Single-cell sequencing identifies a novel stem-like T-lymphocyte cluster in sporadic amyotrophic lateral sclerosis patients

Albert Ziye Ma[1#], Brian D. Kirk[2#], Deyu Yan[3], Qinghua Wang[4], Weihai Ying[5*] and Yan Han[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]School of Biomedical Engineering and Med-X Institute, Shanghai Jiao Tong University, Shanghai, China; [6]Department of Neurology, Yueyang Hospital of Integrated Traditional Chinese and Western Medicine, Shanghai University of Traditional Chinese Medicine, Shanghai, China 200437. [#]Co-first authors. [*]Senior authors.

Correspondence to: Dr. Weihai Ying, School of Biomedical Engineering and Med-X Institute, Shanghai Jiao Tong University, Shanghai, China. Email: weihaiy@sjtu.edu.cn, Phone: (86)130-5221-8008.

The authors have declared that no conflict of interest exists.

January 6, 2023

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 eight sALS patients and depicted the first cellular landscape at single-cell resolution. We discovered a novel stem-like T cell cluster that is located at a critical junction among $CD8^+$ naïve and terminal effector T cells and monocytes in pseudotime analysis. Importantly, the abundance of $CD3^+CD16^+$ cells, accounting for ~32% of this novel cell cluster, is positively correlated with the disease progression in multiplex flow cytometry analyses of five additional sALS patients, suggesting a central role in sALS pathogenesis. We further identified 60 genes that distinguished ALS patients from healthy controls with an unprecedented accuracy of 93% against microarray data from an independent cohort of 792 specimens. Our findings have important implications for understanding sALS pathogenesis, and provide new strategies for accurate diagnosis and potential therapeutic development.

Introduction

Sporadic amyotrophic lateral sclerosis (sALS) 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.

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[2,3]. 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[4-7]. Furthermore, decreased blood $CD4^+$ and $CD8^+$ T cells and B cells have been shown to predict cognitive impairment in ALS patients[8]. 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 peripheral blood biomarkers of ALS could enable early and accurate diagnosis, improve patient management and survival, and probably provide novel therapeutic targets. Towards this end, 14 serum proteins with significant changes in ALS patients compared to controls were identified using quantitative discovery proteomics[9]. 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[10]. Furthermore, 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 an overall 87% accuracy (sensitivity of 86% and specificity of 87%)[11], the highest accuracy in the literature.

Single-cell sequencing technology is a revolutionary tool that, different from bulk 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[12]. Most recently, a scRNA-seq study of ALS4 (caused by mutations in senataxin gene) uncovered an immunological signature consisting of clonally expanded, terminally differentiated effector memory $CD8^+$ T cells in the blood and central nervous system of knock-in mice[13].

In the current study, by combining scRNA-seq together with single-cell T cell receptor (TCR) sequencing (scTCR-seq), we unraveled for the first time the cellular landscape of PBMC samples from sALS patients at single-cell resolution, mapped out the detailed abundance changes of different PBMC cell types, discovered a novel stem-like T cell cluster, and identified 60 genes that collectively distinguished ALS patients from healthy controls at 93% accuracy. Our findings have important implications for the pathogenesis of sALS and other immune-mediated diseases, as well as provide new strategies for accurate diagnosis and potential therapeutic development.

Results

Single-cell transcriptome and TCR analysis of PBMC cells from sALS patients and healthy volunteers

We collected peripheral blood from three sets of clinically confirmed sALS patients. The first set (ALS1) was comprised of pooled samples from six patients (n=6, P1~P6, pooled as equal portion, **Supplementary Table 1**) with an average age of 61.3±11.7. The second set (ALS2) was from a single patient of age 55~60 (n=1, P7, **Supplementary Table 1**). The third set (ALS3) was from a single patient of age 61~65 (n=1, P12, **Supplementary Table 1**). In order to provide adequate controls on age and health conditions, we selected three healthy volunteers as controls, C1~C3, with an average age of 68.7±11.4 years old. ScRNA-seq and scTCR-seq were performed for ALS1~3 and C1 in parallel, while the data for C2 and C3 were obtained from the Gene Expression Omnibus data repository under accession code GSE157007[14].

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

We obtained high-quality scRNA-seq data for a total of 40,922 PBMC cells from the combined Control and ALS groups. After multiple quality-control steps including DoubletFinder to detect and remove doublets, we performed clustering analysis using SingleR and detected 47 clusters (Cluster 0~46) (**Fig.1a~c**). The UMAPs for the PBMC cells from the Control group and the ALS group have a similar distribution (**Fig.1a, Fig.S1a**). These clusters were annotated as 22 different cell types (**Fig.1a~c**) based on their marker genes (**Fig.1d, Fig.S1c, Supplementary Table 2**). The number of genes observed in each cell type was averaged at around 2,000 (**Fig.S1b**).

Clusters 0, 20, 28, 34 and 39 high in *CD8A, CD8B* and *GZMH* were assigned as $CD8^+$ terminal effector T (T_{TE}) cells (**Fig.1d**). Clusters 1, 9 and 10 were designated as $CD4^+$ central

memory T cells. Clusters 2, 3, 11, 14 and 18 (high in *GNLY, SPON2* and *GZMB*) were recognized as natural killer cells. Clusters 4 was defined as naïve B cells (high in *MS4A1, CD79A*, *HLA-DRA* and TCL1A). Clusters 5, 6, 13, 15, 21, 31 and 32 were recognized as classical monocytes, for the varying levels of expression for *S100A8, S100A9 and LYZ*. Clusters 7, 16 and 36 (high in *CCR7, LEF1,* and *TCF7*) were $CD4^+$ naïve T cells, while Cluster 22 was $CD8^+$ naïve T cells. Cluster 8 and 17 were $CD8^+$ effector memory T (T_{EM}) cells. Cluster 12 was classified as memory B cells. Cluster 19 (high in LST1, MS4A7 and SERPINA1) was defined as non-classical monocytes. Clusters 23 was classified as Vd2-gd T cells. Cluster 24 was recognized as $CD4^+$ T_{TE} cells. Cluster 25 was non-Vd2-gd T cells. Cluster 26 (high in *FOXP3*) was regulatory T cells. Cluster 30 was assigned as mucosal-associated invariant T (MAIT) cells. Cluster 33 was recognized as megakaryocytes. Clusters 37 and 42 were myeloid dendritic cells and plasmacytoid dendritic cells, respectively. Cluster 41 was recognized as proliferating T cells (high in MKI67). Cluster 45 was progenitor cells. Clusters 46 was assigned as plasma cells. Clusters 27, 29, 35, 38, 40, 43 and 44 displayed an unusual combination of cytotoxicity genes *S100A8, S100A9, LYZ, GZMH* and *GNLY*, T cell marker genes *CD3D, CD3E, CD8A,* and *CD8B*, as well as major histocompatibility complex (MHC) class II antigen *HLA-DRA* (**Fig.1d**). Therefore, they were collectively labeled as novel ALS cells and further analyzed.

The abundance of each major cell type was compared (**Fig.1e~g, Supplementary Table 3**). The distributions of most cell types in the three controls were very close to each other (**Supplementary Table 3**). Relative to the Control group, the distributions of major cell types in the ALS group with greater than $\log_2 Fc>0.7$ increase included the novel ALS cell cluster, $CD4^+$ T_{TE} cells, and classical monocytes, while those with greater than $|\log_2 Fc|>0.7$ decrease were Vd2-gd T cells, naïve $CD8^+$ and $CD4^+$ T cells, and MAIT cells (**Fig.1e~g**, megakaryocytes,

myeloid dendritic cells, proliferating T cells, plasmacytoid dendritic cells, progenitor cells and plasma cells are excluded in **Fig.1g** due to their small number of cells). Of note, the novel ALS cell cluster exhibited the largest abundance increase in the ALS group comparing to the Control group, from an average of 0.55% (84 cells) in the Control group to 5.03% (1,325 cells) in the ALS group (**Supplementary Table 3**).

A novel stem-like T lymphocyte cluster in sALS patients

The novel ALS cell cluster (1,409 cells) was further analyzed at a higher resolution, yielding a total of 10 subclusters (subclusters 0~9) (**Fig.2a, Fig.S2a**). Most subclusters exhibited high expression in cytotoxicity marker genes such as *S100A8, S100A9, LYZ,* and T cell marker genes (*CD3D, CD3E, CD3G, CD8A, CD8B*) (**Fig.2b, Supplementary Table 4**). The DE genes with large fold changes among different subclusters were *TRBV, IgkV* and *IgLV* genes (**Fig.2b, Fig.S2b, Supplementary Table 4**), suggesting that the reason of the novel ALS cell cluster to be heterogenous and separated into 10 subclusters was at least partially due to their differences in TCR and antibody expression. Out of the 1,409 cells in the novel ALS cell cluster, 55.9% (787 cells) of which had at least one *CD3* genes expressed (labeled as $CD3^+$ cells). Among all these $CD3^+$ cells, 60.9% (479 cells) had at least one *CD8* gene, 25.3% (199 cells) had *CD4* gene, and 9.94% (140 cells) had both *CD8* and *CD4* genes (**Fig.2c**). In addition, 93.1% (1,312 cells) of the novel ALS cell cluster had at least one *S100* (*S100A8* or *S100A9*) expressed, 38.0% (499 cells) of which expressed *FCGR3A* (*CD16*), 48.2% (633 cells) expressed *CD14*, and 15.1% (198 cells) expressed both *CD16* and *CD14* (**Fig.2d**). Of all the novel ALS cell cluster, 51.2% (721 cells) had at least one *CD3* and one *S100* genes (**Fig.2e**), and 32.2% (453 cells) had both *CD3* and *CD16* (**Fig.2f**).

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, and 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 novel ALS cell 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 cluster (**Fig.2b, 2g**). Among the 721 cells that co-expressed *CD3* and *S100* genes, only 5.96% (43 cells) co-expressed *ITGB2* and *ICAM1* genes (**Fig.2g**). Moreover, the average number of genes for the novel ALS cell cluster was about 1500, similar to that of other cell types (**Fig.s1b**), while doublets tend to have a much higher average number of genes. These multiple lines of evidence collectively provided strong support that the novel ALS cell cluster is unlikely T cell-monocyte complex doublet as described previously[15], instead, it may represent a novel cell type with unique features.

