nature neuroscience

Resource

https://doi.org/10.1038/s41593-024-01764-7

A cross-disease resource of living human microglia identifies disease-enriched subsets and tool compounds recapitulating microglial states

Received: 14 December 2023

Accepted: 13 August 2024

Published online: 15 October 2024

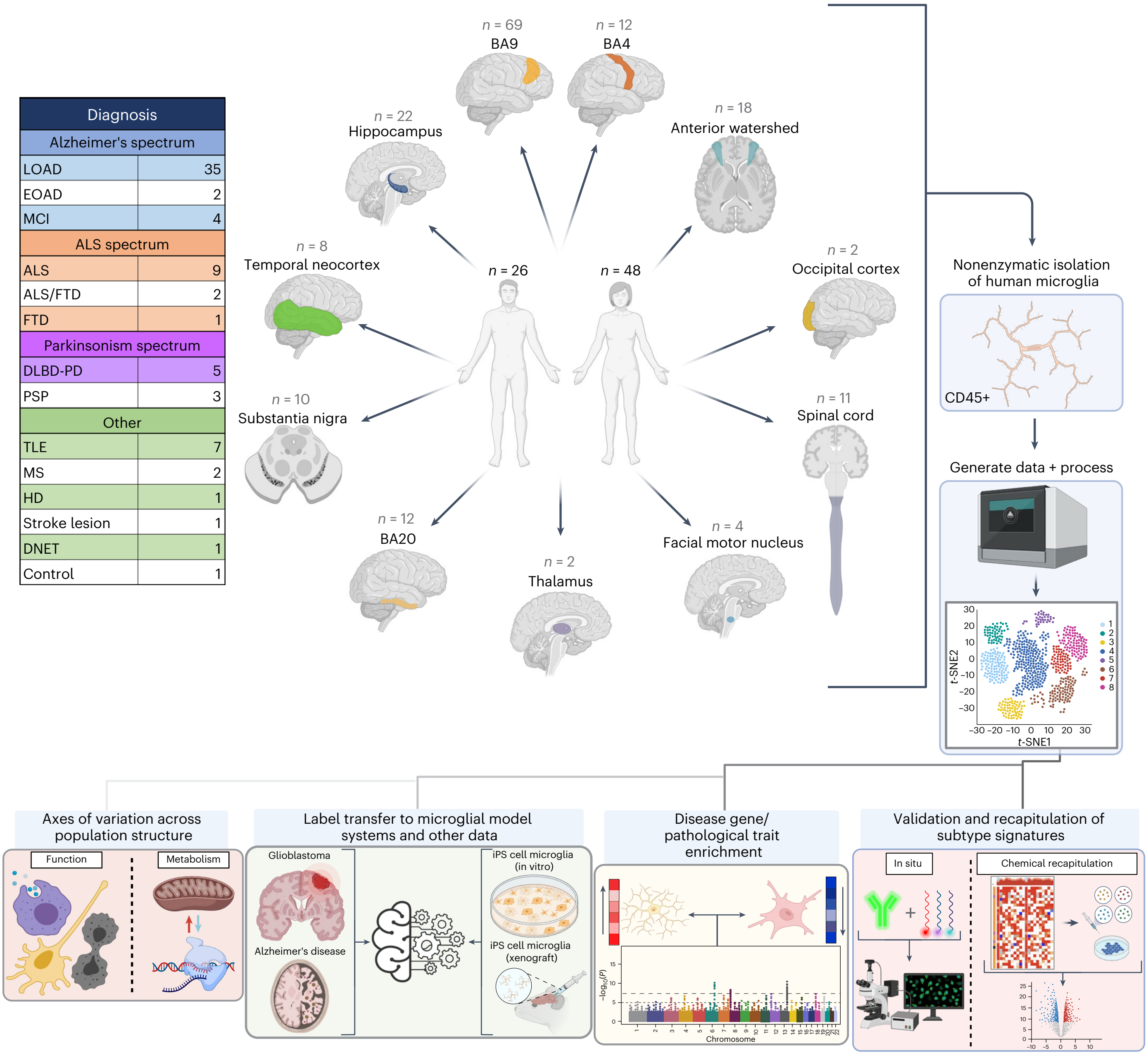
Check for updates

John F. Tuddenham[1,2,3,20], Mariko Taga[1,4,20], Verena Haage[1,20], Victoria S. Marshe[1], Tina Roostaei[1], Charles White[1], Annie J. Lee[1], Masashi Fujita[1], Anthony Khairallah[1], Ya Zhang[1], Gilad Green[5], Bradley Hyman[6], Matthew Frosch[7], Sarah Hopp[8,9], Thomas G. Beach[10], Geidy E. Serrano[10], John Corboy[11], Naomi Habib[5], Hans-Ulrich Klein[1,4], Rajesh Kumar Soni[12], Andrew F. Teich[4,13,14], Richard A. Hickman[15], Roy N. Alcalay[14,16], Neil Shneider[14,17], Julie Schneider[18], Peter A. Sims[2,19], David A. Bennett[18], Marta Olah[1,20], Vilas Menon[1,20] & Philip L. De Jager[1,20] ✉

Human microglia play a pivotal role in neurological diseases, but we still have an incomplete understanding of microglial heterogeneity, which limits the development of targeted therapies directly modulating their state or function. Here, we use single-cell RNA sequencing to profile 215,680 live human microglia from 74 donors across diverse neurological diseases and CNS regions. We observe a central divide between oxidative and heterocyclic metabolism and identify microglial subsets associated with antigen presentation, motility and proliferation. Specific subsets are enriched in susceptibility genes for neurodegenerative diseases or the disease-associated microglial signature. We validate subtypes in situ with an RNAscope–immunofluorescence pipeline and high-dimensional MERFISH. We also leverage our dataset as a classification resource, finding that induced pluripotent stem cell model systems capture substantial in vivo heterogeneity. Finally, we identify and validate compounds that recapitulate certain subtypes in vitro, including camptothecin, which downregulates the signature of disease-enriched subtypes and upregulates a signature previously associated with Alzheimer's disease.

Microglia, the resident parenchymal myeloid population of the CNS[1], can rapidly disengage from key homeostatic functions to fulfill different specialized roles, such as antigen presentation, pathogen response and synaptic pruning[2,3]. They play pivotal roles in CNS development[2] and diseases including Alzheimer's disease (AD)[4] and multiple sclerosis (MS)[5]. We are only beginning to understand their spatial, temporal and functional complexity, particularly in humans, as much of the published profiling work has been performed in mice[6,7]. Structured evaluations of human microglial heterogeneity at the single-cell level have only recently been applied to a limited set of contexts[4,8–15]. Further, most of these studies analyzed only a modest number of samples or used single-nucleus profiling, which may have differential sensitivity to capture genes when compared to single-cell approaches, especially in microglia[16,17]. As a result, our understanding of the range of states that live human microglia can attain, as well as their trajectories of state transition, remains limited. Analyzing data captured in different

A full list of affiliations appears at the end of the paper. ✉e-mail: pld2115@cumc.columbia.edu



Diagnosis	
Alzheimer's spectrum	
LOAD	35
EOAD	2
MCI	4
ALS spectrum	
ALS	9
ALS/FTD	2
FTD	1
Parkinsonism spectrum	
DLBD-PD	5
PSP	3
Other	
TLE	7
MS	2
HD	1
Stroke lesion	1
DNET	1
Control	1

Fig. 1 | Overview of our cross-disease sample collection, data generation approach, downstream analyses and validation. We sampled a wide array of neurological diseases and CNS regions (Supplementary Table 1) from a mix of autopsy samples and surgical resections. We isolated live brain CD45$^+$ cells from a total of 74 donors of both sexes. Single-cell suspensions were loaded directly onto the 10x Chromium controller. Resulting libraries were sequenced on an Illumina HiSeq 4000. The lower part of the figure outlines our analyses and validation efforts, including disease and functional relevance of microglial subtypes, in situ validation, in vitro recapitulation of subtype phenotypes, and annotation of other datasets using our data as a reference.

contexts in a single framework is essential to interpret results across diseases and studies.

In this study, we aimed to (1) generate a broad reference of microglial transcriptional profiles across neurodegenerative diseases that would capture as much of the diversity of microglial states as possible and (2) illustrate the utility of this resource to annotate model systems and identify tool compounds for modulating human microglial states. The study is not designed to identify microglial populations associated with a given disease; that type of effort would require a different study design involving a single brain region and profiling of only one set of participants with a diagnosis and one set of reference participants without that diagnosis. Using cold, enzyme-free, mechanical dissociation, which has been demonstrated to optimally preserve native microglial transcriptional profiles[18,19], we purified live CD45$^+$ cells and collected single-cell RNA sequencing (scRNA-seq) data from a diverse set of CNS regions and clinicopathologic states affecting both men and women. We identified 12 microglial subpopulations represented across all diseases and regions; importantly, our study was not designed to characterize microglia in a particular disease, but rather to sample as many different conditions as possible and to profile them using a single experimental and analytic pipeline. We propose trajectories of cell-state transitions between microglial subsets in our dataset, identifying a central metabolic shift between oxidative and heterocyclic metabolism, microglial subsets enriched for disease genes, microglial subsets that express high levels of the disease-associated microglial (DAM) transcriptional program[6] and subsets associated with immune activation. Given the plasticity of microglia, we suspect that they can differentiate into these populations in adults, after the cell is committed to a microglial fate, and that they can switch signatures depending on the changes occurring in their microenvironment. Using our new subset signatures, we optimized a joint protein–RNA staining protocol to localize microglial subsets in situ and demonstrate morphological shifts associated with expression of selected hallmark genes; in parallel, we also used a multiplexed MERFISH approach to independently validate our subsets in brain tissue sections using a larger number of genes. We also used our new resource to classify microglia profiled in previous studies and evaluated the degree of microglial diversity found in induced pluripotent stem (iPS) cell-derived microglial model systems. Finally, we leveraged the Connectivity Map (CMAP)[20,21] to identify chemical perturbations predicted to drive subtype-specific signatures and cell-state transitions, and we validated these predictions in vitro at the RNA and protein levels. Ultimately, we provide a resource that explores human microglial heterogeneity across regions and diseases and a series of tools for classifying, evaluating and manipulating microglial model systems, bringing us closer to the goal of microglial modulation in humans.

Overview of our samples and analytical approach

Our sample collection encompasses fresh autopsy samples from individuals with both early-onset and late-onset AD, mild cognitive impairment (MCI), amyotrophic lateral sclerosis (ALS), frontotemporal dementia (FTD), Parkinson's disease (PD), progressive supranuclear palsy (PSP), diffuse Lewy body disease (DLBD), MS, Huntington's disease (HD) and stroke, as well as samples from an individual without a diagnosis of neurological disease (Fig. 1). Our cohort also includes surgical resections from individuals with temporal lobe epilepsy and a dysembryoplastic neuroepithelial tumor. These samples were derived from a wide array of brain regions: anterior watershed white matter, frontal cortex (BA9/46), primary motor cortex (BA4), temporal cortex (BA20/21), occipital cortex (BA17/18/19), hippocampus, thalamus, substantia nigra, facial motor nucleus and spinal cord. As our workflow limited the number of regions that could be processed in parallel for each brain, we chose BA9 as a reference in most cases and sampled other regions where possible. As individuals without any diagnosed pathology ('controls') rarely come to autopsy, only one was available for sampling during the study. Our workflow included the use of a previously reported cold, enzyme-free mechanical dissociation approach for isolation of live human microglia and leukocytes[4,22] followed by scRNA-seq of the freshly sorted live cells using the droplet-based 10x Genomics Chromium platform. Further details on the demographic and clinical characteristics of our donors, as well as details regarding our cell hashing strategy, can be found in Supplementary Table 1.

After rigorous preprocessing and quality control (QC), we retained 225,382 individual transcriptomes from 74 donors. As the samples in our dataset encompassed a broad set of disease conditions, brain regions and chemistry versions for the 10x Genomics Chromium platform, we applied algorithms for batch correction and normalization with the goal of identifying microglial states that are conserved across conditions. To separate microglia from non-microglial populations, we used the Seurat package in R[23] to perform Louvain clustering, choosing a resolution where distinct cell types could be identified by canonical markers. In this initial clustering step (Extended Data Fig. 1), small numbers (<5%) of adaptive immune cells, monocytes, erythrocytes and other nonimmune populations segregated from our microglia and were not further analyzed.

Microglial subpopulations and signature genes

After isolating the myeloid cells in silico, we then subclustered them to generate a shared reference model across all regions and diseases. After selecting a model where all pairs of clusters had less than 20% of ambiguous assignment of cells using multilayer perceptron classification (Methods) and retaining clusters with >100 cells, we arrived at a population structure consisting of 12 distinct clusters (Fig. 2a). The mean number of unique molecular identifiers (UMIs) and genes detected in microglia was similar across batches, technologies and clusters (Extended Data Fig. 2a–f and Supplementary Table 1), and post hoc computational cluster validation supported the stability of this cluster structure (Extended Data Fig. 2g). We first confirmed the microglial identity of our clusters by evaluating a set of core microglial genes[5,24–26] as well as monocyte (*S100A8*, *VCAN*) and macrophage (*SELL*, *EMILIN2* and *GDA*) genes (Fig. 2b). Notably, all 12 clusters expressed *AIF1* and *C1QA*, well-validated markers of microglial identity in the brain; however, some microglia-specific murine marker genes, such as *HEXB*[27], are expressed at low levels or are inconsistent in our human data. We also examined proposed markers (*LYVE1*, *MS4A7* and *CD163*) of the border-associated macrophage (BAM) subset recently reported in mice[28,29], and no cluster appears to be predominantly composed of BAM-like cells, although it is impossible to rule out the possibility that BAM-lineage cells may have entered the CNS microenvironment and downregulated lineage-defining genes during the infiltration.

Next, we performed pairwise differential expression analyses to define the genes that best differentiated our microglial subgroups from each other (Methods and Supplementary Table 2). Representative distinguishing genes are shown in Fig. 2c. The clusters are numbered in descending order based on their size, with cluster 1 having the largest number of cells and cluster 12 the least. Clusters 1, 2 and 3 are the most abundant clusters in most individuals (Extended Data Fig. 3). Genes upregulated in cluster 1 (the largest cluster) include disease genes such as *ITPR2* and *SORL1*, as well as transcription factors and RNA-binding proteins, such as those encoded by *MEF2A*, *RUNX1* and *CELF1*. Cluster 6 is, transcriptionally, the closest to cluster 1, expressing high levels of *SRGAP2* and *QKI*, which encodes an RNA-binding protein that regulates microglial phagocytosis in the context of demyelination[30,31]. In contrast, clusters 4 and 9, which are transcriptionally adjacent to cluster 2, have an overlapping set of enriched genes, including *C1QA*, *TYROBP*, *ITM2B*, *GPX1* and *FCER1G*. Thus, the broadest division in microglial subtypes appears to be between clusters 2, 4 and 9 (represented on the left side of our low-dimensional embedding for visualization) and 1, 5, 6 and 7 (on the right side of the same embedding; Fig. 2a). The marker expression profile of cluster 3 suggests that it is intermediate between clusters 1 and 2, and clusters 2 and 3 are more enriched in genes associated with classical homeostatic-active states[5] (*CX3CR1*, *FCGR1A* and *P2RY12*). This suggests that clusters 2 and 3 may be closest to the classic description of 'homeostatic' microglia, while cluster 1 and its closely related family are a divergent branch of microglial differentiation. Notably, clusters 2 and 3 have relatively few differentially upregulated genes compared to all other microglial clusters. Interestingly, cluster 5 appears to be an alternative intermediate state between clusters 1 and 2, as it expresses *CX3CR1* alongside *QKI* and *MEF2A*.

Clusters 8 and 10 are located between these broad families but are more homologous to 2, 4 and 9. Cluster 8 is enriched in *CXCR4* and *SRGN*, while cluster 10 is enriched in *HLA-C*, *CD74* (encoding a protein that plays an important role in antigen presentation) and *CYBA*. Cluster 11 also shares some transcriptomic homology with 2, 4 and 9 but is distinguished by enrichment in *SPP1* and *LGALS1*. Finally, cluster 12 expresses *MKI67* and *PCNA*, suggesting a proliferative phenotype.

The DAM state[6] has been clearly defined in mouse models, but results in human studies have been mixed[4,13,32,33]. The lack of clarity around the possible presence and/or role of DAM genes in humans may stem from technical differences between studies and the relatively small numbers of microglia profiled in studies to date. In addition, the proposed transition from homeostatic to DAM1, an initial TREM2-independent state, then to DAM2, a later, TREM2-dependent state, has been under-explored in humans. We reasoned that separately examining the enrichment of signatures associated with both DAM sub-states might allow us to delineate the distribution of human

https://doi.org/10.1038/s41593-024-01764-7

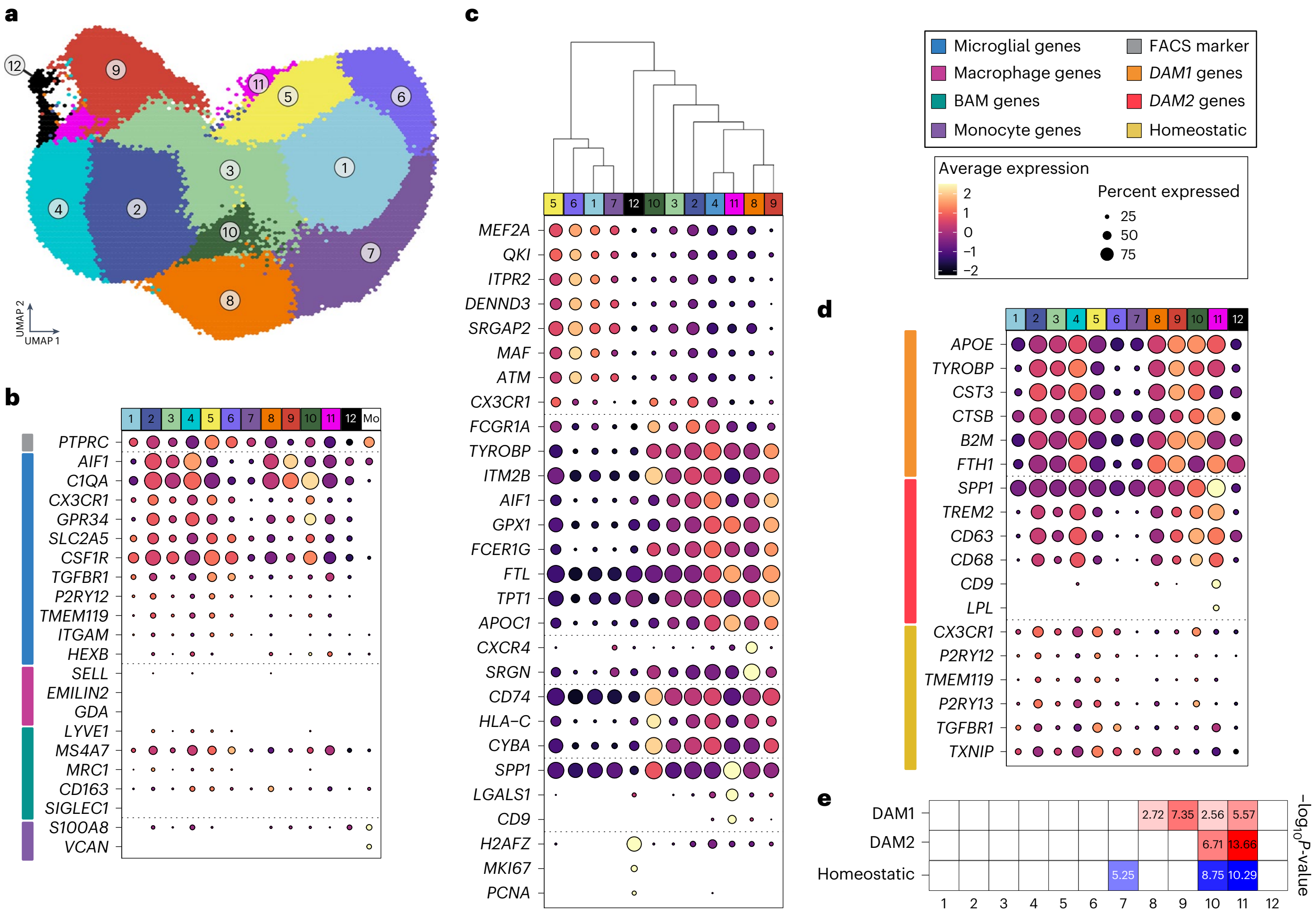


Fig. 2 | Microglial subtypes are defined by distinct marker genes and shared expression programs. a, Visual representation of the 12 microglial subtypes. A hex-binned uniform manifold approximation and projection (UMAP) plot presents microglial subsets: other cells are shown in Extended Data Fig. 1. Each hexagon is colored by the majority cluster identity among all cells aggregated (mean of 50 cells per hexagon). **b**, Expression levels of genes delineating different myeloid identities. The legend (above **d**) summarizes the selected gene sets, which are color coded on the left side. In **b**–**e**, each column presents data from a cluster of cells (microglial subtypes colored as in **a** and monocytes (Mo)), and each row represents the level of expression of a gene. The size of the circle represents the percentage of cells in each cluster that express the gene. The color of the circle represents *z*-scored gene expression. Genes were chosen for association with microglial, macrophage, BAM or monocytic identity. **c**, Subtype-enriched marker genes. Marker genes, selected by pairwise differential expression testing with MAST, delineate broad microglial families with overlapping gene expression programs and small clusters with strongly distinguishing marker genes. Hierarchical clustering with complete linkage on the expression of genes is shown by the dendrogram at the top of the figure. **d**, Expression level of DAM gene sets and homeostatic genes across microglial subsets. **e**, Heat map of DAM gene-set enrichment. Enrichment of DAM subtype signature genes in upregulated (for DAM1/DAM2, in red) or downregulated (homeostatic, in blue) genes associated with each cluster is shown. Each column is one microglial subtype. Enrichment was tested by false discovery rate (FDR)-corrected hypergeometric test. See also Supplementary Table 2. FACS, fluorescence-activated cell sorting.

microglia along this proposed DAM trajectory. Markers for both DAM subsets, as well as homeostatic genes downregulated in the DAM progression, are shown in Fig. 2d. We hypothesized that, if a DAM subset existed in our dataset, it would likely be a small, distinct subset primarily enriched in the DAM2 signature due to the predominance of autopsy tissue from late-stage neurodegenerative disease among our data. Indeed, cluster 11, representing 1% of our microglia, showed strong enrichment for the DAM2 signature. However, as seen in Fig. 2e, the situation is complex: the DAM signature genes are expressed in four different microglial clusters showing different combinations of DAM1 and DAM2 enrichment, with the DAM1 signature being most enriched in cluster 9. We note that cluster 10, the *CD74*^{high} cluster that we had highlighted in our prior report as showing DAM enrichment[4], also showed significant enrichment in the DAM2 signature, albeit at lower level than cluster 11. Notably, clusters 10 and 11 both showed substantial downregulation of the homeostatic microglial signature identified in the original DAM publication[6], while cluster 9 did not demonstrate significant downregulation of the homeostatic gene set, suggesting that clusters 10 and 11 are further along the trajectory of divergence from the homeostatic microglial phenotype. Our data suggest that, in humans, there may be distinct microglial subgroups with different combinations of DAM-related transcriptional programs that fulfill different functions, rather than a single linear DAM trajectory. Finally, it is important to note that the different populations associated with the DAM-like response are minor fractions of the total number of microglia that we have profiled, which provides a possible reason for the difficulty in conclusively identifying these subpopulations in previous human studies.

Axes of metabolic and functional variation across microglial subtypes

After characterizing our microglial subgroups, we next evaluated inter-cluster relatedness using a post hoc machine learning approach leveraging a multilayer perceptron classifier to examine the homology

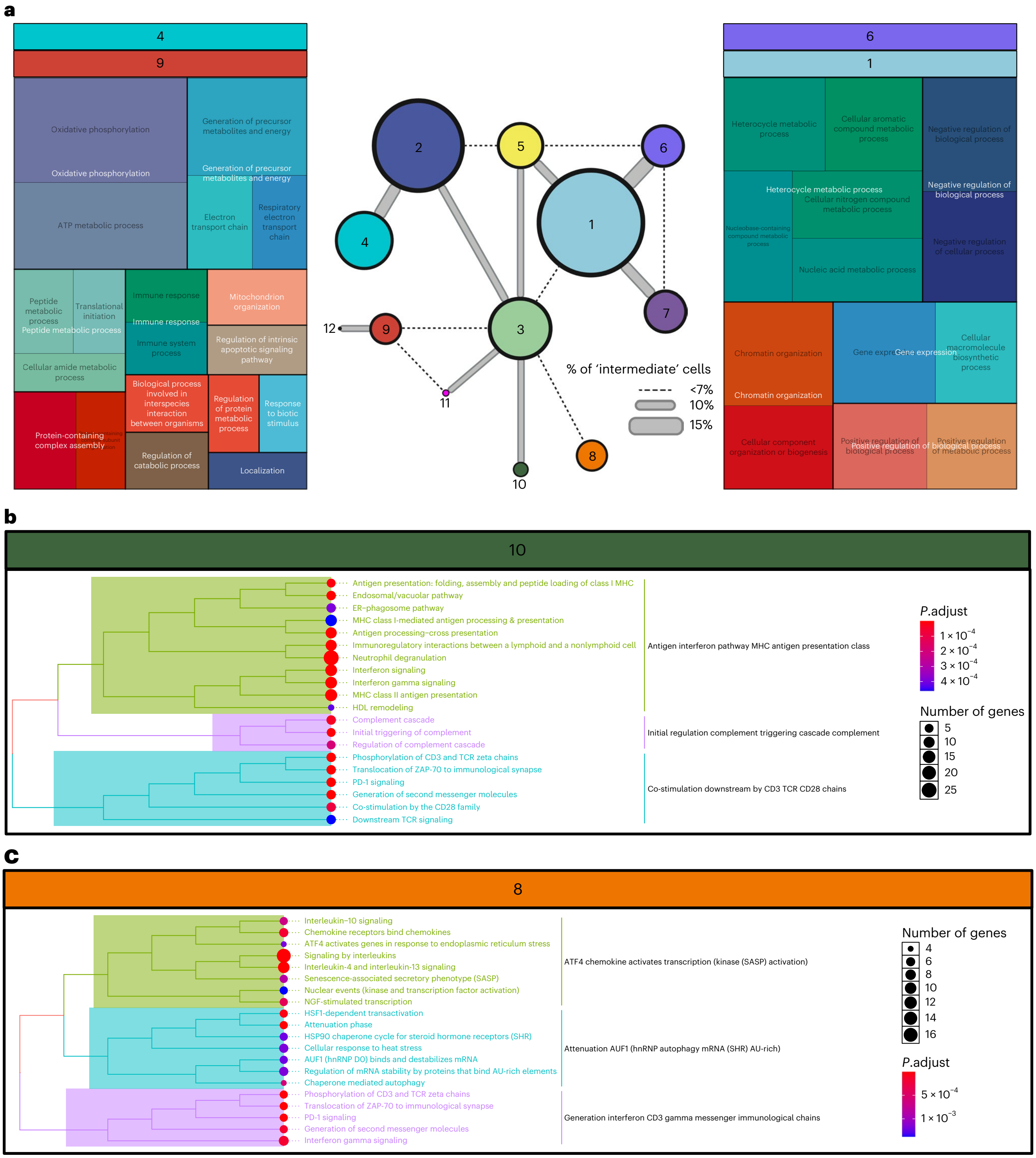


Fig. 3 | Microglia display a complex trajectory of state transition with several primary axes. a, A central metabolic divide separates divergent subtype families. Constellation diagram demonstrates relationships between clusters by way of post hoc classification. Each pair of distinct clusters was used to train a multilayer perceptron 50 times using fivefold cross-validation to obtain a classification for every cell. Cells that were classified to the same cluster less than 40 times were considered ambiguous. The fraction of ambiguous cells determines the width of the connecting lines in the diagram. Each node is a single cluster, with size scaled in proportion to the number of cells contained therein. Notably, even closely related clusters can be reliably distinguished over 85% of the time. Cluster 3, which has few distinct marker genes, has the most 'central' expression profile, with close relationships to the cluster 2/4 family and the 1/6/7 family. Cluster 5 represents another intermediate step between the 2/4 and 1/6/7 families. GO annotation was performed with topGO and summarized with rrvgo. Parent terms are shown in white, overlaid over child terms. GO annotation for clusters 1/6 and clusters 4/9 revealed a metabolic shift between the two groups: clusters 4/9 showed enrichment of oxidative phosphorylation, catabolism and protein metabolism, as well as general immune response, while clusters 1/6 demonstrated upregulation of heterocyclic and nitrogen-containing compound metabolism alongside transcriptional regulation. **b**,**c**, Clusters 8 and 10 shared a signature of interferon-gamma signaling and antigen presentation but differed in other pathways. Reactome annotation of clusters 8 and 10 aggregated by group highlights shared enrichment for T cell interaction and interferon-gamma signaling (purple in cluster 8 and blue in cluster 10). Cluster 10 showed upregulation of complement signaling (purple) and MHC class I/II antigen presentation (green), while cluster 8 showed upregulation of chaperone and steroid signaling (blue) and interleukin signaling (green). See also Extended Data Fig. 4.

of gene expression programs between cells assigned to different clusters[34]. We visualize these results in a constellation diagram (middle of Fig. 3a), which indicates both the proportion of ambiguously classified cells between every pair of clusters (edge thickness) and the total number of cells in each cluster (size of nodes)[4,34,35]. A central question in microglial biology is how different subtypes may branch off from core homeostatic phenotypes, and whether trajectories of microglial state transition are linear or characterized by critical bifurcation points. Based on our analysis, cluster 3 exhibited the most 'central' gene expression profile, with the greatest overlap with other clusters. As expected, clusters 2 and 4 showed substantial homology to one another, as did clusters 1 and 6. Notably, cluster 5, another prospective intermediate state, showed overlap with both clusters 1 and 2, with stronger similarity to cluster 1. This suggests a degree of continuous transition between extremes of gene expression in either direction along this central division. To explore the functional relevance of this division, we used topGO to conduct Gene Ontology (GO)[36–38] analysis on the top differentially expressed genes in each cluster and summarized results with rrvgo[39]. As shown on the left and right sides of Fig. 3a, the most heavily enriched terms along the left (clusters 4 and 9) side of our population structure are related to metabolism, particularly oxidative phosphorylation, catabolism and peptide metabolism, and immune response and localization. In contrast, the right side (clusters 1 and 6) of our population structure shows enrichment of alternative metabolic pathways, including heterocyclic metabolism and nitrogen-containing compound metabolism as well as transcriptional regulation. Intriguingly, the intermediate cluster 5 showed strong association with motility (Extended Data Fig. 4a). This highlights a central divide in metabolism and function, suggesting a homeostatic-active phenotype in clusters 2, 4 and 9 that transitions to different metabolic and functional phenotypes in clusters 1, 6 and 7 of our structure, with intermediate states that may play different functional roles. Cluster 9, which has a partially overlapping transcriptional signature with cluster 4, is the most closely related to cluster 12, which shows enrichment for proliferation and oxidative phosphorylation (Extended Data Fig. 4b). Cluster 9 also showed substantial transcriptomic similarity to cluster 11, which is enriched for lipid processing and beta-amyloid clearance (Extended Data Fig. 4c), consistent with the proposed role of TREM2 in this signature[33] and confirming the continuum of DAM transitional states identified in our earlier analysis (Fig. 2e).

