Article

Cellular communities reveal trajectories of brain ageing and Alzheimer's disease

https://doi.org/10.1038/s41586-024-07871-6

Received: 22 February 2023

Accepted: 24 July 2024

Published online: 28 August 2024

Check for updates

Gilad Sahar Green[1], Masashi Fujita[2,10], Hyun-Sik Yang[3,4,5,10], Mariko Taga[2], Anael Cain[1], Cristin McCabe[6], Natacha Comandante-Lou[2], Charles C. White[4], Anna K. Schmidtner[1], Lu Zeng[2], Alina Sigalov[2], Yangling Wang[7], Aviv Regev[6,8,9], Hans-Ulrich Klein[2], Vilas Menon[2,11]✉, David A. Bennett[7,11], Naomi Habib[1,7,11]✉ & Philip L. De Jager[2,4,11]✉

Alzheimer's disease (AD) has recently been associated with diverse cell states[1–11], yet when and how these states affect the onset of AD remains unclear. Here we used a data-driven approach to reconstruct the dynamics of the brain's cellular environment and identified a trajectory leading to AD that is distinct from other ageing-related effects. First, we built a comprehensive cell atlas of the aged prefrontal cortex from 1.65 million single-nucleus RNA-sequencing profiles sampled from 437 older individuals, and identified specific glial and neuronal subpopulations associated with AD-related traits. Causal modelling then prioritized two distinct lipid-associated microglial subpopulations—one drives amyloid-β proteinopathy while the other mediates the effect of amyloid-β on tau proteinopathy—as well as an astrocyte subpopulation that mediates the effect of tau on cognitive decline. To model the dynamics of cellular environments, we devised the BEYOND methodology, which identified two distinct trajectories of brain ageing, each defined by coordinated progressive changes in certain cellular communities that lead to (1) AD dementia or (2) alternative brain ageing. Thus, we provide a cellular foundation for a new perspective on AD pathophysiology that informs personalized therapeutic development, targeting different cellular communities for individuals on the path to AD or to alternative brain ageing.

The molecular characterization of the ageing human brain has expanded considerably following advances in high-throughput molecular measurements and analytical methods[12,13]. While analyses of autopsied brain samples using bulk tissue profiles can yield molecular insights into AD[12], they lose critical details of the brain's intricate cellular architecture. Single-cell (scRNA-seq) and single-nucleus (snRNA-seq) RNA-sequencing technology has offered a different perspective[1–11,14–17], highlighting AD-associated changes in expression programs of multiple cell types (glial, neuronal and vascular cells). Specific subpopulations of each cell type have been implicated in AD, and they have been organized into multicellular communities of various cell types that are involved in AD[1].

One major challenge is resolving the sequence of changes in cell populations involved in AD, and distinguishing these changes from those associated with brain ageing. A second challenge is using cross-sectional data captured at autopsy to reconstruct the causal chain of events leading to the preclinical phase of asymptomatic accumulation of pathology, the appearance of cognitive decline and, finally, dementia[18]. Previous studies have avoided these challenges by using a case–control design, focusing on extremes, such as advanced stages of AD, to identify cell populations associated with the disease; however, such study designs do not capture the heterogeneity of clinical symptoms and pathological manifestations found among older individuals, which are critical to resolve the cellular changes and dynamics found in the early, asymptomatic stages of disease. Here we take advantage of the phenotypic diversity of the participants in the Religious Order Study and the Rush Memory and Aging Project (ROSMAP)[19,20]—two studies of cognitive ageing with prospectively collected brains and deep ante- and post-mortem characterization[19,20]. We sample the entire range of brain states found among older individuals, including those with AD and with non-AD brain ageing. This relatively random sample of older brains enables us to both explore the heterogeneity of the older brain and determine the sequence of events leading to the most common form of dementia, AD.

Specifically, we built a cellular atlas from RNA profiles of 1.65 million nuclei from the dorsolateral prefrontal cortex (DLPFC, BA9) of 437 ROSMAP participants; these data provide the necessary statistical power to investigate cellular changes while accounting for human

[1]Edmond & Lily Safra Center for Brain Sciences, The Hebrew University of Jerusalem, Jerusalem, Israel. [2]Center for Translational & Computational Neuroimmunology, Department of Neurology and Taub Institute for Research on Alzheimer's Disease and the Aging Brain, Columbia University Irving Medical Center, New York, NY, USA. [3]Harvard Medical School, Boston, MA, USA. [4]Broad Institute of MIT and Harvard, Cambridge, MA, USA. [5]Center for Alzheimer Research and Treatment, Department of Neurology, Brigham and Women's Hospital, Boston, MA, USA. [6]Klarman Cell Observatory, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [7]Rush Alzheimer's Disease Center, Rush University Medical Center, Chicago, IL, USA. [8]Department of Biology, Koch Institute of Integrative Cancer Research, Massachusetts Institute of Technology, Cambridge, MA, USA. [9]Present address: Genentech, San Francisco, CA, USA. [10]These authors contributed equally: Masashi Fujita, Hyun-Sik Yang. [11]These authors jointly supervised this work: Vilas Menon, David A. Bennett, Naomi Habib, Philip L. De Jager. ✉e-mail: vm2545@cumc.columbia.edu; naomi.habib@mail.huji.ac.il; pld2115@cumc.columbia.edu