To provide orthogonal evidence for the increased presence of the novel ALS cell cluster in PBMC samples of ALS patients, we additionally performed multiplex flow cytometry. For this experiment, we chose to sort for cells with high levels of both CD3 and CD16 proteins, which represented 32.2% of all the novel ALS cell cluster detected in our scRNA-seq analysis (**Fig.2f**). The flow cytometry analysis scheme included the following steps: lymphocytes were first gated and used to select for singlet cells (often at 80~99.8% of selected lymphocytes); after dead cells were discarded, cells with high levels of both CD3 and CD16 proteins (labeled as $CD3^{+}CD16^{+}$ cells) were sorted (**Fig.S2c**). For the flow cytometry experiments, we additionally recruited five ALS patients (P7~P11) of varying degrees of disease progression together with two healthy volunteers (C4 and C5) (**Supplementary Table 1**). Among the five new ALS

patients' lymphocyte specimens that we tested, the $CD3^+CD16^+$ cell population varied in the range of 1.27%~11.81%, about 3~29 folds of those from healthy controls C4 and C5 (at 0.41% and 0.48%, respectively) (**Fig.2h~i**). Intriguingly, P11 with the lowest $CD3^+CD16^+$ cell content also had the mildest ALS symptoms, whilst P7~P10 with higher $CD3^+CD16^+$ cell population (at 3.35~11.81% among singlet, live lymphocytes) had much more severe ALS symptoms.

Next, for the shared markers of the novel ALS cell cluster with $CD8^+$ T cells and monocytes, we performed a pseudotime analysis that included the novel ALS cell cluster (1,325 cells), monocytes (classical and non-classical) (6,612 cells) and $CD8^+$ T cells (4,711 cells) from the ALS group (**Fig.2j**). The pseudo-temporal trajectory mapped $CD8^+$ naïve T cells (Branch state 1) at the beginning of the trajectory, and 71.1% of $CD8^+$ T_{EM} cells (1,072 cells, termed as B1 T_{EM} cells) along the increasing pseudotime scale on Branch 1, then the novel ALS cell cluster at a critical three-way junction (**Fig.2j**). Following this junction, ~28.9% of $CD8^+$ T_{EM} cells (435 cells, termed as B2 T_{EM} cells) were mapped along the increasing pseudotime scale of Branch 2, with $CD8^+$ T_{TE} cells located at the end of this branch (Branch state 2), and the majority of monocytes (Branch states 3 and 4) were mapped to the end of branches 4 and 5 following Branch 3 along the pseudotime axis (**Fig.2j**). The novel ALS cell cluster (Branch state 5) appeared to be intermediates among B1 and B2 T_{EM} cells and monocytes on the pseudotime trajectory, agreeing very well with its shared marker genes with $CD8^+$ T cells and monocytes.

We then analyzed the gene expression profiles along the pseudotime trajectory between $CD8^+$ naïve T cells (Branch state 1) and $CD8^+$ T_{TE} cells (Branch state 2) (**Fig.S2d**) and between $CD8^+$ naïve T cells (Branch state 1) and monocyte cells (Branch states 3 and 4) (**Fig.S2e**). The hallmark genes for naïve T cells, for instance *TCF7, MYC, CCR7,* and *LEF1*, became low or extremely low in expression in $CD8^+$ T_{TE} cells or monocytes, as expected. However, comparing

to $CD8^+$ T_{TE} cells or monocytes, the novel ALS cell cluster (at regions surrounding the three-way junction indicated by ⑤ in **Fig.2j; Fig.S2d~e**) had considerably higher expression levels for these genes, suggesting the stem-like characteristics of this novel cluster (**Fig.S2d~f**). Among the 721 novel ALS cells co-expressing *CD3* and *S100*, 64.1% (462 cells) had detectable expressions for at least one gene of *LEF1, CCR7, MYC,* or *TCF7*, confirming the predominance of the stem-like characteristics in these novel cells (**Fig.S2g**). T-cell activation factors *CD27* and *IL7R*, critical players in V(D)J recombination during T-lymphocyte development, also had considerably higher expression levels in the novel ALS cells than in monocytes (**Fig.S2d~f**). On the other hand, the novel ALS cells had significantly higher level of *S100A9* than $CD8^+$ T cells (**Fig.S2f**).

Clonal expansion of the TCR repertoire in sALS 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 14,135 clonotypes with paired TCR α- and β-chains (**Fig.3a, Fig.S3a~b, Supplementary Table 5**). Among them, 95.8% (13,538 cells) were mapped to T cells identified by scRNA-seq in this study (**Supplementary Table 6**). The detected TCR for non-T cells was only 4.2%, suggesting the high specificity of scTCR-seq and confirming the accuracy of scRNA-seq-based T-cell assignment.

Most T cell subtypes from ALS group had a high percentage (37~92%) of cells with paired TCR detected, some of the exceptions were the novel ALS cells at 6%, and understandably non-Vd2 gd T cells (at 15%) and Vd2 gd T cells (at 3%). The percentage of paired TCR detected in the Control group was relatively lower, in the range of 25~78%, again

with the novel ALS cells, non-Vd2 gd T cells and Vd2 gd T cells and at about 13%, 4%, and 2%, respectively (**Supplementary Table 6**). The low percentages of non-Vd2 gd T cells and Vd2 gd T cells with paired TCR α- and β-chains confirmed the assignment of these cell types, whilst the low percentages of novel ALS cells agreed with the relatively low expression of TCR (**Fig.2b**) and were consistent with their overall stem-like characteristics.

The overall ratio of total TCR/unique TCR for all cell types were 1.42 for the Control group, and slightly higher, at 1.52, for the ALS group (**Supplementary Table 6**). As expected, TCR clonotypes in the less-differentiated T cells including $CD4^+$ and $CD8^+$ naïve T cells, and $CD4^+$ central memory T cells were rarely expanded (**Supplementary Table 6**). On the other hand, the more-differentiated $CD4^+$ and $CD8^+$ T_{TE} cells were moderately expanded, at an average ratio of 4.20 and 7.73, respectively, in the Control group, and of 5.70 and 8.31, respectively, in the ALS group (**Supplementary Table 6**).

We further classified TCR clonotypes by the degrees of clonal expansion: high ($x>20$), medium ($5 < x \leq 20$), low ($1 < x \leq 5$) and single ($x = 1$) (**Fig.3b, Fig.S3a~b**). The portions of TCR clonotypes fell in high, medium, low and single classes were at about 25%, 7%, 10%, 58% for the ALS group, and 20%, 9%, 9% and 62% for the Control group, respectively (**Fig.3c, Supplementary Table 7**), with the high class significantly enriched in the ALS group, confirming the higher expansion for the ALS group than the Control group. In total, the ALS group and the Control group had 6,397 and 2,687 unique TCR clonotypes, respectively (**Fig.3d, left panel**). Among them, 6.6% of all unique TCR clonotypes in the ALS group or the Control group were observed in more than one T cell type (**Fig.3d, 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 specificity groups containing at least 3 unique TCRs from 3 or more individuals infected with *Mycobacterium tuberculosis*[16]. By using GLIPH2 on the total of 13,538 paired TCR clonotypes detected on T cells in this study, we identified 741 GLIPH specificity groups with at least 2 copies in each group (**Supplementary Table 8**). The top 10 most abundant specificity groups were shown in **Fig.3e**, which had copy numbers ranging from 71~421 copies. For the top 10 specificity groups (**Fig.3e**), only Group 1 and Group 4 had two composite sequences. Group 1 was found in ALS2 (one patient), with the two composite sequences differing by three residues, 'I' *vs.* 'A' in the CDR3 sequences of TCR β-chains (TCRb) and 'SF' *vs.* 'CY' in the CDR3 sequences of TCR α-chains (TCRa). Group 4 was found in ALS1, with 'S' *vs.* 'T' in the TCRb for the two composite sequences. Although ALS1 contained PBMCs from six patients, the almost identical CDR3 sequences and the same usage of V and J in Group 4 suggest that this specificity group was likely from one patient. All other 8 GLIPH groups were composed of one single paired TCRb and TCRa sequences (**Fig.3e**). 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 3 were from the Control group, whilst all other groups were found in ALS group, agreeing with the slightly higher expansion in the TCR repertoire of the ALS group (**Fig.3c, 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. Among the top 10 usages, seven TRBV genes (TRBV20-1, TRBV5-1, TRBV4-1, TRBV7-2, TRBV7-9, TRBV28, and TRBV19） and nine TRBJ genes (TRBJ2-1, TRBJ2-7, TRBJ2-3, TRBJ2-5, TRBJ2-2,

TRBJ1-1, TRBJ1-2, TRBJ1-5 and TRBJ1-4) were shared between the ALS and Control groups (**Fig.3f-g**).

