF
Group 7	1521	ALS group	CSAQSPGGKAEAFF	59/59	TRBV20-1	TRBJ1-1	CAAKAGSGYSTLTF
Group 8	2781	ALS group	CASSLAGGDLIFQHF	59/59	TRBV7-2	TRBJ1-5	CAGTLSFSSNTGKLIF
Group 9	928	ALS group	CASSLDRGTKNTEAFF	57/57	TRBV7-9	TRBJ1-1	CAMREVNTGANSKLTF
			CASSGANTEAFF	4/48	TRBV9	TRBJ1-1	CAVKNAGNMLTF
Group 10	353	ALS group	CASSGSNTEAFF	44/48	TRBV9	TRBJ1-1	CAVKNAGNMLTF

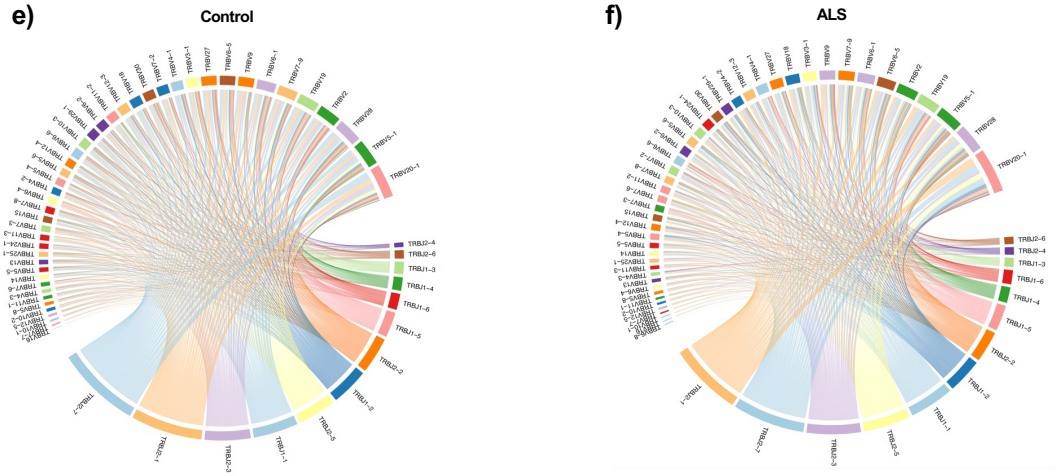


Figure 3

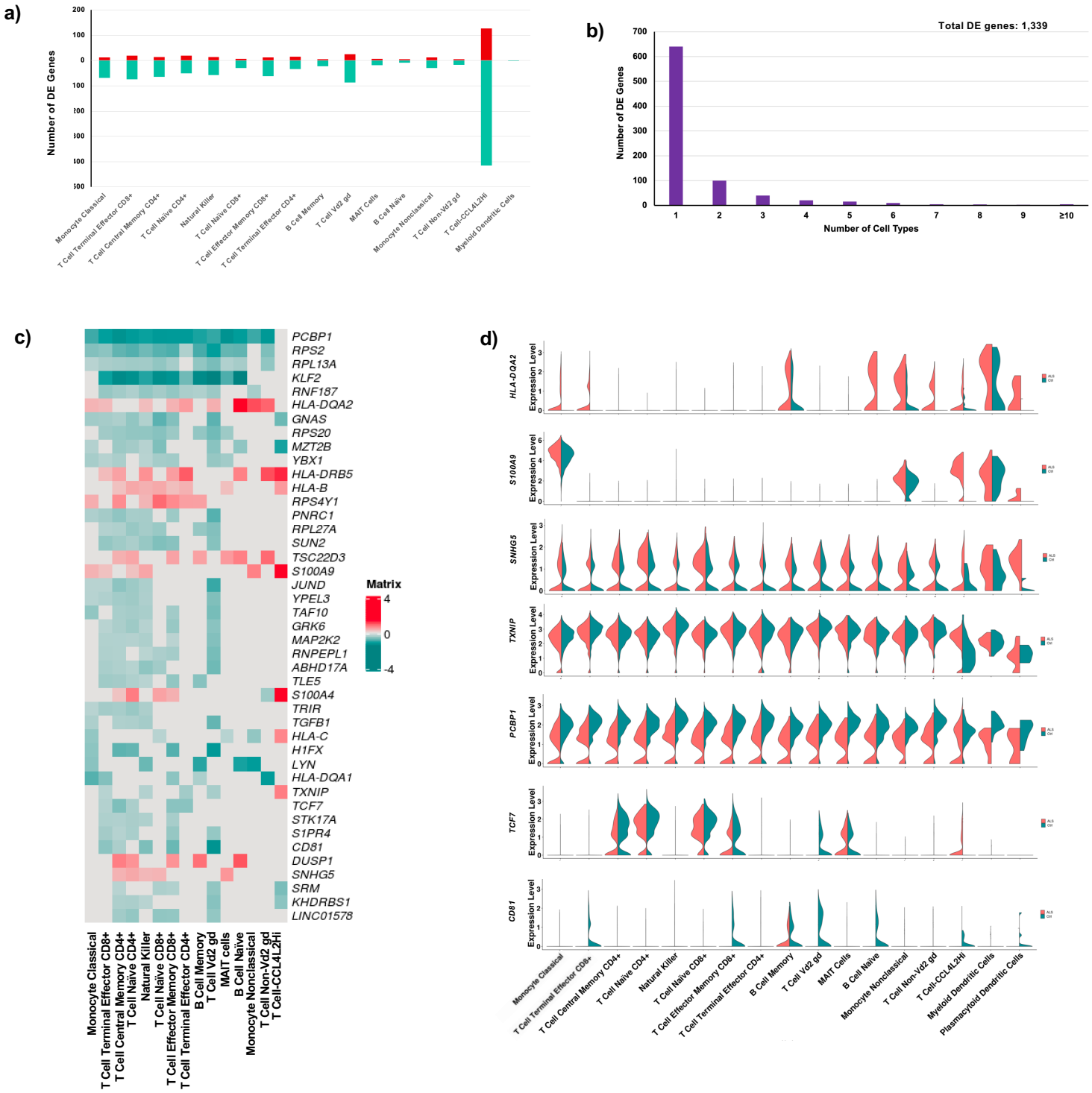

a)
Number of DE Genes
Monocyte Classical
T Cell Terminal Effector CD8+
T Cell Central Memory CD4+
T Cell Naïve CD4+
Natural Killer
T Cell Naïve CD8+
T Cell Effector Memory CD8+
T Cell Terminal Effector CD4+
B Cell Memory
T Cell Vd2 gd
MAIT Cells
B Cell Naïve
Monocyte Nonclassical
T Cell Non-Vd2 gd
T Cell-CCL4L2Hi
Myeloid Dendritic Cells
b)
Total DE genes: 1,339
Number of DE Genes
Number of Cell Types
c)
PCBP1
RPS2
RPL13A
KLF2
RNF187
HLA-DQA2
GNAS
RPS20
MZT2B
YBX1
HLA-DRB5
HLA-B
RPS4Y1
PNRC1
RPL27A
SUN2
TSC22D3
S100A9
JUND
YPEL3
TAF10
GRK6
MAP2K2
RNPEPL1
ABHD17A
TLE5
S100A4
TRIR
TGFB1
HLA-C
H1FX
LYN
HLA-DQA1
TXNIP
TCF7
STK17A
S1PR4
CD81
DUSP1
SNHG5
SRM
KHDRBS1
LINC01578
Matrix
d)
HLA-DQA2
S100A9
SNHG5
TXNIP
PCBP1
TCF7
CD81
Expression Level
Plasmacytoid Dendritic Cells