Clusters 8 and 10, whose signatures suggest substantial microglial activation, are clearly distinguishable from other clusters, although they maintain a relationship to the central cluster 3. To explore this axis, we annotated clusters using ClusterProfiler to perform Reactome[40–42] pathway analysis (Fig. 3b,c). Some overlap was present between these two clusters, as both clusters contained genes associated with antigen presentation, interferon signaling and T cell interaction. However, cluster 10 exhibited stronger association with both class I and class II major histocompatibility complex (MHC) signaling and complement signaling. In contrast, cluster 8, which expressed significant levels of early response genes, showed upregulation of pathways associated with chaperone signaling, steroid response, interleukin signaling (particularly *IL4*/*IL10*/*IL13*), and the senescence-associated secretory phenotype. These phenotypic differences were also recapitulated with GO annotation (Extended Data Fig. 4d).

Thus, this analysis highlights the divergent nature of the microglial differentiation program, suggesting that there are at least three distinct tracks of microglial subtype specification that emerge from the most basal microglial state, including a central metabolic divide, an axis of immunological activation, and a trajectory that contains elements of the DAM signature identified in murine model systems. These tracks appear to be nonlinear and different paths of transition may exist between terminal states, a result that is consistent with an ancillary pseudotime analysis leveraging the Monocle3 algorithm[43–45] that defines a complex trajectory (Extended Data Fig. 4e). In addition, consistent representation of clusters across donors, regions and diseases (Extended Data Fig. 3a,b) supports the conceptual framework of our trajectory analyses, as this shared representation supports the idea that we are examining an actual biological continuity rather than identifying state shifts that result from differences in disease or region. Indeed, our findings here parallel findings from studies of microglia in mice that have hinted at nonlinear trajectories in disease states, either with regard to early branching points[46–48] or partially overlapping terminal phenotypes that arise in similar disease contexts[6,49,50]. However, further dissection of the trajectories in human data that we describe here will be challenging due to the difficulty of profiling human samples along a continuous pathway toward disease and may ultimately require careful cohort design and new tools for tracking patterns of state transition in individual cells. Notably, the results of our analyses further underscore the robust nature of our population structure, as even clusters with substantial overlap of gene expression signatures still comprise cells that are robustly distinguishable with more than 85% accuracy. Starting from this overall framework, further work in vitro and in situ will be required to confirm our observations and to understand the importance of functional and metabolic shifts in health and disease.

Annotating disease and trait associations of distinct microglial subsets

To illustrate variation in microglial composition in our brain tissue samples, we report the proportion of different microglial subtypes in each disease, region and individual (Fig. 4 and Extended Data Fig. 3). First, we note that most clusters are present in each individual, albeit at different frequencies (Extended Data Fig. 3a). The most common microglial subtypes in most individuals were clusters 1 to 6, suggesting that these subtypes capture a homeostatic spectrum and that neurodegenerative diseases involve small shifts toward distinct microglial states. Notably, grouping samples by region (Fig. 4a,b) or diagnosis (Fig. 4c,d), demonstrated that even with different numbers of cells from different regions and diagnoses, we see a similar distribution of cell populations across diseases and regions. As substantial statistical power is needed for comparison of samples with complicated combinations of region, disease, age and sex. Much larger datasets will be needed to directly identify disease associations; our dataset was not designed for this purpose.

However, the depth and quality of our sequencing data enabled us to pursue gene enrichment analyses to implicate certain subsets in different diseases. For MS, we utilized a recent publication from the International Multiple Sclerosis Genetics Consortium that identified a comprehensive set of 551 MS susceptibility genes[51]. We found that clusters on the right side of our microglial cloud, specifically clusters 5 and 6, were significantly enriched in MS susceptibility genes, highlighting a possible role of one arm of our microglial differentiation tree in MS susceptibility (Fig. 5a). Next, we explored the genome-wide association study (GWAS) catalog[52], a curated database containing single nucleotide polymorphism (SNP)–trait associations from GWAS studies. As seen in Fig. 5b, we recapitulated the enrichment of MS in cluster 6 in this database, although cluster 5 did not pass the threshold for significance in this analysis. Clusters 1 and 6 also showed enrichment for other neurodegenerative and neuropsychiatric disease genes, including AD, PD and depression. Similarly, cluster 10 showed enrichment for MS and schizophrenia. The strong complement expression found in cluster 10 aligns with previous reports of the role of complement-related genes in schizophrenia[53]. Finally, cluster 8, the *CXCR4*-enriched cluster, has a set of disease associations that suggest a role in conditions characterized by neuroinflammatory signaling but not neurodegenerative diseases. Interestingly, we found no substantial enrichment of disease genes associated with stroke, seizure, ALS/FTD or glioma. This may be due either to a less important role of microglia in the primary

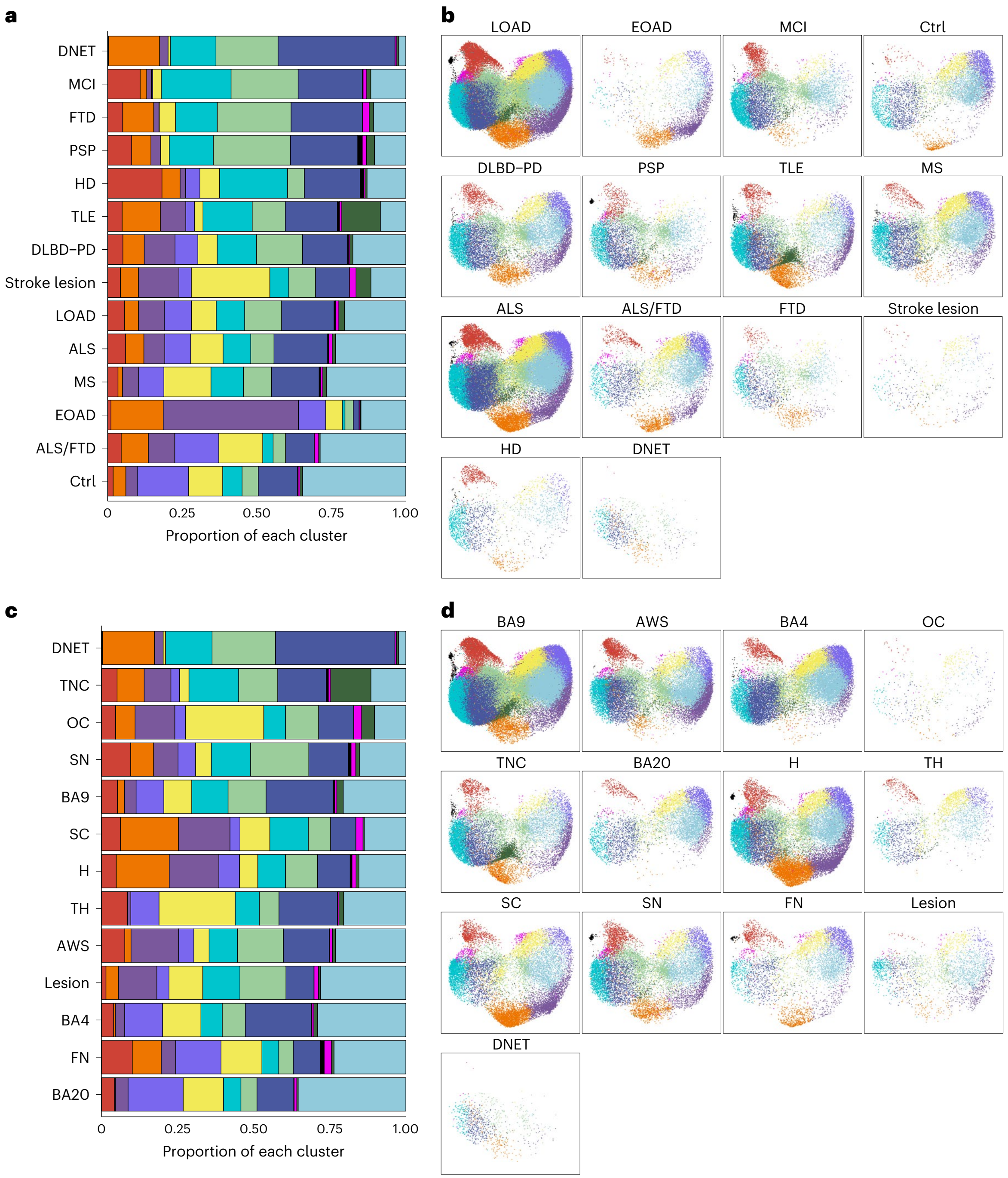


Fig. 4 | Human microglial subsets are found across diseases and regions. **a**–**d**, Microglial subsets are broadly represented across diseases and regions. On the left (**a** and **c**), each bar shows the proportion of each cluster among all microglia from a given disease. On the right (**b** and **d**), UMAP plots are split by disease. Plots are color coded in accordance with Fig. 2a. Most subsets are represented across all diseases and all regions, albeit in slightly different numbers, although larger sample sizes would be required to statistically assess differences in abundance. LOAD, late-onset AD; EOAD, early-onset AD; Ctrl, control; TLE, temporal lobe epilepsy; DNET, dysembryoplastic neuroepithelial tumor; BA, Brodmann area; AWS, anterior watershed; OC, occipital cortex; TNC, temporal neocortex; H, hippocampus; TH, thalamus; SC, spinal cord; SN, substantia nigra; FN, facial nucleus. See also Extended Data Fig. 3.

pathogenesis of these diseases or to less extensive GWAS annotation of these diseases.

To complement these analyses, we repeated our earlier analysis[4] evaluating AD-related traits more thoroughly. We leveraged associations with traits from a large analysis of bulk cortical (BA9) RNA-seq data from 1,092 participants in the ROSMAP[54,55] cohorts (Supplementary Table 3); these individuals do not overlap with the ROSMAP participants included in our microglial dataset. These bulk RNA-seq data contain transcripts from all parenchymal cells, including microglia. To evaluate the enrichment of microglial clusters for genes associated with each trait, we calculated the overlap of cluster-specific signature genes with gene sets that were significantly positively or negatively associated with each of these traits (Fig. 5c). Clusters 1 and 6, the right side of the microglial cloud, were enriched for genes that are positively correlated with amyloid-beta pathology, tau tangle pathology, and both a clinical and pathological diagnosis of AD. Consistent with these results, they were also enriched for genes negatively correlated with the slope of cognitive decline where a larger negative number indicates worsening cognitive

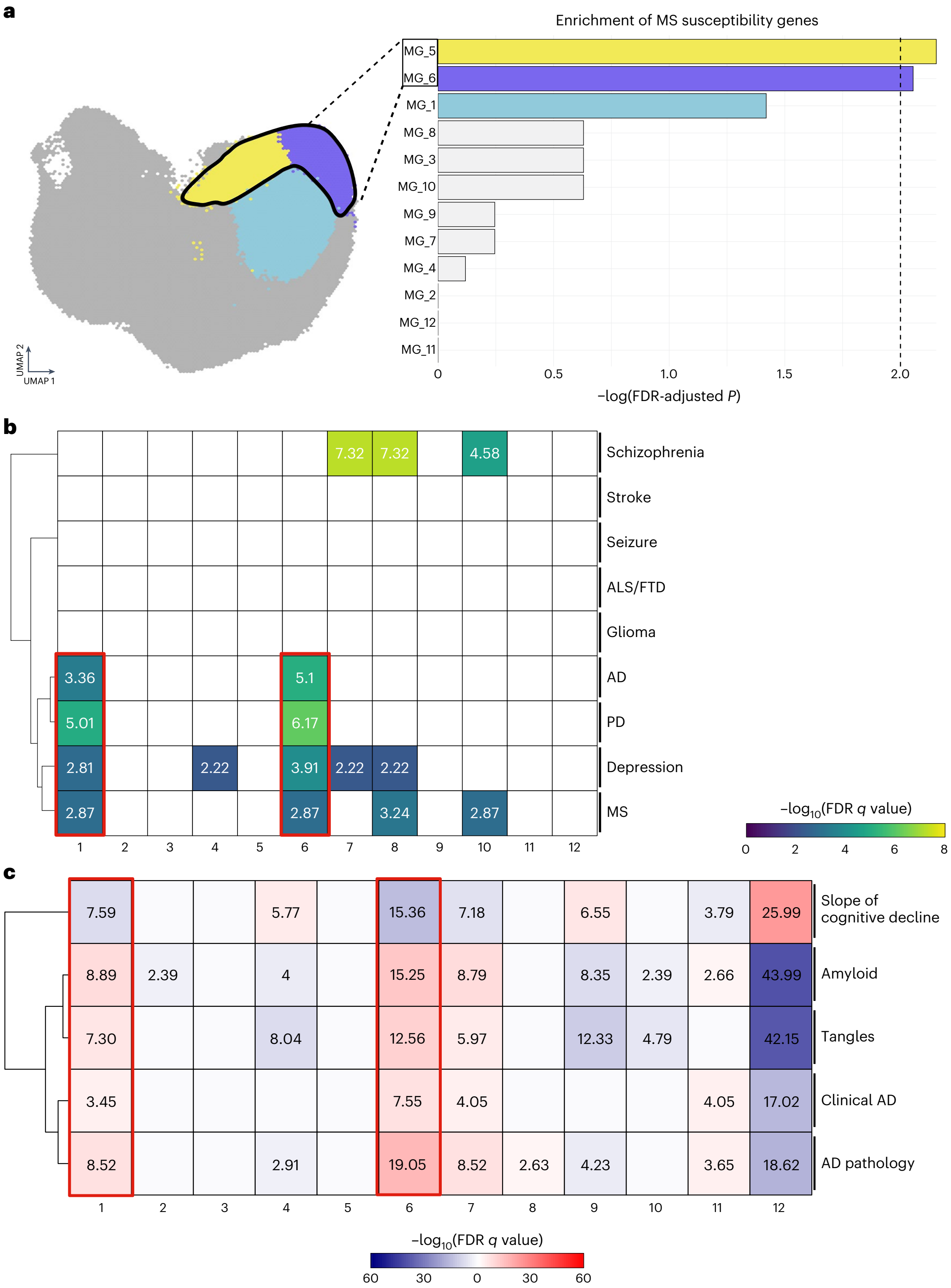


Fig. 5 | Disease annotation implicates specific microglial families in disease. **a**, Clusters 5 and 6 are enriched in GWAS-derived MS susceptibility genes. The *y* axis of the bar plot shows the different clusters, ranked in descending order of the negative log-transformed *P* values on the *x* axis. Enrichment of MS susceptibility genes in upregulated gene lists associated with each cluster was tested with the hypergeometric test using a Benjamini–Hochberg correction. Bars are colored if they have an FDR < 0.01. **b**, Clusters 1 and 6 are enriched in genes associated with neurodegenerative diseases. Enrichment analysis of genes associated with each disease in the GWAS catalog was performed with same parameters. Diseases are listed on the *y* axis, and negative log-transformed *P* values are shown for combinations of clusters and traits where they have an FDR < 0.01. Coloration of squares corresponds to *P*-value magnitude: larger *P* values correspond to darker blue squares, whereas smaller *P* values correspond to yellow coloration. **c**, Clusters 1 and 6 correlate with clinical and pathological traits in AD. In this case, enrichment was performed separately for both the genes positively and negatively correlated with each trait in upregulated genes for each cluster. Coloration of each box relates to the strength and directionality of each association. Red (positive numbers) corresponds to genes upregulated with the trait, while blue corresponds to genes downregulated in relation to the trait. See also Supplementary Table 3.

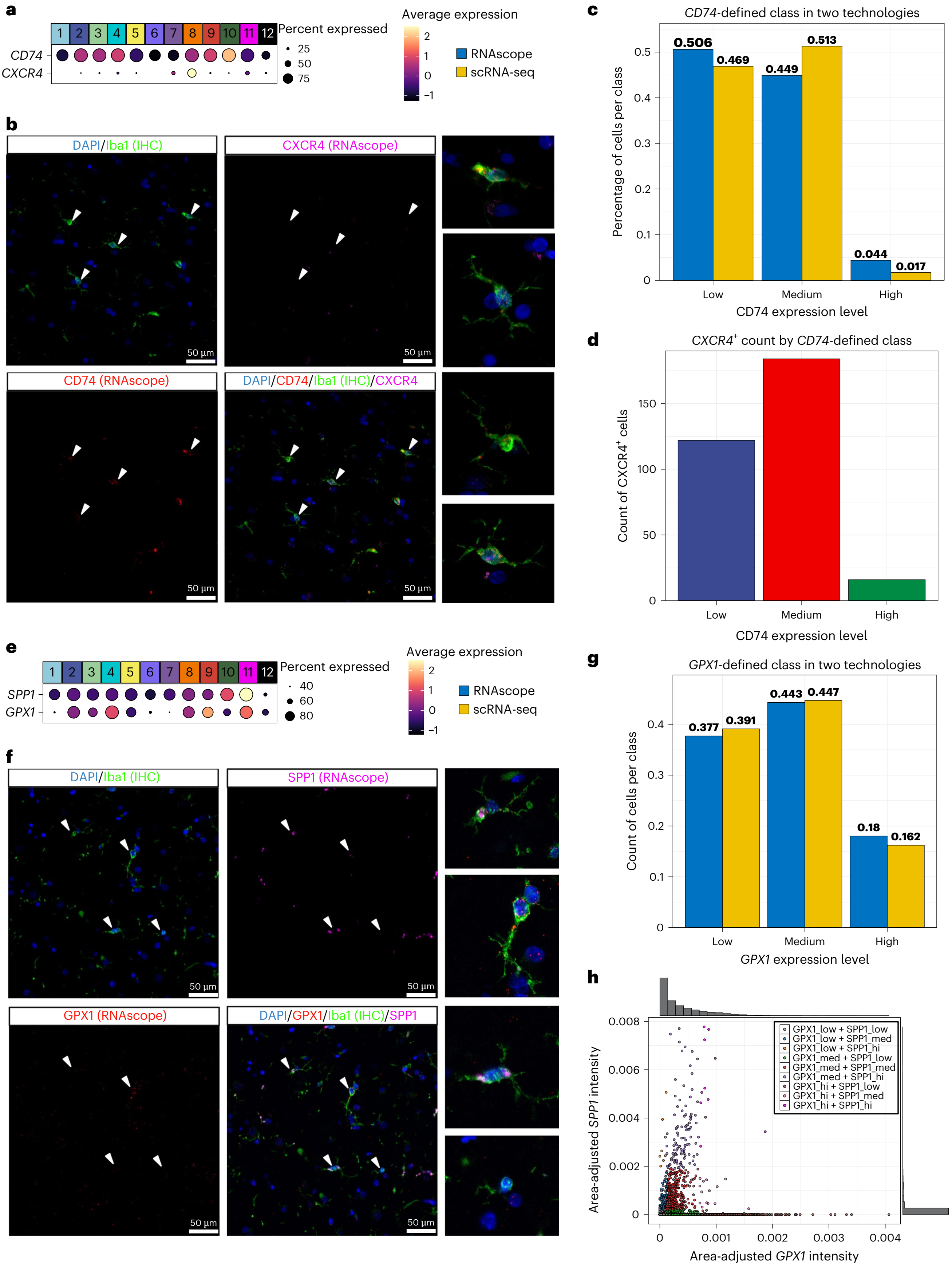

a
1 2 3 4 5 6 7 8 9 10 11 12
CD74
CXCR4
Percent expressed
25
50
75
Average expression
2
1
0
−1
b
DAPI/Iba1 (IHC)
CXCR4 (RNAscope)
CD74 (RNAscope)
DAPI/CD74/Iba1 (IHC)/CXCR4
50 μm
c
CD74-defined class in two technologies
RNAscope
scRNA-seq
Percentage of cells per class
0.506
0.469
0.449
0.513
0.044
0.017
Low
Medium
High
CD74 expression level
d
CXCR4+ count by CD74-defined class
Count of CXCR4+ cells
150
100
50
0
Low
Medium
High
CD74 expression level
e
1 2 3 4 5 6 7 8 9 10 11 12
SPP1
GPX1
Percent expressed
40
60
80
Average expression
2
1
0
−1
RNAscope
scRNA-seq
f
DAPI/Iba1 (IHC)
SPP1 (RNAscope)
GPX1 (RNAscope)
DAPI/GPX1/Iba1 (IHC)/SPP1
50 μm
g
GPX1-defined class in two technologies
Count of cells per class
0.377
0.391
0.443
0.447
0.18
0.162
Low
Medium
High
GPX1 expression level
h
GPX1_low + SPP1_low
GPX1_low + SPP1_med
GPX1_low + SPP1_hi
GPX1_med + SPP1_low
GPX1_med + SPP1_med
GPX1_med + SPP1_hi
GPX1_hi + SPP1_low
GPX1_hi + SPP1_med
GPX1_hi + SPP1_hi
Area-adjusted SPP1 intensity
0.008
0.006
0.004
0.002
0
0.001
0.002
0.003
0.004
Area-adjusted GPX1 intensity

Fig. 6 | In situ confirmation of microglial population structure with joint immunofluorescence–RNAscope with automated segmentation. a, *CD74* demarcates a small, immunologically active subset, while *CXCR4* delineates a distinct immunologically active subset. The size of the circle represents the percentage of cells in each cluster that express the gene, and the color of the circle represents *z*-scored gene expression. *CD74* is overexpressed in cluster 10, while *CXCR4* is primarily expressed in cluster 8. **b**, Representative images showing *CD74* and *CXCR4* in IBA1$^+$ microglia. RNAscope staining for *CD74* (red) and *CXCR4* (pink) in IBA1$^+$ microglial cells (green) in human cortical brain slices, with nuclear DAPI staining (blue). In the same field of view, microglia with different levels of *CD74* and with or without expression of *CXCR4* can be observed (arrowheads point to representative microglia). **c**, Separating single in situ cells using *CD74* expression thresholds adapted from scRNA-seq identified similar proportions across technologies. The proportion of cells, along the *y* axis, that express low, medium or high levels of *CD74*, along the *x* axis, in scRNA-seq is shown in yellow, while in situ results (area-adjusted *CD74* expression binned on thresholds from scRNA-seq) are shown in blue. **d**, *CXCR4*$^+$ cells matched the expected distribution within *CD74* expression classes. *CD74* expression class, as described in **c**, is shown on the *x* axis, and count of CXCR4$^+$ cells is shown on the *y* axis. CXCR4$^+$ microglial cells are identified in situ and most fall into the CD74$^{\text{int}}$ class, confirming our scRNA-seq findings. **e**, *GPX1* and *SPP1* delineate the DAM axis and extremes in homeostatic-active families. **f**, Representative images from our joint staining protocol for *GPX1* and *SPP1*. Staining as in **b**, except that RNAscope *SPP1* is pink and *GPX1* is red. **g**, Separating single in situ cells on the basis of *GPX1* expression thresholds borrowed from scRNA-seq also identified similar proportions across technologies. Analysis performed as in **c** but using *GPX1* expression data. **h**, Gradated expression of both *SPP1* and *GPX1*. Individual cells are plotted as single dots, where the axes represent area-adjusted expression of *GPX1* (*x*) or *SPP1* (*y*). See also Extended Data Fig. 5 and Supplementary Tables 4 and 5. IHC, immunohistochemistry.

dysfunction. In contrast, clusters 2, 4, 9 and 10 were enriched in genes negatively correlated with tau and amyloid pathology, and clusters 4 and 9 were enriched in genes positively correlated with the slope of cognitive decline. Further, our DAM2$^{\text{hi}}$ cluster 11 was enriched for genes positively associated with AD and amyloid pathology, but not tau pathology. This cluster also showed enrichment for genes negatively correlated with cognitive decline. Our results strongly suggest that the human cortex in AD is enriched for genes defining clusters 1 and 6; this could occur either from an increase in the proportion of these subtypes in AD cortical tissue or from the enhanced expression of these signature genes in the microglia of the AD cortex. It also implicates cluster 11 more modestly, suggesting a narrower contribution to amyloid rather than tau proteinopathy. Notably, several molecular pathologies that we evaluated—including cerebral amyloid angiopathy, arteriolar sclerosis, cerebrovascular disease and TDP-43 pathology—showed no enrichment with any of our clusters. Cluster 12 showed enrichment trends akin to clusters 2, 4, 9 and 10, perhaps because of the heavy overrepresentation of genes associated with oxidative phosphorylation. Overall, by leveraging indirect disease annotation, we identify a family of microglial subtypes that are strongly enriched in genes associated with neurodegenerative diseases and associated with AD traits in an independent ROSMAP cohort.

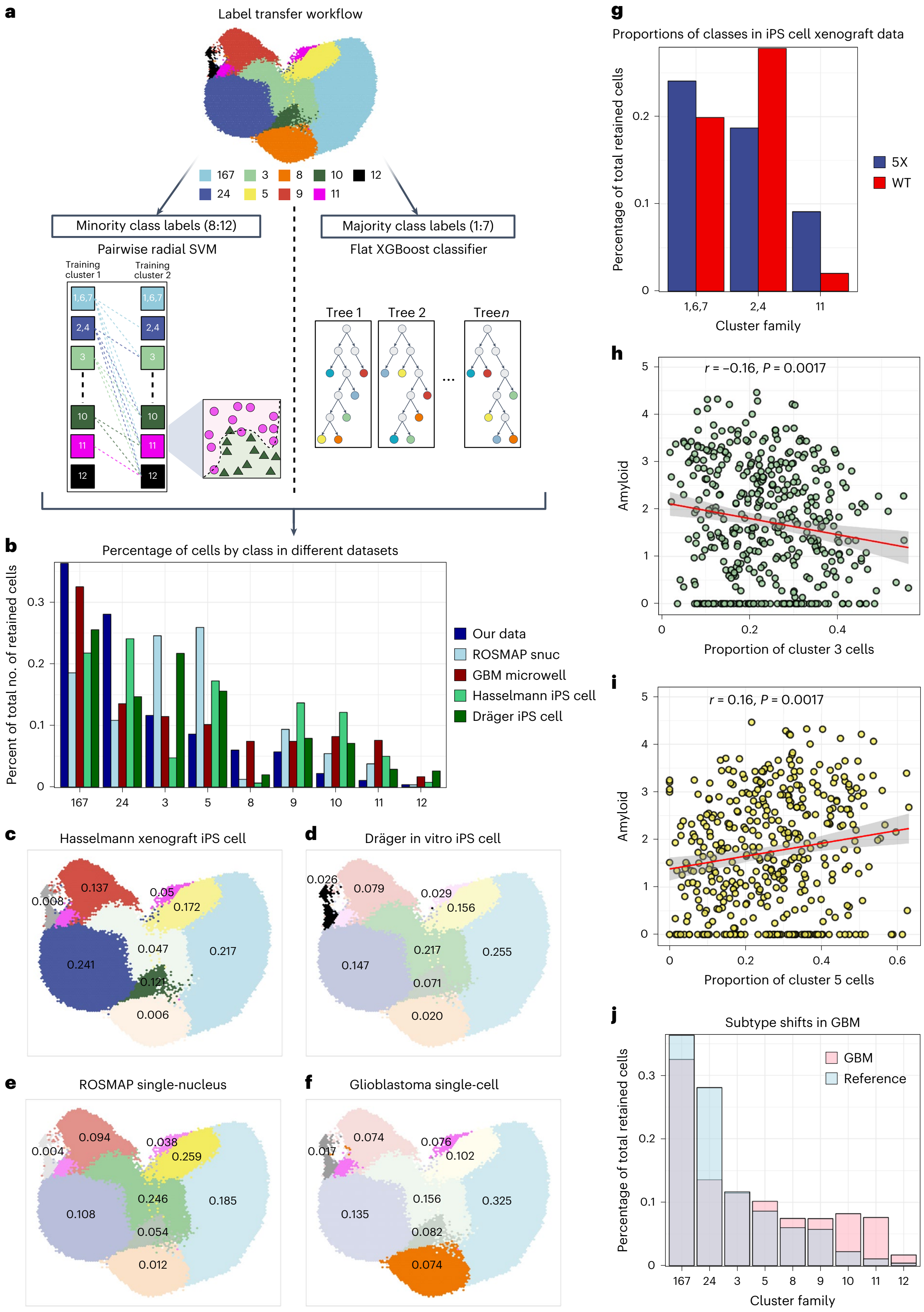
Identifying microglial subsets in situ

Having identified microglial subgroups from dissociated cells, we sought to validate their existence in situ. We had previously done this with immunofluorescence[4,56], but the range of potential antibody markers is limited. Thus, we first optimized a co-detection workflow that merged anti-IBA1 staining (IBA1 is a ubiquitous marker of myeloid cells in the brain at the protein level), with Advanced Cell Diagnostic's RNAscope protocol[57] for fluorescence in situ hybridization to allow for single-molecule RNA detection. We coupled this experimental pipeline to CellProfiler (v.4.2.1)[58–61] for automated segmentation of image data (Methods). This workflow enables the capture of microglia-specific gene expression, localization of transcripts within microglia and structured assessment of cellular morphology.

To illustrate our approach, we chose two panels of genes to discriminate different microglial subsets in situ. In panel 1 (Fig. 6a–d and Extended Data Fig. 5a), we used probes against transcripts of *CD74*, a gene that we found to be relatively enriched in immunologically active subtypes (clusters 2, 3, 4, 8, 9, and 10) and downregulated in clusters 1, 5 and 6. We have previously reported the existence and relevance of a *CD74*$^{\text{high}}$ microglial subset[4], and the equivalent subset in our current model (cluster 10) had a 1.5 fold greater expression of CD74 relative to all other clusters (Fig. 6a). We also included probes against *CXCR4*, a gene predominantly expressed in cluster 8. This panel separates the 1/5/6 family from other clusters, exploring our primary axis of variation, and allows for the discrimination of the two subsets (8 and 10) in a second axis of variation associated with antigen presentation and immune cell interaction. In panel 2 (Fig. 6e–h and Extended Data Fig. 5b), we included probes against *GPX1*, a gene predominantly expressed in clusters 4, 9 and 11, as well as *SPP1*, a previously proposed DAM marker that is enriched in cluster 11. This second panel enables a more detailed examination of clusters along the DAM-like trajectory that were not captured in the first panel. We applied this combined RNA–protein interrogation technique to tissue sections from individuals with pathological diagnoses of AD, PD, PSP or DLBD and individuals with age-related tauopathy (Supplementary Table 4).