Transcriptome signatures of the PBMC cells from the sALS patients

We analyzed the DE genes between the ALS and the Control groups for each cell type (**Supplementary Table 9**). Overall, between the two groups, there were slightly more genes downregulated (9,223 DE genes) than those upregulated in the ALS group (7,095 DE genes) (**Fig.4a**), with a total of 16,318 DE genes. Classical monocyte had the most DE genes (1,061 upregulated and 696 downregulated genes) between the two groups (**Fig.4a**). The progenitor T cells and plasma cells had no upregulated DE genes (**Fig.4a, Supplementary Table 9**). Gene ontology (GO) and pathway analysis revealed that the upregulated DE genes were mostly involved in GO terms directly related to immune response, regulation of actin cytoskeleton and cellular responses to stress (**Fig.S4a, Supplementary Table 10**), while the downregulated DE genes were largely related to immune response, negative regulation of intracellular signal transduction and mRNA metabolic process (**Fig.S4b, Supplementary Table 10**). A total of 42 DE genes were observed in at least 18 cell types (**Fig.4b**). **Fig.4c** shows the heatmap of the relative expression level of these 42 genes, 13 of them were upregulated while 29 were downregulated in the ALS group. Interestingly, out of the 13 upregulated genes in the ALS group, six of them were related to intracellular motility and cytoskeleton, such as β-actin (*ACTB*), coronin-like protein A (*CORO1A*), cytoplasmic dynein light chain LC8-type 1 (*DYNLL1*), myosin light chain 12A (*MYL12A*), profilin 1 (PFN1) and thymosin beta-4 (TMSB4X) (**Fig.4c~d, Fig.S4c**). Also upregulated DE genes were two proteasome proteins *PSMB8* and *PSMB9*, heat shock protein family A member 8 (*HSPA8*), and S100 calcium binding

protein A11 (*S100A11*) (**Fig.4c~d**). On the other hand, among the 29 downregulated DE genes, four proteins are critical players in the nuclear factor kappa B (NFKB) pathway, including TNF alpha induced protein 3 (*TNFAIP3*), NFKB inhibitor alpha (*NFKBIA*), NFKB inhibitor zeta (*NFKBIZ*), and REL proto-oncogene NFKB subunit (*RELB*), and four proteins *RPS27, RPS29, RPL37A,* and *EIF1* are related to ribosome and translation regulation. Also downregulated were phosphatase 1 regulatory subunit 15A (*PPP1R15A*) that facilitates recovery of cells from stress[17] (**Fig.4c~d**).

Transcriptome signatures of the novel ALS cell cluster

Along the pseudo-temporal trajectory of $CD8^+$ T cells and monocytes (**Fig.2j**), the novel ALS cell cluster adjoined two subgroups of $CD8^+$ T_{EM} cells, B1 and B2 $CD8^+$ T_{EM}, respectively, and monocytes. B1 $CD8^+$ T_{EM} cells were between naïve $CD8^+$ T cells and novel ALS cell cluster, while B2 $CD8^+$ T_{EM} cells were between novel ALS cell cluster and $CD8^+$ T_{TE} cells. We analyzed the DE genes for the following five pairs: cells in the novel ALS cell cluster between the ALS and Control groups; the novel ALS cell cluster with B1 or B2 $CD8^+$ T_{EM} cells, or $CD8^+$ T_{TE} cells, or monocytes from the ALS group (**Supplementary Table 11**). This analysis identified a total of 3,915 DE genes, where the comparison between novel ALS cell cluster and monocytes yielded the largest number of DE genes (988 genes in total) (**Fig.5a**). GO and pathway analysis revealed that the significantly enriched GO terms by upregulated and downregulated DE genes were largely overlapping, including immune response, cell killing, cell activation, response to pathogens, and TCR signaling (**Fig.S5a~b, Supplementary Table 12**).

The expression levels of naïve T cell hallmark genes such as *CCR7, LEF1* and *MYC* were higher in the novel ALS cell cluster than in $CD8^+$ T_{TE} cells or monocytes (**Fig.5b~5d**), agreeing

with the more stem-like characteristics of the novel ALS cell cluster. T cell activation marker *CD27* had a higher expression level in the novel ALS cell cluster than in $CD8^+$ T_{TE} cells, but lower than those in B1 $CD8^+$ T_{EM} cells, and on par with those in B2 $CD8^+$ T_{EM} cells (**Fig.5b~5d, Fig.S5c**), thus positioning the novel ALS cell cluster at a similar level of T cell activation as B2 $CD8^+$ T_{EM} cells. Furthermore, the lower expression levels of *CD8A* and *CCL5* in the novel ALS cells than in $CD8^+$ T_{EM} and T_{TE} cells confirmed the less-differentiated status of the novel ALS cell cluster (**Fig.5b~5d**). Moreover, for cytotoxicity genes such as *GNLY* and *GZMB*, the novel ALS cell cluster had a higher expression level than in monocytes, but was lower than those in $CD8^+$ T_{TE} cells, while in contrast, for *S100A8* and *S100A9* genes, the novel ALS cell cluster had a similar expression to those in monocytes, but was higher than in $CD8^+$ T_{TE} cells (**Fig.5b~5d, Fig.S5c**). These unusual transcriptome signatures of the novel ALS cell cluster collectively further established it as a critical differentiation intermediate with stem-like characteristics.

We further compared the expression levels of important gents within the novel ALS cell cluster between the Control and ALS groups. Relative to the Control group, the ALS group exhibited lower expression levels in naïve T cell hallmark genes such as *CCR7, LEF1* and *MYC*, and higher levels of cytotoxicity genes such as *S100A8, S100A9, GNLY* and *GZMB* (**Fig.5b~d, Fig.S5c**), suggesting the more-differentiated status of this cluster in ALS patients than in healthy controls.

Collectively, the novel ALS cell cluster exhibited an unusual combinatorial transcriptome profile, confirming it as a critical inter-junction connecting the pseudotime trajectory among $CD8^+$ naïve T cells, $CD8^+$ T_{EM} cells, $CD8^+$ T_{TE} cells and monocytes, and established it as a novel stem-like cell type that was significantly enriched and more differentiated in ALS patients.

Prediction of ALS diagnosis using gene lists derived from this study

One of the major goals of this study was to seek reliable blood biomarkers of ALS to facilitate early and accurate diagnosis, thus improving patient survival. 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%)[11], the highest accuracy reported to date.

In order to generate a more accurate prediction model, we first identified three DE gene lists based on our current study. These gene lists were obtained by comparing the novel ALS cells between the ALS and Control groups (1,260 genes, Genelist 1), the novel ALS cells with $CD8^+$ T cells (977 genes, Genelist 2) or the novel ALS cells with monocytes (1,082 genes, Genelist 3) from the ALS group (**Supplementary Table 13**). These Genelists were used for initial feature selection towards more accurate prediction of ALS diagnosis against an independent cohort of microarray data for ALS patients (n=396) and healthy controls (n=396) by William *et al.*[11]. Starting from the 1,260 DE genes in Genelist 1, the top 300 genes with the highest correlation to label (ALS or Control) were selected after sorting by the ReliefF algorithm, a feature weighting algorithm that assigns different weights to genes based on the correlation between the individual gene and the label. Subsequently the appropriate gene set was selected using the RFE algorithm. The algorithm searched for a subset of the above 300 genes, built a prediction model in the training set with the given algorithm, ranked the genes by importance and discarded the least important genes. To avoid over-fitting, the Logistic Regression algorithm was used in this process. Finally, optimal parameters were selected by grid search (**Fig.6b**) and the model performance was evaluated using 10,000 random simulation trials (**Fig.6c**). A total of 792 samples (396 ALS patients and 396 Control subjects) by William *et al.*[11] were used for model

construction, of which 592 subjects were used for training (296 ALS patients and 296 Control subjects) and 200 subjects were used for testing (100 ALS patients and 100 Control subjects).

From the 1,260 DE genes in Genelist 1, the average accuracy across the 10,000 simulation trials was 93% (sensitivity: 92%, specificity: 93%) using 60 input genes for ALS diagnosis (**Fig.6a~c, Supplementary Table 14**). Selection of 110 gens from the 977 DE genes in Genelist 2 resulted in an average accuracy of 91% (sensitivity: 91%, specificity: 90%) (**Fig. S6a~b, Supplementary Table 14**). From 1,082 DE genes in Genelist 3, the average accuracy was 92% (sensitivity: 91%, specificity: 93%) with 80 genes as input (**Fig. S6c~d, Supplementary Table 14**). The final Genelists for prediction had 60 genes from Genelist 1, or 110 genes from Genelist 2, or 80 genes from Genelist 3, among which were genes in critical pathways such as actin cytoskeleton, ribosome and translation regulation, NFKB signaling and T cell regulation (**Supplementary Table 14**). Among them, there were 9 genes shared between Genelists 1 and 2, 20 genes between Genelist 2 and 3, 12 genes between Genelist 1 and 3. A total of eight genes were common to all three Genelists (**Fig.6d**).

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 eight sALS patients and three heathy donors. From a total of 40,922 cells with high-quality data, we identified 22 different cell types, of them 12 cell types belonging to T cells. Compared to the Control group, the ALS group has significantly increased abundance in $CD8^+$ and $CD4^+$ T_{TE} cells as previous reported[4-7], and decreased abundances in $CD8^+$ and $CD4^+$ naïve T cells that were not reported.

The current study discovered a novel ALS cell cluster that only sparsely exists in age-matched healthy volunteers, but significantly increased (by ~9 folds) in the ALS group composed of eight patients. Intriguingly, in following-up multiplex flow cytometry analyses of peripheral blood samples, we not only isolated $CD3^+CD16^+$ cells (accounting for ~32% of the novel ALS cell cluster) from sALS patients at about 3~29-fold enrichment over those from healthy controls, but also discovered a positive correlation of the abundance of $CD3^+CD16^+$ cells with disease progression in ALS patients.

Compared to the Control group, the T cells from the ALS patients exhibited a slightly higher level of expansion. This observation is somewhat distinct from the significant clonal expansion of CD8 T_{EMRA} cells recently reported for ALS4[13]. One possible explanation is that ALS4 patients are juvenile onset, where the age-matched controls are young people with very limited T cell clonal expansion, therefore diseased-related expansion can be more easily detected. While in the current study, the aging factor in both ALS and Control groups may mask diseased-related expansion to a certain degree, preventing a clear conclusion of ALS-related T cell expansion.

Out of the total of 14,135 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 might 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.

The novel ALS cell cluster that was increased in the ALS group by more than 9 folds exhibited an unusual combinatorial transcriptome signature. In a recent 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 adaptive cell immunotherapy against human cancer[18]. Intriguingly, the novel ALS cell cluster in this study shares gene markers such as *CCR7, LEF1, TCF7* with the stem-like progenitor $CD8^+$ T cells in type 1 diabetes[19]. It is interesting to speculate that the novel ALS cells, especially those more-developed novel ALS cells from ALS patients, may have been activated and remodeled in critical pathways such as actin cytoskeleton, ribosome and translation regulation, NFKB signaling and T cell regulation so as to perform a similar autoimmune progenitor role, one where they constantly give rise to autoimmune T_{TE} cells to attack ALS target tissues to cause motor neuron neurodegeneration.