Figure 4

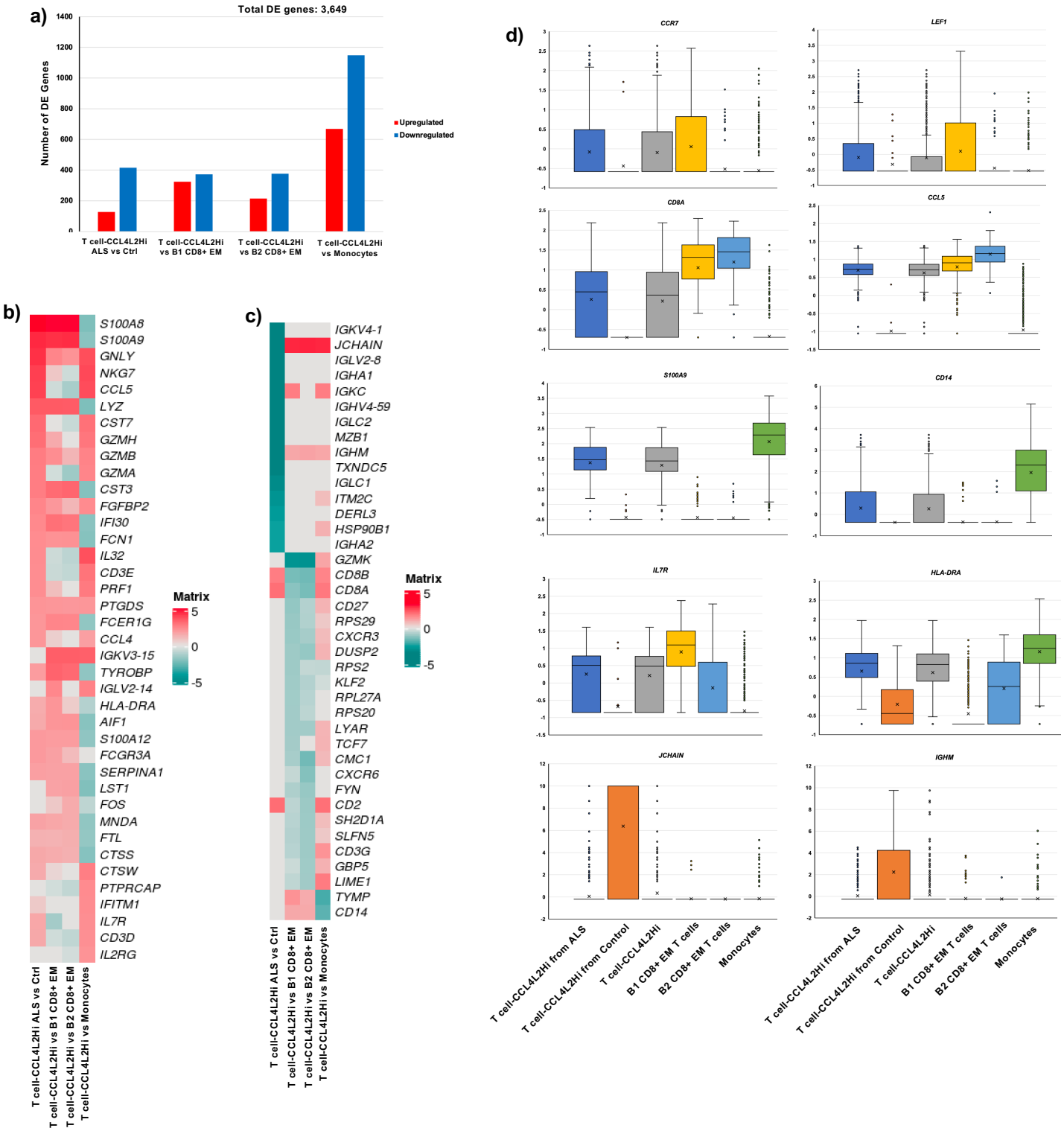

a)
Total DE genes: 3,649
Number of DE Genes
Upregulated
Downregulated
T cell-CCL4L2Hi ALS vs Ctrl
T cell-CCL4L2Hi vs B1 CD8+ EM
T cell-CCL4L2Hi vs B2 CD8+ EM
T cell-CCL4L2Hi vs Monocytes
b)
S100A8
S100A9
GNLY
NKG7
CCL5
LYZ
CST7
GZMH
GZMB
GZMA
CST3
FGFBP2
IFI30
FCN1
IL32
CD3E
PRF1
PTGDS
FCER1G
CCL4
IGKV3-15
TYROBP
IGLV2-14
HLA-DRA
AIF1
S100A12
FCGR3A
SERPINA1
LST1
FOS
MNDA
FTL
CTSS
CTSW
PTPRCAP
IFITM1
IL7R
CD3D
IL2RG
Matrix
5
0
-5
T cell-CCL4L2Hi ALS vs Ctrl
T cell-CCL4L2Hi vs B1 CD8+ EM
T cell-CCL4L2Hi vs B2 CD8+ EM
T cell-CCL4L2Hi vs Monocytes
c)
IGKV4-1
JCHAIN
IGLV2-8
IGHA1
IGKC
IGHV4-59
IGLC2
MZB1
IGHM
TXNDC5
IGLC1
ITM2C
DERL3
HSP90B1
IGHA2
GZMK
CD8B
CD8A
CD27
RPS29
CXCR3
DUSP2
RPS2
KLF2
RPL27A
RPS20
LYAR
TCF7
CMC1
CXCR6
FYN
CD2
SH2D1A
SLFN5
CD3G
GBP5
LIME1
TYMP
CD14
Matrix
5
0
-5
T cell-CCL4L2Hi ALS vs Ctrl
T cell-CCL4L2Hi vs B1 CD8+ EM
T cell-CCL4L2Hi vs B2 CD8+ EM
T cell-CCL4L2Hi vs Monocytes
d)
CCR7
LEF1
CD8A
CCL5
S100A9
CD14
IL7R
HLA-DRA
JCHAIN
IGHM
T cell-CCL4L2Hi from ALS
T cell-CCL4L2Hi from Control
T cell-CCL4L2Hi
B1 CD8+ EM T cells
B2 CD8+ EM T cells
Monocytes