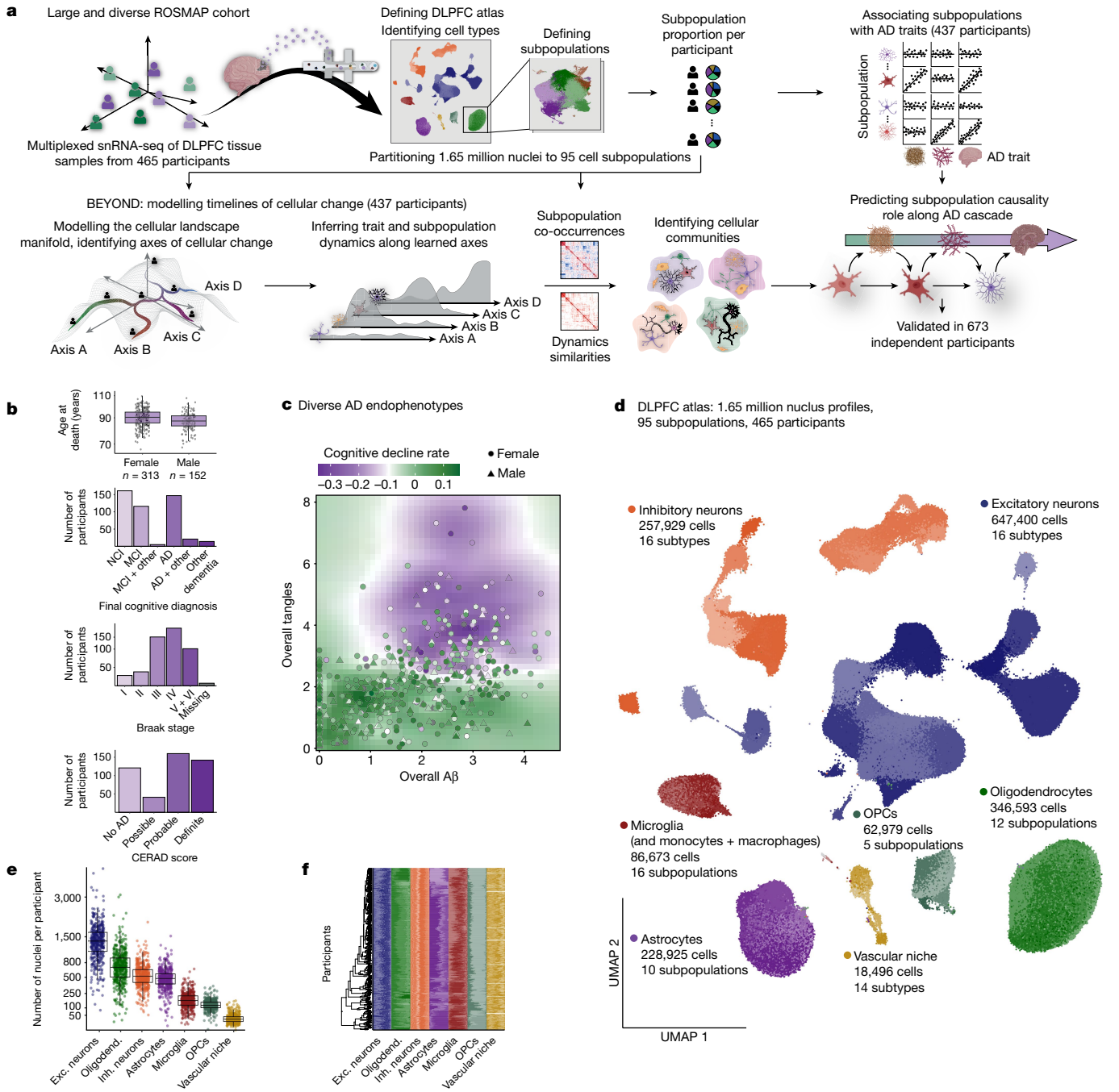


Fig. 1 | Cellular atlas of the human aged DLPFC in older individuals. **a**, Overview of the experimental and analytic steps. **b**,**c**, Clinicopathologic characteristics of the 465 ROSMAP participants. **b**, Participants' age of death, final cognitive diagnosis and distribution of pathologic hallmarks of AD, Aβ (CERAD score) and tau (Braak score) (Methods). Additional details are provided in Supplementary Table 1. **c**, The load of Aβ pathology (*x* axis) compared to the load of tau pathology (*y* axis) among participants. Dots and triangles indicate female and male participants, respectively, coloured by their rate of cognitive decline. **d**, The ageing-DLPFC atlas. UMAP embedding of 1,649,672 single-nucleus RNA profiles from the DLPFC of participants. Major cell types are noted; shades highlight some of the 95 different cell subpopulations. **e**, The atlas scale. The number of nuclei per cell type in each participant is shown. Dots represent individual participants ($n = 465$ per cell type). Additional quality-control graphs are shown in Extended Data Fig. 1. Exc., excitatory; inh., inhibitory; oligodend., oligodendrocytes. **f**, Cellular diversity. The proportions of cell subpopulations across participants are shown. The stacked bar plots show cell subpopulation proportions per participant within each major cell type, colour coded by cell type and shaded by subpopulations. For the box plots in **b** and **e**, the box limits show the first and third quartiles, the centre line shows the median value, and the whiskers extend to the highest and lowest values within 1.5× the distance between the quartiles.

heterogeneity. We devised a computational framework, BEYOND, to align individuals along cellular cascades, enabling us to define two different trajectories that older brains can engage (Fig. 1a): one leading to AD and another to alternative brain ageing. These trajectories are derived from coordinated changes in distinct cellular communities. Finally, within the AD trajectory, we propose a sequence of events in which certain microglial and astrocytic subpopulations are prioritized as having important roles at different stages of the causal chain that starts with the accumulation of amyloid pathology, progresses to tau proteinopathy and, ultimately, leads to cognitive decline and

AD dementia. These insights highlight important therapeutic targets for specific stages of AD.

A cell atlas of the aged and AD cortex

To generate a cell atlas capturing the cellular heterogeneity in the aged neocortex, we profiled frozen DLPFC tissue samples from 465 ROSMAP participants[19,20] using snRNA-seq[21,22]; 437 of the samples had a sufficient number of nuclei for downstream analyses (Fig. 1a and Extended Data Fig. 1a,b). ROSMAP participants do not have dementia at enrolment and undergo annual neuropsychological evaluations; brains are collected prospectively and characterized using a structured neuropathological assessment. Our participants are a relatively representative sample of ROSMAP: they span the full spectrum of clinicopathological stages of AD and brain ageing found among older adults (Fig. 1b,c, Extended Data Fig. 1a,b and Supplementary Table 1). The average age of death was 89.1 years; 67.3% of the participants are women; 62.4% of them fulfilled a pathological diagnosis of AD (NIA Reagan criteria[23,24]); 36.3% received a diagnosis of AD dementia; and 26% had mild cognitive impairment (MCI) at the time of death. Nuclear preparation samples from frozen samples from eight participants were pooled, and individual nuclei were assigned to their participant of origin by genotyping polymorphic sites in the RNA and comparing genotypes to the participants' reference genomes (using demuxlet[25]; Methods). Libraries underwent detailed quality control, including automated cell-type classification and cell-type-specific low-quality and doublet nucleus filtration (Methods and Extended Data Fig. 1c–j). Ultimately, we retained 1,649,672 high-quality nuclear transcriptomes (Fig. 1e,f and Extended Data Fig. 1k–m), capturing all brain cell types. The proportions of the major cell classes were largely maintained across individuals (Fig. 1e and Extended Data Fig. 1m). Within each cell class we further subclustered the nuclei, capturing 95 cell subpopulations (Fig. 1d,f and Extended Data Fig. 1m), characterized by distinct expression profiles, markers and enriched pathways (Fig. 2, Extended Data Figs. 2–5 and Supplementary Table 2). Clusters were tested for coherence and separability, and compared to previous annotations (Fig. 2b,d, Extended Data Figs. 3 and 5).

Microglial nuclei were partitioned into 16 subpopulations (Fig. 2a,b and Extended Data Fig. 2a–c; n = 86,673 nuclei), including proliferative (Mic.1), surveilling (Mic.2–5; expressing *CX3CR1*), reacting (Mic.6–8; *TMEM163*), enhanced-redox (Mic.9–10; *FLT1*), stress response (Mic.11; *HSPH1*, *DNAJB1*, *NLRP1*, expressing heat response and NLRP1 inflammasome genes), interferon response (Mic.14; *IFI6*), inflammatory (Mic.15; *CCL3/4*, *NFKB1*, *IL1B*, *CD83*), *SERPINE1*-expressing (Mic.16) and lipid-associated (Mic.12 and Mic.13; *APOE*) subpopulations. The lipid-associated Mic.12 (*CPM*) and Mic.13 (*PTPRG*) subpopulations both expressed the AD risk genes *APOE* and *GPNMB*, with Mic.13 also expressing high levels of *SPP1* and *TREM2* compared with the other subpopulations (Supplementary Table 2). These annotations capture and extend previous reports (Fig. 2b and Extended Data Fig. 3a). Specifically, Mic.13 and Mic.15 (and, to a lower extent, Mic.12) are enriched for the mouse DAM2 signature genes[11], while Mic.12 and Mic.13 are enriched for the human amyloid-β (Aβ)-associated microglial signature, and Mic.15 expressed human inflammation and stress signatures[1,7,8].

Astrocytes were partitioned into ten subpopulations (Fig. 2c,d and Extended Data Fig. 2d–f; n = 228,925 nuclei)—homeostatic-like (Ast.1–2), enhanced-mitophagy/translation (Ast.3; *PINK1*), reactive-like Ast.4 (*GFAP*, *DPP10*, ECM organization and excitatory synaptic genes) and Ast.5 (*GFAP*, *SERPINA3*, *OSMR*, axonogenesis and wound healing genes), interferon-responding (Ast.7; *IFI6*) and stress response (Ast.8–10): Ast.8, expressing chemical and heat stress, and sterol metabolism genes; Ast.9 (*DNAJB1*, *HSPH1*), heat and oxidative stress response, tau binding and necroptosis; and Ast.10 (*SLC38A2*), oxidative stress, reactive oxygen species, metallothioneins and zinc ion homeostasis. Our astrocyte annotation expands previous reports on the brains of healthy individuals and those with AD, and it links mouse and human signatures: Ast.4 to fibrous-like astrocytes[1]; Ast.5 to mouse disease-associated astrocytes[9]; Ast.7 to interferon-responding astrocytes[1,6]; and Ast.10 to human AD-elevated astrocytes[1] (Extended Data Fig. 3b).

Oligodendrocyte lineage cells were partitioned into 12 subpopulations of mature oligodendrocytes (Oli.1–12; Fig. 2e and Extended Data Fig. 2g–i; n = 346,593 nuclei), such as enhanced-translation (Oli.6) and stress responding (Oli.7 expressing *SLC38A2*, *IGF1R*, *QDPR* and cholesterol biosynthesis genes; Oli.8 expressing *SLC38A2*, heat and oxidative stress response genes, *HSPH1*, *DNAJB1*); three subpopulations of oligodendrocyte precursor cells (OPC.1–3, n = 60,622 nuclei); committed oligodendrocyte precursor (n = 556 nuclei) and myelin-forming oligodendrocyte (n = 1,801 nuclei) subpopulations. Within the OPCs, we also found an enhanced-mitophagy subpopulation (OPC.1; expressing *PINK1*, enriched for oxidative phosphorylation and gene translation, which had higher expression of AD-risk genes such as *APOE* and *CLU*), and an axon projection/regeneration-associated subpopulation (OPC.3; *SERPINA3*, *OSMR*). These annotations capture and further refine previous reports[1,4,6,26] (Fig. 2e and Extended Data Fig. 3c).

Within the vascular niche (Fig. 2f and Extended Data Figs. 3d and 4a–c; n = 18,496 nuclei), we identified arterial, venular and capillary endothelial cells, arterial smooth muscle cells (SMCs), pericytes and meningeal and perivascular fibroblasts; this aligns with and expands recent reports[3,17] (Fig. 2f and Extended Data Fig. 3d). For example, we identified two distinct pericyte subpopulations, Peri.1 and Peri.2, aligning with the recently described extracellular matrix (ECM) and transporter pericytes, respectively. We further revealed diversity within the capillary endothelial cells (End.1–5), such as End.3, which had a higher expression of ECM and angiogenesis genes, and End.5, which expresses genes relating to heat- and oxidative stress responses, tau binding and necroptosis (*HSPH1*), as well as AD-risk genes (such as *APP* and *ADAM10*).