One central goal of this study was to improve accuracy of ALS diagnosis. There have been a number of attempts in the literatures. William *et al.*[11] 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 ALS cell cluster, here we acquired the highest prediction score of ALS diagnosis to date, with an average accuracy of 93% using 60 genes. This ALS-specific expression signatures of human peripheral blood samples not only constitute the foundation for formulating a clinically relevant diagnosis panel for enhanced ALS diagnosis, but also could serve as the starting set for further study to pinpoint causative transcriptome changes that drive ALS development. Furthermore, it is intriguing that a substantially higher population of $CD3^+CD16^+$ cells was constantly detected in lymphocytes from peripheral blood samples of sALS patients by a simple multiplex flow cytometry scheme, and the ratio was positively related to the severity of ALS progression. Larger-scale clinical studies are needed to confirm this finding and further establish whether the multiplex flow cytometry scheme represents a viable new means for early and accurate ALS diagnosis.

Methods and Materials

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 in ALS1, equal volume of fresh whole blood sample from each participant was mixed together prior to PBMC isolation.

Flow cytometry and cell sorting

Multiplex flow cytometry was performed as follows. Blood sample were centrifuged at 1000xg for 5min, after removing supernatant, red blood cells were lysed with 1xRBC lysis buffer (Invitrogen eBioscience). Fluorochrome-conjugated monoclonal antibodies were used (CD3-PC5.5 clone UCHT1 and CD16-APC A750 clone 3G8 from Beckman Coulter) at 1:100 dilution and incubated at room temperature for 30 min protected from light. Dead cells were labeled with SYTOX red (Invitrogen) at 1:1000 dilution at room temperature for 20 min protected from light. Cells were analyzed using CytoFLEX S (Beckman Coulter).

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 (10x Genomics). The libraries were multiplexed and sequenced on Illumina Novaseq 6000.

Single-cell RNA-seq data analysis

scRNA-seq and scTCR-seq data for C2 and C3 were obtained from the Gene Expression Omnibus data repository under accession code GSE157007. We used the 10x Genomics Cellranger 7.0.0 pipeline[20] 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)[21] 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)[22], had mitochondrial gene percentages <7.5%, 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 24,217 genes across 40,922 cells. The top 40 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[23], and subsequently, clustering was performed with FindClusters at a resolution of 2, yielding 47 clusters. Finally, visualization was done with the DimPlot feature.

Cell type annotations were determined with the R package SingleR (v1.4.1)[24], 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 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 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 22 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 42 DE genes which were expressed in eighteen 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 ALS cell cluster

The cells in cluster labelled as novel ALS cells by SingleR were extracted and underwent a second, separate round of clustering. Altogether, 1,409 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 10 clusters (FindClusters; resolution = 0.4). Differential expression analysis (FindAllMarkers; $\log_2$FC threshold = 0.25, min.pct = 0.25) was then reperformed between these 10 clusters, identifying the top upregulated marker genes for each individual cluster. The top 20 upregulated marker genes of each cluster were visualized in a heatmap.

Pseudotemporal analysis

We used the R package Monocle (v2.18.0)[25] to perform pseudotime and branched chain analysis of the $CD8^+$ T cells, novel ALS cells, and monocytes from ALS. 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 $CD8^+$ T_{TE} cell (Branch state 2) branches and between the naïve $CD8^+$ T cell (Branch state 1) and monocyte (Branch state 3 and 4) 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 ALS cell cluster

DE genes were identified with FindMarkers (log_2FC threshold of 0.25) by testing the following expression profiles: 1) novel cells in the ALS group against novel ALS cells in the Control group (labeled as "novel ALS cells from ALS vs Ctrl"); 2) novel ALS cells against $CD8^+$ T_{EM} cells in the ALS group present on branch 1 of the pseudotime plot (labeled as "novel cells vs B1 $CD8^+$ T_{EM} from ALS"); 3) novel ALS cells against $CD8^+$ T_{EM} cells in the ALS group present in branch 2 of the pseudotime plot (labeled as "novel cells vs B2 $CD8^+$ T_{EM} from ALS"); 4) novel ALS cells against $CD8^+$ T_{TE} cells in the ALS group (labeled as "novel cells vs $CD8^+$ T_{TE} from ALS"); 5) novel ALS cells against monocytes in the ALS group (labeled as "novel cells vs monocytes from ALS"). A blanket Bonferroni-adjusted *p*-value threshold of <0.05 was applied to these five DE gene lists. Genes which passed the log_2FC and *p*-value thresholds were considered significant and passed on for subsequent GO enrichment analysis. Additionally, the top 15 upregulated genes and the top 10 downregulated genes of each of the DE gene lists 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 repertoire analysis

Raw scTCR-seq data for each sample was assembled with the 10x Genomics Cellranger V(D)J 7.0.0 pipeline, using human reference dataset GRCh38. Only TCRs with paired chains (TRA and TRB) were considered productive and preserved. This preserved data were then passed to Seurat, thus removing TCR expressing cells that did not belong to the final 40,922 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 were exported for each experimental condition and visualized for both total and unique counts.

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)[26]. The data for Control group and ALS group were analyzed separately.

Gene Ontology analysis

DE gene sets were passed to Metascape[27] 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 44 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[28].

The second DE gene set passed to Metascape for GO enrichment analysis was composed of five gene lists. GO analyses were run separately for each upregulated and downregulated gene list in each comparison, for a total of 10 separate GO runs. The 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, novel cells from ALS vs Ctrl (Genelist 1), novel cells vs $CD8^+$ T cells in the ALS group (Genelist 2) and novel cells vs monocytes in the ALS group (Genelist 3) to uncover sets of genes for ALS diagnosis. We used microarray data from blood samples of ALS patients and healthy controls (GSE112676 and GSE112680) as an independent cohort. 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*[11]. A total of 792 samples (296 ALS patients vs. 296 Control subjects) were used for model construction, of which 592 subjects were used for training (296 ALS patients vs. 296 Control subjects) and 200 subjects were used for testing (100 ALS patients vs. 100 Control 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. 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, A.Z.M. and D.Y. performed the prediction, A.Z.M., Q.W. and W.Y. wrote the manuscript.

Competing interests

The authors declares no competing interest.

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 sALS 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 the novel ALS cell cluster. a). UMAP of novel ALS cell cluster at higher resolution. **b).** Dot plot of marker genes for distinct cell subclusters. **c).** Venn diagram showing the number of cells in the novel ALS cell cluster expressing *CD3 (CD3D, CD3E, CD3G), CD8* (*CD8A, CD8B*) or *CD4* genes. **d).** Venn diagram showing the number of cells in the novel ALS cell cluster expressing at least one *S100* (*S100A8* or *S100A9), CD16* or *CD14* genes. **e).** Venn diagram showing the number of cells in the novel ALS cell cluster expressing at least one *CD3* and/or *S100* genes. **f).** Venn diagram showing the number of cells within in the novel ALS cell cluster expressing at least one *CD3* and/or *CD16*. **g).** Venn diagram showing the number of cells expressing *ICAM1* and/or *ITGB2* within the cells co-expressing *CD3* and/or *S100* genes. **h).** A representative result of multiplex flow cytometry-based sorting of live, singlet human lymphocytes expressing CD3 and/or CD16. **i).** Bar chart showing results from multiplex flow cytometry-based sorting of live, singlet human lymphocytes from four independent experiments. **j).** Pseudotime analysis of novel ALS cells together with monocytes and $CD8^+$ T cells in the ALS group colored by cell types.

Figure 3. Characterization of TCR landscapes. a). Projection of detected TCR landscape on UMAP visualization. **b).** Projection of detected TCR landscape on UMAP visualization colored according to the degrees of clonal expansion. **c).** Donut plots depicting the numbers of TCR clonotypes in each class of clonal expansion. **d).** Bar charts showing the total number of unique TCR clonotypes in the Control or ALS1 group (Left) and the number of unique TCR clonotypes that were not shared among different T cell types (right). **e).** Top 10 Specificity TCR groups from GLIPH2. **f).** TCR beta-chain V and J-gene usage in the Control group. **g).** TCR beta-chain V and J-gene usage in the ALS group. In f) and g), 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 sALS 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 17 cell types. **d).** Violin plots of expression levels of selected marker genes.

Fig.5. Transcriptome signatures of the novel ALS cell cluster. a). Bar chart showing the number of DE genes comparing the novel ALS cell cluster with other cell types. The upregulated genes are shown in red color while the downregulated genes are shown in blue color. **b).** Heatmap of the expression levels of selected upregulated genes. The top 15 most upregulated genes in each list were selected. **c).** Heatmap of the expression levels of the selected downregulated genes. The top 10 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 ALS cell cluster comparing ALS and Control groups. **c**). SVM prediction accuracy using Genelist 1. **d)**. Venn diagram of the relationship among the three gene lists used for prediction.

References

1 Beers, D. R. & Appel, S. H. Immune dysregulation in amyotrophic lateral sclerosis: mechanisms and emerging therapies. *Lancet Neurol* **18**, 211-220, doi:10.1016/S1474-4422(18)30394-6 (2019).

2 Beers, D. R. *et al.* ALS patients' regulatory T lymphocytes are dysfunctional, and correlate with disease progression rate and severity. *JCI Insight* **2**, e89530, doi:10.1172/jci.insight.89530 (2017).

3 Sheean, R. K. *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* **75**, 681-689, doi:10.1001/jamaneurol.2018.0035 (2018).

4 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* **10**, 5941, doi:10.1038/s41598-020-62756-8 (2020).

5 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* **15**, 657465, doi:10.3389/fnins.2021.657465 (2021).