Figure 5

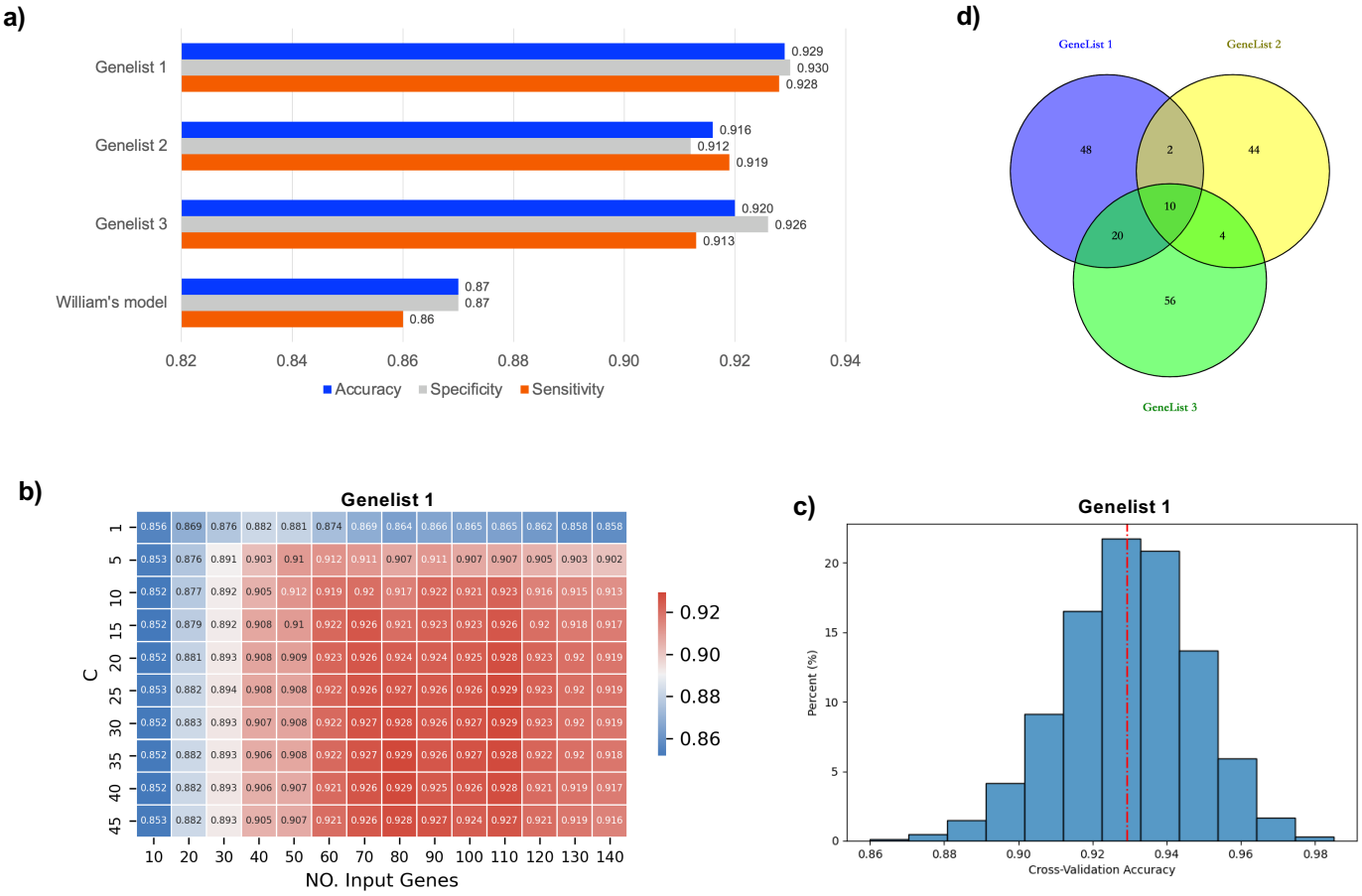

a)
Genelist 1
Genelist 2
Genelist 3
William's model
0.929
0.930
0.928
0.916
0.912
0.919
0.920
0.926
0.913
0.87
0.87
0.86
0.82
0.84
0.86
0.88
0.90
0.92
0.94
Accuracy
Specificity
Sensitivity
d)
GeneList 1
GeneList 2
GeneList 3
48
2
44
10
20
4
56
b)
Genelist 1
C
NO. Input Genes
c)
Genelist 1
Percent (%)
Cross-Validation Accuracy

Figure 6

Supplementary Materials

Supplementary Figures legends

Figure S1. Quality control and processing of single-cell RNA sequencing. a~c). UMAP visualization of cell clusters in three experiments. Each dot represents a single cell, which is colored according to cell clusters. Each cell cluster is labeled. **a).** Control 1; **b).** Control 2; **c).** ALS group; **d).** UMAP visualization of all cells colored by each experiment. **e).** UMAP visualization of transcription level of selected marker genes. **f).** Box plot of the number of expressed genes in each cell type.

Figure S2. Further characterization of the novel T cell-CCL4L2^{Hi} cluster. a). UMAP visualization of T cell-CCL4L2^{Hi} subtypes colored by three experiments. **b).** Heat map of differentially expressed genes for the nine subtypes of T cell-CCL4L2^{Hi}. **c).** Gene expression profile along the pseudotemporal axis between CD8+ naïve T cells and CD8^{+} terminal effector T cells. **d).** Gene expression profile along the pseudotemporal axis between CD8+ naïve T cells and monocytes. The ④ indicates the intersection in the pseudotime analysis in Fig.2c. **e).**

UMAP visualization of transcription levels of selected marker genes. **f**). Occurrence of selected stemness genes in the novel T cell-CCL4L2^{Hi} cluster co-expressing *CD3* and *S100* genes.

Figure S3. TCR and BCR landscapes. a~c) Projection of detected TCR landscape on UMAP visualization colored by experiments. Each dot represents a single cell. The dots in grey represent single cells without detectable TCR, while those in color have detected TCR, which are colored according to the experiments: **a).** Control 1; **b).** Control 2; **c).** ALS group. **d~f)** Projection of detected BCR landscape on UMAP visualization colored by experiments. Each dot represents a single cell. The dots in grey represent single cells without detectable BCR, while those in color have detected BCR, which are colored according to the experiments: **d).** Control 1;

e). Control 2; **f).** ALS group. **g).** Projection of detected BCR landscape on UMAP visualization. **h).** Bar chart of top most abundant BCR clonotypes. Blue color indicates the clonotypes detected in the Control group, while red color indicates the clonotypes found in the ALS group. **i).** Bar charts showing the number of total (top) and unique (bottom) BCR clonotypes detected. In each chart, the clonotypes are compared between the Control group (left, in blue color) and the ALS group (right, in red color).