On the neuronal side, we identified the full range of neuronal classes, including 16 subtypes of excitatory neurons (Exc.1–16, n = 647,400) and 16 subtypes of inhibitory neurons (Inh.1–16, n = 257,929), capturing excitatory diversity across cortical layers and inhibitory classes expressing distinctive neuropeptides, such as somatostatin (*SST*, Inh.1, Inh.5–7) and parvalbumin (*PVALB*, hereafter PV, Inh.13–16) (Fig. 2g,h and Extended Data Fig. 4d–i). Neuronal subtypes were assigned to cortical layers, expressed neuropeptides and markers by matching to previous neuronal annotations from the Allen Brain Map Transcriptomic Explorer (Fig. 2g,h, Methods and Extended Data Fig. 5b,c).

Linking cell subpopulations to AD traits

To identify subpopulations associated with AD, we focused on three quantitative AD-related traits: (1) neocortical Aβ burden, (2) neocortical tau burden and (3) the rate of cognitive decline before death derived from up to 20 years of annual neuropsychologic profiles (Methods). These outcomes, compared with a categorical measure of AD diagnosis, enhanced our statistical power and enabled the assessment of the continuum of AD clinicopathologic severity. The primary association and subsequent downstream analyses were performed on the 437 participants with a sufficient number of nuclei profiled by snRNA-seq, hereafter referred to as the discovery sample (Fig. 3a, Methods and Extended Data Fig. 1b).

For our discovery analysis, we tested the association between the three AD-related traits and the proportions of subpopulations (calculated within each cell class) using linear regression controlling for age, sex, post-mortem interval and library quality (false-discovery rate (FDR)-adjusted $P < 0.05$; Fig. 3b, Methods and Supplementary Table 3). The strongest associations with both Aβ and tau burden were found for the two lipid-associated microglial subpopulations Mic.12 (*APOE*$^+$*GPNMB*$^+$) and Mic.13 (*APOE*$^+$*GPNMB*$^+$*TREM2*$^+$) (Fig. 2a,b). The proportion of Mic.13 was also higher in individuals with a higher rate of cognitive decline. We also found a strong association for the

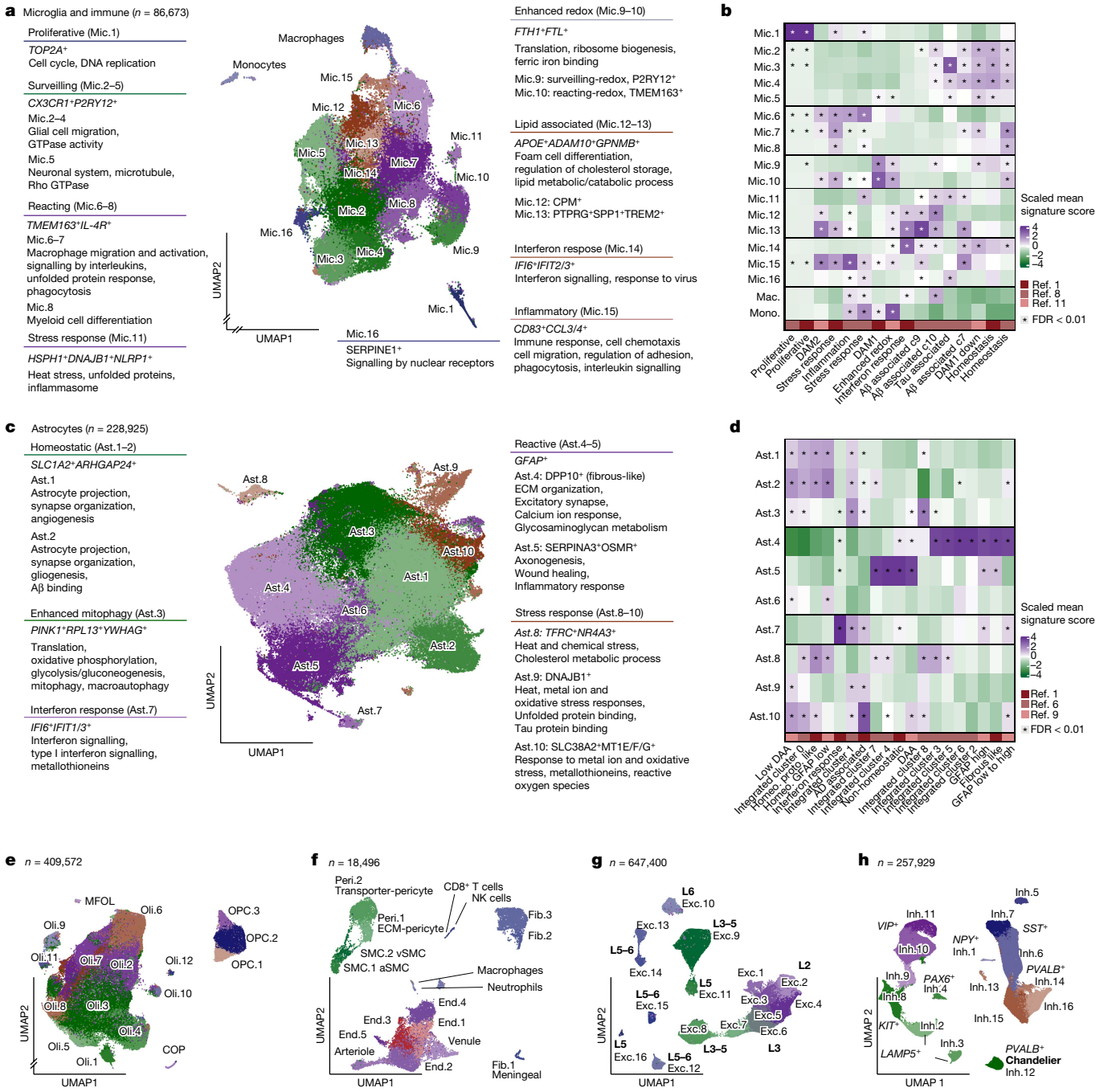


Fig. 2 | Cell subpopulation diversity in DLPFC of aged individuals. **a**,**c**, Subpopulations of microglia (**a**) and astrocytes (**c**). UMAP embeddings of single nuclei profiled using snRNA-seq, coloured by clusters, and annotated for subpopulation, indicating selected marker genes and enriched upregulated pathways. **b**,**d**, Comparison of previous gene signatures to subpopulations of microglia (**b**) and astrocytes (**d**). Scaled mean signature score of published gene signatures (columns) within each subpopulation (rows). Asterisks indicate significantly enriched signatures (*U*-test, FDR < 0.01). Microglial[1,8,11] and astrocytic[1,6,9] signatures were published previously. The ref. 6 signatures are from an integrated atlas including the datasets of refs. 10,14,16. **e**–**h**, The diversity of oligodendroglia (**e**), vascular niche cells (**f**) and excitatory (**g**) and inhibitory (**h**) neuronal cells. UMAP embeddings of nuclei, coloured by subpopulations. Additional annotations and details on the cell atlas are provided in Extended Data Figs. 2–5 and Supplementary Table 2. DAA, disease associated astrocyte; Int, integrated cluster; Proto, protoplasmic astrocytes; MFOL, myelin forming oligodendrocytes; SMC, smooth muscle cell; aSMC, arterial SMC; vSMC ventricular SMC.

proportion of the stress-responding Ast.10 and Oli.7 subpopulations with a greater tau burden and more rapid cognitive decline. Finally, we found two inhibitory neuronal subpopulations that are associated with tau burden: PV$^+$ Inh.16 neurons (*MEPE*$^+$, layers 5–6; Extended Data Fig. 5c) increased in proportion while SST$^+$ Inh.6 neurons (*RSPO3*$^+$*KLF5*$^+$, layers 2–5; Supplementary Table 2) decreased in proportion as the pathology burden increased. Thus, this specific subpopulation of PV$^+$ neurons may be relatively resilient, while the specific subpopulation of SST$^+$ neurons may have enhanced vulnerability in AD relative to all other inhibitory neurons, consistent with previous reports[1,27].