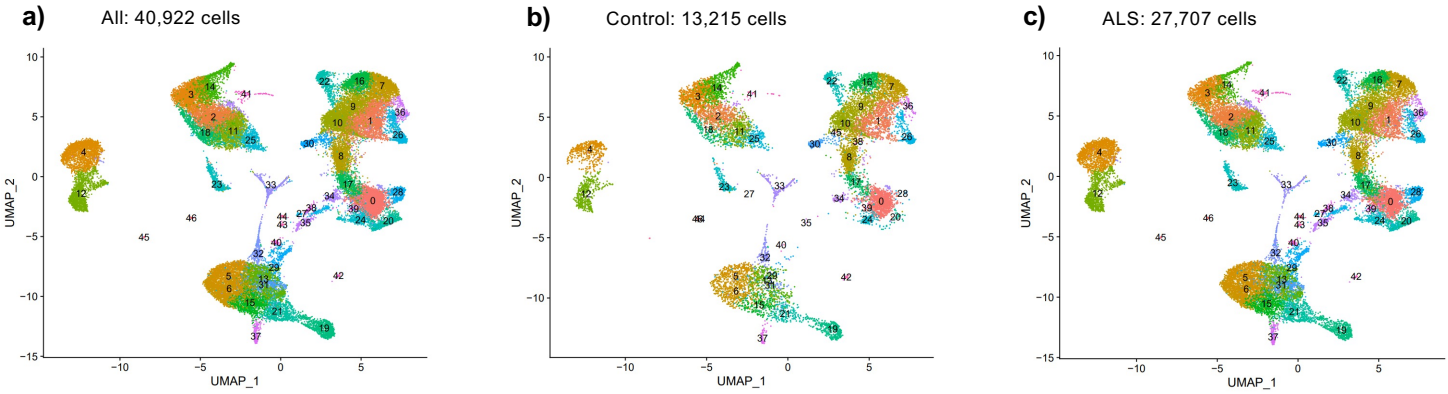
6 Rolfes, L. *et al.* Amyotrophic lateral sclerosis patients show increased peripheral and intrathecal T-cell activation. *Brain Commun* **3**, fcab157, doi:10.1093/braincomms/fcab157 (2021).

7 Gustafson, M. P. *et al.* Comprehensive immune profiling reveals substantial immune system alterations in a subset of patients with amyotrophic lateral sclerosis. *PLoS One* **12**, e0182002, doi:10.1371/journal.pone.0182002 (2017).

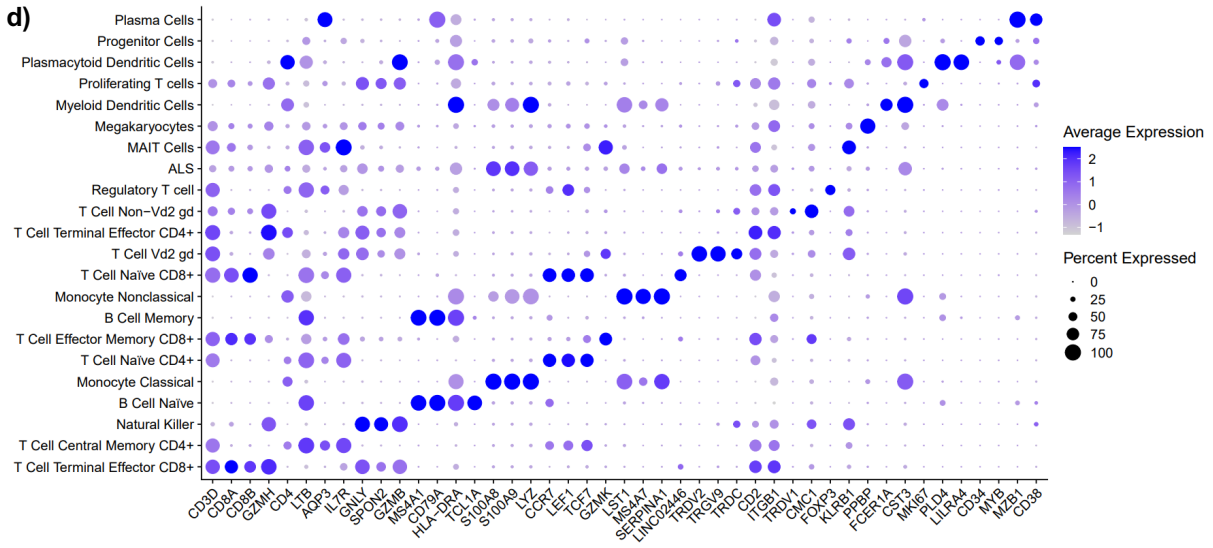
8 Yang, Y. *et al.* Decreased blood CD4+ T lymphocyte helps predict cognitive impairment in patients with amyotrophic lateral sclerosis. *BMC Neurol* **21**, 157, doi:10.1186/s12883-021-02185-w (2021).

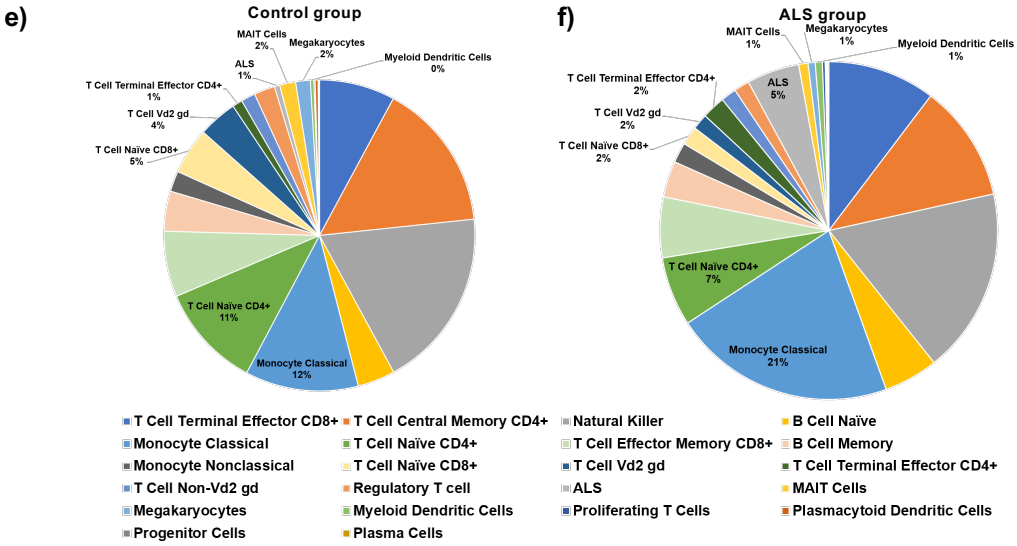
9 Katzeff, J. S. *et al.* Altered serum protein levels in frontotemporal dementia and amyotrophic lateral sclerosis indicate calcium and immunity dysregulation. *Sci Rep* **10**, 13741, doi:10.1038/s41598-020-70687-7 (2020).
10 Gagliardi, S. *et al.* Long non-coding and coding RNAs characterization in Peripheral Blood Mononuclear Cells and Spinal Cord from Amyotrophic Lateral Sclerosis patients. *Sci Rep* **8**, 2378, doi:10.1038/s41598-018-20679-5 (2018).
11 Swindell, W. R., Kruse, C. P. S., List, E. O., Berryman, D. E. & Kopchick, J. J. ALS blood expression profiling identifies new biomarkers, patient subgroups, and evidence for neutrophilia and hypoxia. *J Transl Med* **17**, 170, doi:10.1186/s12967-019-1909-0 (2019).
12 Liu, W. *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* **141**, 104877, doi:10.1016/j.nbd.2020.104877 (2020).
13 Campisi, L. *et al.* Clonally expanded CD8 T cells characterize amyotrophic lateral sclerosis-4. *Nature* **606**, 945-952, doi:10.1038/s41586-022-04844-5 (2022).
14 Luo, O. J., Lei, W., Zhu, G. & al., e. Multidimensional single-cell analysis of human peripheral blood reveals characteristic features of the immune system landscape in aging and frailty. *Nat Aging* **2**, 348-364 (2022).
15 Burel, J. G. *et al.* Circulating T cell-monocyte complexes are markers of immune perturbations. *Elife* **8**, doi:10.7554/eLife.46045 (2019).
16 Huang, H., Wang, C., Rubelt, F., Scriba, T. J. & Davis, M. M. Analyzing the Mycobacterium tuberculosis immune response by T-cell receptor clustering with GLIPH2 and genome-wide antigen screening. *Nat Biotechnol* **38**, 1194-1202, doi:10.1038/s41587-020-0505-4 (2020).
17 Santos, C. X. *et al.* Targeted redox inhibition of protein phosphatase 1 by Nox4 regulates eIF2alpha-mediated stress signaling. *EMBO J* **35**, 319-334, doi:10.15252/embj.201592394 (2016).
18 Krishna, S. *et al.* Stem-like CD8 T cells mediate response of adoptive cell immunotherapy against human cancer. *Science* **370**, 1328-1334, doi:10.1126/science.abb9847 (2020).
19 Gearty, S. V. *et al.* An autoimmune stem-like CD8 T cell population drives type 1 diabetes. *Nature* **602**, 156-161, doi:10.1038/s41586-021-04248-x (2022).
20 Zheng, G. X. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat Commun* **8**, 14049, doi:10.1038/ncomms14049 (2017).
21 Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587 e3529, doi:10.1016/j.cell.2021.04.048 (2021).

22 McGinnis, C. S., Murrow, L. M. & Gartner, Z. J. DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors. *Cell Syst* **8**, 329-337 e324, doi:10.1016/j.cels.2019.03.003 (2019).
23 Zappia, L. & Oshlack, A. Clustering trees: a visualization for evaluating clusterings at multiple resolutions. *Gigascience* **7**, doi:10.1093/gigascience/giy083 (2018).
24 Aran, D. *et al.* Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat Immunol* **20**, 163-172, doi:10.1038/s41590-018-0276-y (2019).
25 Qiu, X. *et al.* Single-cell mRNA quantification and differential analysis with Census. *Nat Methods* **14**, 309-315, doi:10.1038/nmeth.4150 (2017).
26 Shugay, M. *et al.* VDJtools: Unifying Post-analysis of T Cell Receptor Repertoires. *PLoS Comput Biol* **11**, e1004503, doi:10.1371/journal.pcbi.1004503 (2015).
27 Zhou, Y. *et al.* Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat Commun* **10**, 1523, doi:10.1038/s41467-019-09234-6 (2019).
28 Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*. (Springer-Verlag New York, 2016).