Fig.S4. Characterization of Transcriptional profiles of ALS patients. Changes in transcriptional profiles and TF-mediated regulation between healthy volunteers and ALS patients. a) GO terms and pathway enrichment analysis of upregulated DE genes for all cell types. **b).** GO terms and pathway enrichment analysis of downregulated DE genes for all cell types. **c).** Violin plots of expression levels of additional selected marker genes.

Fig.S5. Characterization of transcriptome signatures of the novel T cell-CCL4L2^{Hi} cluster. a) GO terms and pathway enrichment analysis of upregulated DE genes. **b).** GO terms and pathway enrichment analysis of downregulated DE genes. **c).** Box plots of expression levels of additional selected marker genes.

Fig.S6. Classifiers for ALS diagnosis using Genelist 2 and Genelist 3. **a**). SVM cost and input genes parameters using Genelist 2. **b).** SVM prediction accuracy using Genelist 2. **c**). SVM cost and input genes parameters using Genelist 3. **d**). SVM prediction accuracy using Genelist 3.

Supplementary Tables

Supplementary Table 1. Information of participants

Supplementary Table 2. DE genes of cell types

Supplementary Table 3. Cell type statistics

Supplementary Table 4. TCR clonotypes

Supplementary Table 5. Statistics of TCR clonotypes

Supplementary Table 6. Analysis of TCR clonotypes by GLIPH2

Supplementary Table 7. BCR clonotypes

Supplementary Table 8. Statistics of BCR clonotypes

Supplementary Table 9. DE genes of T cell CCL4L2Hi subtypes

Supplementary Table 10. DE genes of ALS (P1-P6) vs Control 1 (C1-C6) in bulk RNA sequencing

Supplementary Table 11. DE genes between Control and ALS groups

Supplementary Table 12. GO and pathway analysis of DE genes between Control and ALS groups

Supplementary Table 13. DE genes comparing the novel T cell CCL4L2-Hi cluster and its neighbors on the pseudotemporal trajectory

Supplementary Table 14. GO and pathway analysis of DE genes between the novel T cell-CCL4L2Hi cluster and its neighbors on the pseudotemporal trajectory

Supplementary Table 15. DE genes comparing the novel T cell-CCL4L2^{Hi} cluster and CD8+ cells and monocytes

Supplementary Table 16. Selected genelists for prediction of ALS diagnosis

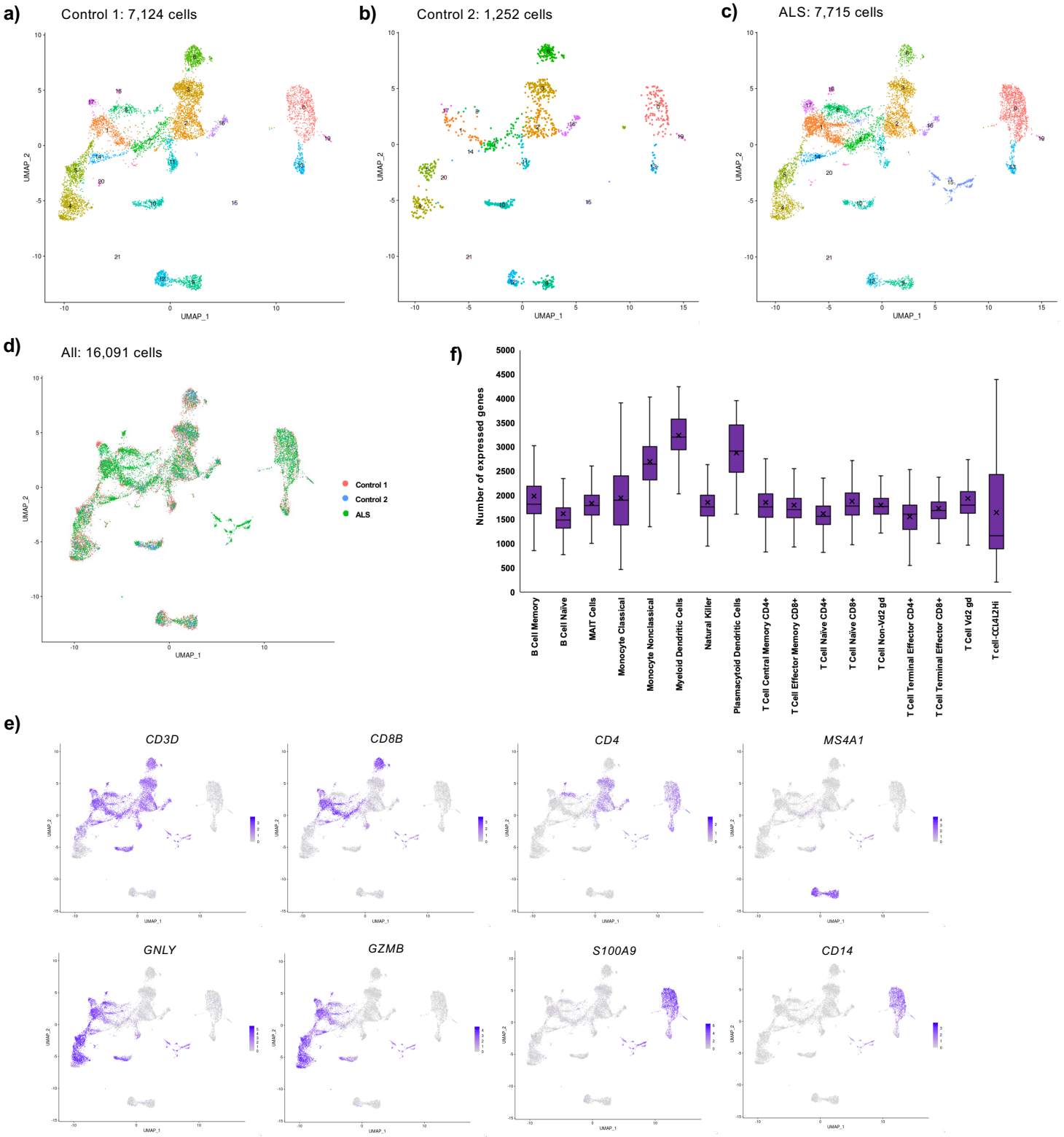

a)
Control 1: 7,124 cells
b)
Control 2: 1,252 cells
c)
ALS: 7,715 cells
d)
All: 16,091 cells
Control 1
Control 2
ALS
f)
Number of expressed genes
B Cell Memory
B Cell Naïve
MAIT Cells
Monocyte Classical
Monocyte Nonclassical
Myeloid Dendritic Cells
Natural Killer
Plasmacytoid Dendritic Cells
T Cell Central Memory CD4+
T Cell Effector Memory CD8+
T Cell Naïve CD4+
T Cell Naïve CD8+
T Cell Non-Vd2 gd
T Cell Terminal Effector CD4+
T Cell Terminal Effector CD8+
T Cell Vd2 gd
T cell-CCL4L2Hi
e)
CD3D
CD8B
CD4
MS4A1
GNLY
GZMB
S100A9
CD14
UMAP_1
UMAP_2