Representative images for panel 1 are shown in Fig. 6b, demonstrating the capture of microglia-specific transcripts, including those in the distal processes of single microglia. By taking area-normalized *CD74* expression per cell and binning cells into low expression, medium expression and high expression based on fold-change thresholds derived from our single-cell data (Methods), we found that the proportions of cells in each bin are similar to the proportions of cells in our single-cell dataset (Fig. 6c). The *CD74*-low grouping included clusters 1, 5, 6, 7 and 12, *CD74*-medium represented clusters 2, 3, 4, 8, 9 and 11, and *CD74*-high identified cluster 10. Moreover, concurrently examining expression of *CXCR4* across all three bins demonstrated peak *CXCR4* expression in the *CD74*-medium populations, in agreement with our scRNA-seq data (Fig. 6d). Thus, panel 1 validates clusters 8 and 10

Fig. 7 | Live microglial population structure enables annotation of datasets from model systems and data produced with different technologies. a, Overview of our label transfer workflow. Similar classes were aggregated (2 and 4 or 1, 6 and 7) to simplify the classification problem, and classifications from two types of models were merged to assign final class labels for all cells in query data. **b**, Distribution of subset proportions across different datasets in comparison to our reference. **c**–**f**, Mapping of query datasets onto our reference model. UMAP colors for each cluster family were shaded by the proportion of cells assigned to each family in each dataset. Numbers are the proportion of cells in each query dataset that were assigned to each cluster. **g**, Xenografted human iPS cell microglia shifted away from homeostatic-active phenotypes and toward disease-associated phenotypes in 5XFAD mice. Bar plot showing the proportion of iPS cell-derived microglia-like cells (*y* axis) in each of three cluster families (*x* axis) from either 5X (blue) or WT (red) mice. *n* = 2 per condition. **h**, GBM induced depletion of homeostatic myeloid cells and shifted microglia toward more inflammatory subtypes. Bar plot showing proportions of cells per group from the reference (blue), or the classified GBM data (red). Between the two datasets, the higher proportion is shown in its corresponding color, and the lower proportion is delineated in gray. **h**, Cluster 3 abundance correlated negatively with amyloid pathology in ROSMAP single-nucleus data. In the dot plot, each dot is a single donor. Axes are amyloid burden (*y*) and proportion of cells classified as cluster 5 (*x*). **h**,**i**, Conversely, cluster 5 abundance correlated positively with amyloid pathology. See also Extended Data Fig. 6 and Supplementary Table 6. **j**, Projection of a GBM dataset into our model; there is a shift in the proportion of microglial subtypes away from homeostatic subtypes and toward activated subtypes in GBM-derived cells (pink) relative to our reference data (blue).


a
Label transfer workflow
167
3
8
10
12
24
5
9
11
Minority class labels (8:12)
Pairwise radial SVM
Training cluster 1
Training cluster 2
1,6,7
2,4
3
10
11
12
Majority class labels (1:7)
Flat XGBoost classifier
Tree 1
Tree 2
Tree n
...
b
Percentage of cells by class in different datasets
Percent of total no. of retained cells
0
0.1
0.2
0.3
167
24
3
5
8
9
10
11
12
Our data
ROSMAP snuc
GBM microwell
Hasselmann iPS cell
Dräger iPS cell
c
Hasselmann xenograft iPS cell
0.008
0.137
0.05
0.172
0.047
0.217
0.241
0.121
0.006
d
Dräger in vitro iPS cell
0.026
0.079
0.029
0.156
0.217
0.255
0.147
0.071
0.020
e
ROSMAP single-nucleus
0.004
0.094
0.038
0.259
0.246
0.185
0.108
0.054
0.012
f
Glioblastoma single-cell
0.017
0.074
0.076
0.102
0.156
0.325
0.135
0.082
0.074
g
Proportions of classes in iPS cell xenograft data
Percentage of total retained cells
0
0.1
0.2
1,6,7
2,4
11
Cluster family
5X
WT
h
r = −0.16, P = 0.0017
Amyloid
0
1
2
3
4
5
0
0.2
0.4
Proportion of cluster 3 cells
i
r = 0.16, P = 0.0017
Amyloid
0
1
2
3
4
5
0
0.2
0.4
0.6
Proportion of cluster 5 cells
j
Subtype shifts in GBM
Percentage of total retained cells
0
0.1
0.2
0.3
GBM
Reference
167
24
3
5
8
9
10
11
12
Cluster family

in situ and demonstrates our ability to discriminate multiple distinct microglial subsets in each field of view. As shown in Fig. 6e,f, panel 2 was designed to discriminate the homeostatic-active family and concurrently identify clusters on the DAM axis, particularly cluster 11, that our first panel cannot capture. We first evaluated *GPX1*, and, as with panel 1 markers, we observed similar proportions of cells in the different *GPX1* categories between the in situ and scRNA-seq approaches, confirming the translatability of our scRNA-seq-derived cluster definitions (Fig. 6g). Notably, we identified cells coexpressing different levels of *GPX1* and *SPP1* and subpopulations that appear similar to clusters 9 and 11, making them good markers for future investigations targeting the DAM-like axis (Fig. 6h). This second panel thus offers an independent validation of a different aspect of our population structure, providing markers for future study design.

Establishing the colocalization and/or segregation of cluster-specific marker genes then allowed us to investigate morphological differences between microglial subgroups. We evaluated morphological measures captured from each microglia by CellProfiler. We found that the partially overlapping *CD74*, *SPP1* and *GPX1*-medium classes had the highest level of ramification (as measured by compactness scores; Extended Data Fig. 5c). Similarly, we found that eccentricity—a feature ranging from 0 (perfectly circular) to 1 (perfectly linear)—was lowest in both the CD74^{high} and GPX1^{high} classes and highest in the CD74^{low} and GPX1^{low} classes (Extended Data Fig. 5d), suggesting that the latter subgroups have a more elongated morphology. Microglial activation induces process retraction and an amoeboid morphology[62], and in our scRNA-seq data, cells expressing high levels of CD74 and GPX1 expressed markers of activation. Interestingly, CXCR4$^+$ cells exhibited higher CD74 radial distance on average (Extended Data Fig. 5e) and higher ramification scores (Extended Data Fig. 5f), suggesting that, unlike other activated classes, CXCR4$^+$ cells are likely to be more ramified.

To complement the RNAscope approach described above, we also assessed our subtype markers using the MERSCOPE platform, which enables highly multiplexed MERFISH assessment of RNA species; specifically, we designed a panel to detect all cortical cell subtypes[63] and included 39 markers to capture our 12 microglial subtypes. Two tissue sections were profiled (one from an AD donor cortex and one from a non-AD donor cortex) and, after preprocessing (Methods), we identified a total of 2,381 microglia (6.4% of all cells). When we projected these microglial cells into our model (Extended Data Fig. 6a), we found that even with a relatively small number of cells and a non-transcriptome-wide signature, we were still able to unambiguously identify 9 of the 12 microglial subtypes. Further work will be needed to enhance the panel and profile a much larger sample of tissue sections to uncover the remaining microglial subtypes.

We, therefore, validated our cross-disease scRNA-seq resource by identifying the dissociated microglial signatures in situ and highlighting morphological differences between subgroups. We also showed the existence of cluster 8 signatures using both RNAscope and MERFISH in the tissue, supporting our hypothesis that the stress response observed in certain ex vivo preparations of microglia is also present in intact human brain tissue. However, it is likely that this stress response may be enhanced by certain microglial manipulations[18,19].

Extending the use of our dataset as a reference: in vitro model systems, single-nucleus data and other diseases

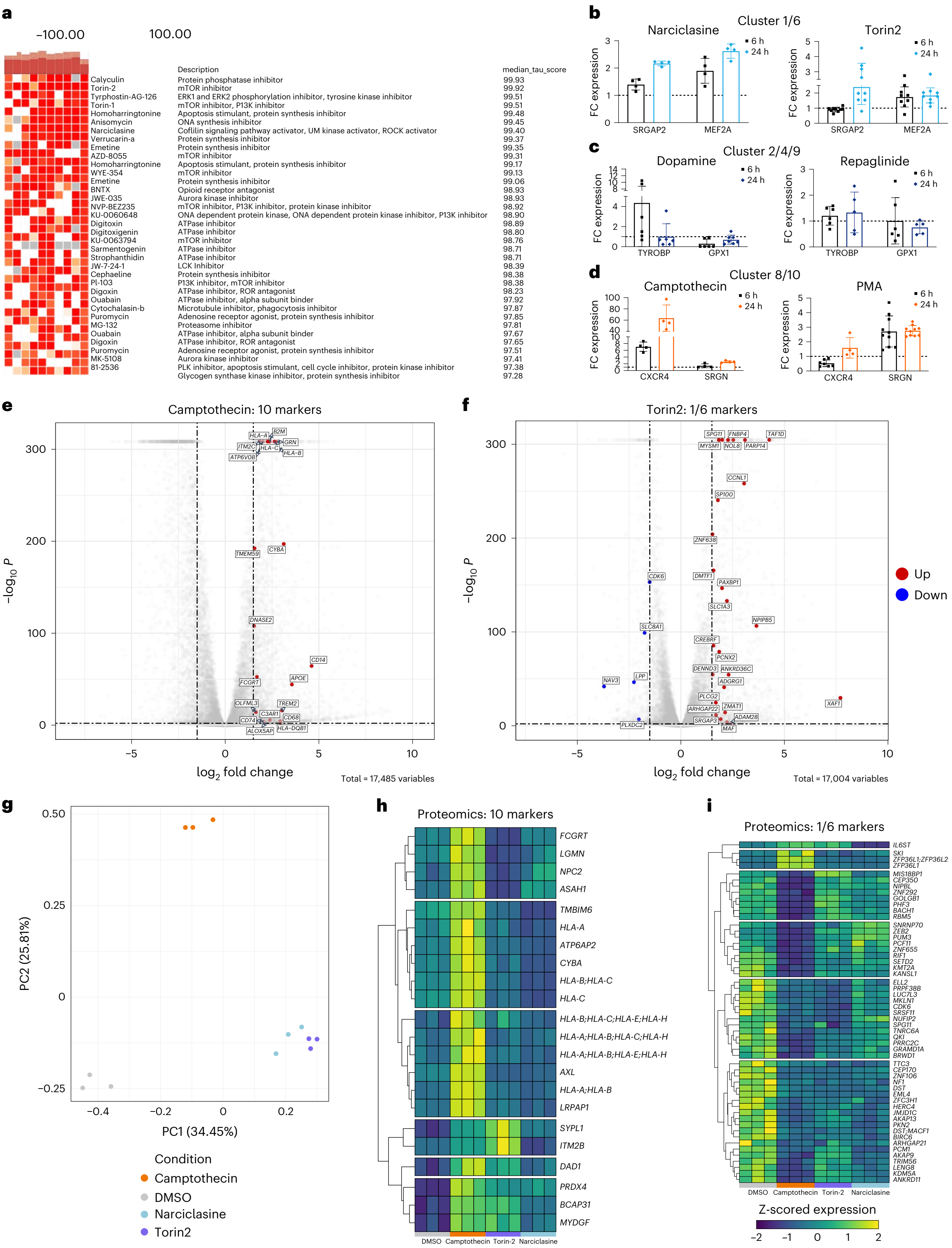
This data resource, which was designed to identify a shared, stable microglial population structure across human diseases and brain regions, can be used to annotate other microglial datasets and to evaluate how much human microglial diversity is captured by model systems. To illustrate these uses, we selected two primary tissue datasets: a single-nucleus RNA-sequencing (snRNA-seq) dataset (Methods) from the ROSMAP[54,55] cohorts of older individuals with and without AD and a single-cell myeloid dataset[64] from surgical resections of glioblastoma multiforme (GBM), a disease where microglia are thought to play a central role[65]. In parallel, we also selected two human iPS cell-derived microglia-like cell datasets: a murine xenograft system[66,67] and an in vitro system used for CRISPR screening[68]. We then applied a label transfer approach on these four datasets; to simplify the classification task, transcriptionally similar groups (that is, clusters 2 and 4 or 1, 6 and 7) in the reference data were grouped. Our approach consisted of a consensus voting of pairwise support vector machine (SVM) classifiers to classify the smaller, more transcriptionally unique subtypes and a flat XGBoost[69] (XGB) classifier retaining only classifications with confidence of 50% or higher for the larger clusters (Fig. 7a). Quality metrics for our label transfer pipeline are shown in Extended Data Fig. 7 and Supplementary Table 6. In short, our approach is sensitive and specific, with joint accuracy averaging 83.8% across all models, and the lowest-confidence assignments emanate from clusters with inherently fluid boundaries, validating the efficacy of this pipeline.

This approach revealed differences in the percentages of cells mapped to each cluster across the different datasets (Fig. 7b–f). In both iPS cell-derived datasets (Fig. 7c,d), cells mapped to most of the states that we identified in our microglial dataset, suggesting that these model systems recapitulate a substantial amount of human in vivo heterogeneity. However, the tissue-derived data (Fig. 7c) showed higher fractions of microglia at terminal points in our axes of differentiation, such as the DAM-like axis of clusters 9, 10 and 11. Notably, we observed a trend toward increased levels of cluster 11 in the 5X FAD mice ($P = 0.087$; Fig. 7g), consistent with our AD enrichment analyses (Fig. 5b,c). In contrast, the in vitro iPS cell-derived dataset contained high numbers of proliferating cells (Fig. 7d), 'intermediate' cluster 3 cells and inflammatory cluster 8 cells. These results highlight the utility of both model systems for modeling microglial diversity.

With regards to the snRNA-seq datasets on primary tissue, annotation of ROSMAP snRNA-seq data (Methods) showed prominent

Fig. 8 | Chemical perturbation recapitulates in vivo human microglial subtype signatures in vitro. a, Representative example of CMAP analysis. The CMAP was used to identify compounds that might drive transcriptional signatures found in different microglial subsets. The cell ID column identified the nine cell lines used in CMAP. Drugs were ranked by the tau score, which quantifies homology between the perturbagen and the query. Scores greater than 90 were considered as candidates for further study. **b**–**d**, qPCR hits by grouping: 1/6, 4/9 or 8/10. Drugs were tested in the HMC3 microglial model system at 6-h and 24-h intervals, and two marker genes were assayed by qPCR per cluster group (1/6, *SRGAP2* and *MEF2A*; 4/9, *TYROBP* and *GPX1*; 8/10, *CXCR4* and *SRGN*). CT values were normalized to *HPRT1*. Bars represent fold-change expression in relation to DMSO control. **e**, Camptothecin upregulated cluster 10 markers. Volcano plot showing log fold change (LFC, *x*), and $-\log_{10} P$ value (*y*) from bulk RNA-seq generated from HMC3 cells treated with camptothecin for 24 h. Data were analyzed with DESeq2. FDR threshold was set to 0.01 and LFC threshold was set at 1.5. The top 20 cluster 10 genes in the differentially expressed gene list, irrespective of direction, were plotted. **f**, Torin-2 upregulated most cluster 1/6 markers. **g**, PCA revealed convergence of narciclasine and Torin-2 at the proteomic level. PCA was calculated on log-normalized proteomic data. At the proteomic level, Torin-2 and narciclasine are similar and divergent from both control and camptothecin. **h**, Camptothecin upregulated cluster 10 markers at the proteomic level. Heat map showing the row-scaled, zero-centered expression values of proteomic data derived from compound-treated HMC3 microglia (24 h; $n = 3$ per treatment). Each column is a single sample, and each row is a single gene. Pairwise differential testing between DMSO control and each of our treated conditions was conducted using Welch's *t*-test with the Benjamini–Hochberg correction (FDR alpha < 0.05, LFC < 1). **I**, Camptothecin downregulates cluster 1/6 markers at the proteomic level. See also Extended Data Fig. 7 and 8 and Supplementary Tables 6 and 7. PMA, phorbol 13-myristate 12-acetate.

https://doi.org/10.1038/s41593-024-01764-7



	Description	median_tau_score
Calyculin	Protein phosphatase inhibitor	99.93
Torin-2	mTOR inhibitor	99.92
Tyrphostin-AG-126	ERK1 and ERK2 phosphorylation inhibitor, tyrosine kinase inhibitor	99.51
Torin-1	mTOR inhibitor, P13K inhibitor	99.51
Homoharringtonine	Apoptosis stimulant, protein synthesis inhibitor	99.48
Anisomycin	ONA synthesis inhibitor	99.45
Narciclasine	Coflilin signaling pathway activator, UM kinase activator, ROCK activator	99.40
Verrucarin-a	Protein synthesis inhibitor	99.37
Emetine	Protein synthesis inhibitor	99.35
AZD-8055	mTOR inhibitor	99.31
Homoharringtonine	Apoptosis stimulant, protein synthesis inhibitor	99.17
WYE-354	mTOR inhibitor	99.13
Emetine	Protein synthesis inhibitor	99.06
BNTX	Opioid receptor antagonist	98.93
JWE-035	Aurora kinase inhibitor	98.93
NVP-BEZ235	mTOR inhibitor, P13K inhibitor, protein kinase inhibitor	98.92
KU-0060648	ONA dependent protein kinase, ONA dependent protein kinase inhibitor, P13K inhibitor	98.90
Digitoxin	ATPase inhibitor	98.89
Digitoxigenin	ATPase inhibitor	98.80
KU-0063794	mTOR inhibitor	98.76
Sarmentogenin	ATPase inhibitor	98.71
Strophanthidin	ATPase inhibitor	98.71
JW-7-24-1	LCK Inhibitor	98.39
Cephaeline	Protein synthesis inhibitor	98.38
Pl-103	P13K inhibitor, mTOR inhibitor	98.38
Digoxin	ATPase inhibitor, ROR antagonist	98.23
Ouabain	ATPase inhibitor, alpha subunit binder	97.92
Cytochalasin-b	Microtubule inhibitor, phagocytosis inhibitor	97.87
Puromycin	Adenosine receptor agonist, protein synthesis inhibitor	97.85
MG-132	Proteasome inhibitor	97.81
Ouabain	ATPase inhibitor, alpha subunit binder	97.67
Digoxin	ATPase inhibitor, ROR antagonist	97.65
Puromycin	Adenosine receptor agonist, protein synthesis inhibitor	97.51
MK-5108	Aurora kinase inhibitor	97.41
81-2536	PLK inhibitor, apoptosis stimulant, cell cycle inhibitor, protein kinase inhibitor	97.38
	Glycogen synthase kinase inhibitor, protein synthesis inhibitor	97.28

representation of cells from the intermediate clusters 3 and 5 (Fig. 7e). Although snRNA-seq can be applied to frozen human brain tissue, it does not capture the cytoplasmic compartment[16,17]. Thus, as the microglial subtype families 2/4 and 1/6/7 represent more polarized and differentiated branches of the homeostatic trajectory, albeit in distinct directions, technical differences may impair our ability to resolve some of the more distal phenotypes on our differentiation trajectory in nuclear data. Nonetheless, cluster 5, part of the same family that exhibits association with AD pathology (Fig. 5c), shows a positive association with amyloid burden (Fig. 7h), while cluster 3 shows a negative association with the same trait (Fig. 7l) in the ROSMAP snRNA-seq data. Finally, annotation of the GBM dataset reveals high numbers of cells that map to the proliferative cluster 12, DAM2^{high} cluster 11, and cluster 8 (Fig. 7f). Upregulation of SPP1, a gene defining cluster 11, has previously been reported in GBM-associated myeloid cells and has been shown to correlate with worse survival in humans with GBM[70]. Comparison of subtype proportions with our reference dataset from neurodegenerative diseases reinforces the presence of a dramatic shift away from the core homeostatic gradient and toward more inflammatory myeloid subtypes in GBM (Fig. 7j). This is consistent with prior observations reporting shifts away from homeostatic phenotypes among myeloid cells found in the glioma microenvironment[65].

We thus demonstrated the utility of our dataset for annotating datasets from a wide variety of sources, such as diseases not captured in our dataset, snRNA-seq data and human iPS cell-derived microglial model systems. We identified shifts in phenotype that accord with those previously reported in GBM and demonstrate that iPS cell datasets, both in vitro and in vivo, capture an impressive amount of the microglial heterogeneity that we have identified in isolated, living human microglia.

Prediction and validation of compounds driving cluster-specific transcriptional signatures and subtype recapitulation in vitro

Next, we sought to leverage our data to understand how to use chemical perturbation to direct state transitions toward specific subtypes. This would enhance (1) in vitro modeling using iPS cell-derived microglia-like cells or monocyte-derived microglia-like cells[71] and (2) drug discovery as such tool compounds could provide leads for therapeutic development of in vivo microglial modulation. Here, we used the V1 database of the CMAP[20,21], a dataset that contains transcriptomic data associated with thousands of chemical perturbations across a wide array of cell lines. To increase the power of our initial analysis, we grouped related microglial subtypes together, querying CMAP using RNA signatures for three groups of microglial subtypes: clusters 1/6, clusters 2/4/9 and clusters 8/10, chosen because they capture two of the primary axes of variation among our microglial subtypes. An overview of our workflow is shown in Extended Data Fig. 8a. From our in silico CMAP analysis (representative example in Fig. 8a; Supplementary Table 7), we prioritized 14 compounds for validation. For our initial screen, we exposed the human microglial cell 3 (HMC3) line[72] to each of the 14 compounds guided by the dosage used in CMAP. Since HMC3 cells were not used in CMAP, we optimized concentrations of each compound to minimize effects on survival and morphology. We then tested the effects of each compound after 6-h and 24-h treatment by assessing the expression levels of two selected marker genes for each of the three groups of microglial subtypes using quantitative PCR with reverse transcription (RT–qPCR), repeating this experiment at least three times with different batches of HMC3 cells (Fig. 8b–d and Extended Data Fig. 8b–d). Four compounds met our predetermined criteria for the screen: Torin-2 and narciclasine both drove upregulation of marker genes associated with clusters 1 and 6, while camptothecin and phorbol 13-myristate 12-acetate drove upregulation of cluster 8 and cluster 10 marker genes. Our results for compounds associated with clusters 2, 4 and 9 were inconclusive, as the marker genes that we chose did not show significant upregulation with these compounds.

To assess the effects of our selected compounds at a broader scale, we profiled cells with bulk RNA-seq and shotgun proteomics after 24 h of treatment in a separate set of experiments. At the transcriptional level, camptothecin induced cluster 8 and 10 genes, such as *HLA-C*, *CXCR4* and *CYBA* (Fig. 8e). Interestingly, camptothecin also downregulated cluster 1/6 genes such as *QKI* and *ATM* (Extended Data Fig. 9a), supporting the transcriptional divergence of clusters 8/10 from 1/6 (Fig. 2c). As predicted, Torin-2 robustly drives the cluster 1/6 signature (Fig. 8f). Narciclasine does not appear to upregulate this signature (Extended Data Fig. 9b); however, GO annotation of genes differentially upregulated with narciclasine suggests a strong upregulation of metabolic pathways, such as nitrogen-compound containing metabolism and heterocyclic metabolism (Extended Data Fig. 9c), that we previously found to be strongly enriched in clusters 1/6 by GO annotation. Moreover, examining cluster 1/6 genes upregulated in Torin-2 and narciclasine suggests that the two compounds engage complementary, but separate aspects of the cluster 1/6 signature (Extended Data Fig. 9d), with narciclasine inducing genes such as *MEF2A* and *NUFIP5*, while Torin-2 upregulated genes such as *DENND3* and *ATM*. In contrast, at the proteomic level, principal component analysis (PCA) suggests that narciclasine and Torin-2 yield similar changes in proteomic profiles relative to both the dimethylsulfoxide (DMSO) control and camptothecin (Fig. 8g), suggesting the engagement of a different proteomic state. Interestingly, neither narciclasine nor Torin-2 clearly drive cluster 1/6 marker genes at the proteomic level (Fig. 8h). This may be because RNA-derived markers may be suboptimal to resolve proteome changes for these microglial subtypes given the known divergence between RNA and protein in microglia[73] and/or the short time course of our perturbation. On the other hand, camptothecin does drive strong upregulation of ten genes such as *HLA*-C and *CYBA* (Fig. 8l) and downregulation of cluster 1/6 genes such as *QKI* (Fig. 8h) at the proteomic level.

We thus identified and validated three tool compounds that polarize a human microglial model system (HMC3 cells) into different targeted states, presenting an approach that can be extended to develop a broader toolkit with which to manipulate microglial differentiation in vitro and potentially in vivo. Notably, one of our compounds, camptothecin, drives a robust signal toward cluster 10 that is detectable at both transcriptomic and proteomic levels; a compound with this property could conceptually be useful therapeutically to shift the distribution of human microglia in vivo away from clusters 1/6 that are strongly enriched in genes and traits associated with AD, MS and other diseases (Fig. 5) and toward the CD74^{hi} cluster 10 subset that we have previously reported to be associated with AD[4].

Discussion

Our understanding of human microglial heterogeneity has been transformed over the last 5 years by single-cell studies from many groups[4,8–10,13–15,74–79]. In this study, we aimed to maximize sampling of microglial diversity and to present a cross-disease microglial population structure derived from 215,658 live human microglia sampled from a wide array of CNS regions and conditions (Fig. 2). Our analysis explores the interconnection of different microglial subtypes, proposing divergent routes of differentiation with exclusive terminal endpoints as well as possible functional and metabolic shifts associated with these trajectories. We identify families of microglial subsets with enrichment of neuropsychiatric disease susceptibility genes. We validate our subtypes using a joint immunofluorescence and in situ hybridization staining protocol with automated image segmentation to evaluate morphological characteristics. Further, we demonstrate the utility of our reference by interpreting external data, including GBM data, a disease that we did not sample. Finally, we demonstrate how our dataset can be leveraged to identify tool compounds that recapitulate microglial subtypes, providing a path toward targeted therapeutic immunomodulation of microglia.

Consistent with other studies of ex vivo human microglia[4,13], we found that shifts in function, metabolism and association with disease genes fall along continuous axes radiating outwards from a central state. The primary axis of variation in our dataset lies between clusters 1 and 6 and clusters 2, 4 and 9. Clusters 1 and 6 are enriched for diseases genes and upregulate heterocyclic metabolism, while clusters 2, 4 and 9 are associated with oxidative metabolism and present a homeostatic-active phenotype. The second major axis of differentiation leads to clusters 8 (interleukin signaling) and 10 (complement and antigen presentation). These clusters may represent two tracks of microglial activation directed toward adaptive immune interaction. They are related to cluster 11 (DAM2^{high} microglia), which we had not observed in our prior study[4], probably because it represents a small minority of microglia. This is in marked contrast to murine brain data; this difference could stem from the accelerated kinetics of murine models compared to human disease, from the ~90 million years of evolution between mice and humans, or from the exhaustion of the DAM response in the protracted course of human disease. We can now define human-centric versions of the DAM that may be more informative in human studies. Thus, we have captured three primary axes of variation in our model: (1) a 1/6 versus 2/4/9 metabolic cline, (2) the 8/10 axis of immune response specialization, and (3) the DAM-like axis of activation that terminates in cluster 11. Of course, it remains unclear whether these different tracks of differentiation arise from different progenitor pools, or whether any given microglia are fully capable of attaining all possible states. Notably, our in situ validation efforts confirm the existence of these axes using key marker genes and suggest functional differences among these subsets. For example, CD74^{high} cells (cluster 10) and SPP1^{high} cells (cluster 11) are both less ramified than other microglia, while CXCR4^{+} (cluster 8) cells exhibit more ramification. Our in situ pipeline also offers opportunities to explore spatial localization of RNA, as we detect transcripts even in distal microglial processes (Fig. 6).

Label transfer offers the opportunity to use our existing structure to analyze external datasets, and our results suggest that snRNA-seq data, xenografted human iPS cell microglia[66] and even an in vitro human iPS cell microglial model system[68] recapitulate an impressive amount of the heterogeneity found among primary microglia. The reference also recovers the role of cluster 11 microglia in GBM[64,70], a disease not sampled in the resource. Ultimately, our resource can facilitate the annotation of smaller datasets, enhancing the analysis of the less common microglial subtypes.

Our reference also forms a foundation that can be leveraged to identify new tools that enable functional studies of subtypes by recapitulating them in vitro. Our prioritized compounds include Torin-2, an mTOR inhibitor that improves survival in animal models of GBM[80] and has neuroprotective effects[81], camptothecin, a topoisomerase inhibitor with neuroprotective effects in murine PD[82], and narciclasine, a pleiotropic drug that inhibits the NF-κB pathway[83]. Camptothecin is particularly interesting as it enhances its target cluster 10 signature while also suppressing the cluster 1/6 signature at both the transcriptomic and proteomic levels. These effects may be particularly relevant to therapeutic development in AD as clusters 1/6 are enriched for AD susceptibility genes (Fig. 5), and we have previously reported an association of the CD74^{high} cluster 10 in AD[4]. This compound prioritization effort offers a generalizable strategy for identifying compounds that may drive distinct microglial subtypes and provide a path toward targeted immunomodulation therapies.

Our work has limitations. First, 'control' donors who have no clinical manifestation of a neurological disorder at the time of death and do not fulfill pathological criteria for a disease are rare and only one came to autopsy during this study. Second, because our study design prioritized uncovering diversity in microglial subtypes across diseases, the samples do not enable association studies for disease, region, sex or other variables. Third, our efforts to mitigate technical and batch effects likely suppressed some true biological variation among samples. This was a necessary tradeoff to achieve our goal, but it also means that we were unable to capture more subtle heterogeneity that may be present in different diseases or regions. However, the differences in single-cell chemistry did not affect the cell purification process, so we do not expect any effects on viability. Fourth, our cross-sectional data rely on algorithmic inference in proposing trajectories, but they recapitulate and extend similar findings seen in studies of murine microglial heterogeneity[6,46–49]. Fifth, our moderate sample size resolved the proliferative cluster 12 (<1% of microglia); however, less frequent subtypes may exist and will require larger sample sizes to be discovered. Finally, our HMC3 model system may be limiting; however, we mitigate this somewhat by reporting stable expression of microglial markers across passages of HMC3 (Supplementary Fig. 1). Indeed, this limitation is shared with iPS cell-derived in vitro and xenograft models that do not recapitulate the full breadth of human microglial heterogeneity (Fig. 7b).

Here, we have created a new cross-disease resource that describes human microglial heterogeneity. Through our validation efforts, we have used our reference to expand the community's microglial toolkit with robust approaches (1) to identify microglial subtypes in situ and, more broadly, to capture morphology and targeted RNA expression from individual microglia in human tissue, (2) to transfer the reference labels to multiple data types, and (3) to recapitulate certain subtypes in vitro via chemical perturbation. The latter results outline a path for targeted development of immunomodulatory strategies that leverages our understanding of microglial subtypes implicated in different diseases.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41593-024-01764-7.