0,20,28,34,39: T Cell Terminal Effector CD8+
1,9,10: T Cell Central Memory CD4+
2,3,11,14,18: Natural Killer
4: B Cell Naïve
5,6,13,15,21,31,32: Monocyte Classical
7,16,36: T Cell Naïve CD4+
8,17: T Cell Effector Memory CD8+
12: B Cell Memory
19: Monocyte Nonclassical
22: T Cell Naïve CD8+
23: T Cell Vd2 gd
24: T Cell Terminal Effector CD4+
25: T Cell Non-Vd2 gd
26: Regulatory T cell
27,29,35,38,40,43,44: ALS
30: MAIT Cells
33: Megakaryocytes
37: Myeloid Dendritic Cells
41: Proliferating T Cells
42: Plasmacytoid Dendritic Cells
45: Progenitor Cells
46: Plasma Cells





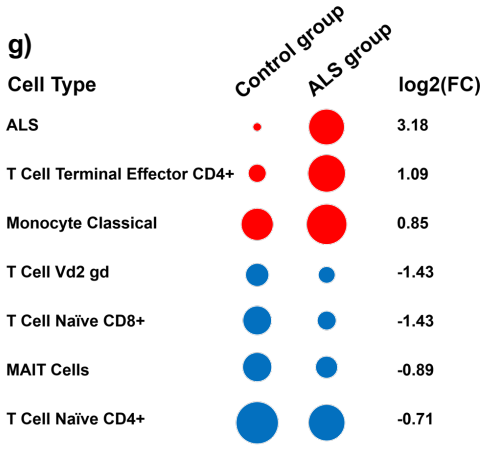


Figure 1

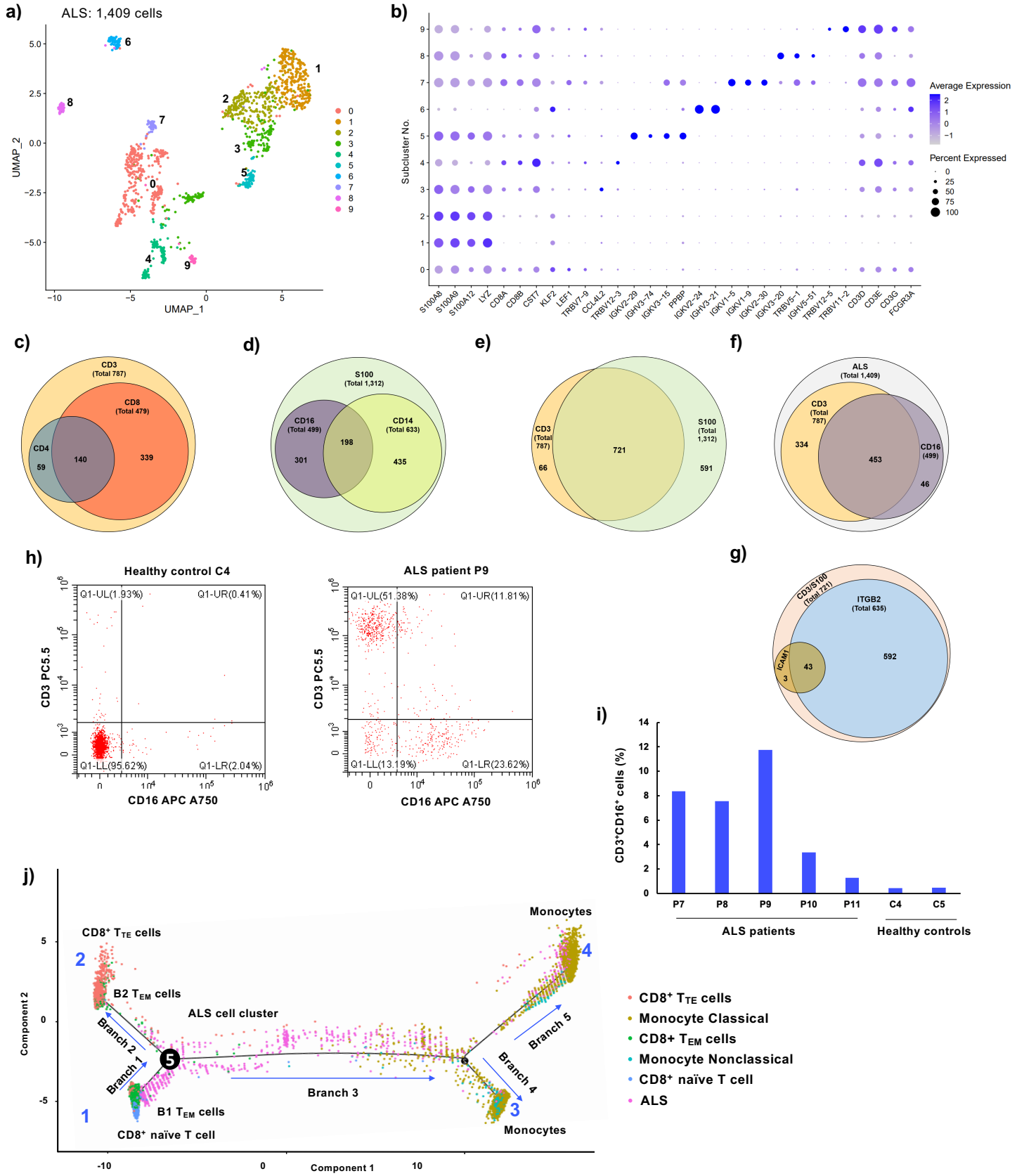
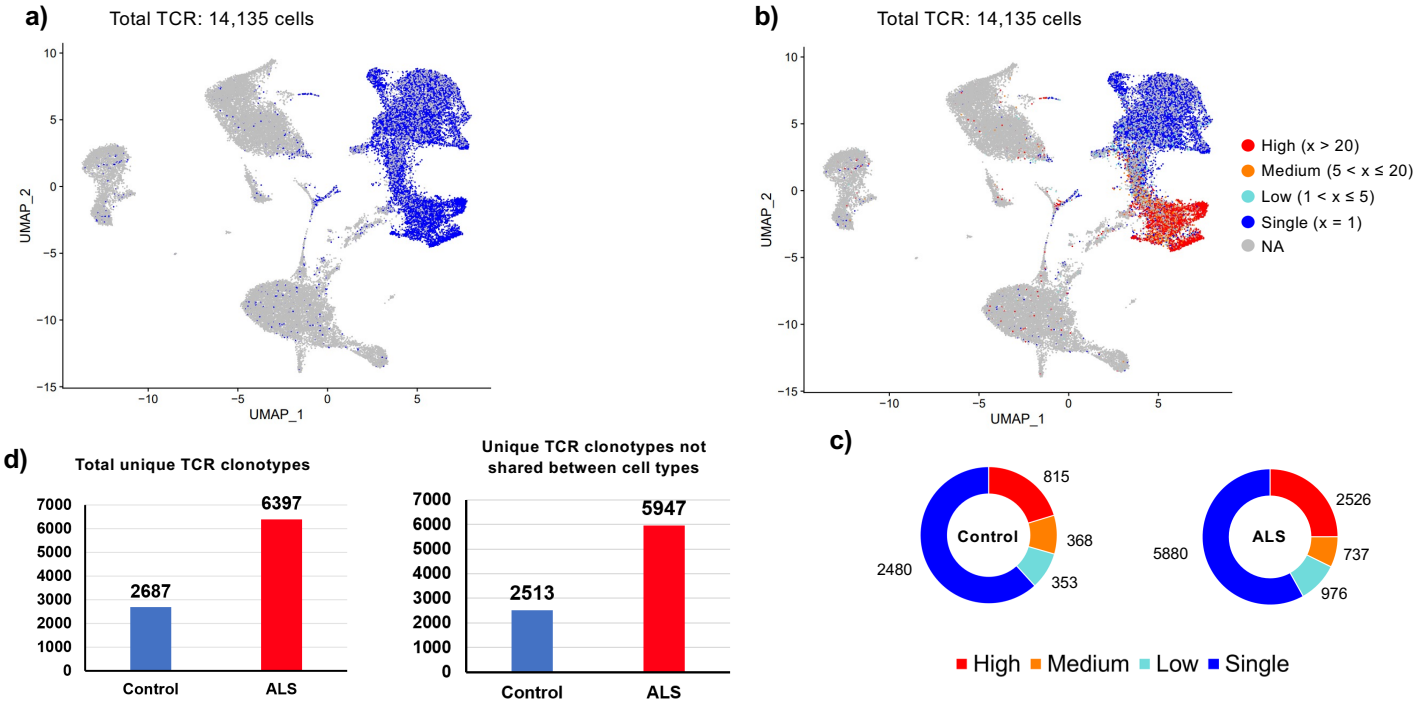

a)
ALS: 1,409 cells
UMAP_2
UMAP_1
b)
Subcluster No.
Average Expression
Percent Expressed
S100A8
S100A9
S100A12
LYZ
CD8A
CD8B
CST7
KLF2
LEF1
TRBV7-9
CCL4L2
TRBV12-3
IGKV2-29
IGHV3-74
IGKV3-15
PPBP
IGKV2-24
IGHV3-21
IGKV1-5
IGKV1-9
IGKV2-30
IGKV3-20
TRBV5-1
IGHV5-51
TRBV12-5
TRBV11-2
CD3D
CD3E
CD3G
FCGR3A
c)
CD3 (Total 787)
CD8 (Total 479)
CD4
59
140
339
d)
S100 (Total 1,312)
CD16 (Total 499)
CD14 (Total 633)
198
301
435
e)
CD3 (Total 787)
S100 (Total 1,312)
66
721
591
f)
ALS (Total 1,409)
CD3 (Total 787)
CD16 (499)
334
453
46
h)
Healthy control C4
Q1-UL(1.93%)
Q1-UR(0.41%)
Q1-LL(95.62%)
Q1-LR(2.04%)
ALS patient P9
Q1-UL(51.38%)
Q1-UR(11.81%)
Q1-LL(13.19%)
Q1-LR(23.62%)
CD3 PC5.5
CD16 APC A750
g)
CD3/S100 (Total 727)
ITGB2 (Total 635)
ICAM1
3
43
592
i)
CD3+CD16+ cells (%)
P7
P8
P9
P10
P11
C4
C5
ALS patients
Healthy controls
j)
CD8+ TTE cells
B2 TEM cells
ALS cell cluster
Branch 1
Branch 2
Branch 3
Branch 4
Branch 5
B1 TEM cells
CD8+ naïve T cell
Monocytes
Component 1
Component 2
CD8+ TTE cells
Monocyte Classical
CD8+ TEM cells
Monocyte Nonclassical
CD8+ naïve T cell
ALS