Figure S1

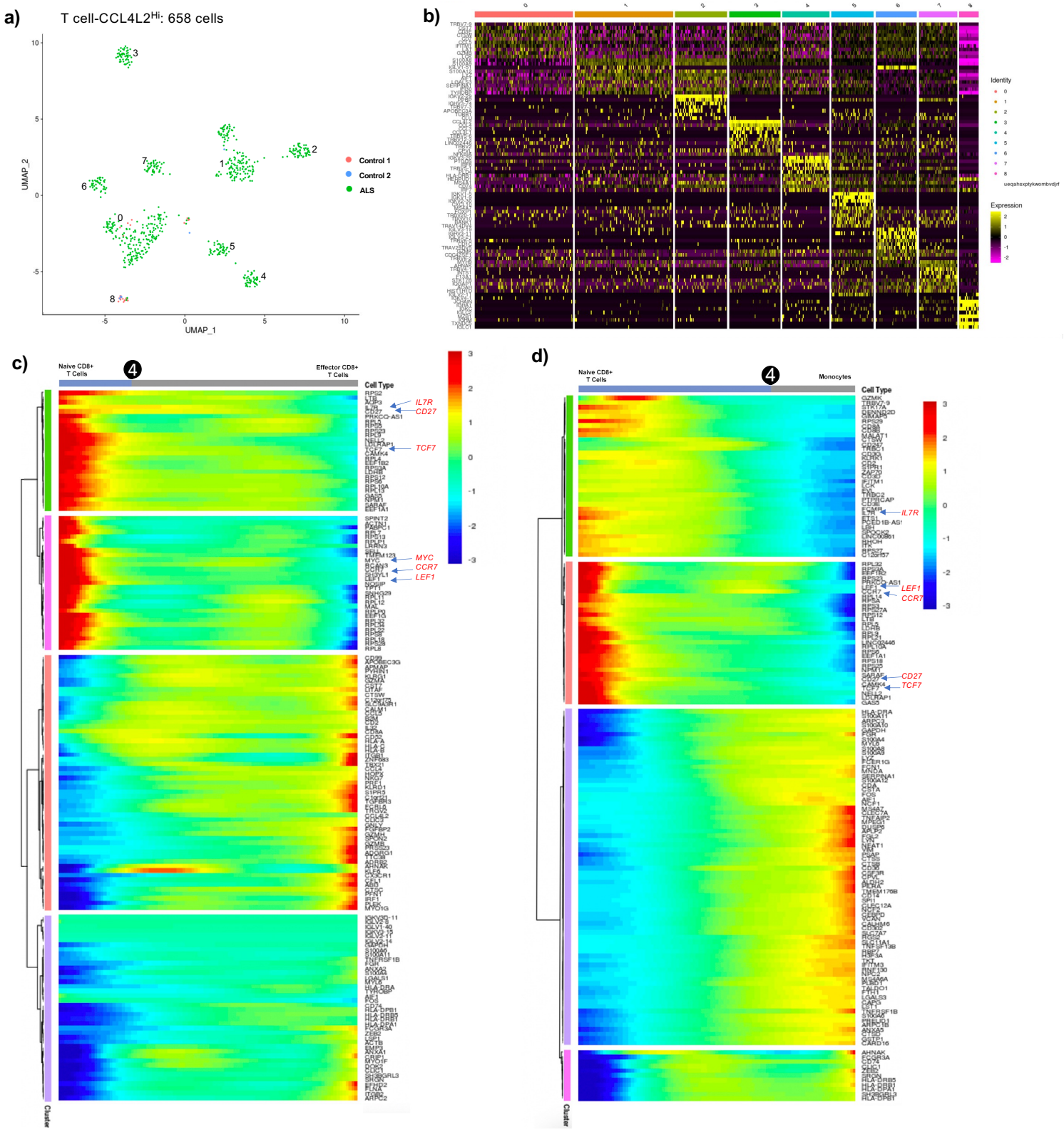
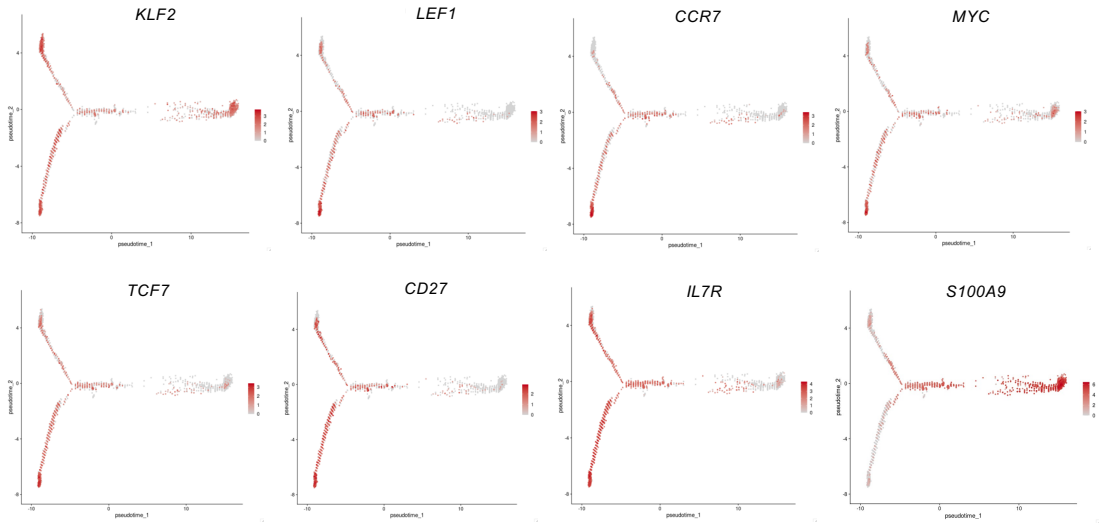

a)
T cell-CCL4L2Hi: 658 cells
UMAP_2
UMAP_1
Control 1
Control 2
ALS
b)
Identity
Expression
c)
Naive CD8+ T Cells
Effector CD8+ T Cells
Cell Type
IL7R
CD27
TCF7
MYC
CCR7
LEF1
Cluster
d)
Naive CD8+ T Cells
Monocytes
Cell Type
IL7R
LEF1
CCR7
CD27
TCF7
Cluster