References

1. Ginhoux, F., Lim, S., Hoeffel, G., Low, D. & Huber, T. Origin and differentiation of microglia. *Front. Cell. Neurosci.* **7**, 45 (2013).
2. Li, Q. & Barres, B. A. Microglia and macrophages in brain homeostasis and disease. *Nat. Rev. Immunol.* **18**, 225–242 (2018).
3. Liddelow, S. A. et al. Neurotoxic reactive astrocytes are induced by activated microglia. *Nature* **541**, 481–487 (2017).
4. Olah, M. et al. Single cell RNA sequencing of human microglia uncovers a subset associated with Alzheimer's disease. *Nat. Commun.* **11**, 6129 (2020).
5. Butovsky, O. & Weiner, H. L. Microglial signatures and their role in health and disease. *Nat. Rev. Neurosci.* **19**, 622–635 (2018).
6. Keren-Shaul, H. et al. A unique microglia type associated with restricting development of Alzheimer's disease. *Cell* **169**, 1276–1290 (2017).
7. Ayata, P. et al. Epigenetic regulation of brain region-specific microglia clearance activity. *Nat. Neurosci.* **21**, 1049–1060 (2018).
8. Gerrits, E. et al. Distinct amyloid-β and tau-associated microglia profiles in Alzheimer's disease. *Acta Neuropathol.* **141**, 681–696 (2021).
9. Masuda, T., Sankowski, R., Staszewski, O. & Prinz, M. Microglia heterogeneity in the single-cell era. *Cell Rep.* **30**, 1271–1281 (2020).
10. Chen, Y. & Colonna, M. Microglia in Alzheimer's disease at single-cell level. Are there common patterns in humans and mice? *J. Exp. Med.* **218**, e20202717 (2021).
11. Colonna, M. & Brioschi, S. Neuroinflammation and neurodegeneration in human brain at single-cell resolution. *Nat. Rev. Immunol.* **20**, 81–82 (2020).

12. Dumas, A. A., Borst, K. & Prinz, M. Current tools to interrogate microglial biology. *Neuron* **109**, 2805–2819 (2021).
13. Masuda, T. et al. Spatial and temporal heterogeneity of mouse and human microglia at single-cell resolution. *Nature* **566**, 388–392 (2019).
14. Kracht, L. et al. Human fetal microglia acquire homeostatic immune-sensing properties early in development. *Science* https://doi.org/10.1126/science.aba5906 (2020).
15. Young, A. M. H. et al. A map of transcriptional heterogeneity and regulatory variation in human microglia. *Nat. Genet.* **53**, 861–868 (2021).
16. Thrupp, N. et al. Single-nucleus RNA-seq is not suitable for detection of microglial activation genes in humans. *Cell Rep.* **32**, 108189 (2020).
17. Bakken, T. E. et al. Single-nucleus and single-cell transcriptomes compared in matched cortical cell types. *PLoS ONE* **13**, e0209648 (2018).
18. Marsh, S. E. et al. Dissection of artifactual and confounding glial signatures by single-cell sequencing of mouse and human brain. *Nat. Neurosci.* **25**, 306–316 (2022).
19. Mattei, D. et al. Enzymatic dissociation induces transcriptional and proteotype bias in brain cell populations. *Int. J. Mol. Sci.* **21**, 7944 (2020).
20. Subramanian, A. et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell* **171**, 1437–1452 (2017).
21. Lamb, J. et al. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* **313**, 1929–1935 (2006).
22. Olah, M. et al. A transcriptomic atlas of aged human microglia. *Nat. Commun.* **9**, 539 (2018).
23. Stuart, T. et al. Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902 (2019).
24. Gerrits, E., Heng, Y., Boddeke, E. W. G. M. & Eggen, B. J. L. Transcriptional profiling of microglia; current state of the art and future perspectives. *Glia* **68**, 740–755 (2020).
25. Haage, V. et al. Comprehensive gene expression meta-analysis identifies signature genes that distinguish microglia from peripheral monocytes/macrophages in health and glioma. *Acta Neuropathol. Commun.* **7**, 20 (2019).
26. Jurga, A. M., Paleczna, M. & Kuter, K. Z. Overview of general and discriminating markers of differential microglia phenotypes. *Front. Cell. Neurosci.* **14**, 198 (2020).
27. Masuda, T. et al. Novel Hexb-based tools for studying microglia in the CNS. *Nat. Immunol.* **21**, 802–815 (2020).
28. Van Hove, H. et al. A single-cell atlas of mouse brain macrophages reveals unique transcriptional identities shaped by ontogeny and tissue environment. *Nat. Neurosci.* **22**, 1021–1035 (2019).
29. Kierdorf, K., Masuda, T., Jordão, M. J. C. & Prinz, M. Macrophages at CNS interfaces: ontogeny and function in health and disease. *Nat. Rev. Neurosci.* **20**, 547–562 (2019).
30. Lee, J. et al. QUAKING regulates microexon alternative splicing of the Rho GTPase pathway and controls microglia homeostasis. *Cell Rep.* **33**, 108560 (2020).
31. Ren, J. et al. Qki is an essential regulator of microglial phagocytosis in demyelination. *J. Exp. Med.* **218**, e20190348 (2021).
32. Nguyen, A. T. et al. APOE and TREM2 regulate amyloid-responsive microglia in Alzheimer's disease. *Acta Neuropathol.* **140**, 477–493 (2020).
33. Zhou, Y. et al. Human and mouse single-nucleus transcriptomics reveal TREM2-dependent and -independent cellular responses in Alzheimer's disease. *Nat. Med.* **26**, 131–142 (2020).
34. Tasic, B. et al. Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nat. Neurosci.* **19**, 335–346 (2016).
35. Tasic, B. et al. Shared and distinct transcriptomic cell types across neocortical areas. *Nature* **563**, 72–78 (2018).
36. Ashburner, M. et al. Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).
37. The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res.* **47**, D330–D338 (2019).
38. Alexa, A. & Rahnenfuhrer, J. topGO: enrichment analysis for Gene Ontology. Bioconductor version: release 3.13. https://doi.org/10.18129/B9.bioc.topGO (2021).
39. Sayols, S. rrvgo: a Bioconductor package to reduce and visualize Gene Ontology terms. https://doi.org/10.18129/B9.bioc.rrvgo (2020).
40. Jassal, B. et al. The reactome pathway knowledgebase. *Nucleic Acids Res.* **48**, D498–D503 (2020).
41. Wu, G. & Haw, R. Functional interaction network construction and analysis for disease discovery. *Methods Mol. Biol.* **1558**, 235–253 (2017).
42. Yu, G., Wang, L. -G., Han, Y. & He, Q. -Y. clusterProfiler: an R Package for comparing biological themes among gene clusters. *OMICS.* **16**, 284–287 (2012).
43. Trapnell, C. et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32**, 381–386 (2014).
44. Cao, J. et al. The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566**, 496–502 (2019).
45. Qiu, X. et al. Reversed graph embedding resolves complex single-cell trajectories. *Nat. Methods* **14**, 979–982 (2017).
46. Mathys, H. et al. Temporal tracking of microglia activation in neurodegeneration at single-cell resolution. *Cell Rep.* **21**, 366–380 (2017).
47. Sala Frigerio, C. et al. The major risk factors for Alzheimer's disease: age, sex, and genes modulate the microglia response to Aβ plaques. *Cell Rep.* **27**, 1293–1306 (2019).
48. Ellwanger, D. C. et al. Prior activation state shapes the microglia response to antihuman TREM2 in a mouse model of Alzheimer's disease. *Proc. Natl Acad. Sci. USA* **118**, e2017742118 (2021).
49. Krasemann, S. et al. The TREM2–APOE pathway drives the transcriptional phenotype of dysfunctional microglia in neurodegenerative diseases. *Immunity* **47**, 566–581 (2017).
50. Marschallinger, J. et al. Lipid-droplet-accumulating microglia represent a dysfunctional and proinflammatory state in the aging brain. *Nat. Neurosci.* **23**, 194–208 (2020).
51. International Multiple Sclerosis Genetics Consortium. Multiple sclerosis genomic map implicates peripheral immune cells and microglia in susceptibility. *Science* **365**, eaav7188 (2019).
52. Buniello, A. et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019).
53. Sekar, A. et al. Schizophrenia risk from complex variation of complement component 4. *Nature* **530**, 177–183 (2016).
54. David, A. B., Julie, A. S., Zoe, A. & Robert, S. W. Overview and findings from the Religious Orders Study. *Curr. Alzheimer Res.* **9**, 628–645 (2012).
55. David, A. B. et al. Overview and findings from the Rush Memory and Aging Project. *Curr. Alzheimer Res.* **9**, 646–663 (2012).
56. Patrick, E. et al. A cortical immune network map identifies distinct microglial transcriptional programs associated with β-amyloid and Tau pathologies. *Transl. Psychiatry* **11**, 50 (2021).
57. Wang, F. et al. RNAscope. *J. Mol. Diagn.* **14**, 22–29 (2012).
58. Carpenter, A. E. et al. CellProfiler: image analysis software for identifying and quantifying cell phenotypes. *Genome Biol.* **7**, R100 (2006).
59. McQuin, C. et al. CellProfiler 3.0: next-generation image processing for biology. *PLoS Biol.* **16**, e2005970 (2018).

60. Kamentsky, L. et al. Improved structure, function and compatibility for CellProfiler: modular high-throughput image analysis software. *Bioinformatics* **27**, 1179–1180 (2011).
61. Stirling, D. R. et al. CellProfiler 4: improvements in speed, utility and usability. *BMC Bioinf.* **22**, 433 (2021).
62. Stence, N., Waite, M. & Dailey, M. E. Dynamics of microglial activation: a confocal time-lapse analysis in hippocampal slices. *Glia* **33**, 256–266 (2001).
63. Green, G. S. et al. Cellular communities reveal trajectories of brain ageing and Alzheimer's disease. *Nature* **633**, 634–645 (2024).
64. Yuan, J. et al. Single-cell transcriptome analysis of lineage diversity in high-grade glioma. *Genome Med.* **10**, 57 (2018).
65. Hambardzumyan, D., Gutmann, D. H. & Kettenmann, H. The role of microglia and macrophages in glioma maintenance and progression. *Nat. Neurosci.* **19**, 20–27 (2016).
66. Hasselmann, J. et al. Development of a chimeric model to study and manipulate human microglia in vivo. *Neuron* **103**, 1016–1033 (2019).
67. Claes, C. et al. Plaque-associated human microglia accumulate lipid droplets in a chimeric model of Alzheimer's disease. *Mol. Neurodegener.* **16**, 50 (2021).
68. Dräger, N. M. et al. A CRISPRi/a platform in human iPSC-derived microglia uncovers regulators of disease states. *Nat. Neurosci.* **5**, 1149–1162 (2022).
69. Chen, T. & Guestrin, C. XGBoost: a scalable tree boosting system. in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794 https://doi.org/10.1145/2939672.2939785 (Association for Computing Machinery, 2016).
70. Szulzewsky, F. et al. Glioma-associated microglia/macrophages display an expression profile different from M1 and M2 polarization and highly express Gpnmb and Spp1. *PLoS ONE* **10**, e0116644 (2015).
71. Ryan, K. J. et al. A human microglia-like cellular model for assessing the effects of neurodegenerative disease gene variants. *Sci. Transl. Med.* **9**, eaai7635 (2017).
72. Dello Russo, C. et al. The human microglial HMC3 cell line: where do we stand? A systematic literature review. *J. Neuroinflammation* **15**, 259 (2018).
73. Boutej, H. et al. Diverging mRNA and protein networks in activated microglia reveal SRSF3 suppresses translation of highly upregulated innate immune transcripts. *Cell Rep.* **21**, 3220–3233 (2017).
74. Miedema, A. et al. Brain macrophages acquire distinct transcriptomes in multiple sclerosis lesions and normal appearing white matter. *Acta Neuropathol. Commun.* **10**, 8 (2022).
75. Absinta, M. et al. A lymphocyte–microglia–astrocyte axis in chronic active multiple sclerosis. *Nature* **597**, 709–714 (2021).
76. Patel, T. et al. Transcriptional landscape of human microglia implicates age, sex, and *APOE* -related immunometabolic pathway perturbations. *Aging Cell* **21**, e13606 (2022).
77. Alsema, A. M. et al. Profiling microglia from Alzheimer's disease donors and non-demented elderly in acute human postmortem cortical tissue. *Front. Mol. Neurosci.* **13**, 134 (2020).
78. Li, Y. et al. Decoding the temporal and regional specification of microglia in the developing human brain. *Cell Stem Cell* **29**, 620–634 (2022).
79. Kumar, P. et al. Single-cell transcriptomics and surface epitope detection in human brain epileptic lesions identifies pro-inflammatory signaling. *Nat. Neurosci.* **25**, 956–966 (2022).
80. Amin, A. G. et al. Targeting the mTOR pathway using novel ATP-competitive inhibitors, Torin1, Torin2 and XL388, in the treatment of glioblastoma. *Int. J. Oncol.* **59**, 83 (2021).
81. Johnson, S. C. et al. mTOR inhibition alleviates mitochondrial disease in a mouse model of Leigh syndrome. *Science* **342**, 1524–1528 (2013).
82. He, D. et al. Camptothecin regulates microglia polarization and exerts neuroprotective effects via activating AKT/Nrf2/HO-1 and inhibiting NF-κB pathways in vivo and in vitro. *Front. Immunol.* **12**, 619761 (2021).
83. Stark, A. et al. Narciclasine exerts anti-inflammatory actions by blocking leukocyte-endothelial cell interactions and down-regulation of the endothelial TNF receptor 1. *FASEB J.* **33**, 8771–8781 (2019).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

[1]Center for Translational & Computational Neuroimmunology, Department of Neurology, Columbia University Irving Medical Center, New York, NY, USA. [2]Department of Systems Biology, Columbia University Irving Medical Center, New York, NY, USA. [3]Medical Scientist Training Program, Columbia University Irving Medical Center, New York, NY, USA. [4]Taub Institute for Research on Alzheimer's Disease and the Aging Brain, Columbia University Irving Medical Center, New York, NY, USA. [5]Edmond & Lily Safra Center for Brain Sciences, The Hebrew University of Jerusalem, Jerusalem, Israel. [6]Department of Neurology, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA. [7]Neuropathology Service, C.S. Kubik Laboratory for Neuropathology, Massachusetts General Hospital/Harvard Medical School, Boston, MA, USA. [8]Department of Pharmacology, UT Health San Antonio, San Antonio, TX, USA. [9]Glenn Biggs Institute for Alzheimer's and Neurodegenerative Diseases, UT Health San Antonio, San Antonio, TX, USA. [10]Banner Sun Health Research Institute, Sun City, AZ, USA. [11]Department of Neurology, University of Colorado, and Rocky Mountain Multiple Sclerosis Center at the University of Colorado, Aurora, CO, USA. [12]Proteomics and Macromolecular Crystallography Shared Resource, Herbert Irving Comprehensive Cancer Center, New York, NY, USA. [13]Department of Pathology and Cell Biology, Columbia University Irving Medical Center, New York, NY, USA. [14]Department of Neurology, Columbia University Irving Medical Center, New York, NY, USA. [15]Department of Pathology, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [16]Movement Disorders Division, Neurological Institute, Tel Aviv Sourasky Medical Center, Tel Aviv, Israel. [17]Eleanor and Lou Gehrig ALS Center, Columbia University Medical Center, New York, NY, USA. [18]Rush Alzheimer's Disease Center, Rush University Medical Center, Chicago, IL, USA. [19]Department of Biochemistry and Molecular Biophysics, Columbia University Irving Medical Center, New York, NY, USA. [20]These authors contributed equally: John F. Tuddenham, Mariko Taga, Verena Haage, Marta Olah, Vilas Menon, Philip L. De Jager. ✉e-mail: pld2115@cumc.columbia.edu

Methods

Source of CNS specimens

Details of the acquisition of autopsy samples from Rush University Medical Center/Rush Alzheimer's Disease Center (RADC)[54,55] in Chicago (Dr. Bennett) and Columbia University Medical Center/New York Brain Bank in New York (Drs. Vonsattel and Teich)[84], as well as surgically resected brain specimens from Brigham and Women's Hospital in Boston (Drs. Sarkis, Cosgrove, Helgager, Golden and Pennell) are detailed in our prior publication[4]. In addition, samples were obtained from donation programs at Massachusetts General Hospital, Boston (Drs. Bradley T. Hyman and Matthew Frosch), Banner Sun Health Research Institute (Dr. Thomas G Beach) and Rocky Mountain MS Center (Dr. John Corboy). All brain specimens were obtained through informed consent and/or a brain donation program at the respective organizations. All procedures and research protocols were approved by the corresponding ethical committees of our collaborator's institutions as well as the Institutional Review Board of Columbia University Medical Center (protocol AAAR4962). For a detailed description of the brain regions sampled, clinical diagnosis, age and sex of the donors, see Supplementary Table 1.

Shipping of brain specimens

After weighing, the tissue was placed in ice-cold transportation medium (Hibernate-A medium (Gibco, A1247501) containing 1% B27 serum-free supplement (Gibco, 17504044) and 1% GlutaMax (Gibco, 35050061)) and shipped overnight at 4 °C with priority shipping.

Microglia isolation, cell hashing and sorting

The isolation of microglia was performed according to our published protocol[22], with minor modifications. In case of the cortical autopsy samples (BA9/46, BA4, BA17/18/19), the cortex (gray matter) and the underlying white matter (subcortical white matter) were dissected under a stereomicroscope. The subcortical white matter samples were not used in this study. The epilepsy surgery samples of temporal lobe (BA20/21) were processed without dissection as in this case the cortical white and gray matter were not always distinguishable due to the surgical procedure. The substantia nigra and the thalamus were dissected by separation from the surrounding white matter tracts. The hippocampus samples contained the dentate gyrus, CA4/CA3/CA2 and CA1 regions, both white and gray matter. The spinal cord sample was sampled at the level of the lumbar section and included both white and gray matter. The anterior watershed area deep white matter did not need any further dissection. All steps of the protocol were performed on ice. The dissected tissue was placed in HBSS (Lonza, 10-508F) and weighed. Subsequently, the tissue was homogenized in a 15-ml glass tissue grinder, using 0.5 g at a time. The resulting homogenate was filtered through a 70-µm filter and spun down at 300 rcf for 10 min. The pellet was resuspended in 2 ml staining buffer (RPMI (Fisher, 72400120) containing 1% B27) per 0.5 g of initial tissue and incubated with anti-myelin magnetic beads (Miltenyi, 130-096-733) for 15 min according to the manufacturer's specification. The homogenate was then washed once with staining buffer, and the myelin was depleted using Miltenyi large separation columns (Miltenyi, 130-042-202). The cell suspension was spun down and was then incubated with anti-CD11b Alexa Fluor 488 (BioLegend, 301318) and anti-CD45 Alexa Fluor 647 (BioLegend, 304018) antibodies as well as 7AAD (BD Pharmingen, 559925) and cell hashing antibodies (for catalog numbers of cell hashing antibodies, see Supplementary Table 1) for 20 min on ice. Subsequently, the cell suspension was washed twice with staining buffer, filtered through a 70-µm filter and the CD11b$^+$/CD45$^+$/7AAD$^-$ cells or CD45$^+$/7AAD$^-$ cells were sorted on a BD FACS Aria II or BD Influx cell sorter. Cells from each brain region were sorted in a separate A1 well of a 96-well PCR plate (Eppendorf, 951020401) containing 100 µl of PBS buffer with 0.3% BSA. Following sorting, cells from different brain regions were combined and immediately submitted to single-cell capture, reverse transcription and library construction on the 10x Chromium platform. All sorting was performed using a 100-µm nozzle. The sorting times varied according to the quality of the sample but was usually between 10 min and 20 min per sample. The sorting speed was kept between 8,000 and 10,000 events per second.

10x Genomics Chromium single-cell 3′ library construction

Cell capture, amplification and library construction on the 10x Genomics Chromium platform were performed according to the manufacturer's publicly available protocol. Briefly, viability was assessed by trypan blue exclusion assay, and cell density was adjusted to 175 cells per microliter. In total, 7,000 cells were then loaded onto a single channel of a 10x Chromium chip for each sample. The 10x Genomics Chromium technology enables 3′ digital gene expression profiling of thousands of cells from a single sample by separately indexing each cell's transcriptome. First, thousands of cells are partitioned into nanoliter-scale Gel Bead-In-EMulsions (GEMs). Within one GEM, all generated cDNA share a common 10x barcode. Libraries were generated and sequenced from the cDNA, and the 10x barcodes were used to associate individual reads back to the individual partitions. To achieve single-cell resolution, the cells were delivered at a limiting dilution. Upon dissolution of the Single Cell 3′ Gel Bead in a GEM, primers containing (i) an Illumina R1 sequence (read 1 sequencing primer), (ii) a 16-nucleotide 10x Barcode, (iii) a 10-nucleotide UMI, and (iv) a poly-dT primer sequence were released and mixed with cell lysate and Master Mix. Incubation of the GEMs then produced barcoded, full-length cDNA from poly-adenylated mRNA. After incubation, the GEMs were broken and the pooled fractions were recovered. Full-length, barcoded cDNA was then amplified by PCR to generate sufficient mass for library construction. Enzymatic fragmentation and size selection were used to optimize the cDNA amplicon size before library construction. R1 (read 1 primer sequence) was added to the molecules during GEM incubation. P5, P7, a sample index and R2 (read 2 primer sequence) were added during library construction via end repair, A-tailing, adaptor ligation and PCR. The final libraries contained the P5 and P7 primers used in Illumina bridge amplification. The described protocol produced Illumina-ready sequencing libraries. A single-cell 3′ library comprises standard Illumina paired-end constructs that begin and end with P5 and P7. The single-cell 3′ 16-bp 10x Barcode and 10-bp UMI are encoded in read 1, while read 2 is used to sequence the cDNA fragment. Sample index sequences were incorporated as the i7 index read. Read 1 and read 2 are standard Illumina sequencing primer sites used in paired-end sequencing. Sequencing the library produced a standard Illumina BCL data output folder. The BCL data include the paired-end read 1 (containing the 16-bp 10x Barcode and 10-bp UMI) and read 2 and the sample index in the i7 index read.

Batch structure and sequencing

Tissue specimens were processed upon receipt. The different brain regions from the same donor were processed and hashed in parallel and loaded in a single well of a 10x Chromium 3′ chip as described above. Accordingly, each sample (containing multiple brain regions from the same donor) constitutes one batch for all three procedures (microglia isolation, cell capture and library construction). All sequencing was performed on either an Illumina HiSeq 4000 or a NovaSeq 6000 machine. For specifics on the sequencing machines and QC metrics regarding the generated reads, see Supplementary Table 1.

Single-cell RNA-seq data processing, alignment and hashtag deconvolution

The majority of our downstream analysis was conducted using the R programming language (v.4.0.5 for harmonization and clustering, v.4.1.0 for annotation and downstream visualization)[85] and the RStudio[86] integrated development environment. Cell Ranger v.3.1.0 with default parameters was used to demultiplex and align our barcoded reads with the Ensembl transcriptome annotation (downloaded March 2019, GRCh38.91). A recent report[87] suggested that filtering cells with greater than 10% mitochondrial reads is the preferred baseline for human tissue, and that for brain tissue a higher threshold may even be optimal. Thus, a mitochondrial percentage that was the higher of either 10% of reads or the two absolute deviations above the median for mitochondrial reads within the sample was chosen as a threshold. Cells below this threshold with between 500 and 10,000 UMIs were retained for downstream analysis. All ribosomal genes, mitochondrial genes and pseudogenes were removed, as they interfered with the downstream differential gene expression. For samples where we used cell hashing to combine regions or subjects in a single sequencing run, droplets were demultiplexed using the following workflow. For each hashtag oligonucleotide (HTO), a mixture model with two components was fitted to the HTO counts using an expectation–maximization algorithm. The component with the smaller mean (negative component) represents droplets that were not tagged with the HTO, whereas the component with the larger mean (positive components) represents droplets that were tagged. We then assign each droplet to either the negative or positive component based on its posterior probability. Droplets that were assigned to the negative component for all HTOs as well as multiplets were discarded. Singlets with high uncertainty, that is, without confident assignment to either the negative or positive component, were discarded as well, leaving only high-certainty singlets for downstream analysis. The method is implemented in the R package demuxmix[88]. Some of our hashtag data had lower overall counts, and thus, the demuxmix model was unable to effectively segregate distributions for some hashtags in several samples. These samples were identified as having high percentages of negative/uncertain cells with demuxmix. In these cases, to try and recover cells for further analysis, the problematic hashtags were reclassified using one of two different algorithms, a demixing algorithm developed for MULTI-seq[89] or HTOdemux from Seurat (v.3.2.0)[23]. Hashtag classifications were merged, and doublet/negative/uncertain cell removal proceeded as described earlier.

Batch correction

Striking differences were observed in the distributions of UMI counts between 10x v2 and v3 chemistry. As this was driving differential clustering, count matrices from v3 samples were downsampled by 50% using the DropletUtils[90] package in R to achieve comparable UMI distributions across the two technologies (Supplementary Table 1 and Extended Data Fig. 2d). Next, after testing a series of recently published batch correction tools, SCTransform[91] combined with mNN[92] was chosen to mitigate batch effect in our dataset. A range of numbers of differentially expressed genes (3,000–6,000) and components (20–40) were tested, and 4,500 differentially expressed genes and 40 components were used for downstream analysis. Using these parameters, the full pipeline is as follows: SCTransform, which normalizes for minor differences in sequencing depth, was performed on each individual batch, then corrected counts were log normalized with the NormalizeData function in the Seurat package. Subsequently, processed datasets were merged on the corrected count matrix using the fastmNN algorithm accessed through the RunFastMNN function in the SeuratWrappers package. Library preparation batch is confounded with samples, and diseases are confounded with technical variables in our dataset, due to the necessity to process all tissue immediately upon receipt and the irregular schedule by which samples are received. As mNN is a harsher integration approach when compared to other commonly used tools, our pipeline is likely to have removed relevant biological signal in integrating the datasets. However, our priority was to obtain a robust cluster structure across the diverse brain regions and diseases found in our dataset while avoiding the issue of spurious signal from batch effects driving separate clustering, which motivated the approach that we have described here.

Clustering

The graph-based clustering approach implemented in Seurat (v.3)[23] was used to cluster our cells. In brief, a *k*-nearest neighbors graph based on Euclidean distance in our corrected mNN space was calculated and used to derive refined edge weights based on Jaccard similarity. The Louvain algorithm was then applied to iteratively delineate a population structure on our dataset. This was implemented with the FindNeighbors and FindClusters functions in Seurat. A UMAP projection of our dataset was computed with the RunUMAP function for visualization (Extended Data Fig. 1a). Contaminating red blood cells from our dataset were removed using classical markers (*HBB*/*HBA*), and microglia were subclustered using an identical integration and clustering pipeline (Fig. 2a). Any microglial subsets with fewer than 100 cells were discarded. Basic quality metrics are shown in Extended Data Fig. 2a–f and reported in Supplementary Table 1.

Validation of cluster stability

To evaluate cluster stability, a post hoc pairwise machine learning approach was used to evaluate the similarity between clusters. Logically, one would expect that operating with a simple classifier, separation of cells from divergent clusters would be simpler and lead to higher accuracy of prediction, while separation of cells from clusters that are transcriptionally overlapping would be more difficult, and would thus lead to a lower accuracy of prediction. As such, we trained simple machine learning models on pairs of different clusters, using accuracy of prediction as a proxy for homology of individual pairs of clusters. The top 10,000 variable genes in the dataset were identified by applying the FindVariableFeatures function from Seurat on the log-normalized RNA expression data from our dataset. Using these features, keras[93] was used to train a multilayer perceptron classifier to distinguish each pair of clusters. After basic hyperparameter optimization, the following parameters were chosen for the model: the rmsprop optimizer, a one-layer structure with 500 nodes in the hidden layer, the tanh activation function for the dense layer and a sigmoid activation function for the output layer. Ten epochs and a batch size of 100 were used for training. As our only concern was the raw accuracy of classification, mean squared error was used as the loss function. For each pair of clusters, the data were split fivefold, then the classifier was trained on 80% of the dataset and the remaining 20% was classified. This process was repeated five times to classify every cell in our dataset once. This entire pipeline was then repeated 50 times for each pair of clusters. Cells that were ambiguously classified (<40 times of 50) to the same cluster were designated intermediate cells. A threshold of <20% overlap between clusters was chosen as the threshold for merging clusters; in our model, no clusters met this parameter and as such, the original 12-cluster model was retained. As such, this analysis leverages the ability of simple classifiers to separate distinct clusters as a direct metric for the transcriptional similarity of said clusters, and conducted across all combinations of clusters, it provides an overarching metric for the similarity of each cluster to every other cluster in our dataset. The constellation diagram shown in Fig. 3a depicts the results of this analysis: edges between clusters represent the percentage of total cells classified as intermediate, and the area of each node is scaled to correspond to the overall size of the corresponding cluster. As orthogonal validation of our microglial clustering parameters, a resampling clustering approach was also used to assess cluster robustness. Over 100 iterations, 75% of the cells from our dataset were randomly sampled and our clustering pipeline was run with identical parameters. This was also done in a pairwise fashion to examine fluidity between individual pairs of clusters. For each pair of clusters, the frequency with which cells assigned to one cluster in our original clustering were reclustered into the cluster that contained fewer cells with the same original classification was recorded. Clusters remain generally stable, with most of the overlap being found between adjacent, closely related clusters or the intermediate clusters in our dataset. The results are shown in

Extended Data Fig. 2g, visualized using the corrplot package[94]. Each of our microglial subtypes was also subclustered, but no stable, distinct subclusters were identified.

Identification of cluster-defining gene sets

To identify cluster-defining gene sets, the FindMarkers function in Seurat was used to implement a pairwise testing approach. We prioritized differentially expressed genes that could best delineate a given cluster from each other cluster in our dataset. To do so, MAST[95] was applied to normalized count data from the 'RNA' assay of the Seurat object to find differentially expressed genes between every combination of pairs of clusters. Within each cluster, all the differentially expressed genes that were identified with this approach were filtered to only include those that were only found to be differentially expressed in one direction (either up or down). Any genes that were found to be upregulated in comparison to some clusters but downregulated in comparisons to other clusters or vice versa were removed from our downstream analysis. Furthermore, to ensure that the specific cluster-defining genes were prioritized, upregulated genes were ranked by the number of comparisons in which they were upregulated, and only those upregulated in three or more comparisons were used for downstream analyses. An identical process was applied for downregulated genes. Full marker gene lists are reported in Supplementary Table 2.

Enrichment of DAM signature gene sets

The DAM signature gene sets for 'DAM1', 'DAM2' and 'homeostatic microglia' described in ref. 6 were examined separately in our analysis. These sets consist of two sets of genes upregulated in the DAM trajectory, as well as homeostatic microglial genes known to be downregulated in DAM microglia. The overlap of 'DAM1' and 'DAM2' gene sets with upregulated cluster-defining genes, and the overlap of the 'homeostatic microglia' gene set with downregulated cluster-defining genes ('Identification of cluster-defining gene sets') was examined using a hypergeometric test with an FDR-corrected threshold for significance at $q = 0.01$ (ref. 96). The results of this analysis were visualized in a heat map where the color intensity corresponds to the $-\log_{10} P$ value of the FDR q value for enrichment of DAM1/DAM2 genes in upregulated genes or homeostatic genes in downregulated genes from our clusters (Fig. 2e).