To validate our findings in a larger and independent set of individuals, we used 673 ROSMAP participants with bulk RNA-seq data from the DLPFC. They are independent of the discovery sample participants, and we refer to them as the replication sample (Fig. 3a). We applied

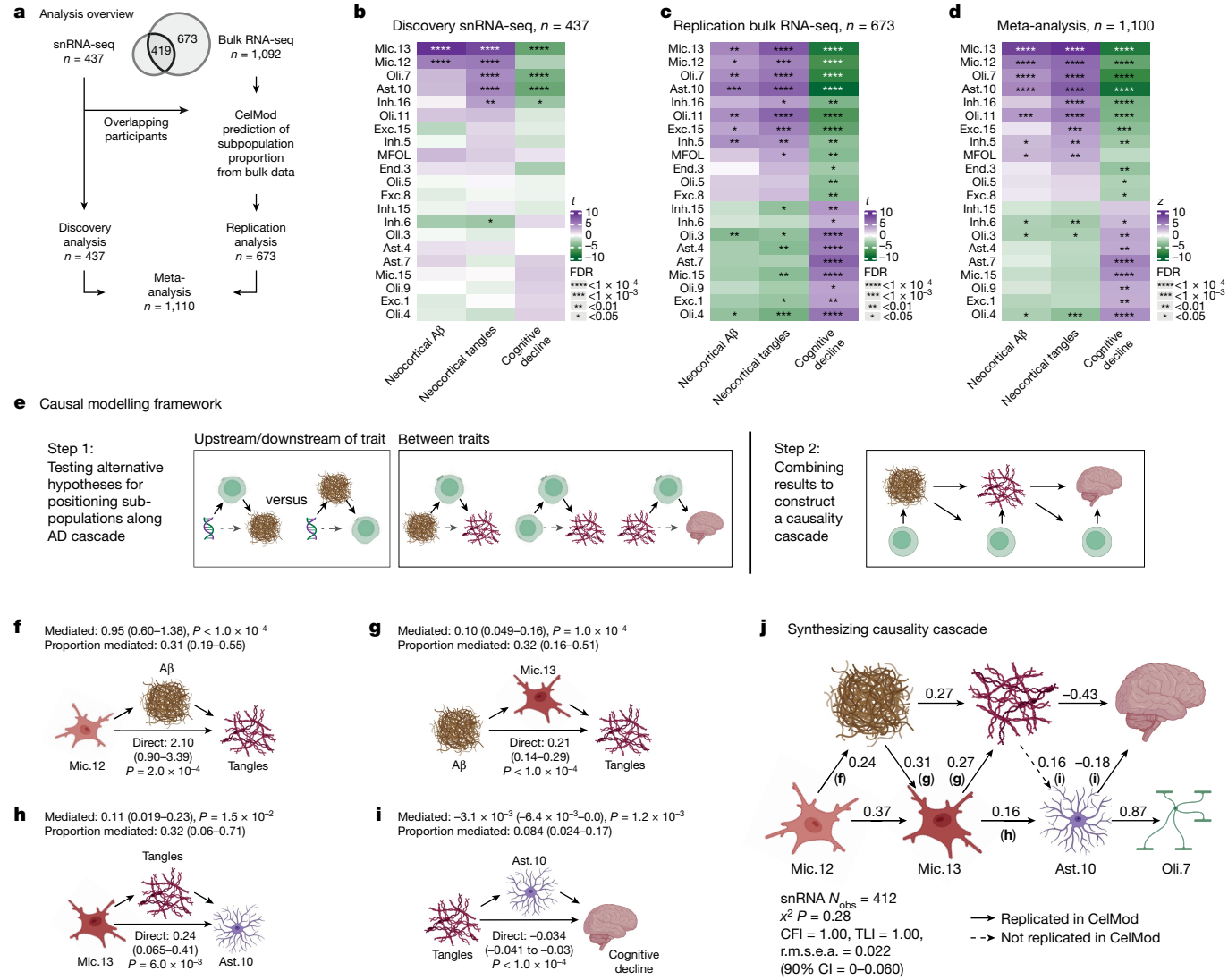


Fig. 3 | Associating subpopulations with AD-related endophenotypes and causality modelling along the AD cascade. a, Overview of analysis and cohorts: the discovery analysis includes 437 participants with snRNA-seq data; the replication analysis includes 673 non-overlapping participants with CelMod-estimated subpopulation proportions from bulk RNA-seq data; and a meta-analysis of both datasets. **b**–**d**, The association between subpopulation proportions and endophenotypes—neocortical Aβ load, neocortical tau load and the rate of cognitive decline (linear regression controlled for confounders, FDR < 0.05; Methods); we find subpopulations that are significantly associated with at least one of the tested traits in one on the cohorts: discovery sample (snRNA-seq measurements, $n = 437$ participants; **b**); replication sample (bulk RNA estimations, $n = 673$ participants; **c**); and meta-analysis of discovery and replication samples ($n = 1{,}110$ participants; **d**). The colour scale shows the association effect size, indicating the direction and strength from negative (green) to positive (purple) associations. **e**, The causal modelling framework for positioning a subpopulation's effect upstream of an endophenotype, mediating the effect of one endophenotype on another, or mediating the effect of a different subpopulation on an endophenotype. **f**–**i**, Causal mediation models positioning Mic.12, Mic.13 and Ast.10 within the Aβ→tau→cognitive decline AD cascade, indicating direct and mediated effects, as well as the proportion of effect mediated. The numbers of participants were as follows: $n = 432$ (**f**–**h**) and $n = 433$ (**i**). **j**, SEM positioning Mic.12, Mic.13, Ast.10 and Oli.7 within the AD cascade. Integration of all of the independent mediation results (**f**–**i**; Extended Data Fig. 6d–h) in a SEM. The arrows show the association directionality and relative strength, and indicate whether association was replicated in the replication sample (solid, $P < 0.05$) or not (dashed). The letter indicates the guiding mediation model. N_{obs}, number of observations (participants); r.m.s.e.a., root mean square error of approximation.

the recently developed CelMod method[1] to infer the proportion of our cell subpopulations in the bulk RNA-seq data, fitting CelMod using profiles of the subset of 419 ROSMAP participants (from our 437 participants in the discovery sample; training sample) with both snRNA-seq and bulk RNA-seq data; the proportions of 62 subpopulations were predicted with high confidence and were retained for downstream analyses (Fig. 3a, Methods and Extended Data Fig. 6a). We replicated our top results and found high correlations between the estimated effect sizes over all of the subpopulations tested in both the discovery ($n = 437$ snRNA-seq) and replication ($n = 673$ bulk-inferred) analyses for all three traits (Fig. 3c and Extended Data Fig. 6b,c).

Finally, to maximize our statistical power, we performed a meta-analysis on the full set of 1,110 participants ($n = 437$ (discovery) and $n = 673$ (replication); Fig. 3d, Methods and Supplementary Table 3). With the increased statistical power, lipid-associated Mic.12, Mic.13, stress response Ast.10 and Oli.7, and PV$^+$ Inh.16 and SST$^+$ Inh.6 neurons had the same or stronger evidence of association with the AD-related traits (Fig. 3d). The meta-analysis also identified multiple additional subpopulations to be significantly associated with one or more traits, particularly oligodendrocyte and excitatory neurons cell subpopulations (Fig. 3d and Extended Data Fig. 6c). We therefore obtained a robust set of results prioritizing

specific neocortical cell subpopulations across different cell types in relation to AD.

Causal modelling of glia in the AD cascade

Next, we used statistically rigorous methods with an established set of priors to propose a putative causal chain of events that relates specific subpopulations to specific stages of the AD cascade. We used mediation modelling, leveraging the widely accepted cascade of AD progression: Aβ accumulation occurs first, followed by pathogenic tau aggregation and finally cognitive decline[28–30]. We used the strong genetic risk factor *APOE* ε4 as an anchor as it is present at birth. We first performed causal mediation analysis to quantify the direct and mediated (indirect) effects of subpopulation proportions on our AD traits; we then followed with structural equation modelling to assess the entire proposed AD cascade (Fig. 3e and Methods). We focused on the top robustly associated subpopulations (FDR < 0.01) with two or more AD-related traits: lipid-associated Mic.12, Mic.13 and stress response-related Ast.10 and Oli.7.

Our model placed Mic.12, Mic.13, Ast.10 and Oli.7 subpopulations at specific points along the AD cascade: (1) Mic.12 upstream of Aβ (Fig. 3f); (2) Mic.13 downstream of Aβ and Mic.12, but upstream of tau and cognitive decline (Fig. 3g and Extended Data Fig. 6d–f); (3) Ast.10 downstream of tau and Mic.13, but upstream of cognitive decline (Fig. 3h,i and Extended Data Fig. 6g); and (4) Oli.7 downstream of Ast.10 (Extended Data Fig. 6h). We found that Mic.12 had a strong association with age ($P = 2.8 \times 10^{-5}$) and might therefore represent an age-dependent microglial dysfunction leading to poor Aβ clearance, given its position in the model. The proportion of Mic.13 mediated 32% of the strong association between Aβ and tau (Fig. 3g), yet tau mediated most of the association between Mic.13 and cognitive decline (62%). This suggests that the strong influence of Mic.13 on cognitive decline occurs mostly through tau-dependent mechanisms (Extended Data Fig. 6f). In contrast to Mic.12, Mic.13 was not associated with age but was associated with the genetic risk factor *APOE* ε4 (Extended Data Fig. 6e). Furthermore, there was an association between Mic.13 and Ast.10 even after adjusting for tau, suggesting that these two populations may interact separately from tau-related processes (Fig. 3h). Finally, Ast.10 mediated 8.4% of the strong tau–cognitive-decline association (Fig. 3i); notably, Ast.10 explained an additional 2.5% of the variance in cognitive decline not related to AD pathology. This suggests that Ast.10 might serve as a point of convergence for tau-dependent and tau-independent mechanisms leading to cognitive decline.