Figure S2-1

e)



f)

Gene	Occurrence in 582 T cell-CCL4L2Hi expressing *CD3* and *S100*
KLF2	159
LEF1	156
CCR7	199
MYC	134
TCF7	185
Any one of the above	411

Figure S2-2

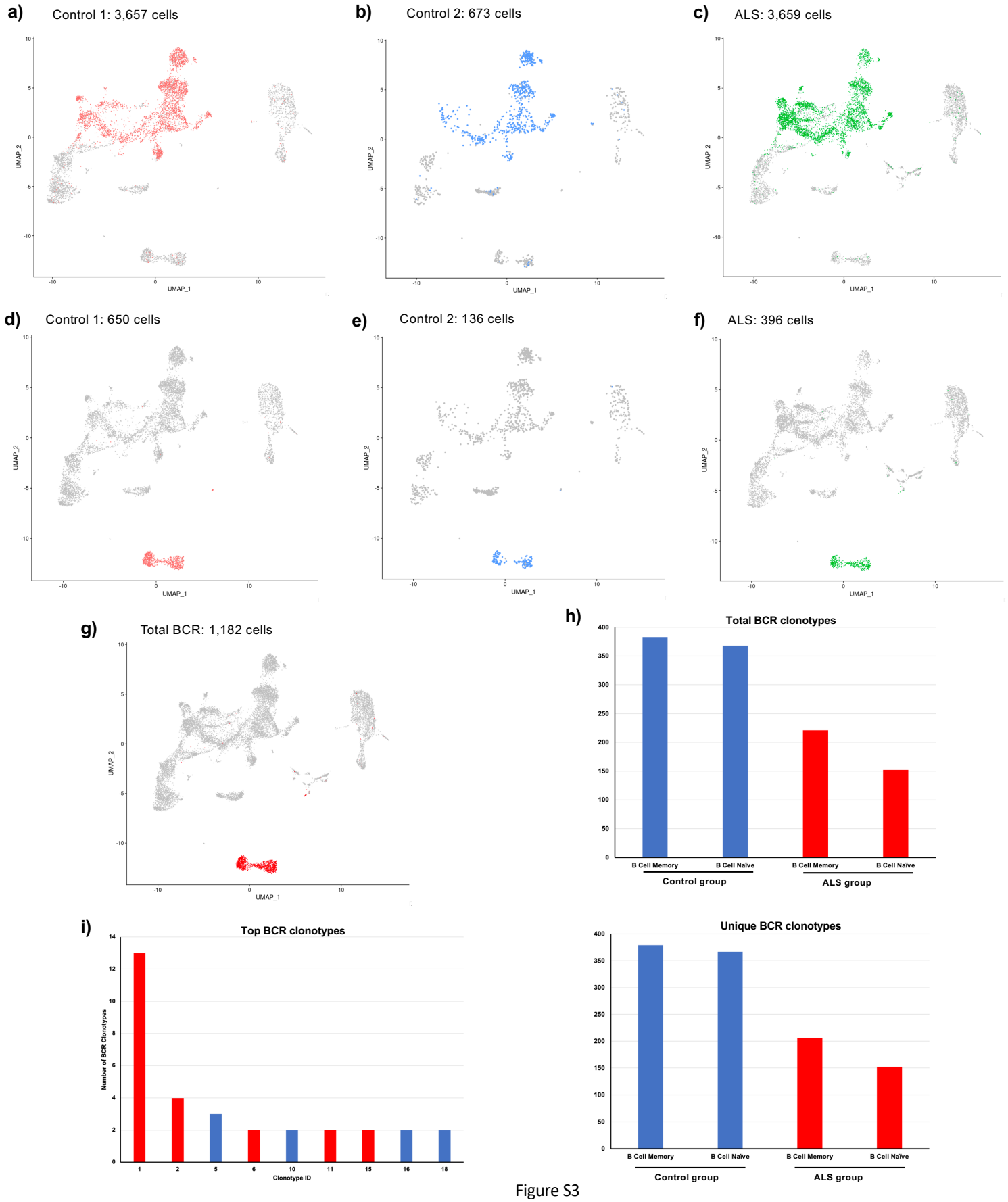

a)
Control 1: 3,657 cells
b)
Control 2: 673 cells
c)
ALS: 3,659 cells
d)
Control 1: 650 cells
e)
Control 2: 136 cells
f)
ALS: 396 cells
g)
Total BCR: 1,182 cells
UMAP_1
UMAP_2
h)
Total BCR clonotypes
B Cell Memory
B Cell Naïve
Control group
ALS group
i)
Top BCR clonotypes
Number of BCR Clonotypes
Clonotype ID
Unique BCR clonotypes