Monocle3 pseudotime analysis

As an orthogonal method of evaluating the continuity of different microglial states in our cluster structure, the Monocle3 algorithm was used to build a pseudotime trajectory across our dataset (Extended Data Fig. 4e). Using the Seurat interface to Monocle3 found in SeuratWrappers, the Seurat object was converted into a Monocle data object, and a pseudotime trajectory was derived using the 'learn_graph' function, retaining the clustering assignments from our original clustering pipeline (see 'Clustering').To establish an originating point, the pseudotime root was placed on the border of clusters 2 and 3, as these cells had the strongest gene expression signature for classic homeostatic microglia and few differentiating genes, suggesting that they could form the basal state for microglia from which they would differentiate into other states. Interestingly, this state was best captured by choosing cells with maximal *AVP* expression, a marker of hematopoietic stem cells[97] that is frequently used to mark the root cells in hematopoietic pseudotime tracing.

Functional annotation of microglial clusters

To perform functional annotation of microglial clusters, upregulated or downregulated gene lists for each cluster were defined as genes upregulated in three or more comparisons or downregulated in two or more comparisons. Annotation of these gene lists was performed with several resources: GO by way of TopGO[38] as well as Reactome[40] pathway analysis with clusterProfiler[42]. For GO analysis, we conducted analysis with biological process annotation. For all functional analysis, the Benjamini–Hochberg correction[98] was used to correct P values for multiple testing. Corrected P values below a threshold of 0.01 were chosen as significant for both GO and Reactome results. GO results were aggregated and summarized by use of the rrvgo[39] package. Aggregated results of pathways are shown in Fig. 3 and Extended Data Fig. 4. For Fig. 3a, terms were filtered to include only those terms that were simultaneously upregulated in both clusters 4 and 9 and downregulated in both clusters 1 and 6, or vice versa to best highlight differences between these families.

Examining enrichment of MS susceptibility genes

The enrichment of MS susceptibility genes was evaluated separately from other diseases due to the availability of a recent publication by the International Multiple Sclerosis Genetics Consortium extensively mapping genomic risk loci in MS[51]. A hypergeometric test was used to evaluate the enrichment of 551 putative MS susceptibility genes identified as targets of MS variants in genes upregulated in our clusters ('Identification of cluster-defining gene sets'). The FDR-corrected threshold for significance was set at $q = 0.01$.

Examining enrichment of disease genes from the GWAS Catalog

To confirm the results of our MS analysis and to examine enrichment of genetic risk from other neurodegenerative or neuroinflammatory diseases, we were interested in using a more comprehensive source of disease–gene associations. Thus, the GWAS Catalog[52,99,100], a curated database that focused on SNP–trait association, was used for further analysis. This dataset contains select studies that include a primary GWAS analysis (per the GWAS catalog website: 'array-based genotyping and analysis of 1000,000 pre-QC SNPs selected to tag variation across the genome and without regard to gene content') or an imputation analysis with sufficient genome-wide coverage to meet the definition of a GWAS catalog mentioned previously. This catalog is updated on a weekly basis by curators, and eligible studies are generally added within 1–2 months of publication. The 2021-08-16 data release was used for this study. For our analysis, we chose to focus on specific disease entities where microglia are proposed to play relevant roles. To narrow our scope, GWAS catalog entries were filtered by a specific disease name. For example, to examine AD associations, all records containing 'Alzheimer' in the 'DISEASE_TRAIT' column were retained. Similarly, for stroke and cerebrovascular disorders, we filtered for all records containing the keywords 'stroke', 'brain ischemia', 'cerebral ischemia', 'cerebral artery' and 'cerebrovascular'. We carried out a similar approach for all other diseases we investigated in this analysis. After obtaining sets of disease–gene associations for each disease entity of interest, we applied a similar testing approach to that described in our MS disease gene analysis (see 'Examining enrichment of MS susceptibility genes').

Examining association of ROSMAP traits with clusters

The ROSMAP RNA-seq cohort used in our analysis contains a total of 1,092 samples, with a total of 18,629 genes captured across all samples. Using the DEseq2 R package[101], the DESeq function was used to perform differential expression analysis in association with one of 12 traits. The model for differential testing was:

$$\text{gene expression} \sim \text{phenotype} + \text{age at death} + \text{sex} + \text{technical variables}$$

Technical variables included: Batch, LOG_ESTIMATED_LIBRARY_SIZE, LOG_PF_READS_ALIGNED, PCT_CODING_BASES, PCT_INTERGENIC_BASES, PCT_PF_READS_ALIGNED, PCT_RIBOSOMAL_BASES, PCT_UTR_BASES, PERCENT_DUPLICATION, MEDIAN_3PRIME_BIAS, MEDIAN_5PRIME_TO_3PRIME_BIAS, MEDIAN_CV_COVERAGE, pmi, and study. For slope of cognitive decline, additional adjustment was performed for years of education in the model. Lists of positively and negatively associated genes were derived with an FDR-adjusted *P* value with an alpha level of 0.01. Enrichment of positively and negatively associated genes for each trait in the upregulated and downregulated genes for each cluster, respectively, was evaluated with a similar testing approach and threshold to our disease annotation analyses (see 'Examining enrichment of disease genes from the GWAS Catalog'). Detailed descriptions of the ROSMAP traits used in this analysis can be found at https://www.radc.rush.edu/docs/var/varIndex.htm, and they have been described in detail elsewhere[22,54,55,56,102]. Results of this analysis are reported in Supplementary Table 4.

In situ confirmation of microglia subset identity and abundance

As data do not always correlate between the transcriptomic and protein levels, a phenomenon that has been noted to be pronounced specifically in activated microglia[73], Advanced Cell Diagnostic's RNAscope was used to confirm our scRNA-seq results. A cohort of samples from the New York Brain Bank (details on donor cohort can be found in Supplementary Table 5) consisting of prefrontal cortex (BA9) tissue sections from 16 donors was chosen for validation efforts. After extensive optimization of a co-detection workflow to merge immunofluorescence and RNAscope, our final pipeline is described below. Initial optimization was performed with positive and negative RNAscope 4-plex controls (ACD; positive, 321831; negative, 321801), and once an optimal pipeline was identified, it was run with probes of interest.

All reagents from the RNAscope Multiplex Fluorescent Reagent Kit v2 (ACD, 323100) were prepared for use in accordance with the manufacturer's instructions. All wash buffers were prepared immediately before performing experiments.

Tissue sections cut at a thickness of 6 µm were deparaffinized with CitriSolv Clearing Agent (Decon Laboratories, 1601) for 20 min at room temperature (RT). This was followed by an ethanol series (100%, 100%, 70%; 100%, Fisher Scientific, BP2818; 70%, Fisher Scientific, BP8203) for 30 s per bath with agitation and rehydration in distilled water for 1 min at RT. Subsequently, 4–6 drops of hydrogen peroxide were used to cover the tissue section, and slides were incubated for 10 min at RT. Hydrogen peroxide (ACD, 322381) solution was removed by tapping on absorbent paper, and slides were washed with distilled water twice. Antigen retrieval was performed with pH 6.0 citrate (Sigma-Aldrich, C9999) and heating with a microwave for 25 min at 400 watts. Slides were then placed in tap water for 5 min, then moved to 100% ethanol for 1 min. Slides were allowed to dry fully at RT, then hydrophobic barriers were drawn around the tissue section with Super Pap Pen Liquid Blocker (Newcomer Supply, 6505). Slides were then blocked for 30 min at RT with RNAscope Co-Detection Antibody Diluent (ACD, 323160). Diluent was removed by tapping on absorbent paper, and slides were treated with primary antibody diluted in RNAscope Co-Detection Antibody Diluent for 2 h at RT. Slides were washed three times with PBS (Corning, 46-013-CM) containing 0.1% Tween-20 (Sigma-Aldrich, P9416; PBS-T), then submerged in fresh 10% Neutral Buffered Formalin (Sigma-Aldrich, HT5011) for 1 h at RT. Slides were washed with PBS-T three times, then four drops of RNAscope protease plus (ACD, 322381) were added to the slide and spread to fully cover the tissue. After incubation for 40 min at 40 °C in the RNAscope HybEz II oven (ACD, 321710), slides were washed once with distilled water. In total, 125 µl of pre-mixed RNAscope probe mix was then added to each slide, and then slides were incubated for 2 h at 40 °C. Slides were removed from the oven and washed twice with RNAscope wash buffer (ACD, 310091). Slides were then covered with 5× SSC (Sigma-Aldrich, S6639-1L) buffer and left overnight until the morning, when the protocol was resumed.

On the second day of the protocol, the slides were washed twice with RNAscope wash buffer, then four drops of RNAscope AMP1 (ACD, 323101) were added to each slide. After 30 min of incubation at 40 °C, slides were washed twice with RNAscope buffer, then four drops of AMP2 (ACD, 323102) were added per slide. After 30 more minutes of incubation at 40 °C, slides were washed twice with RNAscope buffer, then four drops of AMP3 (ACD, 323103) were added per slide. After a final 15 min of incubation, slides were washed twice with RNAscope buffer.

Next, four drops of HRP-C1 (ACD, 323104) were added per slide. After 15 min of incubation at 40 °C, slides were washed twice with RNAscope buffer, then 150 µl of Opal 570 (Akoya, FP1488001KT) dye diluted in RNAscope TSA diluent (ACD, 322809) was added per slide. Slides were incubated for 30 min, then washed twice with RNAscope wash buffer. Finally, four drops of HRP blocker (ACD, 323107) were added, followed by a 15-min incubation period and two washes with RNAscope buffer. This HRP-TSA-block process was repeated identically with either HRP-C2 (ACD, 323105) or HRP-C3 (ACD, 323106) depending on the channel of the original probes, and Opal 690 (ACD, FP1497001KT) dye diluted in TSA diluent. Finally, this HRP-TSA-block process was repeated once more with HRP-C4 (ACD, 323121), and the following modifications: TSA-DIG (Akoya, FP1501001KT) diluted in RNAscope TSA diluent for the second step and a 30-min incubation at RT instead of 40 °C, and swapping the second wash after HRP blocking from RNAscope wash buffer to PBS-T.

After all HRP steps were completed, counterstaining was performed with 200 µl of secondary antibody diluted in RNAscope Co-Detection Antibody Diluent. After incubation for 30 min at RT, slides were washed three times with PBS-T. Slides were then incubated with 150 µl of Opal Polaris 780 dye (Akoya, FP1501001KT) diluted in Antibody Diluent/Block (Akoya, ARD1001EA) for 30 min at RT. After three washes with PBS-T, slides were incubated with 200 µl of 1× Trueblack (Biotium, 23007) for 2 min at RT to quench lipofuscin autofluorescence. Slides were washed three times with PBS and counterstained with four drops of DAPI (ACD, 323108) per slide incubated at RT for 30 s. After removing DAPI by tapping slides on absorbent paper, the hydrophobic barrier was removed, and the slides were mounted with one drop of Prolong Gold (Thermo Fisher Scientific, P36934) and coverslips (Fisher Scientific, 12545F). Bubbles were removed from the mounting medium using gentle pressure from a pipette tip. Slides were dried for 30 min at RT in the dark, then transferred to 4 °C for imaging the following day.

The primary antibody used in staining was goat anti-human Iba1 (Wako, 01127991; dilution of 1:50). The secondary antibody used in staining was donkey anti-goat IgG (H + L) highly cross-adsorbed secondary antibody conjugated to Alexa Fluor Plus 488 (Thermo Fisher Scientific, A11055; 1:500 dilution). The RNAscope probes used in our experiments were: *CD74* (ACD, 477521), *CXCR4* (ACD, 310511-C2), *GPX1* (ACD, 492881) and *SPP1* (ACD, 420101-C4). Two additional probes were used but provided insufficient signal for downstream analysis: *MEF2A* (ACD, 452891-C3) and *CX3CR1* (ACD, 411251-C3).

Fields of view were captured using the ×40 objective of a Nikon Eclipse Ni-E immunofluorescence microscope. For each donor, 15 images were obtained from the gray matter with same exposure time, then loaded into CellProfiler software where automated segmentation and downstream analysis was performed as described below. Representative images can be found in Fig. 6b,f and Extended Data Fig. 6a,b.

Automated image analysis using CellProfiler

To automatically segment images and localize transcripts within microglia, we developed an extensive pipeline in CellProfiler v.4.2.1. First, the IdentifyPrimaryObjects module was used to segment based on DAPI, the EnhanceOrSuppressFeatures module was used to enhance the Iba1 signal (using 'Neurites' as the feature type) using the 'Line structures' filter, and segmented Iba1 signal was identified using IdentifyPrimaryObjects. After another round of enhancement of Iba1 signal by applying the 'Tubeness' filter (again using 'Neurites' as the feature type), the RelateObjects module was used to relate segmented DAPI and Iba1 objects using the DAPI as the parent objects and the Iba1 (after enhancement) as the child objects. Next, morphological parameters of each joint segmentation-defined cell were measured with the MeasureObjectSizeShape module. Then, for each of the channels with RNAscope imaging data, the following set of steps were run: enhancement of the signal for the given channel with EnhanceOrSuppressFeatures specifying 'Speckles' as the feature type, identification of RNA puncta using IdentifyPrimaryObjects, setting a mask for the channel with MaskObjects, then relating identified RNA puncta to the parent cells with RelateObjects. Subsequently, the intensity in each of the channels for all RNA that were associated with a parent microglial cell was measured with the MeasureObjectIntensity module, and all data were exported with the ExportToSpreadsheet module. Our full pipeline is available in Supplementary Table 5.

Analyzing CellProfiler-processed RNAscope data

Downstream of CellProfiler processing, all RNAscope puncta in any channel were filtered to exclude those found outside microglia identified by the joint segmentation pipeline. Using the microglia-localized puncta, an overall area-adjusted score was computed for each channel (cy3, cy5 or cy7) by dividing the summed intensity of puncta detected within a given microglial cell by the computed area for each microglial cell segmented by CellProfiler. The distribution of area-adjusted intensity for the different images in our cohort was evaluated, showing that most of our sections had similar distributions of cells, with rare outliers that had abnormally high signal in all channels. These outliers were excluded from further analysis. For three of our markers—*CD74*, *GPX1* and *SPP1*—expression levels for each of these markers were thresholded into three bins—low, medium and high—as we found substantial detection of these markers across all of our tissue samples. In contrast, *CXCR4* was detected only in a small fraction of cells at similar levels, so cells were segmented as either *CXCR4* positive or negative.

Thresholds were determined based on the distributions of the RNAscope data, as well as the levels of expression for different subtypes in our scRNA-sequencing dataset. For example, in our scRNA-seq data, the 'high' subtype for *CD74*, cluster 10, has a median fold change of 2.899 for *CD74* compared to other subsets. Thus, the threshold for *CD74* in the RNAscope data was set as 2.899 times the median of *CD74* expression. Similarly, to derive the low threshold for *CD74*, the fold change in *CD74* expression was compared between the set of families with low *CD74* expression, which included the closely related clusters 1, 5, 6 and 7, as well as the proliferative cluster 12, and all other clusters where the difference in expression of *CD74* was found to be significant by our pairwise testing approach ('Identification of cluster-defining gene sets'). In this case, the median fold change in expression for our low classes versus all other clusters was 0.394. As such, the *CD74*-low class threshold was set as 0.394 times the median of *CD74* expression. This process was repeated for *SPP1* and *GPX1*; for example, the *GPX1*-high classes were clusters 2, 4 and 9, while the low classes were 1, 6, 7 and 12, and median fold change of these two groups of clusters versus other clusters was used to determine the high and low threshold, respectively. A small fraction of cells with abnormally high signal that no longer appeared punctate in form, but rather diffuse and sometimes extending beyond the boundary of the cells was identified. Although these could represent real cells, these might also represent cells with high levels of background in our specific channels. Thus, for *CD74* and *SPP1*, the two markers where these types of cells were observed, cells that were 1.5 times the interquartile range above the 75th percentile for expression for all channels were excluded. This excluded a small number of cells (49 of a total of 7,364 cells for panel 1 and 13 of a total of 3,710 cells for panel 2).

To provide the most accurate comparison of numbers of cells between RNAscope and scRNA-seq, cells in scRNA-seq coming from AD diagnoses (EOAD, LOAD), PD diagnoses (PD-DLBD, PSP) or our control sample, were chosen for comparison, as these were the diagnoses represented in the samples that we obtained for our in situ analysis. Proportions of cells per binned class (that is, *CD74*$^{\text{lo}}$, *CD74*$^{\text{int}}$, *CD74*$^{\text{hi}}$) were then compared between the two datasets. The numbers of cells binned into the low, medium and high *CD74* classes were 3,756, 3,333 and 329, respectively. The numbers of cells binned into the *CXCR4* negative and positive classes were 7,096 and 322, respectively. The numbers of cells binned into the low, medium and high *GPX1* classes were 1,404, 1,653 and 671, respectively. The numbers of cells binned into the low, medium and high *SPP1* classes were 3,216, 388 and 125, respectively.

For analyses leveraging various features output by our CellProfiler pipeline, including the 'Compactness' and 'Eccentricity' features, the output of CellProfiler for each of these features was used. For others, such as 'median distance', the median distance of puncta for a given channel (for example, cy3, cy5 or cy7) was manually computed from the centroid of single segmented microglial cells. In all cases involving median distance of puncta from cellular centroids, we excluded all cells in the 'low' class for all channels in question to only include cells with real data. Significance of differences in morphological features between expression classes was tested with Welch's *t*-test with the Holm–Bonferroni correction[103,104], setting a significance threshold for an adjusted *P* value of 0.05.

MERFISH data generation

Human postmortem frozen brain tissue was embedded in Optimum Cutting Temperature medium (VWR, 25608-930) and sectioned on a Leica cryostat at −20 °C at 10 µm onto MERSCOPE coverslips (Vizgen, 2040003). These sections were then processed for MERSCOPE imaging according to the manufacturer's instructions. Once adhered to the coverslip, the tissue was fixed followed by three washes with 1× PBS. After aspiration, 70% ethanol was added to permeabilize the tissue for at least 24 h. After a wash with Formamide Wash Buffer, the sample was incubated with a custom MERFISH probe library and left to hybridize for 36–48 h. The sample was then washed and incubated at 47 °C with Formamide Wash Buffer twice, and then the tissue was embedded in a polyacrylamide gel followed by incubation with tissue clearing solution overnight at 37 °C. After the tissue became transparent, samples were washed with the wash buffer (Vizgen, 20300001) and incubated with DAPI and polythymine (polyT) staining reagent (Vizgen, 20300021) for 15 min with agitation. After washing, the coverslip was assembled into the imaging chamber and placed into the microscope for imaging. Each section was imaged using MERSCOPE 500 Gene Imaging Kit (Vizgen, 0400006) on a MERSCOPE (Vizgen). Briefly, the sample was loaded into a flow chamber connected to the MERSCOPE Instrument. A low-resolution mosaic was acquired using a ×10 objective, and the regions of interest were selected for high-resolution imaging with a ×60 lens. For the high-resolution imaging, the focus was locked to the fiducial fluorescent beads on the coverslip. Cell segmentation was performed using the Watershed algorithm, using DAPI nuclear seeds and PolyT total RNA staining basins. Images were decoded to RNA spots with xyz and gene ID using Vizgen's Merlin software.

Validation with MERFISH

We excluded cell entities suggesting failed segmentation (zero transcripts) and retained cells with 25–2,500 transcripts, more than five unique genes, and a size of 40–2,500 µm.

For downstream integration of cells across tissues, count matrices were merged and normalized using SCTransform[91] in the *R* package, Seurat (v.4)[105]. We identified major cell types based on 40 extracted PCs using a *k*-nearest neighborhood of 30 cells via FindNeighbors and clustering using the Louvain algorithm via FindClusters (Seurat).

For differential expression and downstream projection, we used log-normalized, downsampled counts. The AD tissue showed higher median UMI counts than the non-AD tissue. We downsampled UMI counts in the AD tissue to a proportion of 0.796 using downsampleMatrix from DropletUtils[106] to ensure a similar distribution to the non-AD tissue. Cell types were annotated based on known cell-type markers and differentially expressed genes identified using MAST[95] implemented in FindMarkers (Seurat).

To identify microglial subtypes, we extracted the microglia followed a similar preprocessing and integration protocol. To optimize the number of clusters, we used a subsampling-based approach (chooseR[107]) to calculate silhouette scores as a metric of cluster robustness. Across eight resolutions (0.3 to 1), we iteratively derived clusters ($B = 100$) in 80% subsets of the AD-derived microglia. Silhouette scores were averaged per cluster in each resolution, and optimal resolution was selected based on a median per-cluster silhouette score greater than the bootstrapped median silhouette score ($B = 25{,}000$) across the resolution parameter set. We chose a resolution of 0.4, which showed the second-highest median silhouette score averaged across clusters and an overall higher range than a resolution of 0.3.

We projected the microglia from the AD and the non-AD tissue into the existing UMAP using 50 projected mNN dimensions. The log-normalized, downsampled counts were used to identify anchor cells based on 51 overlapping genes in FindTransferAnchors and labels were transferred using TransferData.

Training machine learning models for label transfer to other single-cell microglial datasets

In the initial evaluation of our query datasets, substantial batch effects were evident. As this was likely to confound our downstream label transfer workflow, a version of our mNN integration pipeline was adapted for upstream removal of batch effects. To do so, our reference data were concatenated with the query data in a single Seurat object, and the unique differentially expressed genes from our pairwise differential expression testing (Identification of cluster-defining gene sets) were used for cross-batch merging with the fastmNN algorithm. For all these analyses, we used 40 components. The normalized 'mnn.reconstructed' assay, which represents per-gene corrected log-expression values, was used for downstream analysis.

After testing a number of different models in our label transfer pipeline, a combinatorial workflow leveraging two distinct models for different clusters showed the best accuracy: a set of pairwise SVM classifiers using consensus voting to assign labels for the smaller clusters (8–12) and a flat XGB[69] classifier to assign labels for the larger clusters (1–7) with higher transcriptional homology. This set of models was chosen because the SVM achieved highest accuracy in initial testing with smaller classes, but lower-than-average accuracy on larger classes, whereas the XGB results followed the exact opposite trend. Predictions from these two models were thus integrated to achieve higher predictive accuracy. The overall workflow for both methods was similar: as a few of our classes are transcriptionally similar, similar classes are condensed (clusters 1/6/7 and clusters 2/4), then a subset of the cells in our dataset are selected for training. Next, the differentially expressed genes from our pairwise differential expression testing ('Identification of cluster-defining gene sets') were selected as the features for training, and PCA was performed on the resulting subset of the data.

For the SVM, the training subset was 0.2 for classes 1–9, and 0.5 for classes 10–12. A separate classifier was trained for each unique pair of clusters (that is, a classifier to compare clusters 1/6/7 and 2/4, 1/6/7 and 3….1/6/7 and 12, then 2/4 and 3, 2/4 and 5….2/4 and 12) using only the genes found to be differentially expressed (both up and down) between that specific pair of clusters. Data classes were then rebalanced using combined over/under resampling to reduce class imbalance for smaller classes. Caret[108] was used to perform PCA and hyperparameter optimization of a SVM model using a radial kernel and tenfold cross-validation repeated three times. PCA was conducted independently during each fold. Conversely, for XGB, the training subset was 0.33, and the model trained only on cells from groupings 1/6/7, 2/4, 3 and 5. Similarly, PCA was performed upstream on the subset of scaled data consisting of all genes found to be differentially expressed between any clusters. Hyperparameter optimization with fivefold validation was performed in a stepwise fashion: tree number was first optimized, then tree-specific parameters were tuned with a restrictive grid search, then regularization parameters were tuned with a restrictive grid search, then final optimization was conducted with grid search in a narrow range around prior optimal parameters.

To construct a validation subset, a subset of 50% of the dataset was sampled exclusively from cells not used for training of either the SVM or XGB models. The same scaling and subsetting operations described above were applied to these data. Optimized SVM and XGB models were used to classify the data. For SVM models, final classifications were obtained with hard consensus voting, as the class with the majority of votes was chosen as the final class of the SVM voting ensemble. Similarly, for XGB, which outputs a probability for each class summing to 1 across all classes, the highest probability was used to choose the assigned label. However, the class probabilities for XGB also provided the opportunity to evaluate the confidence of the classifier and drop lower-confidence assignments. As such, cells were only retained for SVM classifications in classes 10–12 or for XGB classifications in classes 1–7 that had higher than 50% classification probability for the assigned probability. Final classifications were merged across datasets, and accuracy was evaluated by examining sensitivity, specificity and congruence of marker gene expression patterns of cells assigned to each class with marker gene expression patterns seen in our original data. Identical procedures were performed for query datasets.

This approach demonstrates high sensitivity and specificity on test data, with joint accuracy averaging 85% across models trained for different query datasets. Notably, uniformly high specificity is observed, even for clusters with lower sensitivity, such as clusters 3 and 5. These two clusters are also associated with lower confidence scores from our XGB model, an expected result given the transcriptionally intermediate nature of these clusters. Thus, the model's greatest difficulties with classification come in cases where the true classification boundary is not well defined, which provides a vote of confidence for the reliability of the model. Notably, for marker genes detected in query datasets, the transcriptional profiles of cells assigned to our distinct microglial clusters closely match the profiles of cells in those clusters in our original dataset (Extended Data Fig. 6).

To analyze association of mapped proportion numbers with continuous traits in the ROSMAP single-nucleus data, a linear model from the stats package in R with the formula 'proportion ~ trait' was used to examine the relationship of amyloid burden to cluster proportion. *P* values were adjusted with the Benjamini–Hochberg correction[98]. All ROSMAP donors with single-nucleus data were used for this analysis (described below).

Single-nucleus library preparation and sequencing of single nuclei

Dorsolateral prefrontal cortex tissue specimens were received frozen from the RADC. We observed variability in the morphology of these tissue specimens with differing amounts of gray and white matter and presence of attached meninges. Working on ice throughout, we carefully dissected to remove white matter and meninges, when present. The following steps were also conducted on ice: about 50–100 mg of gray matter tissue was transferred into the dounce homogenizer (Sigma, D8938) with 2 ml of NP40 Lysis Buffer (0.1% NP40, 10 mM Tris, 146 mM NaCl, 1 mM $CaCl_2$, 21 mM $MgCl_2$, 40 U ml^{-1} of RNase inhibitor (Takara, 2313B)). Tissue was gently dounced while on ice 25 times with pestle A followed by 25 times with pestle B, then transferred to a 15-ml conical tube. Then, 3 ml of PBS + 0.01% BSA (NEB, B9000S) and 40 U ml^{-1} of RNase inhibitor were added for a final volume of 5 ml and then immediately centrifuged with a swing bucket rotor at 500*g* for 5 min at 4 °C. Two samples were processed at a time, the supernatant was removed and the pellets were set on ice to rest while processing the remaining tissues to complete a batch of eight samples. The nuclei pellets were then resuspended in 500 ml of PBS + 0.01% BSA and 40 U ml^{-1} of RNase inhibitor. Nuclei were filtered through 20-µm pre-separation filters (Miltenyi, 130-101-812) and counted using the

Nexcelom Cellometer Vision and a 2.5 μg μl^{-1} DAPI stain at a 1:1 dilution with cellometer cell counting chamber (Nexcelom CHT4-SD100-002). In total, 5,000 nuclei from each of eight participants were then pooled into one sample, and 40,000 nuclei in a volume of 15–30 μl were run on the 10x single-cell RNA-seq platform using the Chromium Single Cell 3′ Reagent Kits version 3. Libraries were made following the manufacturer's protocol, briefly, single nuclei were partitioned into nanoliter-scale GEMs in the Chromium controller instrument where cDNA share a common 10x barcode from the bead. Amplified cDNA was measured by Qubit HS DNA assay (Thermo Fisher Scientific, Q32851) and quality assessed by BioAnalyzer (Agilent, 5067-4626). This whole-transcriptome-amplified material was diluted to <8 ng ml^{-1} and processed through a v3 library construction, and resulting libraries were quantified again by Qubit and BioAnalzyer. Libraries from four channels were pooled and sequenced on one lane of Illumina HiSeqX by the Broad Institute's Genomics Platform, for a target coverage of around one million reads per channel.

Processing of snRNA-seq reads

For each batch of snRNA-seq FASTQ files, Cell Ranger software (v.6.0.0; 10x Genomics) was used to map reads onto the reference human genome GRCh38, to collapse reads by UMI, and to count UMIs per gene per droplet. As a transcriptome model, the 'GRCh38-2020-A' file set distributed by 10x Genomics was used. The '--include-introns' option was set to incorporate reads mapped to intronic regions of nuclear pre-mRNA into UMI counts. To call cells among the entire droplets, the 'remove-background' module of CellBender[109] was applied to raw UMI count matrices with command line parameters. The admixture of ambient RNA was estimated and subtracted from UMI counts by CellBender. These filtered UMI count matrices were used in the subsequent analyses.

Demultiplexing

Because our snRNA-seq library consisted of nuclei from eight individuals, original individuals of each droplet were inferred by harnessing SNPs in snRNA-seq reads. We used two different procedures, depending on whether all eight individuals had been genotyped with whole-genome sequencing (WGS). When eight individuals were genotyped, we used demuxlet[110] software. From the WGS-based VCF file of 1,196 ROS/MAP individuals, we extracted SNPs that were in transcribed regions, passed a filter of GATK, and at least one of the eight individuals had its alternate allele. The extracted SNP genotype data were fed to demuxlet along with BAM files generated by Cell Ranger. When less than eight individuals were genotyped, we used freemuxlet (https://github.com/statgen/popscle/), which clusters droplets based on SNPs in snRNA-seq reads and generates a VCF file of snRNA-seq-based genotypes of the clusters. The number of clusters was specified to be eight. The snRNA-seq-based VCF file was filtered for genotype quality > 30 and compared with available WGS genotypes using the bcftools gtcheck command. Each WGS-genotyped individual was assigned to one of the droplet clusters by visually inspecting a heat map of the number of discordant SNP sites between snRNA-seq and WGS. The above two procedures converged to a table that mapped droplet barcodes onto inferred individuals. Each BAM file generated by Cell Ranger was split into eight per-individual BAM files, each of which contained reads from distinct individuals, using subset-bam (https://github.com/10XGenomics/subset-bam/). UMI count matrices filtered by CellBender were split into eight per-individual UMI count matrices.

QC

To identify and exclude potential sample swaps, we assessed concordance of genotypes between snRNA-seq and WGS. LOD scores, a metric of genotype concordance, were computed by comparing the per-individual BAM files with WGS genotypes of matched individuals using Picard CrosscheckFingerprints (v.2.25.4). We used a haplotype map downloaded from https://github.com/naumanjaved/fingerprint_maps/. After inspecting a histogram of LOD scores, ten individuals whose LOD scores were less than 50.0 were filtered out. These individuals received few cells by the demultiplexing procedure. As another measure to detect sample swaps, we checked RNA expression levels of the *XIST* gene and confirmed that they were consistent with clinical sex. Five individuals were further excluded because they failed QC of WGS. Four were marked as potential sample swaps among WGS, and the other was marked as an outlier of genotype principal component analysis.