Synthesizing the above findings, we constructed a structural equation model (SEM) integrating Mic.12, Mic.13, Ast.10, Oli.7 and AD endophenotypes into a single model of the AD cascade (comparative fit index (CFI) = 1.00, Tucker Lewis index (TLI) = 1.00; Fig. 3j and Methods). We then validated the SEM using the replication sample ($n = 605$ individuals) and reproduced all of the relationships outlined in our snRNA-seq model except for one (Extended Data Fig. 6i), demonstrating the robustness of our results and indicating that our modelling strategy has derived plausible causal pathways aligning specific cell subpopulations with key AD endophenotypes. We endorse the limitation of such causal inference models derived from brain autopsy samples; thus, we do not exclude the possibility that other, as yet unmeasured, factors contribute to the AD cascade. Nonetheless, the mediation and SEM analyses provide a strong set of hypotheses for future mechanistic testing.

Two disease-associated microglial states

Our model proposes that Mic.12 and Mic.13 have causal roles in the proteinopathies that define AD. Compared with other microglial cells, both subpopulations have a higher expression of the AD susceptibility genes *APOE* and *GPNMB*, and share multiple enriched pathways, such as increased expression of foam-cell differentiation, cholesterol storage, lipid metabolic/catabolic process and decreased glial cell migration genes (Fig. 4a). However, these two microglial subpopulations are distinct: Mic.12 had higher expression of genes linked to the regulation of endocytic vesicles (for example, *PELI1* and *PELI2*) and major histocompatibility complex (MHC) class II genes (such as *HLA-DRA*), while Mic.13 upregulated pathways related to cell junction, adhesion and extracellular matrix (ECM) organization (for example, *ADAM10*, *TGFBR1*, *SMAD3* and *PPARG*), genes involved in negative regulation of immune system processes, and the AD susceptibility genes *ADAM10* and *TREM2* (Fig. 4b,c, Extended Data Fig. 7a and Supplementary Table 2). We also found an upregulation of genes linked to exocytosis (for Mic.13, *SCIN*, *WIPF3*) and phagocytosis (for Mic.12 and 13, *PRKCE* and *MSR1*; for Mic.13, *WIPF3* and *PPARG*; Fig. 4c).

To validate the existence of the Mic.12 and Mic.13 subpopulations, we performed single-molecule RNA fluorescence in situ hybridization (smFISH) in an independent sample of 15 DLPFC brain slices from the New York Brain Bank (8 AD, 1 MCI and 6 non-impaired individuals). We used selected RNA markers—Mic.12 (*CPM*), Mic.13 (*TPRG1*) and macrophages (*MRC1*)—along with immunofluorescence staining for the myeloid protein IBA1, DAPI to stain nuclei and the anti-phospho Tau antibody AT8 (Methods). Using automated image segmentation (Methods), we found that $DAPI^+IBA1^+$ cells expressing high levels of the Mic.12 marker *CPM* are distinct from those with high expression of the Mic.13 marker *TPRG1*, validating that Mic.12 and Mic.13 are distinct microglial populations (Fig. 4d,e, Extended Data Fig. 7b–d and Supplementary Table 4). Next, we evaluated the association of Mic.12 and Mic.13 with tau burden. The AT8-based tau burden was associated with AD status within our samples (Fig. 4f), and we found that it was associated with the proportion of Mic.12 (CPM^+IBA1^+ cells; $t = 3.2$, $P = 0.0066$) and Mic.13 ($TPRG1^+IBA1^+$ cells; $t = 2.8$, $P = 0.014$) (Fig. 4f and Methods), consistent with our transcriptomic results (Fig. 3).

Finally, to test for functional differences between these two microglial subpopulations, we analysed morphological measures captured from our images (CellProfiler[31]; Methods). We observed significantly reduced compactness—a measure of cell membrane ramification—among both the Mic.12 and Mic.13 subpopulations compared with all other microglia (FDR < 1×10^{-4}; Fig. 4g, Methods and Extended Data Fig. 7c). We also found increased eccentricity—a measure of cell elongation—for Mic.12 only (FDR < 0.05, Fig. 4g and Methods). To further explore the morphological features of these two subpopulations, we repurposed scores derived from a standard neuropathologic microglial classification system that are available for ROSMAP participants: microglia are manually scored from stage I (highly ramified) to stage III (dense cytoplasm and retracted processes, deemed to be activated)[32]. The proportion of activated microglia (PAM) was previously found to be associated with tau burden and other AD traits[32]. In total, 91 participants had both snRNA-seq data and a PAM score for the midfrontal cortex, and we found that both Mic.12 and Mic.13 have a strong positive association with the PAM score, while other microglia have a negative association (Fig. 4h and Methods). Yet, the association between Mic.12 and PAM is attenuated when adjusting for Mic.13, suggesting that Mic.13 may be the primary driver of this association (Fig. 4h). We therefore link the Mic.13 subpopulation, expressing AD risk genes (such as *APOE*, *GPNMB* and *TREM2*) and the DAM2 signature[11], to morphologically defined activated (stage-III, PAM) microglia[32], consistent with our morphological analysis of Mic.13 cells (Fig. 4g).

Distinct cellular paths of brain ageing

We next used our data to address the heterogeneity observed among older individuals. While we are limited by the lack of longitudinal molecular brain measurements (as specimens are obtained at autopsy and we cannot capture temporal dynamics within an individual),