Figure 2



e)

	GLIPH ID	Donor Group	TCRb sequence	Frequency	V	J	TCRa sequency
			CASKISGTSGYNEQFF	385/421	TRBV27	TRBJ2-1	CALSFAGGTSYGKLTF
Group 1	114	ALS group	CASKASGTSGYNEQFF	36/421	TRBV27	TRBJ2-1	CALCYAGGTSYGKLTF
Group 2	1283	ALS group	CASSLKAYEQYF	298	TRBV7-9	TRBJ2-7	CATDAYGEETSGSRLTF
Group 3	4	Control group	CASSQVRRGAVAEQFF	159	TRBV4-1	TRBJ2-1	CALNGNTDKLIF
			CASSPPSGSYEQYF	38/143	TRBV7-9	TRBJ2-7	CAVDNFNKFYF
Group 4	237	ALS group	CASSPPTGSYEQYF	105/143	TRBV7-9	TRBJ2-7	CAVDNFNKFYF
Group 5	2564	ALS group	CASSQETGLASGELFF	115	TRBV4-2	TRBJ2-2	CAMRAASDSGSARQLTF
Group 6	5413	ALS group	CASSKLGNIQYF	89	TRBV27	TRBJ2-4	CAETYTNAGKSTF
Group 7	1133	ALS group	CASSVSHYGEKFF	84	TRBV6-1	TRBJ2-1	CAASVTGNQFYF
Group 8	321	ALS group	CASSPGTVNTEAFF	76	TRBV12-3	TRBJ1-1	CLVASDNAGNNRKLIW
Group 9	2176	ALS group	CASTQDRNYEQYF	71	TRBV11-2	TRBJ2-7	CAAAPNFGNEKLTF
Group 10	4878	ALS group	CASSLGPPMTEAFF	71	TRBV19	TRBJ1-1	CAANGRDQAGTALIF

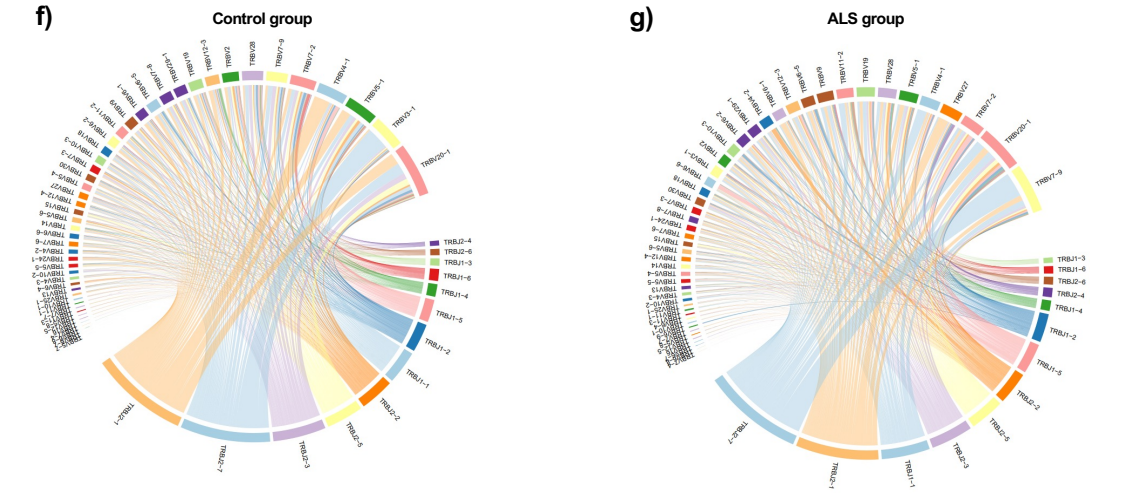


Figure 3

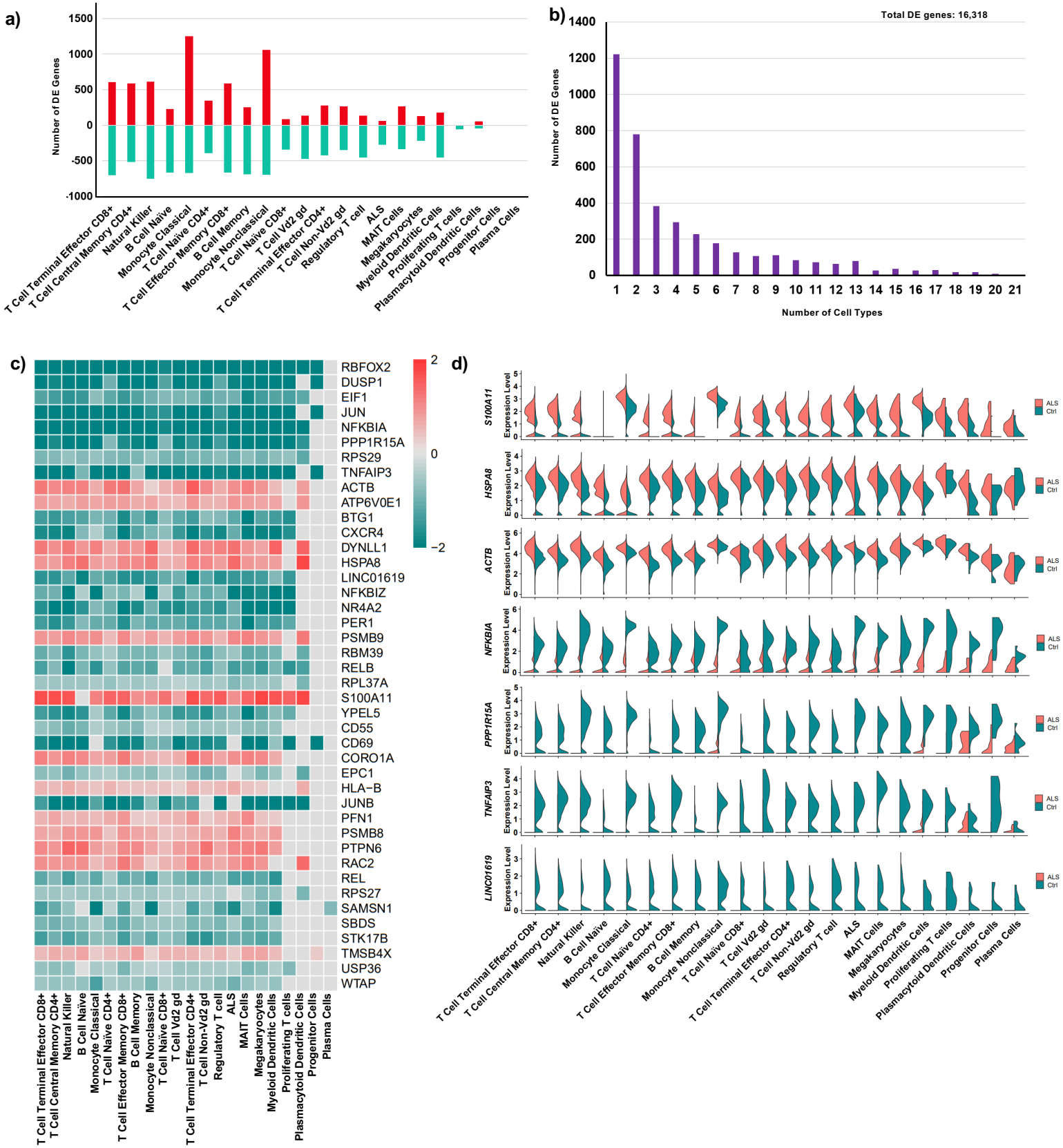

a)
Number of DE Genes
1500
1000
500
0
-500
-1000
T Cell Terminal Effector CD8+
T Cell Central Memory CD4+
Natural Killer
B Cell Naïve
Monocyte Classical
T Cell Naïve CD4+
T Cell Effector Memory CD8+
B Cell Memory
Monocyte Nonclassical
T Cell Naïve CD8+
T Cell Vd2 gd
T Cell Terminal Effector CD4+
T Cell Non-Vd2 gd
Regulatory T cell
ALS
MAIT Cells
Megakaryocytes
Myeloid Dendritic Cells
Proliferating T cells
Plasmacytoid Dendritic Cells
Progenitor Cells
Plasma Cells
b)
Total DE genes: 16,318
Number of DE Genes
1400
1200
1000
800
600
400
200
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21
Number of Cell Types
c)
RBFOX2
DUSP1
EIF1
JUN
NFKBIA
PPP1R15A
RPS29
TNFAIP3
ACTB
ATP6V0E1
BTG1
CXCR4
DYNLL1
HSPA8
LINC01619
NFKBIZ
NR4A2
PER1
PSMB9
RBM39
RELB
RPL37A
S100A11
YPEL5
CD55
CD69
CORO1A
EPC1
HLA−B
JUNB
PFN1
PSMB8
PTPN6
RAC2
REL
RPS27
SAMSN1
SBDS
STK17B
TMSB4X
USP36
WTAP
2
0
−2
d)
S100A11
HSPA8
ACTB
NFKBIA
PPP1R15A
TNFAIP3
LINC01619
Expression Level
ALS
Ctrl