Figure S3

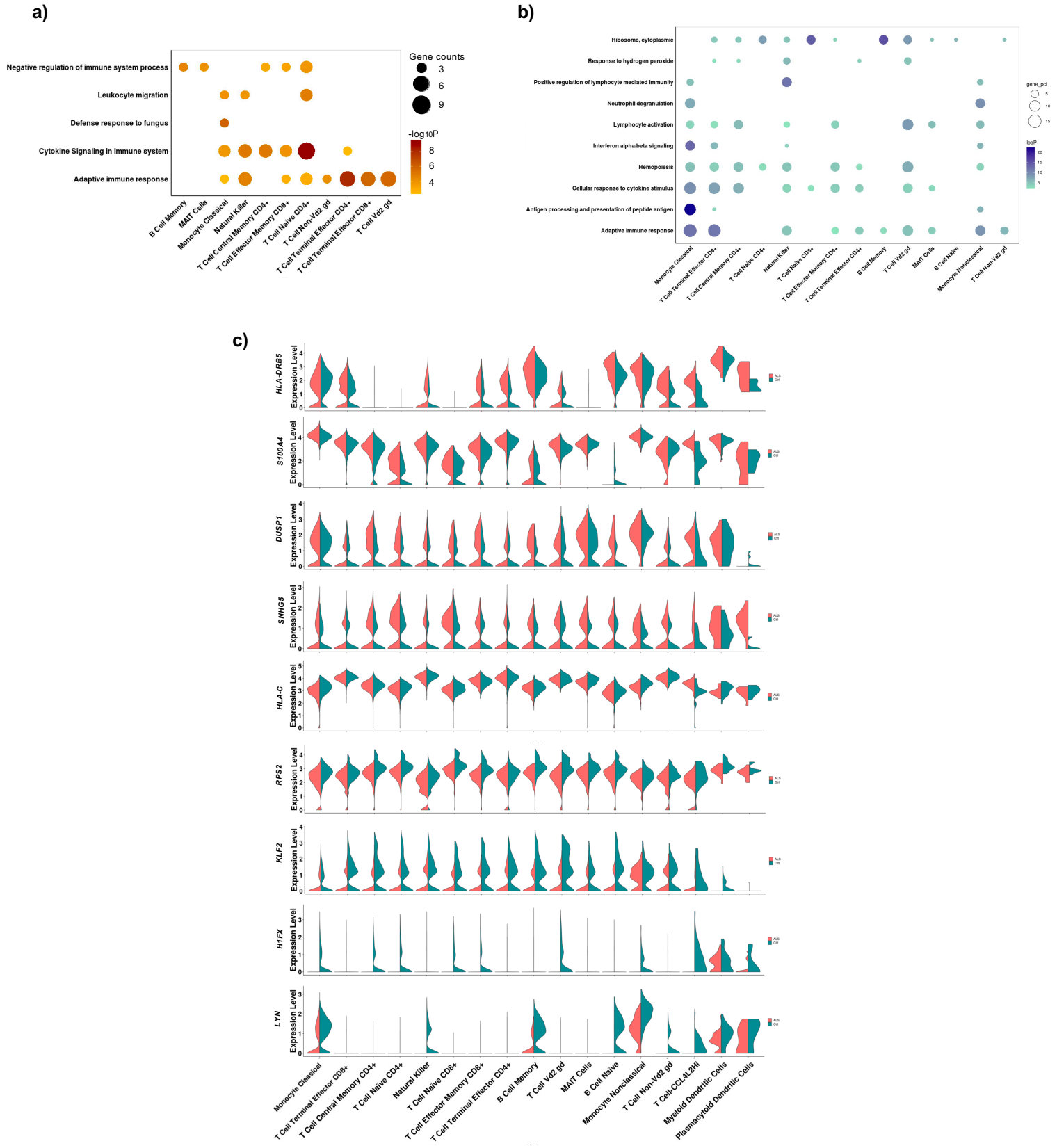

a)
Negative regulation of immune system process
Leukocyte migration
Defense response to fungus
Cytokine Signaling in Immune system
Adaptive immune response
B Cell Memory
MAIT Cells
Monocyte Classical
Natural Killer
T Cell Central Memory CD4+
T Cell Effector Memory CD8+
T Cell Naive CD4+
T Cell Non-Vd2 gd
T Cell Terminal Effector CD4+
T Cell Terminal Effector CD8+
T Cell Vd2 gd
Gene counts
3
6
9
-log10P
8
6
4
b)
Ribosome, cytoplasmic
Response to hydrogen peroxide
Positive regulation of lymphocyte mediated immunity
Neutrophil degranulation
Lymphocyte activation
Interferon alpha/beta signaling
Hemopoiesis
Cellular response to cytokine stimulus
Antigen processing and presentation of peptide antigen
Adaptive immune response
Monocyte Classical
T Cell Terminal Effector CD8+
T Cell Central Memory CD4+
T Cell Naive CD4+
Natural Killer
T Cell Naive CD8+
T Cell Effector Memory CD8+
T Cell Terminal Effector CD4+
B Cell Memory
T Cell Vd2 gd
MAIT Cells
B Cell Naive
Monocyte Nonclassical
T Cell Non-Vd2 gd
gene_pct
logP
c)
HLA-DRB5
S100A4
DUSP1
SNHG5
HLA-C
RPS2
KLF2
H1FX
LYN
Expression Level
ALS
Ctrl
Monocyte Classical
T Cell Terminal Effector CD8+
T Cell Central Memory CD4+
T Cell Naive CD4+
Natural Killer
T Cell Naive CD8+
T Cell Effector Memory CD8+
T Cell Terminal Effector CD4+
B Cell Memory
T Cell Vd2 gd
MAIT Cells
B Cell Naive
Monocyte Nonclassical
T Cell Non-Vd2 gd
T Cell-CCL4L2Hi
Myeloid Dendritic Cells
Plasmacytoid Dendritic Cells