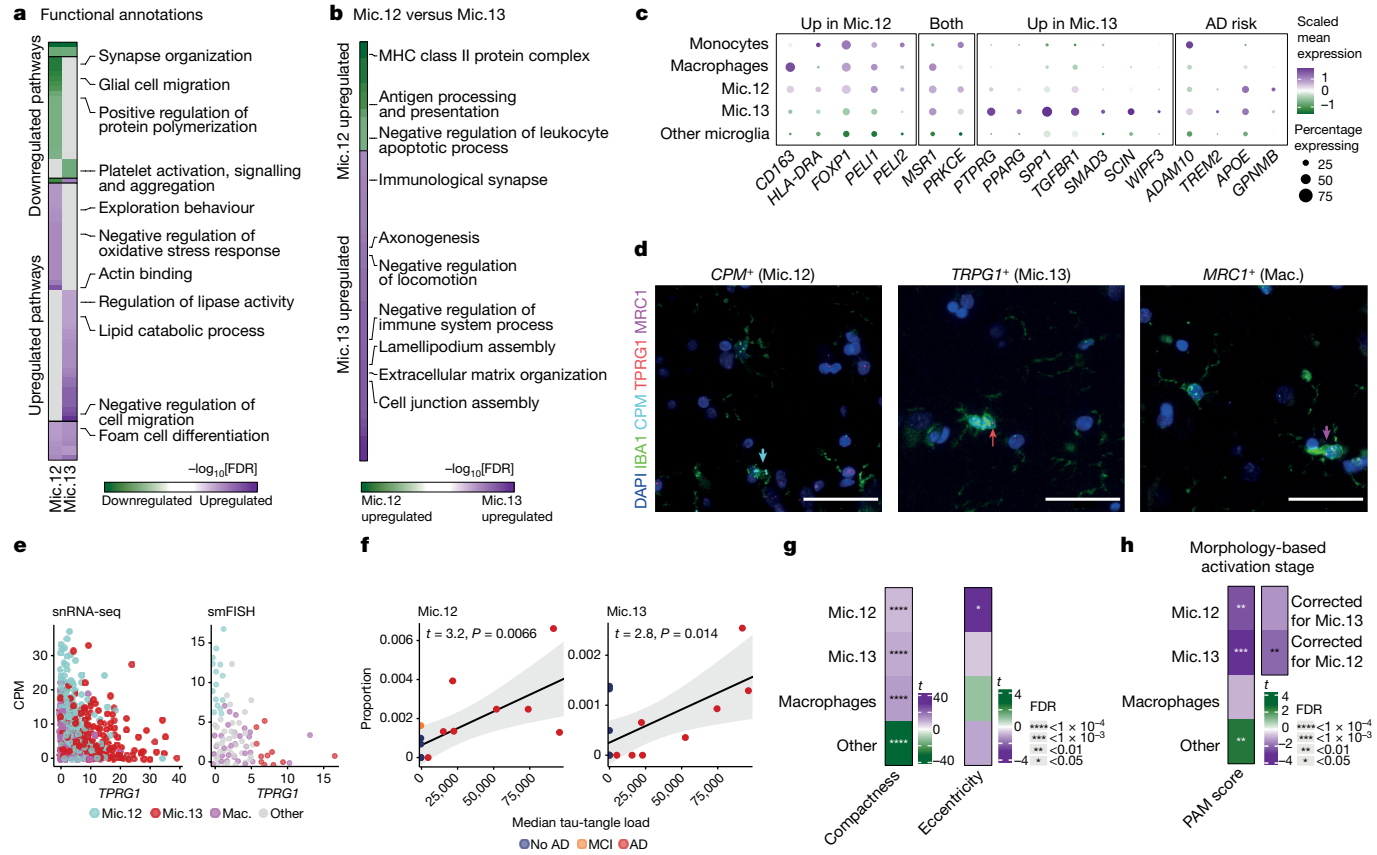


Fig. 4 | Distinct AD associated Mic.12 and Mic.13 subpopulations. **a**,**b**, Enriched pathways by Mic.12 and Mic.13 subpopulations. Upregulated (purple) or downregulated (green) pathways of Mic.12 or Mic.13 compared with all microglia (**a**) or Mic.13 compared with Mic.12 (**b**). The colour scale shows the $-\log_{10}$[FDR] of enrichment (hypergeometric test, FDR < 0.05; Methods). **c**, Distinct gene expression patterns for Mic.12 and Mic.13. The dot colour shows the mean expression in expressing cells (column scaled); the dot size shows the percentage of cells expressing the gene. All microglial subpopulations are shown in Extended Data Fig. 7a. **d**–**g**, Validation using RNAscope and immunohistochemistry images of independent DLPFC brain samples. **d**, Validation of distinct microglial cells expressing the Mic.12 or Mic.13 markers. Representative RNAscope images (out of 27,892 images) showing microglia/myeloid cells (green, anti-IBA1 immunofluorescence), nuclei (blue, DAPI), and RNAscope targeting *CPM* (cyan, Mic.12 marker), *TPRG1* (red, Mic.13 marker) and *MRC1* (magenta, macrophage marker). The arrows indicate examples of CPM^{high}, $TPRG1^{high}$ and $MRC1^{high}$ cells positive for a single marker. snRNA-seq marker expression and additional images are provided in Extended Data Fig. 7b,c. **e**, The expression distributions of marker genes for Mic.12 and Mic.13 measured using snRNA-seq (left) and smFISH (right) analysis of $IBA1^{+}$ cells. The dot colour shows the subpopulation annotation (Methods). **f**, Association of Mic.12 and Mic.13 RNAscope proportions with tau-tangle pathology load quantified by immunohistochemistry using anti-phosphorylated Tau antibody AT8 (as the total area occupied; Methods). Dots are coloured by participant clinical diagnosis. The error bands show the 95% confidence intervals. **g**, Association of Mic.12, Mic.13 and macrophage RNAscope proportions with morphological features captured in the same smFISH images (Methods). The colour scale indicates the effect size. **h**, Association of snRNA-seq Mic.12 and Mic.13 proportions to previous neuropathologic activated microglia classification (PAM)[32]. The PAM score is the square root of stage III activated macrophage-appearance microglial density proportion (Methods). The colour scale shows the association effect size between snRNA-seq proportions to PAM score (histopathology). Right, associations with corrected proportions of Mic.12 and Mic.13 (corrected to Mic.13 and Mic.12 proportions, respectively).

rigorous methodologies can infer dynamics by considering that each participant represents one timepoint along one of the trajectories that an ageing brain can follow. We therefore devised a conceptual-analytic framework that we call BEYOND (Fig. 5a): an integrative approach to identify trajectories of cellular changes in high-dimensional data, to relate these trajectories to clinicopathological outcomes and to accurately align participants along each trajectory. BEYOND builds a cellular landscape manifold that captures the diversity of observed cellular environments, with each participant as a single point in the high-dimensional manifold (that can be visualized in low dimensions; Fig. 5b and Extended Data Fig. 8a). It then reconstructs trajectories of change along the manifold using similarities in the cellular environments across participants (Fig. 5b,c). Next, it assesses whether the trajectories are associated with AD-related traits, matching them to potential disease states (Fig. 5d–f). Finally, BEYOND assigns cell subpopulations to multicellular communities with similar dynamics and abundance by integrative clustering (Fig. 5a).

When applying BEYOND, we identified two distinct trajectories of cellular change in the ageing DLPFC, starting from a shared point in the cellular manifold (Fig. 5b and Supplementary Table 5). BEYOND assigned a trajectory probability and pseudotime ordering for each participant along these two trajectories that can be used to quantify the dynamics of various traits and subpopulation proportions. Along each axis, we found that certain subpopulations appear to underlie the trajectories (Fig. 5c, Extended Data Fig. 8b and Supplementary Table 5). For example, the disease-associated subpopulations Mic.12, Mic.13, Ast.10 and Oli.7 all increase in proportion along one axis, while other subpopulations of reactive glial cells, such as Ast.5 and OPC.3, increase along a different axis. These divergent cellular patterns are robust to both the embedding algorithm and to the number of cell subpopulations used as input (Extended Data Fig. 8c). The two trajectories of cellular change and pseudotime assignment were robust to different algorithms, comparing Palantir[33] and VIA[34] (Methods, Extended Data Fig. 8d–g and Supplementary Table 5).