Four individual-level sequencing metrics were computed from the per-individual UMI count matrices: estimated number of cells, median UMI counts per cell, median genes per cell and total genes detected. After inspecting these metrics, individuals whose median UMI counts per cell were less than 1,500 were excluded. Thirteen individuals were found to be sequenced twice in distinct batches. After comparing sequencing metrics, one of these duplicates was excluded from further analyses. After these QC processes, 424 individuals remained.

Cell-type classifications

To annotate for cell type, we fitted a weighted ElasticNet-regularized logistic regression classifier over the data of our previous work[111], predicting one of the eight major cell types for every nucleus: excitatory neurons, inhibitory neurons, astrocytes, microglia, oligodendrocytes, oligodendrocyte precursor cells, endothelial and pericytes. The gene expression matrix was log normalized (using NormalizeData method, Seurat package) and scaled over the top 700 variable features excluding noncoding RNA (using the FindVariableFeatures method, setting the selection method to vst and ScaleData method, in Seurat package).

We trained five different models with a mixing parameter of alpha = 0 (Ridge), 0.25, 0.5, 0.75 and 1 (Lasso), over a randomly selected 75% of the data (n = 139,311). Samples were weighted as 1/ for the number of nuclei of cell type present in the training set. This step ensured that even lowly represented cell types such as endothelial and pericytes will be properly learned. To select the models' regularization parameters, we applied tenfold cross-validation (using cv.glmnet method, glmnet package). Fitted models were evaluated using the held-out 25% of the data (n = 43,428), and their accuracy with respect to the misclassification error was calculated. As all models achieved very high accuracies, with misclassifications mostly between excitatory and inhibitory neurons, we selected the ElasticNet model with a mixing parameter using an alpha level of 0.25 to induce sparsity to the model. Fitted models used only 121 of the 700 available features and achieved a test accuracy of 99.95, with most misclassified nuclei being between inhibitory and excitatory neurons. The nuclei assigned to the microglial cluster were extracted and used in our analyses.

Leveraging the CMAP to identify chemical and genetic targets for in vitro recapitulation

The CMAP[20,21] is a catalog of gene expression signatures for a series of different genetic and pharmacologic perturbations across a wide variety of different cell lines. To identify chemical targets that might drive signatures associated with our distinct microglial subsets in vitro, upregulated gene lists were assembled for each cluster corresponding to genes upregulated in comparison to three or more clusters. The web interface found at clue.io was used to interface with the CMAP database, and the ListMaker tool was used to assemble lists, which were then submitted as inputs to the Query tool. The v.1.0 L1000 gene expression data compendium was used for all analyses. Output lists were downloaded and ranked by 'median_tau_score'. Results were aggregated into families: 1 and 6, 4 and 9, and 8 and 10. Chemical perturbagens of interest were selected from those with a 'median_tau_score' above 90 and chosen based on prior knowledge and the pathways they targeted. Full output lists from CMAP separated by cluster can be found in Supplementary Table 7.

Drug screening in the HMC3 model system

Compounds of interest were obtained from a wide range of reputable vendors and resuspended in DEPC-treated water (Invitrogen, AM9915G), PBS (Corning, 21-040-CV) or DMSO (Sigma-Aldrich, 472301). To keep the design of our experiment as similar as possible to the CMAP study, the target stock concentration was 10 mM, but this was adjusted depending on the solubility of each compound. Extensive dose titration with doses ranging from 0.01 μM to 0.1 mM was conducted to determine the highest tolerable dose for each compound. Each concentration of drug was plated in triplicate with early-passage HMC3 cells (American Type Culture Collection, CRL-3304), and the viability was read out using Calcein AM (Invitrogen, C1430) and propidium iodide (Invitrogen, P3566) using a Celigo plate (Nexcelom Bioscience) reader at 6 h and 24 h. An optimal dose of each drug was then chosen based on cell morphology and viability. Subsequently, optimal doses were applied to plated HMC3s and collected for RNA extraction after 6 h and 24 h. In-well lysis was performed with RLT buffer (QIAGEN, 74136) containing 2-mercaptoethanol (Thermo Fisher Scientific, 63689), and RNA extraction was performed with the QIAGEN RNEasy mini plus kit (QIAGEN, 74136) following the manufacturer's instructions. gDNA eliminator columns were used to remove contaminating genomic DNA. Initial RNA quality and quantity were assessed using Nanodrop (Thermo Fisher Scientific) followed by cDNA preparation using the iScript cDNA Synthesis kit (Bio-Rad, 1708891). cDNA was subsequently purified with AMPure XP beads (Thermo Fisher Scientific, A63880) using a 1:1.8 ratio of cDNA:beads.

RT–qPCR analysis

Real-time qPCR reactions to amplify 1 ng of total cDNA were performed in a QuantStudio 3 Real-Time PCR Cycler (A28132, Applied Biosystems) using the Applied Biosystems Fast SYBR Green Master Mix (Thermo Fisher Scientific, 4385612). CT values were normalized using hypoxanthine phosphoribosyltransferase 1 (*Hprt1*) as the housekeeping gene. Primers were tested for their efficiency beforehand, and the $\Delta\Delta C_t$ method was applied for analysis of relative gene expression. The melting curves of each product were analyzed to ensure the specificity of the PCR product. The following primers were used: *HPRT1* - fw: CCTGGCGTCGTGATTAGTGAT, rev: AGACGTTCAGTCCTGTCCATAA; *SRGAP2* - fw: GTTGTGACTTAGGCTACCATGC, rev: TGCTTCGACTGTTCCAGGTTT; *MEF2A* – fw: GGTCTGCCACCTCAGAACTTT, rev: CCCTGGGTTAGTGTAGGACAA; *TYROBP* – fw: ACTGAGACCGAGTCGCCTTAT, rev: ATACGGCCTCTGTGTGTTGAG; *GPX1* – fw: CAGTCGGTGTATGCCTTCTCG, rev: GAGGGACGCCACATTCTCG; *CXCR4* – fw: ACGCCACCAACAGTCAGAG, rev: AGTCGGGAATAGTCAGCAGGA; *SRGN* – fw: GGACTACTCTGGATCAGGCTT, rev: CAAGAGACCTAAGGTTGTCATGG. For visualization, the mean for each gene is shown with error bars that denote the standard deviation. Individual points are plotted to visualize the distribution of the data.

Bulk RNA-seq of compound-treated microglia

Around 0.5 × 10^6 HMC3 microglial cells were seeded into a six-well plate and incubated overnight. The next day, microglia were treated with the respective concentrations of camptothecin (1 μM; EMD Millipore, 390238), narciclasine (0.1 μM; MilliporeSigma, SML2805), Torin-2 (10 μM; Cayman Chemical Company, 14185) or DMSO (Sigma-Aldrich, 472301) as control and incubated for 24 h before collection. Cells were trypsinized (Gen Clone, 25-510 F), counted, the cell viability was assessed and cells were then resuspended in 350 μl RLT Lysis buffer (QIAGEN, 74136) containing 2-mercaptoethanol (Thermo Fisher Scientific, 63689), and isolated using a QIAGEN Plus Mini kit (QIAGEN, 74136). RNA quality was assessed using 2100 Bioanalyzer G2938C using an Agilent RNA 6000 Nano Kit (Agilent, 5067-1511) and Qubit 4 Fluorometer (Invitrogen) using Qubit 1X dsDNA HS Assay kit (Thermo Fisher Scientific, Q33231) before further processing for RNA-seq.

mRNA libraries were prepped using the Illumina TruSeq Stranded mRNA Library prep (Illumina, 20020595), in accordance with manufacturer recommendations, and using IDT for Illumina TruSeq DNA UD Indices (Illumina, 20022370) for adaptors. Briefly, 500 ng of total RNA was used for purification and fragmentation of mRNA. Purified mRNA underwent first-strand and second-strand cDNA synthesis. cDNA was then adenylated, ligated to Illumina sequencing adaptors and amplified by PCR (using ten cycles). The cDNA libraries were quantified using the Fragment Analyzer 5300 (Advanced Analytical) kit FA-NGS-HS (Agilent, DNF-474-1000) and Spectramax M2 (Molecular Devices) kit Picogreen (Life Technologies, P7589). Libraries were sequenced on an Illumina NovaSeq sequencer, using 2 × 100-bp cycles.

Sequencing QC was performed using Picard v.1.83 and RSeQC v.2.6.1. STAR v.2.5.2a was used to align reads to the GRCh38 genome, using Gencode v.25 annotation. Bowtie2 v.2.1.0 was used to measure rRNA abundance. Annotated genes were quantified with featureCounts v.1.4.3-p1.

To analyze the data, a generalized linear model within DESeq2 (ref. 101) was used to test for differentially expressed genes across each of our three treatment conditions in comparison to control. The DESeq object was constructed with a standard one-factor model, using '~treatment' as the model for analysis, and genes with less than ten overall counts across all samples were discarded before analysis. For analysis of similarity between samples, we used the variance stabilizing transformation in DESeq2, then computed PCA on the resultant matrix. Differential expression was performed with the DESeq function, and thresholds for significance were set as an FDR alpha of less than 0.01 and a LFC of 1.5. Shrinkage of LFC was performed with the ashr package[112], and shrunk LFC values were used for downstream visualization. GO annotation was performed with TopGO[38], and GO results were summarized with rrvgo[39]. To examine specific genes associated with given cluster families in each treatment condition, the top 20 nonoverlapping markers for each member of the grouped clusters (that is, the top 20 genes for cluster 1, the top 20 genes for cluster 6 that are not in the top 20 gene list for cluster 1) that were present in the differentially expressed gene list for that given condition, regardless of the direction of change (up or down) were chosen for visualization.

Generation and analysis of global quantitative proteomic data

For global quantitative proteomics of compound-treated HMC3 microglia cells, diaPASEF[113] (data independent acquisition)-based proteomics was used. In brief, 0.5 × 10^6 HMC3 microglial cells were seeded into a six-well plate and incubated overnight. The next day, cells were treated with the respective concentrations of camptothecin (1 μM; EMD Millipore, 390238), narciclasine (0.1 μM; MilliporeSigma, SML2805), Torin-2 (10 μM; Cayman Chemical Company, 14185) or DMSO (Sigma-Aldrich, 472301) as control and incubated for 24 h before collection. Cells were trypsinized (Gen Clone, 25-510F), counted, and the cell viability was assessed. Cells were then washed with ice-cold PBS (Corning, 21-040-CV) and cellular pellets were snap frozen and stored at −80 °C until further processing.

Subsequently, cells were lysed in lysis buffer[114] (1% SDC, 100 mM Tris-HCl, pH 8.5, and protease inhibitors; MilliporeSigma, D6750, 9290-OP) and boiled for 15 min at 60 °C, at 1,500 rpm. Protein reduction and alkylation of cysteine was performed with 10 mM TCEP (MilliporeSigma, C4706) and 40 mM 2-chloroacetamide (MilliporeSigma, C0267) at 45 °C for 15 min followed by sonication in a water bath, cooled down to RT. Protein digestion was processed for overnight by adding LysC and trypsin in a 1:50 ratio (μg of the enzyme to μg of protein; Promega, V5071) at 37 °C and 1,400 rpm. Peptides were acidified by adding 1% trifluoroacetic acid (TFA) (Thermo Fisher Scientific, 28904), vortexed, and subjected to StageTip clean-up via styrenedivinylbenzene-reversed-phase sulfonate[114]. Peptides were loaded on one 14-gauge StageTip plug. Peptides were washed two times with 200 μl 1% TFA 99% ethyl acetate (Thermo Fisher Scientific, 28904; MilliporeSigma, 270989) followed 200 μl 0.2% TFA/5% acetonitrile (ACN; Thermo Fisher Scientific, 28904; Thermo Fisher Scientific, A955) in a centrifuge at 3,000 rpm, followed by elution with 60 μl of 1% ammonia/50% ACN (Honeywell-Fluka, 4427310X1ML; Thermo Fisher Scientific, A955) into microcentrifuge tubes and dried at 45 °C in a SpeedVac centrifuge. Samples were resuspended in 10 μl of LC buffer (3% ACN/0.1% formic acid; Fisher Scientific, A11710X1-AMP). Peptide concentrations were determined using NanoDrop (Thermo Fisher Scientific) and 200 ng of each sample were used for diaPASEF analysis on a timsTOFPro (Bruker). Peptides were separated within 120 min at a flow rate of 400 nl min^{-1} on a reversed-phase C18 column with an integrated CaptiveSpray Emitter (25 cm × 75 μm, 1.6 μm, IonOpticks). Mobile phases A and B were with 0.1% formic acid in water and 0.1% formic acid (Fisher Scientific, A11710X1-AMP) in ACN (Thermo Fisher Scientific, A955). The fraction of B was linearly increased from 2% to 23% within 90 min, followed by an increase to 35% within 10 min and a further increase to 80% before re-equilibration. The timsTOF Pro (Bruker) was operated in diaPASEF mode[113] and data were acquired at defined 32 × 25-Th isolation windows from 400 to 1,200 *m/z*. To adapt the MS1 cycle time in diaPASEF, we set the repetitions to 2 in the 16-scan diaPASEF scheme. The collision energy was ramped linearly as a function of the mobility from 59 eV at 1/K0 = 1.6 Vs cm^{-2} to 20 eV at 1/K0 = 0.6 Vs cm^{-2}. The acquired diaPASEF raw files were searched with the UniProt Human proteome database in the DIA-NN search engine with default settings of the library-free search algorithm[115]. The FDR was set to 1% at the peptide precursor and protein level.

Results obtained from DIA-NN were further analyzed in R. To preliminarily filter the data, peptides without a valid matching gene symbol, as well as peptides that were detected in a fourth of our samples or fewer were removed. For further analyses, total intensity log-normalized protein abundances were used. PCA was performed on the dataset in its entirety to assess relative similarity of treatment conditions. Next, pairwise differential testing between DMSO control and each of our treated conditions was conducted using a Welch's[103] *t*-test with the Benjamini–Hochberg correction[98], setting a threshold of 0.05 for the corrected *P* value and a threshold of 1 for the $\log_2$ fold change. Top differentially expressed genes were then used for GO annotation with topGO ('Bulk RNA-seq of compound-treated microglia'). As there were fewer differentially expressed genes overall, all genes associated with each cluster family that overlapped with the differentially expressed gene list for each condition, irrespective of direction (up or down) were selected for plotting.

Visualizing gene expression across clusters with DotPlots

Seurat's DotPlot function was used to concurrently visualize gene expression and percentage of cells in each cluster expressing said genes. Using this function, a single circle is plotted for each cluster for each given gene. The size of this circle represents the percentage of cells within a cluster that express the gene, and it is absent entirely if fewer than 10% of cells in a given cluster expressed a gene. Conversely, the color of the circle represents the average expression of the gene. This is computed by computing the mean of expression for each cluster, then scaling and zero-centering the average expression level for each discrete cluster. The viridis 'magma' color palette was used for this visualization. Legends for the size and color scheme for each dot plot accompany each figure. In addition, for Fig. 2c, the 'cluster.idents' parameter was used to hierarchically cluster our different clusters by the marker genes involved using complete linkage, enabling clearer visualization of broad differences. The cluster dendogram was manually recomputed and added to the dot plot with the ggtree[116] package. Notably, data visualization was performed with Seurat v.4.0.4 instead of v.3.2.0.

Statistical analysis and data visualization

Statistical analysis was conducted as described in the associated methods sections above. Specific *P* values (both significant and not), if not found in the figures, may be found in Supplementary Information tables before and after testing for multiple correction. *T* values and degrees of freedom are also provided where relevant. Unless otherwise noted, all measurements are taken from distinct samples. In general, statistical methods were not used to recalculate or predetermine sample sizes. All plots were created in R v.4.1.0 using either base R visualization packages, ggplot2 (ref. 117) with ggrepel[118], ggfortify[119], patchwork[120], cowplot[121] and ggsci[122], or packages mentioned in the methods text. Heat maps were made with the pheatmap[123] package. Volcano plots were made with the EnhancedVolcano[124] package. All boxplots denote the 25th percentile, median and 75th percentile, with whiskers representing 1.5 times the interquartile range in both directions. Outliers, if any, are represented as circles beyond the whiskers.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Raw scRNA-seq data (fastq files) generated from CD45^+ cells isolated from autopsy samples were deposited to the Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/) under accession number GSE204702. Bulk RNA-seq data from compound-treated HMC3 cells were deposited to the GEO under accession number GSE202556. Bulk proteomic data from compound-treated HMC3 cells were deposited to ProteomeXChange (http://www.proteomexchange.org/) under accession number PXD033844. Data repurposed for label transfer was retrieved from the GEO under accession numbers GSE133432, GSE178317 and GSE103224.

Code availability

Code used to perform preprocessing, clustering, cluster validation and label transfer of scRNA-seq data in the current study is available publicly at https://github.com/jtuddenham/single-cell-microglia-v2/. The CellProfiler pipeline used to analyze joint immunofluorescence–RNAscope data is available as Supplementary Information (Supplementary Table 5), and in the aforementioned GitHub repository. Code for visualization, analysis of bulk RNA-seq/proteomic data and downstream analysis of CellProfiler outputs is available from the corresponding author upon request.

References

84. Vonsattel, J. P. G. et al. Twenty-first century brain banking: practical prerequisites and lessons from the past: the experience of New York Brain Bank, Taub Institute, Columbia University. *Cell Tissue Bank.* **9**, 247–258 (2008).
85. R Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing (2021).
86. RStudio Team. RStudio: integrated development for R. RStudio, PBC (2020).
87. Osorio, D. & Cai, J. J. Systematic determination of the mitochondrial proportion in human and mice tissues for single-cell RNA-sequencing data quality control. *Bioinformatics* **37**, 963–967 (2021).
88. Klein, H. -U. demuxmix: demultiplexing oligonucleotide-barcoded single-cell RNA sequencing data with regression mixture models. *Bioinformatics* **39**, btad481 (2023).
89. McGinnis, C. S. et al. MULTI-seq: sample multiplexing for single-cell RNA sequencing using lipid-tagged indices. *Nat. Methods* **16**, 619–626 (2019).
90. Lun, A. T. L. et al. EmptyDrops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. *Genome Biol.* **20**, 63 (2019).
91. Hafemeister, C. & Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol.* **20**, 296 (2019).

92. Haghverdi, L., Lun, A. T. L., Morgan, M. D. & Marioni, J. C. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.* **36**, 421–427 (2018).
93. Allaire, J. & Chollet, F. keras: R interface to 'Keras'. https://CRAN.R-project.org/package=keras (2021).
94. Taiyun, W. & Viliam, S. R package 'corrplot': visualization of a correlation matrix (version 0.90) (2021).
95. Finak, G. et al. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* **16**, 278 (2015).
96. Storey, J. D. & Tibshirani, R. Statistical significance for genome-wide studies. *Proc. Natl Acad. Sci. USA* **100**, 9440–9445 (2003).
97. Zheng, S. et al. Molecular transitions in early progenitors during human cord blood hematopoiesis. *Mol. Sys. Biol.* **14**, e8041 (2018).
98. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Methodol.* **57**, 289–300 (1995).
99. MacArthur, J. et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* **45**, D896–D901 (2017).
100. Li, M. J. et al. GWASdb v2: an update database for human genetic variants identified by genome-wide association studies. *Nucleic Acids Res.* **44**, D869–D876 (2016).
101. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
102. Mostafavi, S. et al. A molecular network of the aging human brain provides insights into the pathology and cognitive decline of Alzheimer's disease. *Nat. Neurosci.* **21**, 811–819 (2018).
103. Welch, B. L. The generalisation of student's problems when several different population variances are involved. *Biometrika* **34**, 28–35 (1947).
104. Holm, S. A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* **6**, 65–70 (1979).
105. Hao, Y. et al. Integrated analysis of multimodal single-cell data. *Cell* https://doi.org/10.1016/j.cell.2021.04.048 (2021).
106. Griffiths, J. A., Richard, A. C., Bach, K., Lun, A. T. L. & Marioni, J. C. Detection and removal of barcode swapping in single-cell RNA-seq data. *Nat. Commun.* **9**, 2667 (2018).
107. Patterson-Cross, R. B., Levine, A. J. & Menon, V. Selecting single cell clustering parameter values using subsampling-based robustness metrics. *BMC Bioinformatics* **22**, 39 (2021).
108. Kuhn, M. et al. caret: classification and regression training. R Core Team. https://CRAN.R-project.org/package=caret (2021).
109. Fleming, S. J. et al. Unsupervised removal of systematic background noise from droplet-based single-cell experiments using CellBender. *Nat. Methods* **20**, 1323–1335 (2023).
110. Kang, H. M. et al. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat. Biotechnol.* **36**, 89–94 (2018).
111. Cain, A. et al. Multi-cellular communities are perturbed in the aging human brain and Alzheimer's disease. *Nat. Neurosci.* **26**, 1267–1280 (2023).
112. Stephens, M. et al. ashr: methods for adaptive shrinkage, using empirical Bayes. R package ashr version 2.2-47 (2022).
113. Meier, F. et al. diaPASEF: parallel accumulation–serial fragmentation combined with data-independent acquisition. *Nat. Methods* **17**, 1229–1236 (2020).
114. Kulak, N. A., Pichler, G., Paron, I., Nagaraj, N. & Mann, M. Minimal, encapsulated proteomic-sample processing applied to copy-number estimation in eukaryotic cells. *Nat. Methods* **11**, 319–324 (2014).
115. Demichev, V., Messner, C. B., Vernardis, S. I., Lilley, K. S. & Ralser, M. DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. *Nat. Methods* **17**, 41–44 (2020).
116. Yu, G. et al. ggtree: an R package for visualization of tree and annotation data. Bioconductor version: release 3.14. https://doi.org/10.18129/B9.bioc.ggtree (2022).
117. Wickham, H. Ggplot2: elegant graphics for data analysis. (Springer-Verlag, 2016).
118. Slowikowski, K. ggrepel: automatically position non-overlapping text labels with 'ggplot2'. R package version 0.9.1. https://CRAN.R-project.org/package=ggrepel (2021).
119. Tang, Y., Horikoshi, M. & Li, W. ggfortify: unified interface to visualize statistical results of popular R packages. *R. J.* **8**, 474–485 (2016).
120. Pedersen, T. L. Patchwork: the composer of plots. https://CRAN.R-project.org/package=patchwork (2020).
121. Wilke, C. O. cowplot: streamlined plot theme and plot annotations for 'ggplot2'. https://wilkelab.org/cowpl (2020).
122. Xiao, N. ggsci: scientific journal and Sci-Fi themed color palettes for 'ggplot2'. R package version 2.9 https://cran.r-project.org/package=ggsci (2018).
123. Kolde, R. pheatmap: pretty heatmaps. R package version 1(2):726 https://cran.r-project.org/package=pheatmap (2019).
124. Blighe, K. et al. EnhancedVolcano: Publication-ready volcano plots with enhanced colouring and labeling. Bioconductor version: release (3.14) https://doi.org/10.18129/B9.bioc.EnhancedVolcano (2022).

Acknowledgements

We thank the individuals and their families who donated the brain samples used in this project. The work was supported by the Chan-Zuckerberg Initiative's Neurodegeneration Challenge Network grant CS-02018-191971. Some of the work also emerged from support from National Institutes of Health (NIH)/National Institute on Aging (NIA) grants R01 AG070438, U01 AG061356, RF1 AG057473 and R01AG048015. Research reported in this publication was supported by the National Institute of General Medical Sciences of the NIH under award number T32GM007367 and by the National Cancer Institute of the NIH under award number F30CA261090. The Parkinsonism brain bank at Columbia University is supported by the Parkinson's Foundation. R.A.H. was supported by grant funding from the Huntington Disease Society of America and Hereditary Disease Foundation and was a Columbia University Irving Medical Center ADRC Research Education Component trainee (P30AG066462). We are grateful to the Banner Sun Health Research Institute Brain and Body Donation Program of Sun City, Arizona for the provision of human brain tissue. The Brain and Body Donation Program has been supported by the National Institute of Neurological Disorders and Stroke (U24NS072026 National Brain and Tissue Resource for Parkinson's Disease and Related Disorders), the NIA (P30AG19610 and P30AG072980 Arizona Alzheimer's Disease Core Center), the Arizona Department of Health Services (contract 211002, Arizona Alzheimer's Research Center), the Arizona Biomedical Research Commission (contracts 4001, 0011, 05-901 and 1001 to the Arizona Parkinson's Disease Consortium) and the Michael J. Fox Foundation for Parkinson's Research. The Rocky Mountain Multiple Sclerosis Tissue Bank is supported by the National Multiple Sclerosis Society. The Genomics Shared Resource is supported by NIH/NCI P30CA013696. ROSMAP is supported by P30AG10161, P30AG72975, R01AG15819, R01AG17917, U01AG46152 and U01AG61356. ROSMAP resources can be requested at https://www.radc.rush.edu/. Research reported in this publication was partially performed in the Columbia Center for Translational Immunology Flow Cytometry Core, supported in part by the Office of the Director, NIH under award S10OD020056. All figures were created using BioRender.com and Figs. 2–8 and Supplementary Fig. 1 were generated in Inkscape (https://inkscape.org).

Author contributions

Conceptualization, J.F.T. and P.L.D.; Methodology, J.F.T., M.T., Y.Z., P.A.S., V.M. and P.L.D.; Software, J.F.T. and M.T.; Validation, J.F.T., M.T., V.H. and P.L.D.; Formal analysis, J.F.T., V.M., H.K., C.W. and P.A.S.; Investigation, J.F.T., M.T., V.H., A.K., M.O. and P.L.D.; Resources, B.H., T.R., A.J.L., G.G., G.E.S., N.H., M.F., S.H., T.G.B., J.C., R.K.S., A.F.T., R.A.H., R.N.A., N.S., J.S. and D.A.B.; Data curation, J.F.T., M.O. and V.M.; Writing–original draft, J.F.T. and P.L.D.; Writing–review and editing, all authors; Visualization, J.F.T. and V.H.; Supervision, P.L.D.; Project administration, P.L.D.; Funding acquisition, P.L.D.

Competing interests

R.N.A. is funded by the NIH, DoD, the Parkinson's Foundation and the Michael. J. Fox Foundation. R.N.A. received consultation fees from Avrobio, Caraway, GSK, Merck, Ono Therapeutics and Genzyme/Sanofi. P.L.D. has served as a consultant for Biogen, Merck-Serono and PureTech. All other authors declare no competing interests.

Additional information

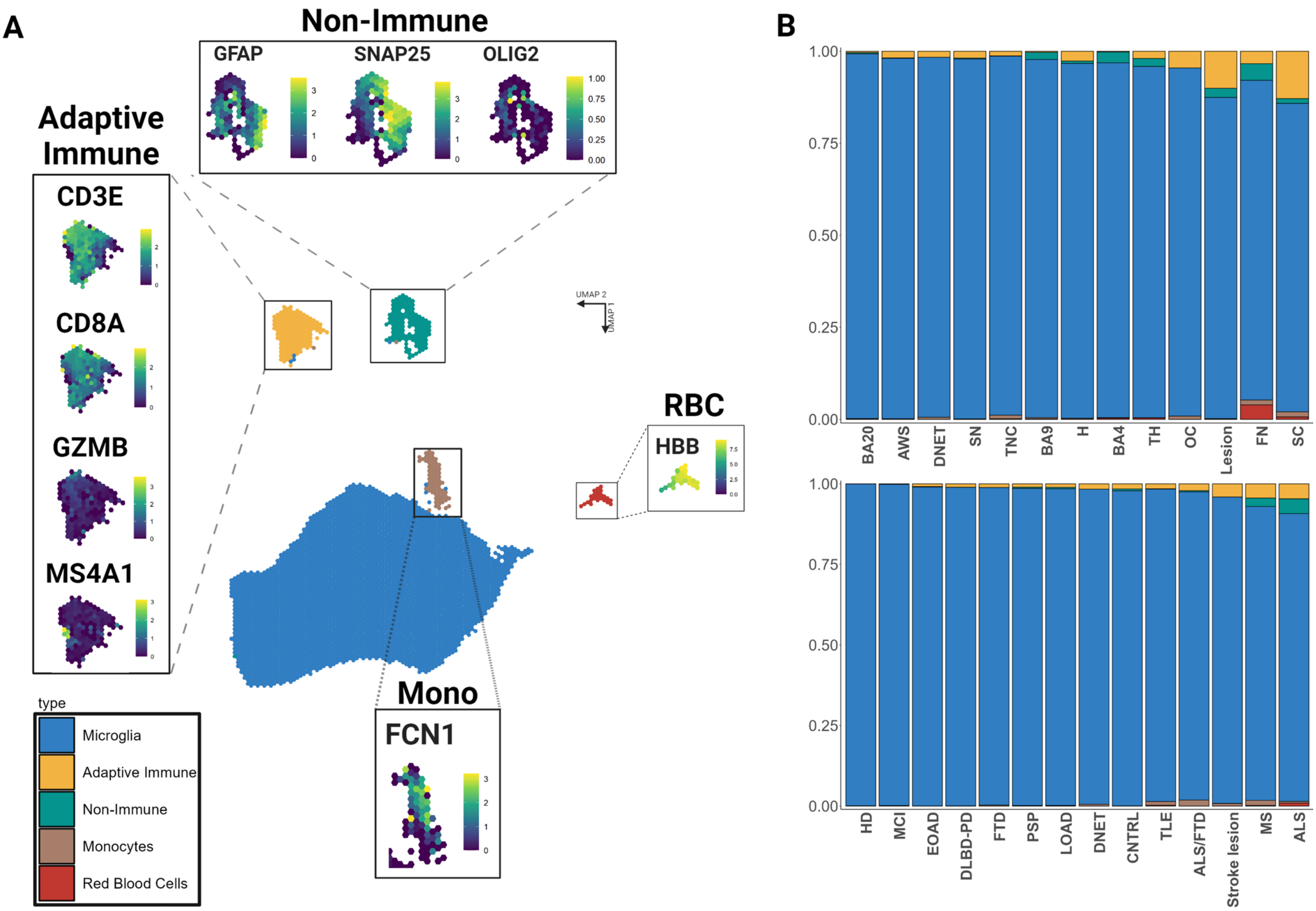
Extended data is available for this paper at https://doi.org/10.1038/s41593-024-01764-7.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41593-024-01764-7.