Figure S4

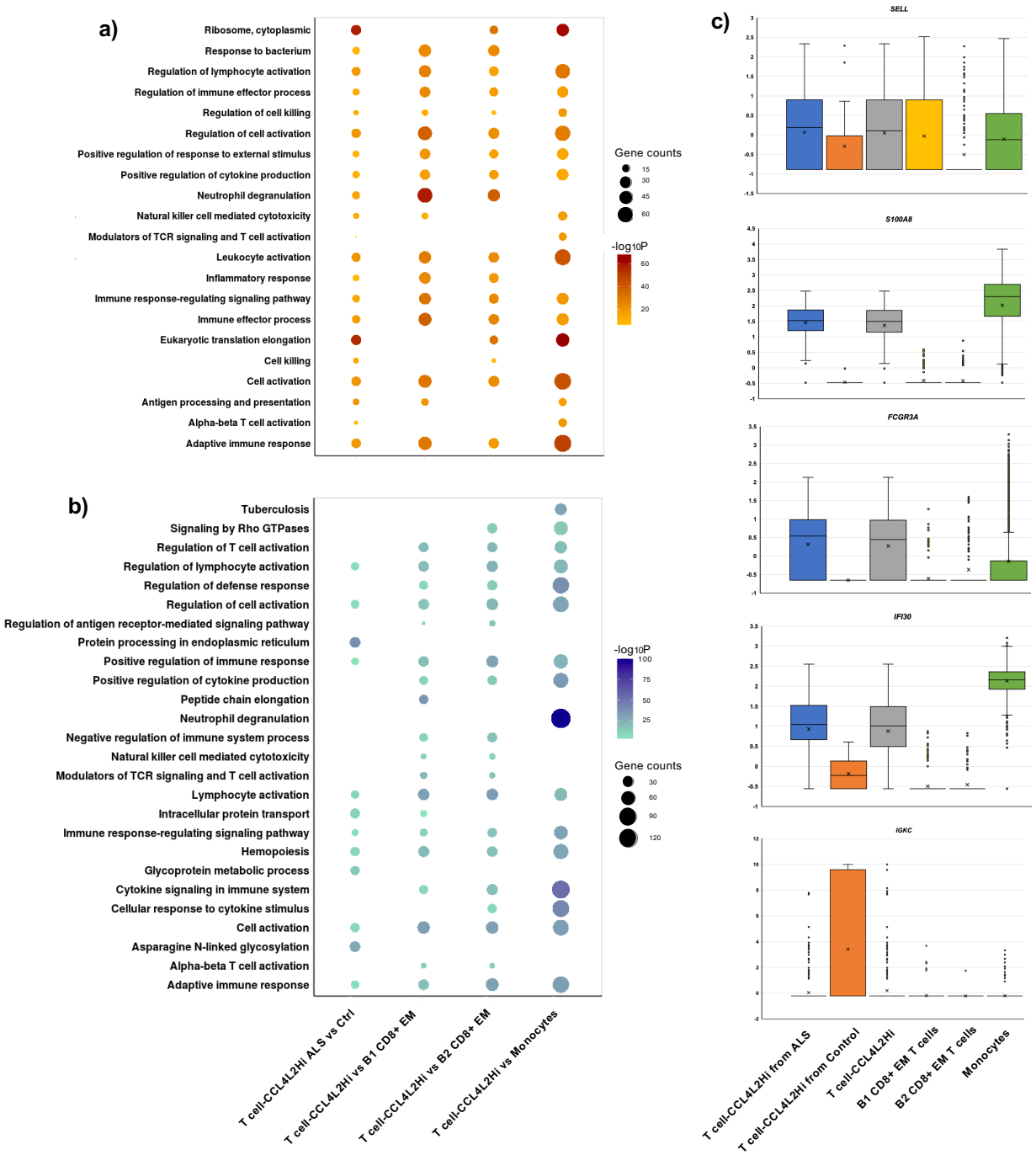

a)
Ribosome, cytoplasmic
Response to bacterium
Regulation of lymphocyte activation
Regulation of immune effector process
Regulation of cell killing
Regulation of cell activation
Positive regulation of response to external stimulus
Positive regulation of cytokine production
Neutrophil degranulation
Natural killer cell mediated cytotoxicity
Modulators of TCR signaling and T cell activation
Leukocyte activation
Inflammatory response
Immune response-regulating signaling pathway
Immune effector process
Eukaryotic translation elongation
Cell killing
Cell activation
Antigen processing and presentation
Alpha-beta T cell activation
Adaptive immune response
Gene counts
15
30
45
60
-log10P
60
40
20
b)
Tuberculosis
Signaling by Rho GTPases
Regulation of T cell activation
Regulation of lymphocyte activation
Regulation of defense response
Regulation of cell activation
Regulation of antigen receptor-mediated signaling pathway
Protein processing in endoplasmic reticulum
Positive regulation of immune response
Positive regulation of cytokine production
Peptide chain elongation
Neutrophil degranulation
Negative regulation of immune system process
Natural killer cell mediated cytotoxicity
Modulators of TCR signaling and T cell activation
Lymphocyte activation
Intracellular protein transport
Immune response-regulating signaling pathway
Hemopoiesis
Glycoprotein metabolic process
Cytokine signaling in immune system
Cellular response to cytokine stimulus
Cell activation
Asparagine N-linked glycosylation
Alpha-beta T cell activation
Adaptive immune response
-log10P
100
75
50
25
Gene counts
30
60
90
120
T cell-CCL4L2Hi ALS vs Ctrl
T cell-CCL4L2Hi vs B1 CD8+ EM
T cell-CCL4L2Hi vs B2 CD8+ EM
T cell-CCL4L2Hi vs Monocytes
c)
SELL
S100A8
FCGR3A
IFI30
IGKC
T cell-CCL4L2Hi from ALS
T cell-CCL4L2Hi from Control
T cell-CCL4L2Hi
B1 CD8+ EM T cells
B2 CD8+ EM T cells
Monocytes

Figure S5

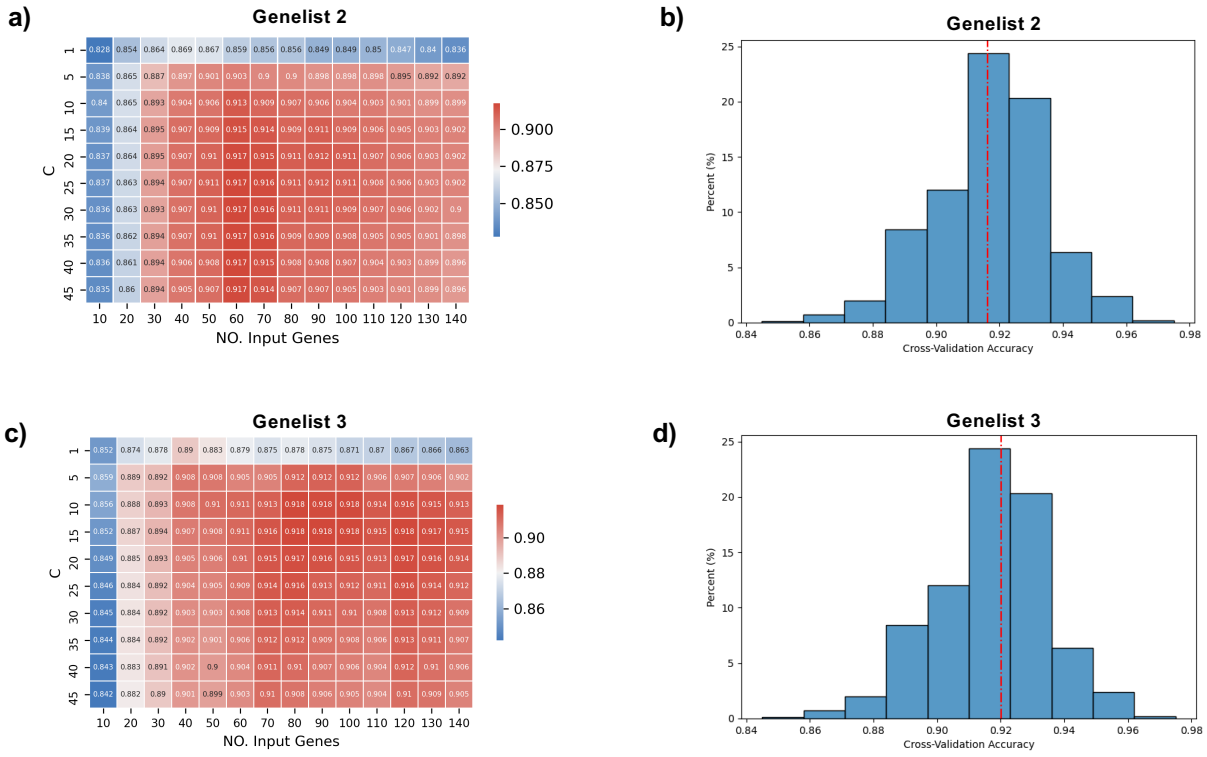

a)
Genelist 2
C
NO. Input Genes
0.900
0.875
0.850
b)
Genelist 2
Percent (%)
Cross-Validation Accuracy
c)
Genelist 3
C
NO. Input Genes
0.90
0.88
0.86
d)
Genelist 3
Percent (%)
Cross-Validation Accuracy

Figure S6