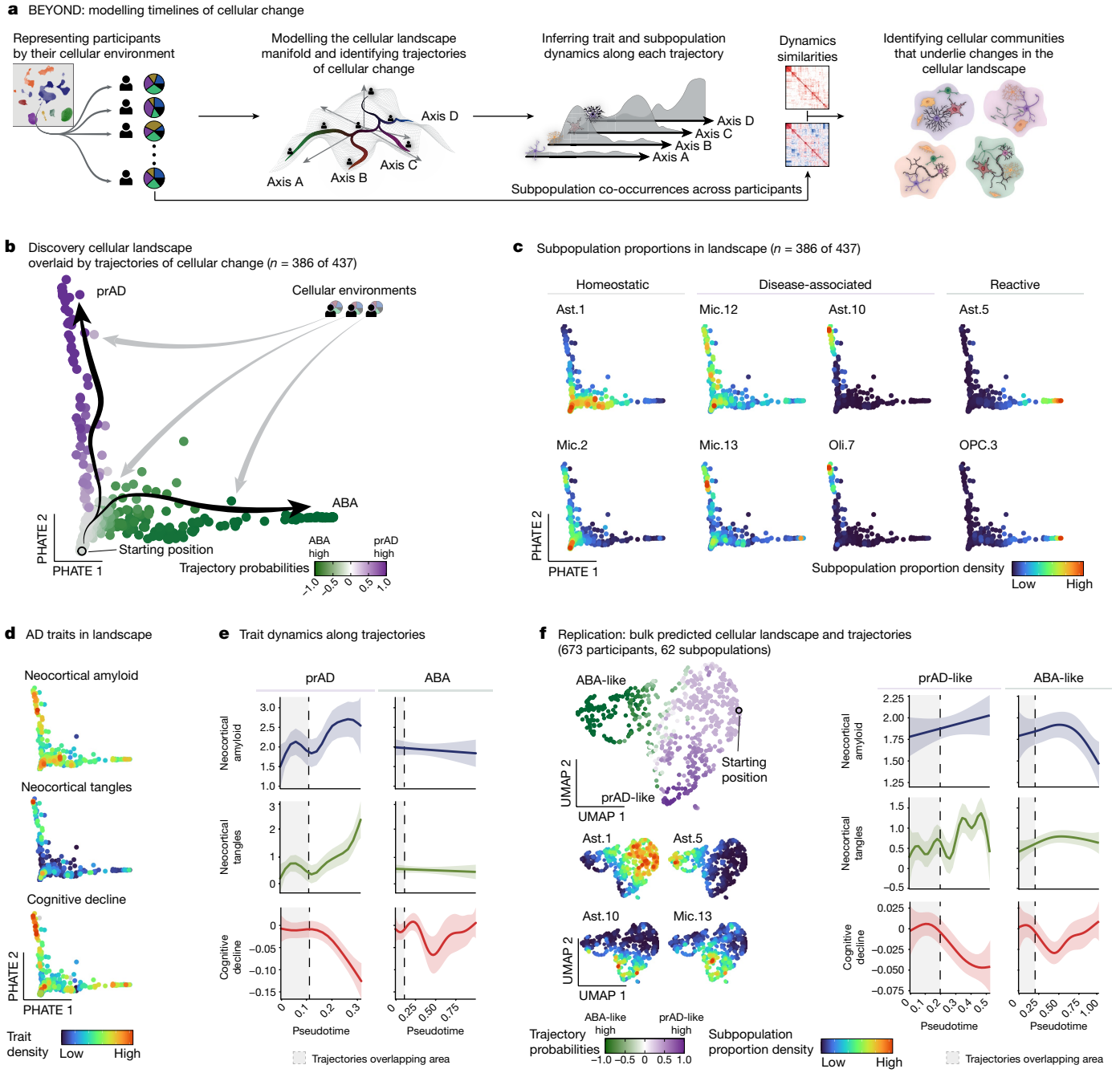


Fig. 5 | Modelling the cellular landscape manifold uncovered distinct trajectories of brain ageing leading to AD or ABA. a, Schematics of the BEYOND algorithm: representing participants by their cellular environments (cellular compositions, step 1), aligning participants along trajectories of cellular change (step 2), inferring subpopulation and trait dynamics (step 3) and grouping of cellular communities (step 4). **b**, The structure of the cellular landscape manifold captured by BEYOND for the discovery sample. A 2D PHATE embedding of each participant (individual dots) is shown based on similarity of their cellular compositions (Methods). Dots are coloured by the difference in participants' assigned trajectory probabilities. See also Extended Data Fig. 8a,d–f. **c**, Distinct patterns of subpopulation proportions along the cellular landscape manifold. The participants are coloured by the locally smoothed proportions of each subpopulation in the 2D embedding, showing distinct patterns for different subpopulations for representative examples along the cellular manifold. See also Extended Data Fig. 8b,c. **d**,**e**, Distinct patterns and dynamics of AD traits along the prAD and ABA trajectories, showing for neocortical Aβ, neocortical tau and cognitive decline distinct patterns in the cellular landscape manifold per AD trait. The plots are coloured by the locally smoothed density values of the trait per participant (embedded as in **b**) (**d**); and distinct trait dynamics along the pseudotime in each inferred trajectory (**e**; Methods). The error bands show the 95% confidence intervals. See also Extended Data Fig. 8g,h. **f**, Validation of the cellular landscape, trajectories and dynamics results by applying BEYOND to the independent replication sample (n = 673 participants) using only the 62 well-predicted subpopulations (FDR < 0.01; Extended Data Figs. 6a and 8i,j). For the replication manifold, two inferred trajectories of cellular changes (as in **b**) (top left); distinct densities for key subpopulations (as in **c**) (bottom left); and distinct AD-trait dynamics along the two directions (as in **e**), similar to the patterns over the discovery manifold (right) are shown. The error bands show the 95% confidence intervals.

VIA predicted an additional minor trajectory suggesting potential additional diversity in cellular trajectories within ageing brains (Extended Data Fig. 8d).

We next related the pseudotime and probabilities of belonging to each of the two trajectories to the participants' clinicopathological characteristics (Fig. 5d,e, Extended Data Fig. 8h and Supplementary Table 5). Matching the observed cellular changes along the trajectories (Fig. 5c), we found one of the two trajectories to be strongly related to AD; we named it progression to AD (hereafter, prAD), capturing a monotonous increase in neocortical Aβ and tau burden as well as an increase in the rate of cognitive decline (Fig. 5d,e). We named the other trajectory alternative brain ageing (ABA), as it captured individuals with a constant level of Aβ, low/no neocortical tau and a slower rate of cognitive decline with a variable pattern (Fig. 5d,e). The variability in the pattern of decline along the ABA trajectory may indicate that it is influenced by other neuropathologies contributing to cognitive decline and/or resistance mechanisms that are, as yet, unappreciated.

We validated our identification of the two distinct trajectories that capture different disease courses by applying the BEYOND methodology to the replication sample (n = 673 independent ROSMAP participants with bulk-inferred proportions for only 62 subpopulations; Fig. 3a and Methods). These data recapitulated the two main trajectories and confirmed the association of the prAD trajectory with AD-related traits (Fig. 5f and Extended Data Fig. 8i–j). Our ability to replicate these findings in an independent set of participants, with only a partial set of subpopulations, highlights the robustness of our findings. BEYOND was therefore able to reconstruct a timeline that captures the continuous changes along a trajectory leading to AD, and a second major trajectory of ABA. Notably, individuals along the ABA trajectory are not completely free from cognitive decline, but the difference in cellular composition between the prAD and ABA trajectories suggests that they may need to be considered separately in preclinical and clinical studies.

Dynamics of multicellular communities

To map coordinated changes in the brain's cellular environment underlying the prAD and ABA trajectories, we defined multicellular communities (Fig. 5a, Methods and Supplementary Table 5); we previously used correlations in the proportion of cell subpopulations to define such communities[1]. Here, we refined the concept by also leveraging the coordinated patterns of change among cell subpopulations along our two trajectories: some subpopulations changed in proportion predominantly along one trajectory or followed a sequence along one trajectory, such as early (Mic.12 and Mic.13), mid (Ast.3, OPC.1 and Oli.6) and late (Oli.7, Ast.10) changes along prAD (Fig. 6a and Extended Data Fig. 9a). This is implemented in BEYOND, which assigns cell subpopulations to multicellular communities by integrative clustering using two properties quantifying different aspects of coordinated change: (1) co-occurrences of subpopulations across individuals (similar to ref. 1) and (2) shared patterns of subpopulation dynamics along all trajectories (Fig. 5a and Methods). The clustering revealed three cellular communities (C1–C3), which further clustered to subcommunities (C1.1–1.2 and C2.1–2.3; Fig. 6b and Extended Data Fig. 9b).

The healthy community C1 was composed of subpopulations that decreased in frequency along both trajectories (Fig. 6b,c and Extended Data Fig. 9c). Subcommunity C1.2 included subpopulations of homeostatic glia (for example, Ast.1, Ast.2, Mic.2, Oli.1, OPC.2 as well as End.2), while C1.1 consisted primarily of neurons. The prAD C2 community increased in proportion along the prAD trajectory but not the ABA trajectory (Fig. 6b,c). This was refined into three subcommunities: a neuronal subcommunity C2.1 (for example, Inh.16), the mid-prAD C2.2 community (for example, the enhanced-mitophagy and translation Ast.3, OPC.1, Oli.6 and Mic.12), and the late-prAD C2.3 community that included the disease-associated subpopulations, Mic.13, Ast.10 and Oli.7, as well as Ast.9, Mic.11 and End.5, among others (Fig. 6b,c). Finally, we identified the ABA reactive C3 community, which included the reactive-like subpopulations Ast.5, OPC.3, Mic.6-7, Oli.4 and Peri.1 (Figs. 2 and Fig. 6b,c and Extended Data Fig. 9c). We validated these results in the replication sample (n = 673 independent participants with bulk-inferred proportions; Fig. 3a and Methods), highlighting the robustness of our proposed cellular communities (Extended Data Fig. 9d).

Within each cellular community, we found enrichment of specific functions. For example, there are upregulated translation, mitochondrial and mitophagy pathways, chaperone-mediated autophagy, oxidative phosphorylation, as well as regulation of neurotransmitter transport and glutamate secretion in the mid-prAD community C2.2 (Ast.3, OPC.1 and partially Oli.6). In the prAD-late community C2.3 (Ast.9, Ast.10, Oli.7, Mic.11 and End.5), there is upregulation of stress-response pathways, such as metal ion and heat and oxidative stress (Fig. 6c). Overall, we note a progression starting with lipid-associated/inflammation, followed by altered glucose metabolism/mitochondrial/synaptic function and ending with reactive oxygen species and stress responses along the prAD trajectory (Fig. 6c).

We next associated AD traits to a summary score for each community and found that the C2 and subcommunity C2.3 proportions increase with the accumulation of neocortical Aβ and tau as well as worsening cognitive decline, whereas C1 and its subcommunity C1.1 become depleted (Fig. 6d and Methods). Thus, the reconstruction of the dynamics of cellular environment in ageing brains expands our perspective beyond the trait-association analysis of individual cell subpopulations (Fig. 3): it highlights possible transitions in the composition of cellular communities at different stages on the trajectory to AD.

To validate our results, we analysed spatial transcriptomics (Visium platform; Methods) data for DLPFC sections from ten ROSMAP participants from our discovery cohort (Extended Data Fig. 9e and Supplementary Table 6). We focused on three distinct cell subpopulations of interest: Mic.13, Ast.10 (community C2) and Ast.5 (community C3). Each subpopulation was localized and quantified across the brain slices by the combined expression of selected marker genes (Fig. 6e and Methods). Mic.13 and Ast.10, of which the proportions are coordinated in the snRNA-seq data (Fig. 6a,b), were found to be co-localized in tissue sections of participants assigned to the prAD trajectory and not in the other participants (FDR < 0.01; Fig. 6f, Methods and Extended Data Fig. 9f). Finally, we evaluated the presence of the two astrocyte subpopulations that increase specifically along one of our projected trajectories, Ast.10 (prAD) and Ast.5 (ABA) (Fig. 6a). While we did not have enough spatial transcriptomics data from participants in the advanced ABA trajectory (Extended Data Fig. 9e), the Ast.10 signature was significantly (FDR < 0.01) higher compared with the expression of the Ast.5 signature among participants from the prAD trajectory (FDR < 0.01; Fig. 6g and Extended Data Fig. 9g). This provides additional validation of our proposed trajectories, which will need to be evaluated further in larger collections of samples.