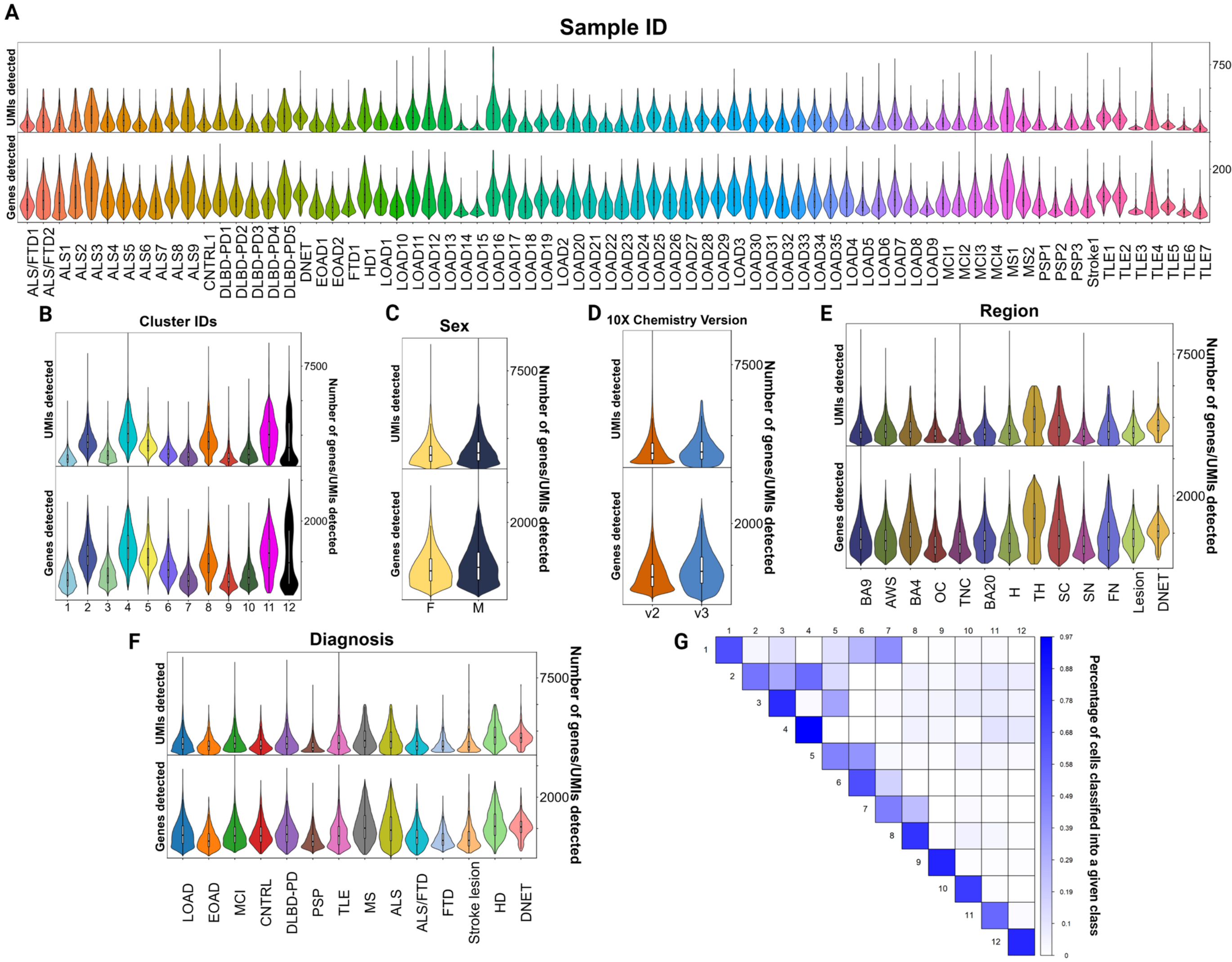
Correspondence and requests for materials should be addressed to Philip L. De Jager.

Peer review information *Nature Neuroscience* thanks Özgün Gökçe and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

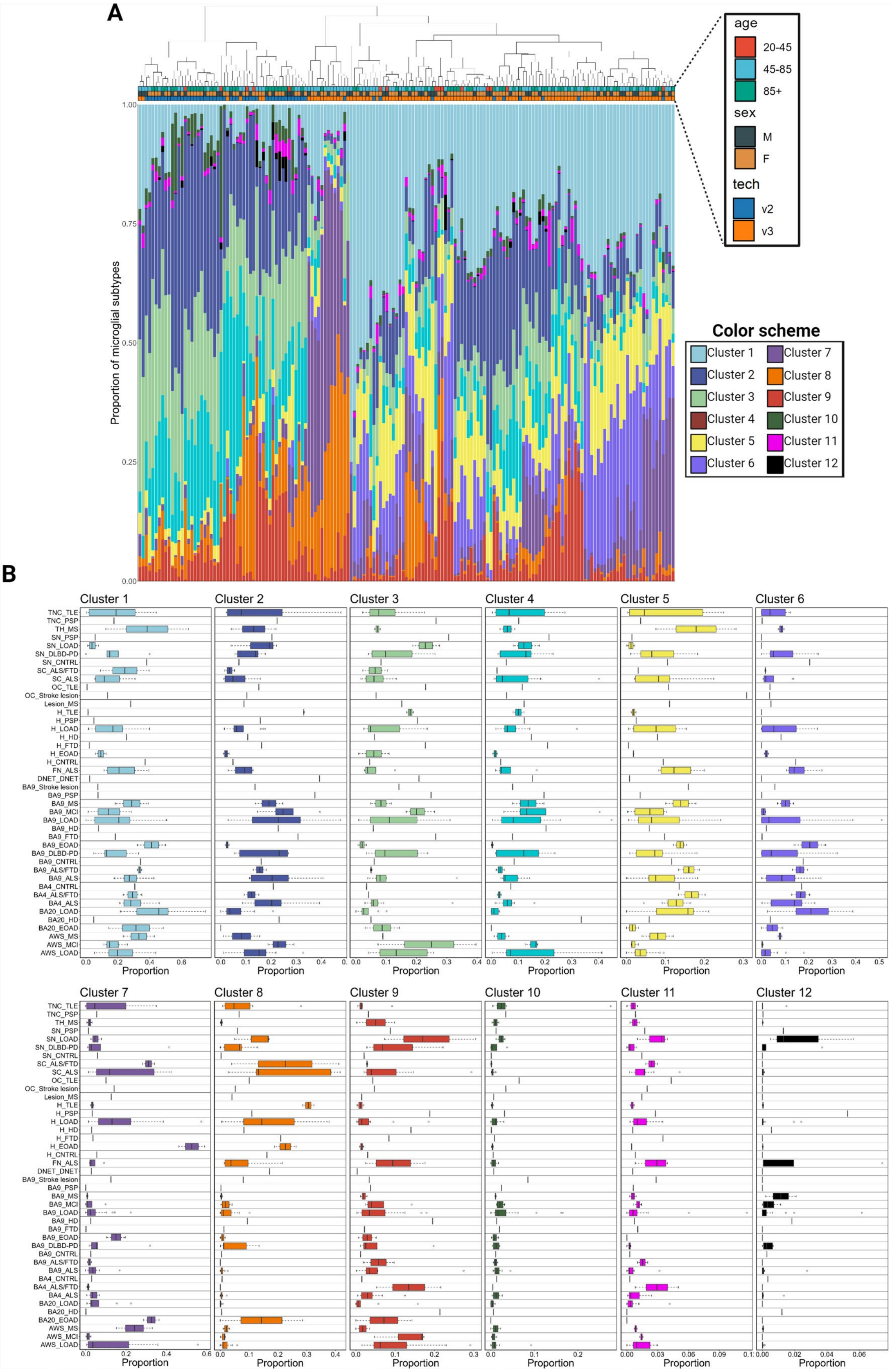
Reprints and permissions information is available at www.nature.com/reprints.



Extended Data Fig. 1 | Proportions of overarching cell types in our dataset. (**A**) **Different cell types are discriminable in UMAP space or by marker genes**. Unsupervised Jaccard-Louvain clustering on a kNN neighbor graph delineates distinct cell types, including adaptive immune cells, monocytes, glial/neuronal cells, and erythrocytes. UMAP plots are binned in hexagons: each single hexagon represents a merged representation of all cells falling within the region. The central UMAP plot is colored by the majority cell type. Different cell types are easily distinguishable in 2-D UMAP plots. The other schex-UMAP plots show gene expression values of selected characteristic marker genes projected onto cells. The color gradient bar represents log-normalized gene expression values. Yellow represents the maximal expressed value, while purple represents the lowest expression values. Markers of distinct immune subpopulations are detected in our data: CD8 T-cells (*CD8A*), NK cells (*GZMB*), B cells (*MS4A1*). Similarly, different non-neuronal cells can be detected in our analysis: astrocytes (*GFAP*), neurons (*SNAP25*), and oligodendrocytes (*OLIG2*). Monocytes (*LYZ*) localize close to our microglial cells and were used for comparative expression of marker genes in Fig. 2b. Red blood cells (*HBB)* were also easily discriminable. (**B**) **Microglia are the predominant cell type recovered across regions and diseases.** Bar plots showing the relative representation of different cell types across different metadata parameters, with each bar summing to 100%. Overall, 95.7% of cells are microglial, 2.2% are adaptive immune, 1.5% are glial/neuronal, 0.4% are monocytic, and 0.3% are erythrocytes. The upper bar plot shows proportion of each overarching cell group across regions, while the lower plot shows the same across diseases. Mono monocytes, RBC red blood cells, LOAD late-onset Alzheimer's disease, EOAD early onset Alzheimer's disease, MCI mild cognitive impairment, CNTRL control, DLBD-PD diffuse Lewy body disease-Parkinson's disease, PSP progressive supranuclear palsy, TLE temporal lobe epilepsy, MS multiple sclerosis, ALS amyotrophic lateral sclerosis, FTD frontotemporal dementia, HD Huntington's disease, DNET dysembryoplastic neuroepithelial tumor, BA Brodmann area, AWS anterior watershed, OC occipital cortex, TNC temporal neocortex, H hippocampus, TH thalamus, SC spinal cord, SN substantia nigra, FN facial nucleus.

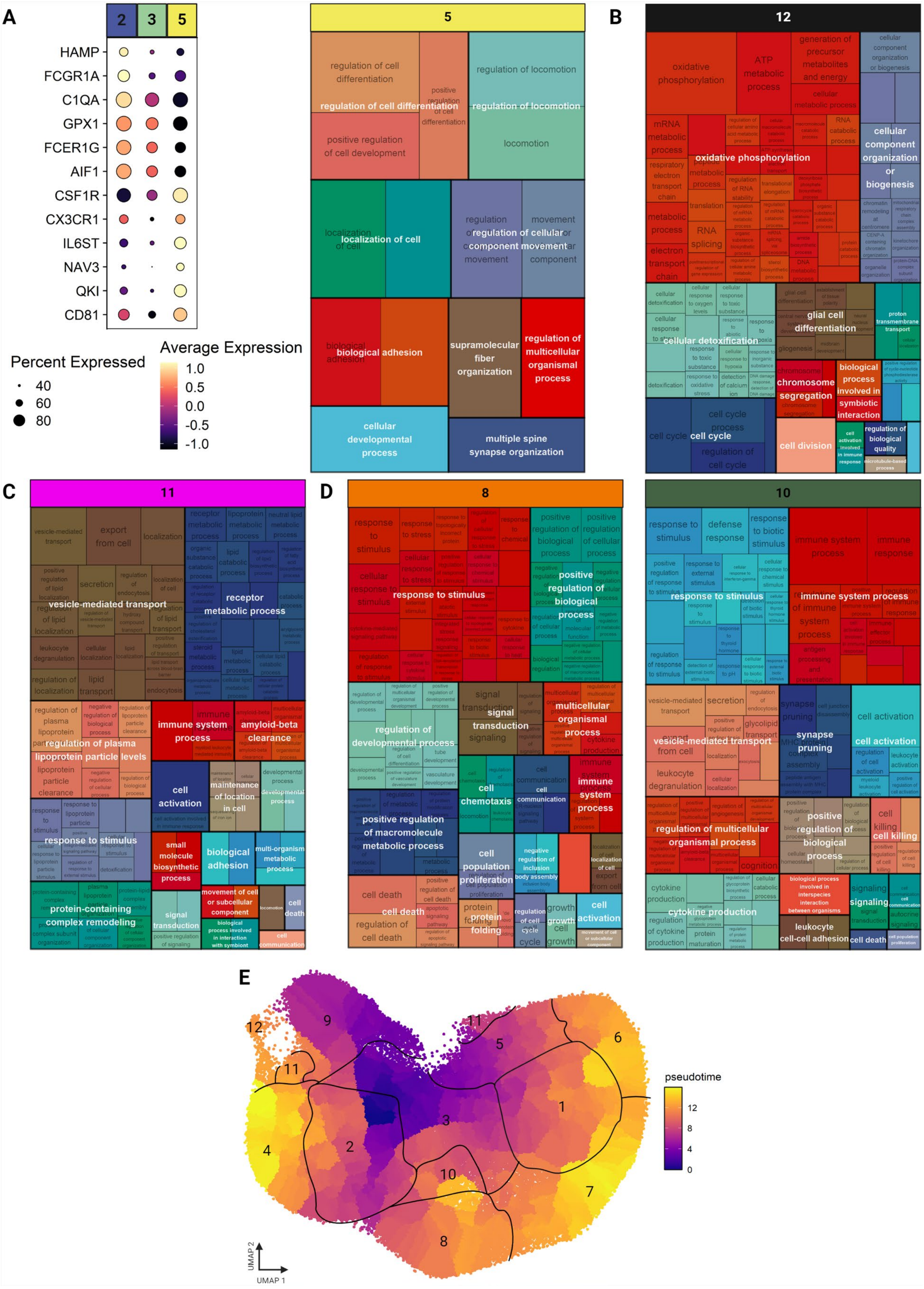


Extended Data Fig. 2 | Quality control metrics across our data after downsampling to account for 10x chemistry differences. (**A-F**). Violin plots showing the distribution of our cellular data with overlaid boxplots. The center of boxplots is the median, and the hinges of the box span the 25% to 75% percentiles. Whiskers represent 1.5 IQR from the nearest hinge. Outliers are not shown in this visualization, nor are minima or maxima. Further information about metadata traits and number of cells included in each violin plot may be found in Supplementary Table 1 under 'QC_' tabs. The distributions of unique molecular identifiers (UMIs) and genes detected on a per-cell level after downsampling are similar across donors (**A**), clusters (**B**), genders (**C**), 10x chemistry versions (**D**), regions, (**E**), and diagnoses (**F**). Notably, after downsampling, differences between 10x chemistry versions in these metrics are largely eliminated. (**G**) **Validation of population stability by resampling and reclustering demonstrates that overlap of gene expression is largely observed for clusters with similarly related families, such as 2 and 4, or for intermediate subsets such as 5 and 3.** To evaluate clustering stability, we randomly sampled ¾ of the cells from our dataset and ran our clustering pipeline with identical parameters. We recorded the frequency of 'misclassification', where cells were re-clustered into clusters different from the one that contained most cells with the same original classification. This process was repeated between pairs of cells, and repeated 50 times for each comparison. Cells were considered to be classified into the 'correct' class if they were assigned correctly in ¾ of classification runs. Otherwise, they were considered 'misclassified' into a different cluster. Classification frequency is visualized in a heatmap here. LOAD late-onset Alzheimer's disease, EOAD early onset Alzheimer's disease, MCI mild cognitive impairment, CNTRL control, DLBD-PD diffuse Lewy body disease-Parkinson's disease, PSP progressive supranuclear palsy, TLE temporal lobe epilepsy, MS multiple sclerosis, ALS amyotrophic lateral sclerosis, FTD frontotemporal dementia, HD Huntington's disease, DNET dysembryoplastic neuroepithelial tumor, BA Brodmann area, AWS anterior watershed, OC occipital cortex, TNC temporal neocortex, H hippocampus, TH thalamus, SC spinal cord, SN substantia nigra, FN facial nucleus.



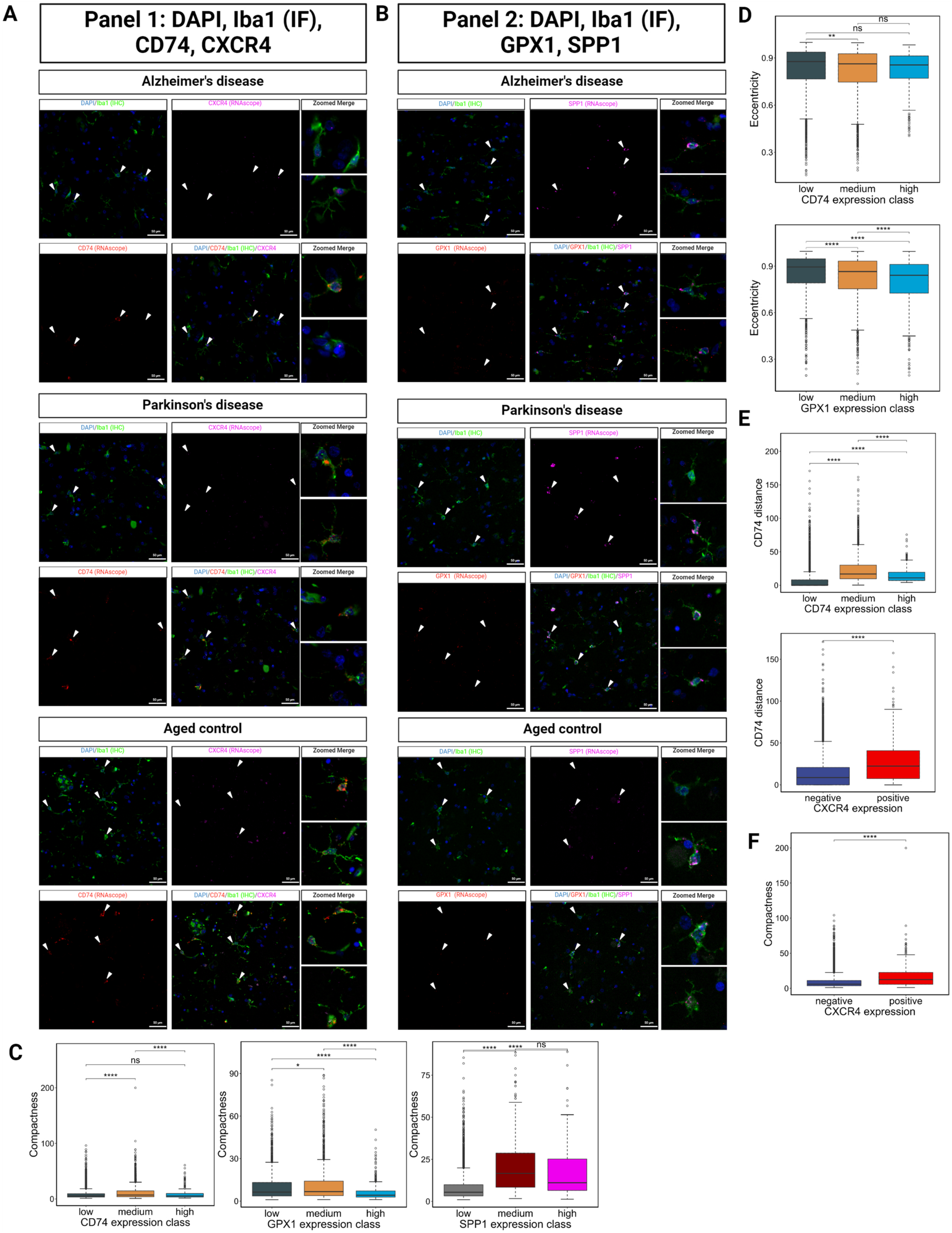
Extended Data Fig. 3 | See next page for caption.

Extended Data Fig. 3 | Microglial proportions across individual donors and donor-region pairings. (**A**) **Proportions of microglial subtypes across single donors**. Proportions of microglial subtypes are plotted by donor, with selected metadata annotated in a header bar above. Each bar represents a single donor and sums to 100%. Samples are clustered hierarchically based on proportions of each subtype. Donors have variability in the exact proportions of different subtypes but exhibit consistent amounts of the most common subtypes in our dataset, clusters 1 through 6. (**B**) **Proportions of microglial subtypes across region-donor pairings**. Samples are aggregated to donor-region pairings (for example, AD1-BA9) to give a proportion of different clusters for each region for each individual. Boxplots are computed for specific region-disease pairings showing the median (center), 25% (left hinge), and 75% (right hinge), for the proportion of cells across all samples for which that combination of disease and region was sampled. Whiskers represent 1.5 IQR from the nearest hinge, and outliers are not shown, nor are minima or maxima. Proportions are shown on the x-axis, and the scale varies depending on the cluster in question. [Number of independent samples per category: TNC_TLE (6), TNC_PSP (1), TH_MS (2), SN_PSP (1), SN_LOAD (3), SN_DLBD-PD (5), SN_CNTRL (1), SC_ALS/FTD (2), SC_ALS (9), OC_TLE (1), OC_Stroke_lesion (1), Lesion_MS (1), H_TLE (2), H_PSP (1), H_LOAD (14), H_HD (1), H_FTD (1), H_EOAD (2), H_CNTRL (1), FN_ALS (4), DNET_DNET (1), BA9_Stroke_lesion (1), BA9_PSP (1), BA9_MS (2), BA9_MCI (4), BA9_LOAD (35), BA9_HD (1), BA9_FTD (1), BA9_EOAD (2), BA9_DLBD-PD (5), BA9_CNTRL (1), BA9_ALS/FTD (2), BA9_ALS (8), BA4_CNTRL (1), BA4_ALS/FTD (2), BA4_ALS (9), BA20_LOAD (9), BA20_HD (1), BA20_EOAD (2), AWS_MS (2), AWS_MCI (3), AWS_LOAD (13)].



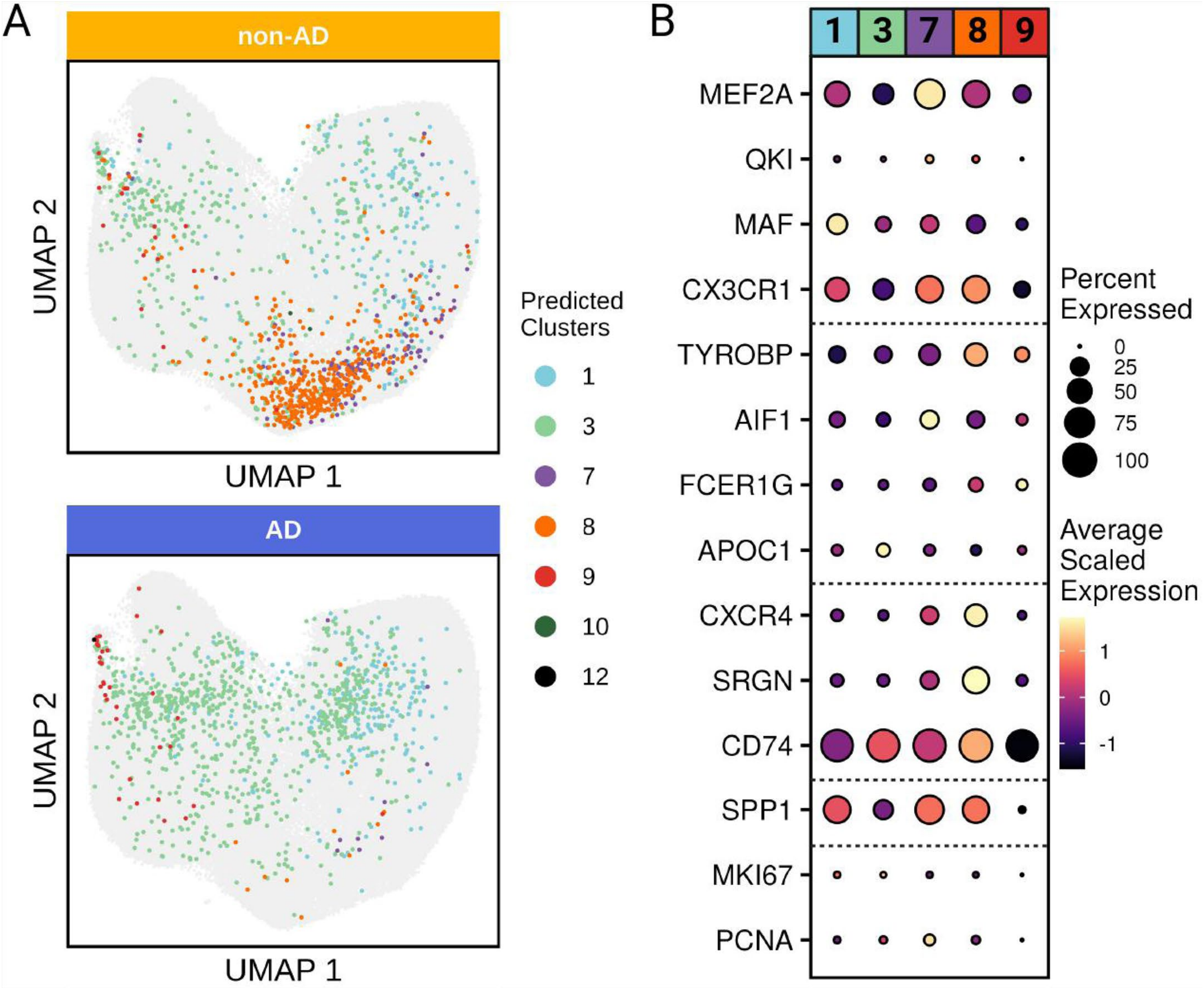
Extended Data Fig. 4 | See next page for caption.

Extended Data Fig. 4 | Further exploration of microglial phenotypes with pseudotime analysis and GO annotation validates our trajectory map and reveals subsets associated with motility, lipid trafficking, and proliferation. (**A**) **Cluster 5, an intermediate cluster, shows association with motility**. On the left, the size of the circle represents the percentage of cells in a cluster that express the gene, with no circle plotted if less than 10% of cells in a cluster express the gene. The color of the circle represents the z-scored expression of the gene. Cluster 5 expresses a transcriptional signature partially overlapping with the core homeostatic or transitional clusters, 2 and 3, but expresses unique sets of genes associated with motility. GO annotation was performed with topGO and summarized with rrvgo. Parent terms are shown in white, overlaid over child terms. Terms associated with motility are enriched in cluster 5. (**B**) **Cluster 12 is associated with oxidative phosphorylation and proliferation**. (**C**) **Cluster 11 interfaces with lipids and beta-amyloid**. (**D**) **GO annotation of clusters 8/10 parallels results of Reactome pathway analysis, highlighting common immunological activation but divergence in other aspects of phenotype**. (**E**) **Trajectories of state shift in pseudotime analysis parallel those seen in other analyses**. Monocle3 was used to build a pseudotime trajectory across our dataset, setting the root point at the boundary of clusters 2 and 3. Shifts in pseudotime from this root point reinforces the directionality laid out in the constellation diagram, suggesting that a broad intermediate gradient between a series of terminal points exists, with pseudotime scores in 6-7, 4, and 10 showing most divergence from the root point. GO gene ontology.

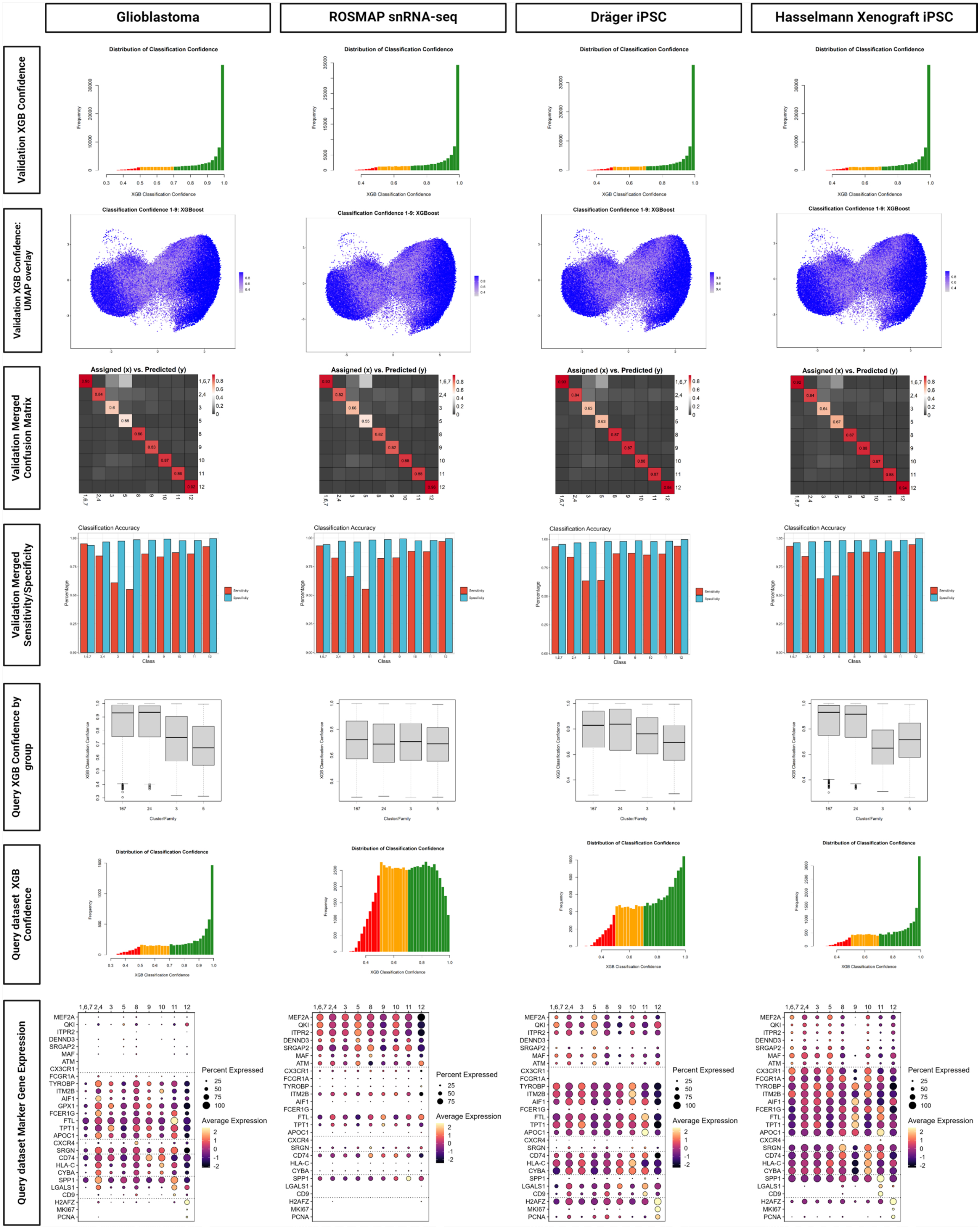


Extended Data Fig. 5 | See next page for caption.