Figure 4

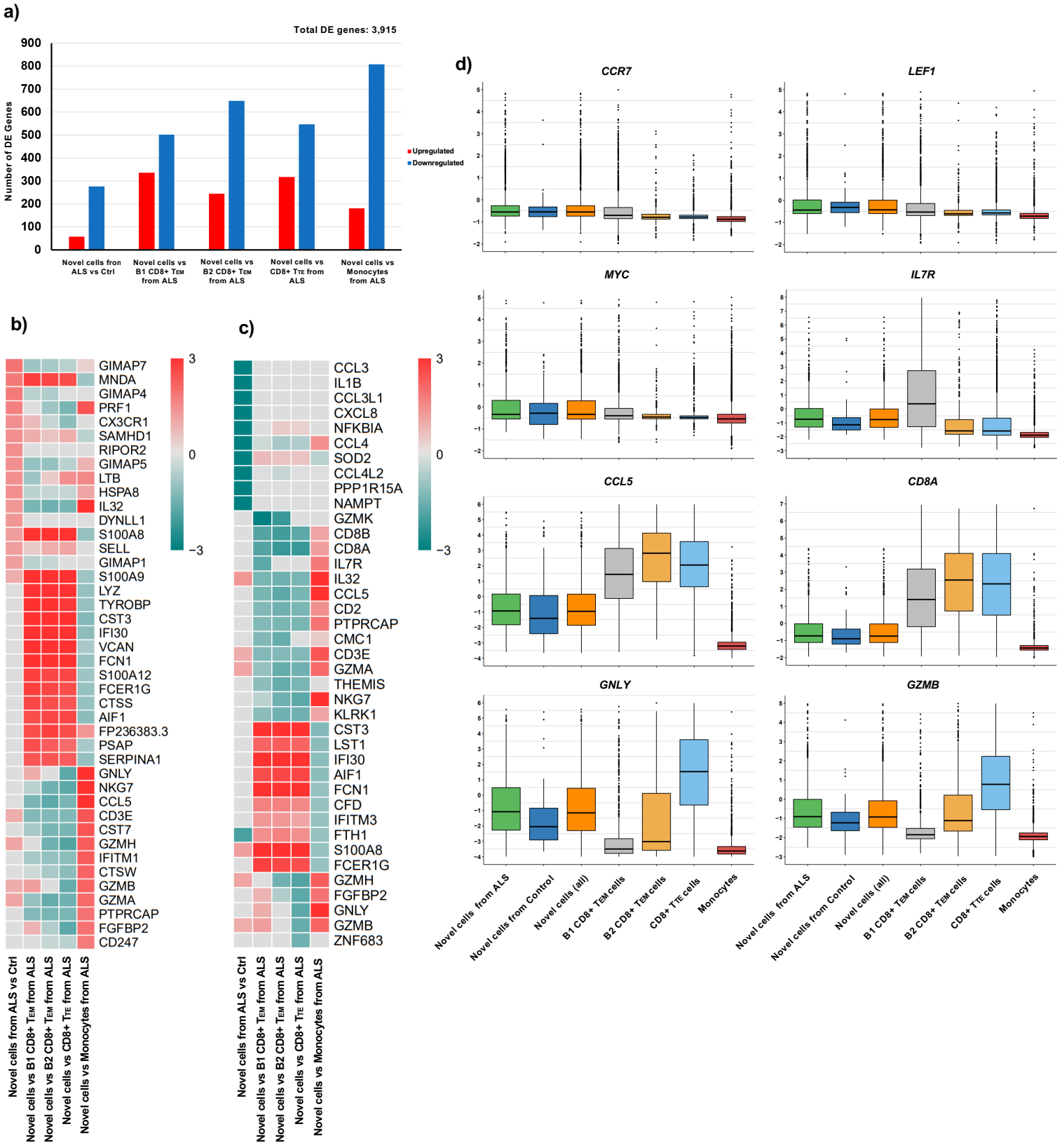

a)
Total DE genes: 3,915
Number of DE Genes
Upregulated
Downregulated
Novel cells from ALS vs Ctrl
Novel cells vs B1 CD8+ TEM from ALS
Novel cells vs B2 CD8+ TEM from ALS
Novel cells vs CD8+ TTE from ALS
Novel cells vs Monocytes from ALS
b)
GIMAP7
MNDA
GIMAP4
PRF1
CX3CR1
SAMHD1
RIPOR2
GIMAP5
LTB
HSPA8
IL32
DYNLL1
S100A8
SELL
GIMAP1
S100A9
LYZ
TYROBP
CST3
IFI30
VCAN
FCN1
S100A12
FCER1G
CTSS
AIF1
FP236383.3
PSAP
SERPINA1
GNLY
NKG7
CCL5
CD3E
CST7
GZMH
IFITM1
CTSW
GZMB
GZMA
PTPRCAP
FGFBP2
CD247
c)
CCL3
IL1B
CCL3L1
CXCL8
NFKBIA
CCL4
SOD2
CCL4L2
PPP1R15A
NAMPT
GZMK
CD8B
CD8A
IL7R
IL32
CCL5
CD2
PTPRCAP
CMC1
CD3E
GZMA
THEMIS
NKG7
KLRK1
CST3
LST1
IFI30
AIF1
FCN1
CFD
IFITM3
FTH1
S100A8
FCER1G
GZMH
FGFBP2
GNLY
GZMB
ZNF683
Novel cells from ALS vs Ctrl
Novel cells vs B1 CD8+ TEM from ALS
Novel cells vs B2 CD8+ TEM from ALS
Novel cells vs CD8+ TTE from ALS
Novel cells vs Monocytes from ALS
d)
CCR7
LEF1
MYC
IL7R
CCL5
CD8A
GNLY
GZMB
Novel cells from ALS
Novel cells from Control
Novel cells (all)
B1 CD8+ TEM cells
B2 CD8+ TEM cells
CD8+ TTE cells
Monocytes

Figure 5

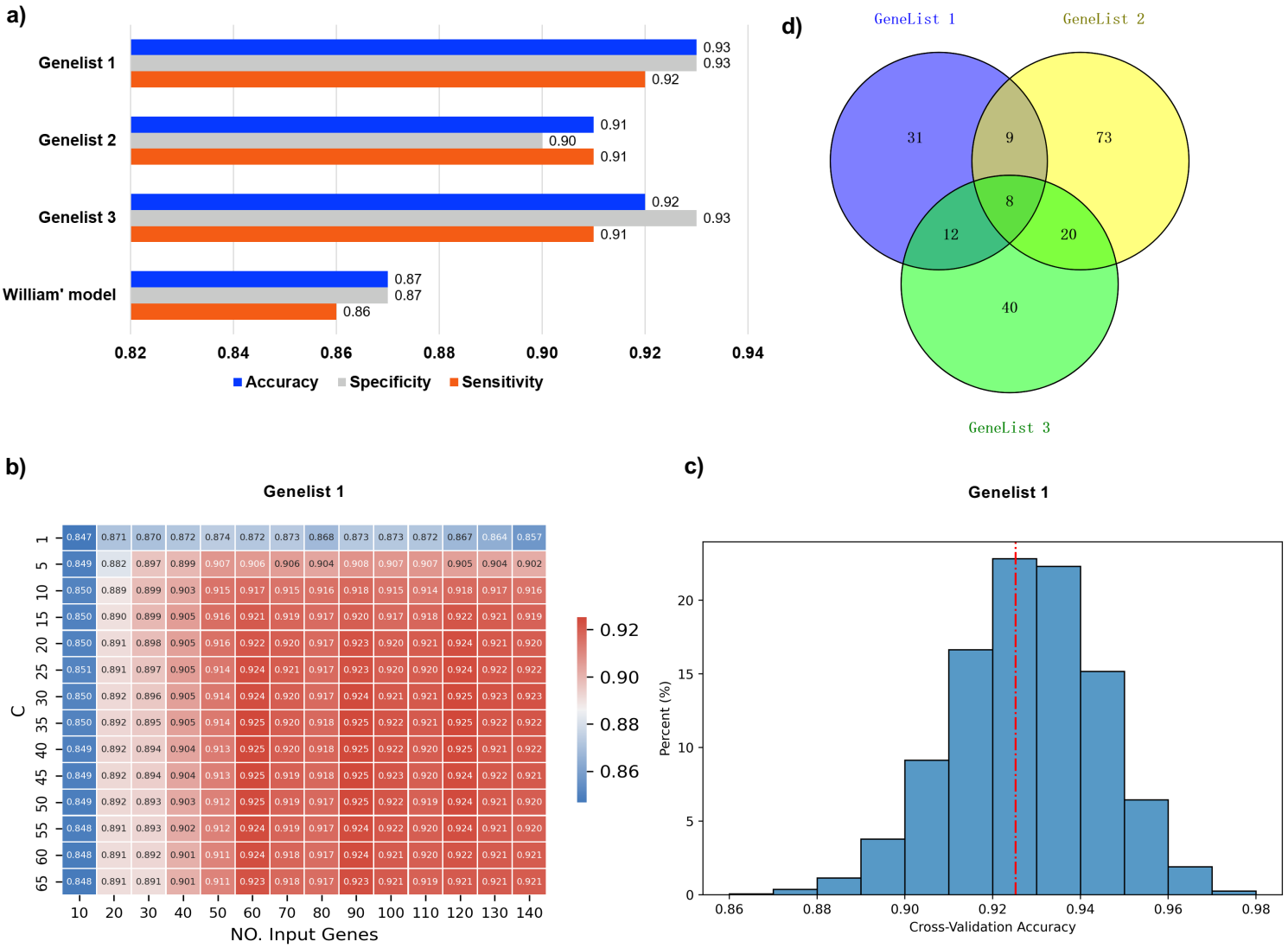

a)
Genelist 1
Genelist 2
Genelist 3
William' model
0.93
0.93
0.92
0.91
0.90
0.91
0.92
0.93
0.91
0.87
0.87
0.86
0.82
0.84
0.86
0.88
0.90
0.92
0.94
Accuracy
Specificity
Sensitivity
d)
GeneList 1
GeneList 2
GeneList 3
31
9
73
8
12
20
40
b)
Genelist 1
C
NO. Input Genes
c)
Genelist 1
Percent (%)
Cross-Validation Accuracy

Figure 6