Extended Data Fig. 5 | Additional representative images from our joint RNAscope/IF and CellProfiler measures highlight morphological differences between expression-defined subtypes. Representative images are shown for both panel 1 (**A**) and panel 2 (**B**) across different diseases. (**C**) **Compactness is highest in the medium classes of *CD74*, *GPX1*, and *SPP1*-defined expression groups**. Compactness (a measure of ramification, where high values indicate high ramification) is shown across *CD74*-, *GPX1*-, and *SPP1*-expressing IBA1+ microglial cells quantified using CellProfiler. For this and following panels, significance was calculated with two-sided, two-sample Welch's t-tests. Multiple testing correction was performed with Holm-Bonferroni correction. For boxplots in these visualizations, the center is the median, and the hinges of the box span the 25% to 75% percentiles. Whiskers represent 1.5 IQR from the nearest hinge. Outliers are shown as circles, but minima and maxima are not explicitly depicted. Significance thresholds for p-values: >0.05 = ns, <0.05 = *, <0.01 = **, <0.005 = ***. (**D**) **Compactness is higher in the *CXCR4*+ class**. (**E**) **Eccentricity is highest in the low classes for *CD74* and *GPX1***. Eccentricity (a measure of shape, where 0 is a circle and 1 is a line), is shown across *CD74*- and *GPX1*- expressing Iba1+ microglia. (**F**) ***CD74* distance is highest in the *CD74* medium group, but also in the *CXCR4*$^{+}$ group**. *CD74* distance (calculated as the median of all puncta for a given cell from the cellular centroid) is shown across *CD74*-, and *CXCR4*-expressing Iba1+ microglia. Number of cells per expression class are as follows. *CD74*: low (3756), medium (3333), high (329), *GPX1*: low (1404), medium (1653), high (329), *SPP1*: low (3216), medium (388), high (125), *CXCR4*: positive (322), negative (7096). 16 tissue sections were stained with panel 1 (*CD74*/*CXCR4*) and eight were stained with panel 2 (*GPX1*/*SPP1*).

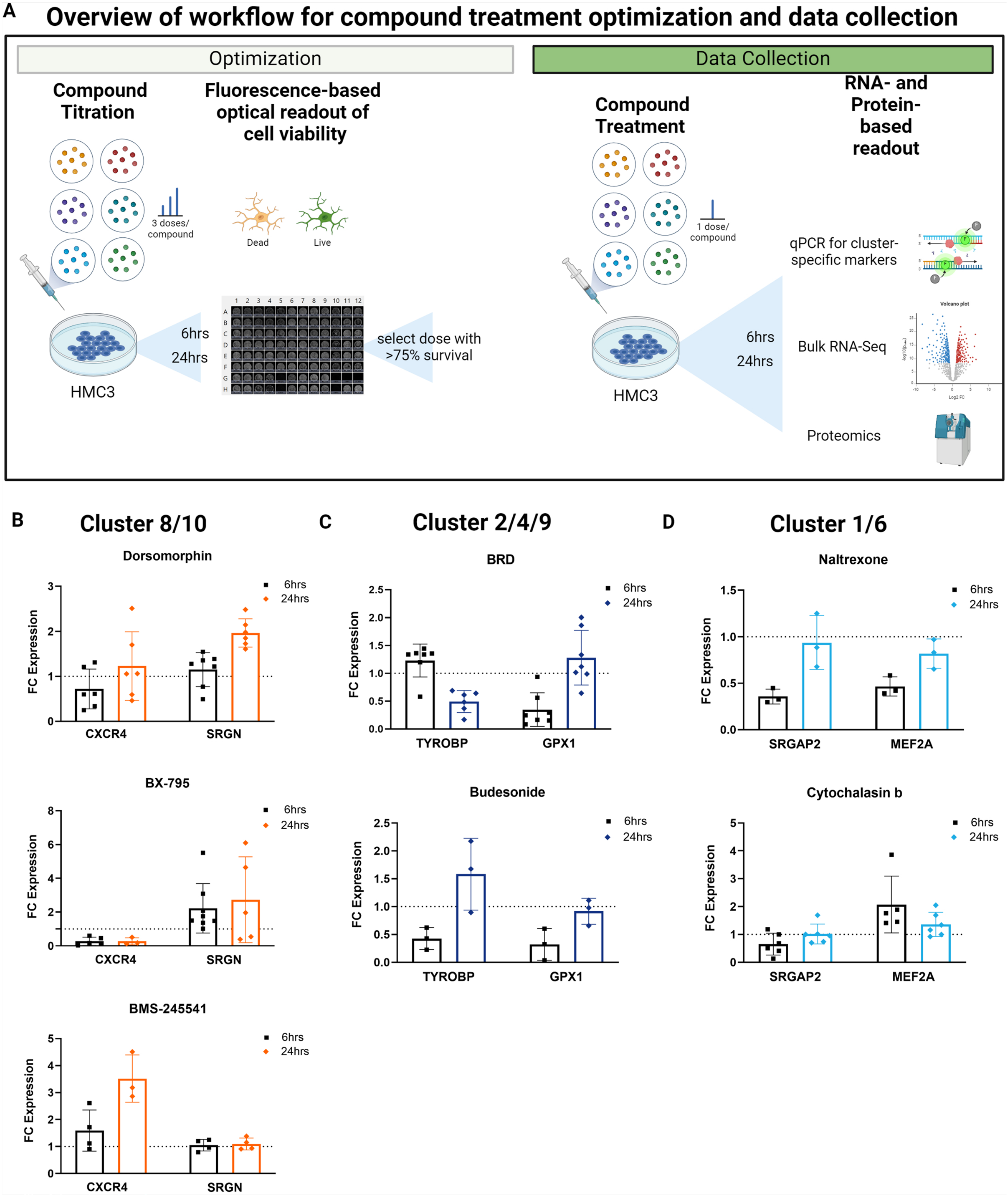


Extended Data Fig. 6 | *In situ* merFISH validation of microglia subtypes. (**A**) **Projection of microglial cells into the established scRNAseq model**. UMAP space showing predicted cluster subtypes within a projected UMAP space (established model shown in greyed-out background). Seven out of twelve microglial subtypes were identified across AD (blue) and non-AD (yellow) cortex tissue, with different observed proportions. Clusters 8/10 show depletion in AD cortex (< 1%) compared to non-AD cortex (35.7%). (**B**) **Expression signatures of predicted clusters *in situ***. Microglia predicted to belong to clusters 8/10 show a greater average expression and percent expression of *CXCR4*, *SRGN*, and *CD74*. Showing clusters with at least 5 predicted microglia.



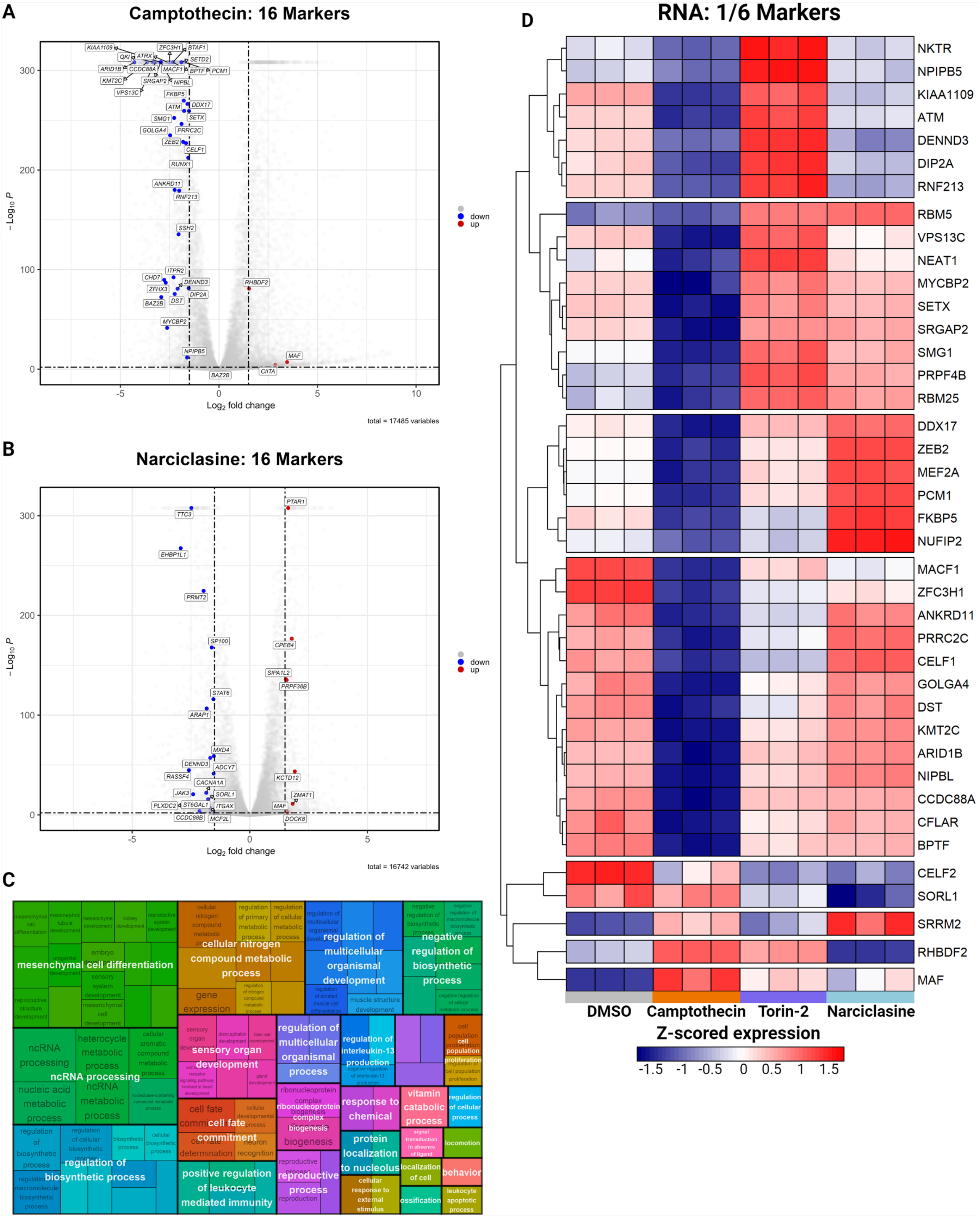
Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | Performance metrics across models trained for different datasets. Each row contains a different performance metric, while each column represents a single dataset. Training and validation sets were identical, but mNN correction incorporates the query dataset, slightly modifying input data. Accuracy metrics are derived from analysis of the holdout validation set, consisting of approximately 50% of the original dataset not used for training either SVM or XGB models (104902 cells). The first row presents histograms of XGBoost classification confidence for cells in the validation set, highlighting cells below 70% confidence in yellow and below 50% in red (the latter cells are dropped). Most cells in the validation set are classified with high confidence. Row 2 contains a UMAP visualization of classification confidence, revealing higher confidence for cells at the UMAP periphery and lower confidence for intermediate cells. Row 3 shows confusion matrices for the validation set. Row 4 presents sensitivity and specificity per class, which are comparable across different datasets. Row 5 shows boxplots for XGB classification confidence across the 4 classes. Boxplots represent the median (center), 25% (lower hinge), and 75% (upper hinge) percentiles. Whiskers extend to 1.5 times the IQR from the nearest hinge, with more extreme values represented as circles. Minima and maxima are not explicitly depicted. Classification confidence varies substantially depending on the data, with the ROSMAP data being the only dataset where classification confidence for families 167 and 24 is generally comparable to that for 3 and 5. Row 6 contains histograms of XGBoost classification confidence for the query cells. Notably, the glioblastoma and xenograft data have similar classification confidence to the validation set, but the ROSMAP data, and to a lesser extent, the Dräger data, diverge noticeably. Finally, row 7 shows marker gene expression across assigned labels in the query datasets. The size of the circle represents the percentage of cells in each cluster expressing the gene (no circle plotted if less than 10% of cells in a cluster express the gene). The color of the circle represents z-scored expression of the gene. Despite systematic differences, label transfer aligns expression profiles effectively.



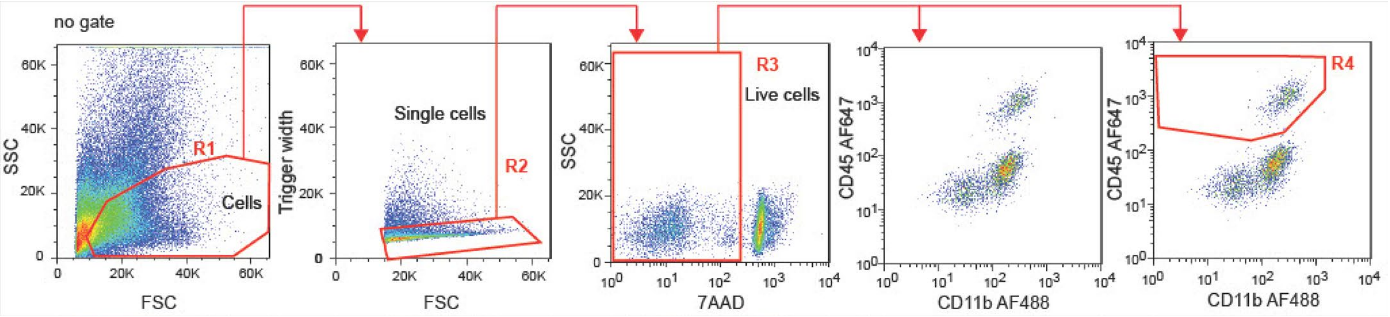
Extended Data Fig. 8 | See next page for caption.

Extended Data Fig. 8 | Screening of *in silico* predictions identifies successful hits and compounds that fail to drive predicted signatures. (**A**) **Schematic overview of workflow for compound treatment**. To explore the correct dosage for downstream studies, we conducted dose titration to examine viability of cells after treatment with varying dosages of our drugs. After choosing optimal concentrations, we conducted initial screening with qPCR to select candidates for final validation, then conducted final validation with bulk RNA-seq and proteomics. (**B**)-(**D**) **qPCR results for different cluster families.** Results not shown in Fig. 8b-d are shown here. Some compounds had effects on specific marker genes, but these did not pass our criteria for further study. Bars represent mean fold change expression, and error bars represent SD. All replicates are biological. Number of replicates per experiment as follows - Dorsomorphin: 6hrs: CXCR4 - n = 6, SRGN – n = 7; 24hrs: both n = 6, BX-795: 6hrs: CXCR4 - n = 5, SRGN – n = 8; 24hrs: CXCR4 - n = 3, SRGN – n = 5, BMS-2455421: 6hrs: both - n = 4; 24hrs: CXCR4 - n = 3, SRGN – n = 4, BRD: 6hrs: both - n = 7; 24hrs: TYROPB - n = 6, GPX1 – n = 7, Budesonide: 6hrs: n = 3; 24hrs: n = 3, Naltrexone: 6hrs: n = 3; 24hrs: n = 3, Cytochalasin b: 6hrs: SRGAP2 - n = 6, MEF2A – n = 5; 24hrs: both n = 6.



Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | Different compounds modulate different aspects of the cluster 1/6 signature at the transcriptomic level. (**A**) **Camptothecin downregulates the cluster 1/6 signature**. Bulk RNA-seq was generated from HMC3 cells treated with our candidate drugs for 24 h. Data was analyzed with DESeq2, which fits a negative binomial model to the data then uses Wald significance tests with Benjamini-Hochberg correction, and fold change shrinkage was performed with ashr. To examine the genes associated with cluster families, we took the top 20 non-overlapping genes for each individual cluster in our overarching groupings that were present in the differentially expressed gene list for each compound, irrespective of directionality and plotted them in volcano plots. FDR threshold was set to 0.01 and fold change threshold was set at 1.5. (**B**) **Narciclasine does not upregulate the cluster 1/6 signature**. (**C**) **Narciclasine upregulates GO processes also found in cluster 1/6**. GO annotation was computed on differentially expressed genes that passed an FDR threshold of 0.01 and a fold change threshold of 1.5. Terms were grouped based on similar etiology and parent terms were overlaid. Notably, Narciclasine drives metabolic shifts such as in nitrogen-containing metabolism, heterocyclic metabolism, and nucleic acid metabolism, that are strongly enriched in clusters 1/6 (Fig. 3a). (**D**) **Narciclasine and Torin-2 drive distinct modules of cluster 1/6 marker genes**. Cluster 1/6 genes were selected and shown in a row-scaled, zero-centered heatmap. Columns are individual replicates, and rows are genes. These two compounds appear to drive separate modules of genes associated with cluster 1/6. Camptothecin downregulates almost all 1/6 associated genes.



Extended Data Fig. 10 | Representative flow gating images. Cells that were stained with anti-CD11b and anti-CD45 antibodies and 7AAD were sorted by flow cytometry. Flow gates demonstrate selection of live singlets that are CD45-positive.

nature portfolio

Corresponding author(s): Philip L. De Jager

Last updated by author(s): Jun 3, 2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	CellRanger software V3.1.0 (from 10x Genomics) was used to align and quantify single-cell RNA-seq transcripts for our single-cell data and CellRanger V6.0.0 was used to align and quantify single-nucleus data. Nikon Elements (NIS-Element AR 5.21.03) was used to acquire images from tissue sections. For bulk RNA-seq, Picard version 1.83 was used for QC, RSeQC version 2.6.1. STAR version 2.5.2a was used to align reads, Bowtie2 version 2.1.0 was used to measure rRNA abundance, and annotated genes were quantified with featureCounts version 1.4.3-p1. For bulk proteomic data, the DIA-NN search engine was used to search the acquired diaPASEF raw files. BD FACS Software 1.2.0.142. was used to collect and gate flow cytometry data for sorting of single microglia.
Data analysis	R statistical software (v4.1.0) was used to analyze single-cell and bulk transcriptomic and proteomic data, and all custom code is available at https://github.com/jtuddenham/single-cell-microglia-v2.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Raw scRNA-seq data (fastq files) generated from CD45+ cells isolated from autopsy samples were deposited to GEO (https://www.ncbi.nlm.nih.gov/geo/) under accession number GSE204702, or Synapse, with ID number syn61001870. Bulk RNA-seq data from compound-treated HMC3 cells were deposited to GEO under accession number GSE202556. Bulk proteomic data from compound-treated HMC3 cells were deposited to ProteomeXChange (http://www.proteomexchange.org/) under accession number PXD033844. Correspondence & material/data requests should be addressed to Philip L. De Jager.
Data repurposed for label transfer was retrieved from GEO, under accession numbers GSE133432, GSE178317, and GSE103224. Bulk RNA-seq data from ROSMAP used to derive associations of gene expression with clinicopathological traits can be accessed on Synapse (syn25741873).
Datasets/databases used in this study included: CMAP (GSE92742)

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	Sex and gender were self-reported in this study. We have a total of 26 male donors and 48 female donors. As our interest was in examining microglial diversity independent of metadata parameters, we did not perform sex- or gender-based analysis.
Reporting on race, ethnicity, or other socially relevant groupings	We did not perform race- or ethnicity-based analysis.
Population characteristics	Details of the acquisition of autopsy samples from Rush University Medical Center/Rush Alzheimer's Disease Center (RADC) in Chicago, IL (Dr. Bennett) and Columbia University Medical Center/New York Brain Bank in New York, NY (Drs. Vonsattel and Teich), as well as surgically resected brain specimens from Brigham and Women's Hospital in Boston, MA (Drs. Sarkis, Cosgrove, Helgager, Golden, and Pennell) were detailed in our prior publication (Olah et al. 2020, Nature Communications). In addition, samples were obtained from donation programs at Massachusetts General Hospital, Boston, MA (Drs. Bradley T. Hyman and Matthew Frosch), Banner Sun Health Research Institute, Sun City (Dr. Thomas G Beach), and Rocky Mountain MS Center, Denver, CO (Dr. John Corboy). All brain specimens were obtained through informed consent and/or brain donation program at the respective organizations. All procedures and research protocols were approved by the corresponding ethical committees of our collaborator's institutions as well as the Institutional Review Board (IRB) of Columbia University Medical Center (protocol AAAR4962). Detailed descriptions of the Religious Orders Study and the Memory and Aging Project (ROS/MAP) can be found in the following publications: PMIDs 29865057, 22471860, 22471867. Further information on the brain donation system at Massachusetts General Hospital can be found at https://www.madrc.org/brain-autopsy-and-donation-information. Further information on the brain donation system at Rocky Mountain Multiple Sclerosis center can be found at: https://www.mscenter.org/research/tissue-bank/information-for-researchers. Description of the brain donation system at Sun Health Research institute can be found here: PMID 18347928. The description of the brain bank at Columbia University Medical Center can be found here: PMID: 29496134. All donors were consented for the use of their tissue for research purposes. Age range for donors ranged between 22 and 90+, specific diagnoses included ALS, ALS/FTD, Control, DLBD-PD, DNET, EOAD, FTD, HD, LOAD, MCI, MS, PSP, Stroke, and TLE. Treatment information and genotypic information were not uniformly available for donors.
Recruitment	See above.
Ethics oversight	This study was approved by ethics committees from 1) Rush University, Chicago, IL, 2) Columbia University, New York, NY, 3) Brigham and Women's Hospital, Boston, MA, 4) Massachusetts General Hospital, Boston, MA, 5) Banner Sun Health Research Institute, Sun City (Dr. Thomas G Beach), and 6) Rocky Mountain MS Center, Denver, CO.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences   □ Behavioural & social sciences   □ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size

Sample sizes were not calculated ahead of time. Moreover, as we do not ask specific questions about association of microglial proportions with specific diseases or regions, we do not require specific sample numbers for this type of analysis. Our primary constraint was ensuring that there was sufficient representation of cell type heterogeneity across our sample pool; as such, we sampled as extensively as possible. Single-cell RNA-seq studies of human tissue also cannot control the number of cells per donor, as it is defined by the quality of the sample.

Data exclusions

First, for samples that used hashing antibodies, we removed unlabeled cells and doublets ascertained using demuxmix (https://github.com/cu-ctcn/demuxmix). Next, we excluded cells with fewer than 500 transcripts or more than 10,000 transcripts (Unique Molecular Identifiers), as well as cells with more than 10% mitochondrial reads. These thresholds accord with the standards of the field and remove doublets and low-quality/dying cells. These standards were established ahead of time, and help ensure that downstream analyses are not polluted by low-quality data.

Replication

When possible, we replicated the results of our analyses using similar databases. For example, for our indirect disease association, we replicated patterns of enrichment of multiple sclerosis susceptibility genes from a recent publication by the International Multiple Sclerosis Genetics Consortium extensively mapping genomic risk loci in MS (PMID: 31604244) using data from the GWAS catalog to evaluate enrichment of GWAS-based risk genes in our clusters.

Similarly, for our compound stimulation work, we sought to replicate the results of our initial qPCR screen showing upregulation of cluster-associated genes in our 3 compounds of interest: Torin-2, Narciclasine, and Camptothecin. We had at least 4 independent replicates for each of our compound treatment conditions for qPCR.

In addition, we replicated the results of our initial screen for Torin-2 and Campthothecin at the transcriptomic level using bulk RNA-sequencing, finding that Torin-2 and Camptothecin are especially congruent with the qPCR results, as Narciclasine drives a slightly different aspect of the broad transcriptomic signature we were targeting. Bulk RNA-seq experiments had 3 independent replicates per treatment condition.

Notably, we also sought to replicate the effect of compound stimulation at the proteomic level, finding that Camptothecin still drives the predicted signature at the proteomic level, although Torin-2 and Narciclasine do not drive the expected result at this level. This latter result is not wholly surprising, as the discordance between RNA and protein is well documented, especially in microglia, and our expected target signature was defined entirely at the RNA level. As with bulk RNA-seq experiments, proteomic experiments were conducted with 3 independent replicates per treatment condition.

Randomization

Aside from compound treatment, where the treatment of interested defined the groupings used for analysis, none of our analyses required specification of sample groups so no randomization was performed. To correct for potential batch effects and effects of different 10x technologies, we downsampled our 10x v3 data to the same depth as our v2 data and used SCTransform and mNN to correct for batch effects, as described in our methods section.

Blinding

In this study, we did not have a hypothesis to test, and thus, blinding of team members to the characteristics of the samples was not necessary. None of the algorithms for clustering or label transfer took sample or donor metadata into account.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study	
□	☒	Antibodies
□	☒	Eukaryotic cell lines
☒	□	Palaeontology and archaeology
☒	□	Animals and other organisms
☒	□	Clinical data
☒	□	Dual use research of concern
☒	□	Plants

Methods

n/a	Involved in the study	
☒	□	ChIP-seq
□	☒	Flow cytometry
☒	□	MRI-based neuroimaging

Antibodies

Antibodies used

TotalSeq™-B0255 anti-human Hashtag 5 Antibody (Biolegend, Cat #: 394639; RRID: AB_2820042) [1 µg per 100 µl staining volume volume]

TotalSeq™-B0256 anti-human Hashtag 6 Antibody (Biolegend, Cat #: 394641; RRID: AB_2820042) [1 µg per 100 µl staining volume volume]

TotalSeq™-B0257 anti-human Hashtag 7 Antibody (Biolegend, Cat #: 394643; RRID: AB_2820043) [1 µg per 100 µl staining volume volume]

TotalSeq™-B0258 anti-human Hashtag 8 Antibody (Biolegend, Cat #: 394645; RRID: AB_2820044) [1 µg per 100 µl staining volume volume]

TotalSeq™-B0259 anti-human Hashtag 9 Antibody (Biolegend, Cat # 394647; RRID: AB_2820045) [1 µg per 100 µl staining volume

volume]
TotalSeq™-B0260 anti-human Hashtag 10 Antibody (Biolegend, Cat #: 394649; RRID: AB_2820046)
Alexa Fluor® 488 anti-mouse/human CD11b Antibody (Biolegend, Cat #: 101217; Lot #: multiple) [0.5 µg per 100 µl staining volume volume]
Alexa Fluor® 647 anti-human CD45 Antibody (Biolegend, Cat #: 304018; Lot #: multiple) [0.5 µg per 100 µl staining volume volume]
Goat anti-Human Iba1 (Wako, Cat #: 01127991; Lot #: SKK1868) [dilution 1:50]
Donkey anti-Goat IgG (H+L) Cross-Adsorbed Secondary Antibody, Alexa Fluor 488 (Invitrogen, Cat #: A11055; Lot #: 2211210) [dilution 1:500]

Validation

As per the manufacturer's website: "Each lot of this antibody [TotalSeq-B] is quality control tested by immunofluorescent staining with flow cytometric analysis and the oligomer sequence is confirmed by sequencing. TotalSeq™-B antibodies are compatible with 10x Genomics Single Cell Gene Expression Solutions". Validation on human PBMCs is available under the application note from BioLegend titled "Efficient Multiplexing of Samples Using TotalSeq™ Hashtag Antibody Oligonucleotide Conjugates for Single-Cell RNA and Proteomics Studies".
Anti-Iba1 has been validated by the manufacturer and multiple subsequent publications demonstrating its utility in immunohistochemistry and western blotting (see manufacturer's website). It has been used by our group and others to detect microglia in the brain (PMID: 33257666). Biolegend's anti-human CD11b and CD45 antibodies have been used by our group for several years, and transcriptomic analysis of cells sorted with these antibodies confirm that they primarily label brain myeloid cells, the majority of which are microglia (PMIDs: 29416036 and 33257666).
The Donkey anti-goat secondary antibody has been extensively validated by the manufacturer for use in flow cytometry, ICC, IF, and IHC. It has over 2000 references.

Eukaryotic cell lines

Policy information about cell lines and Sex and Gender in Research

Cell line source(s)

HMC3 (human, ATCC)

Authentication

No sequencing-based authentication of the identity of this cell line was performed. Conversely, we tested expression of microglial marker genes by qPCR every 3 passages to ensure that we were working with a model system that transcriptionally resembled our cell of interest, microglia.

Mycoplasma contamination

Cell lines were not tested for mycoplasma, but no evidence of infection was observed.

Commonly misidentified lines
(See ICLAC register)

No commonly misidentified lines were used in this study.

Plants

Seed stocks

Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.

Novel plant genotypes

Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.

Authentication

Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.

Flow Cytometry

Plots

Confirm that:

☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☒ All plots are contour plots with outliers or pseudocolor plots.

☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation

Procurement of specimens is described above. The isolation of microglia was performed according to our published protocol (Olah et al. 2020, Nature Communications), with minor modifications. In case of the cortical autopsy samples (BA9/46, BA4, BA17/18/19), the cortex (grey matter and the underlying white matter (subcortical white matter) were dissected under a stereomicroscope. The subcortical white matter samples were not used in this study. The epilepsy surgery samples of temporal lobe (BA20/21) were processed without dissection as in this case the cortical white and grey matter was not always distinguishable due to the surgical procedure. The substantia nigra (SN) and the thalamus (TH) were dissected by separation from the surrounding white matter tracts. The hippocampus samples (H) contained the dentate gyrus, CA4/CA3/CA2 and CA1 regions, both white and grey matter. The spinal cord sample (SC) was sampled at the level of lumbar section and included both white and grey matter. The anterior watershed area (AWS) deep white matter did not need any further dissection. All steps of the protocol were performed on ice. The dissected tissue was placed in HBSS (Lonza, 10-508F) and weighed. Subsequently the tissue was homogenized in a 15 ml glass tissue grinder - 0.5 g at a time. The resulting homogenate was filtered through a 70 um filter and spun down at 300rcf for 10 minutes. The pellet was resuspended in 2 ml staining buffer (RPMI (Fisher, 72400120) containing 1% B27) per 0.5 g of initial tissue and incubated with anti-myelin magnetic beads (Miltenyi, 130-096-733) for 15 minutes according to the manufacturer's specification. The homogenate was than washed once with staining buffer and the myelin was depleted using Miltenyi large separation columns (Miltenyi, 130-042-202). The cell suspension was spun down and was then incubated with anti-CD11b AlexaFluor488 (BioLegend, 301318) and anti-CD45 AlexaFluor647 (BioLegend, 304018) antibodies as well as 7AAD (BD Pharmingen, 559925) and cell hashing antibodies (for catalogue numbers of cell hashing antibodies see Table S1) for 20 minutes on ice. Subsequently the cell suspension was washed twice with staining buffer, filtered through a 70 µm filter and the CD11b+/CD45+/7AAD- cells or CD45+/7AAD- cells were sorted on a BD FACS Aria II or BD Influx cell sorter. Cells from each brain region were sorted in a separate A1 well of a 96 well PCR plate (Eppendorf, 951020401) containing 100 µl of PBS buffer with 0.3% BSA. Following sorting cell from different brain regions were combined and immediately submitted to single cell capture, reverse transcription and library construction on the 10x Chromium platform. All sorting was performed using a 100 µm nozzle. The sorting times varied according to the quality of the sample but was on average between 10 and 20 minutes per sample. The sorting speed was kept between 8000 - 10,000 events per second.

Instrument

BD's Aria IIu and BD Influx sorters were used for fluorescent activated cell sorting of microglial cells from human brain.

Software

BD's FACSDiva version 8.0.1 software was used during fluorescent activated cell sorting of microglial cells from human brain.

Cell population abundance

Microglial cells represented on average 0.4% of all the events. Among the 7AAD- live cells the CD11b+/CD45+ cells represented 50% (RBCs will also show up as 7AAD- since they lack a nucleus). The analysis of sorted cells showed that they were ~96% microglia (CD11b+/CD45+/7AAD-) cells.

Gating strategy

The detailed description of the gating strategy was included in our previous publications (PMIDs: 29416036 and 33257666). Briefly, cells were gated on the FSC/SSC scatter plots (Gate 1), from which the dead cells were excluded based on their 7AAD positivity (Gate 2: 7AAD- events). The third gate was placed on the CD11b/CD45 double positive events (Gate 3: CD11b+/CD45+).

☒ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.