These findings confirm the existence of distinct cell subpopulations in the human cortex, and they suggest that these multicellular communities may be topologically coherent. Overall, we therefore present evidence of distinct cellular cascades in the older brain; some of these are specific to the sequence of events leading to AD (prAD trajectory) and appear to be largely absent from the portion of ageing individuals who are following another cellular path of brain ageing (ABA trajectory).

Discussion

Here, we present insights into the cellular and molecular cascades leading to AD and brain ageing using a large-scale, high-resolution cell atlas of 437 ROSMAP participants (1.65 million cells assigned to 95 cell subpopulations) and an algorithmic approach, BEYOND. Our work addresses fundamental open questions in human ageing and AD. First, we identified two distinct trajectories distinguishing prAD

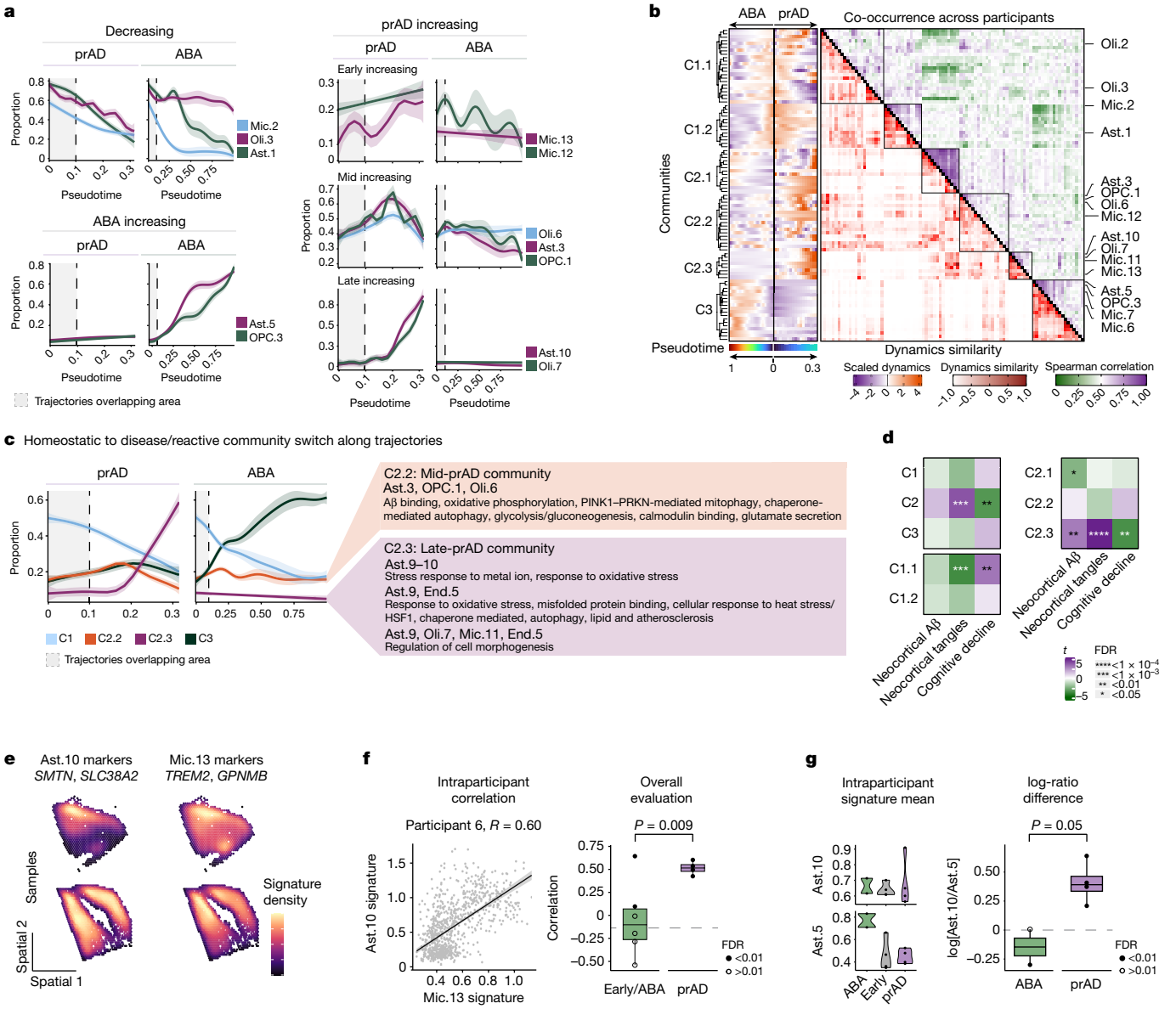


Fig. 6 | Distinct multicellular communities change along each trajectory of brain ageing. a, Subpopulation proportion dynamics along pseudotime in each trajectory. The error bands show the 95% confidence intervals. Similar dynamics were observed in the replication sample (Extended Data Fig. 9d). **b**, BEYOND identified multicellular communities of subpopulations (a schematic of the method is shown in Fig. 5a). Left, scaled dynamics of all subpopulations along both trajectories. Right, subpopulation similarity measures—pairwise correlations of proportions (co-occurrence, top) and dynamics similarity (bottom). Subpopulations are grouped by communities. See also Extended Data Fig. 9b. **c**, Distinct dynamic patterns of communities across trajectories. Community dynamics showing the cumulative change of subpopulation proportions along pseudotime in each trajectory. Right, shared pathways within specific communities. **d**, Associations between cellular community proportions and AD traits, coloured on the basis of the estimated effect size (linear regression *t*-test, FDR < 0.05). **e**–**g**, Validation of cellular communities by spatial transcriptomics (Visium) of DLPFC sections from ten ROSMAP participants from the discovery sample. **e**, Joint densities of Mic.13 and Ast.10 marker gene expression on representative brain slices. See also Extended Data Fig. 9e. **f**, Intraparticipant spatial co-localization of Mic.13 and Ast.10, showing the correlation between signature levels per Visium spot (dot) for an example prAD-assigned participant (left); and all intraparticipant correlations (by trajectory assignment) (right). The dot colour shows the significance of correlation. *P* values were calculated using one-sided *t*-tests, testing higher correlations in prAD participants (FDR corrected; Extended Data Fig. 9f). **g**, Trajectory-specific enrichment of Ast.10 or Ast.5 gene signatures: the average signature level of participants by trajectory assignment (left); and comparison of the interparticipant log-ratio of Ast.10 to Ast.5 signature levels, per trajectory assignment (right). The dot colour indicates the significance of the intraparticipant difference (Extended Data Fig. 9g). *P* values were calculated using one-sided *t*-tests, testing interparticipant higher log-ratio in prAD participants. For the box plots in **f** and **g**, the box limits show the first and third quartiles, the centre line shows the median value, and the whiskers extend to the highest and lowest values within 1.5× the distance between the quartiles.

from ABA. While participants on the prAD trajectory had increasing levels of neocortical Aβ and tau burden as well as accelerated cognitive decline, ABA participants had a low and constant Aβ burden, limited tau pathology and varying rates of cognitive decline. Previous studies have associated cells and pathways with AD[1–11,14–17], consistent with our associations, but their case–control study design could neither infer dynamic changes in cellular subpopulations nor decouple AD from other aspects of brain ageing. Second, our reconstruction of pseudotemporal cellular dynamics advances our understanding of the sequence of events leading to AD and ABA. We highlight a switch from

a homeostatic cellular environment to either a disease-associated environment in prAD (for example, Mic.13, Ast.10, Oli.7) or to a reactive-like environment in ABA (for example, Ast.5, OPC.3). Third, individual subpopulations do not contribute by themselves: we identified multicellular communities within which the subpopulations share unique dynamics and molecular pathways, providing a more comprehensive set of interacting cellular targets for therapeutic development. Fourth, mediation modelling, using the key cellular subpopulations Mic.12, Mic.13 and Ast.10, proposed where, in the cascade of events linking Aβ and tau proteinopathies to cognitive decline, they may exert their putative causal role. All of these results were replicated in an independent sample of 673 ROSMAP participants and further validated by smFISH, immunofluorescence and spatial transcriptomics.

We can now suggest the following sequence of events underlying the prAD trajectory. At the early stages of the sequence of events leading to AD, we note a decrease in a homeostatic cellular community and an increase in the lipid-associated *APOE*$^{+}$*GPNMB*$^{+}$ Mic.12 subpopulation; these cells are influenced by age and predicted to directly contribute to the accumulation of Aβ proteinopathy. Increasing Aβ deposition is known to contribute to the accumulation of tau proteinopathy, and we find that the *APOE*$^{+}$*GPNMB*$^{+}$*TREM2*$^{+}$ Mic.13 subpopulation mediates this effect of Aβ (Figs. 3 and 6). Notably, Mic.13 is also influenced by *APOE* ε4. The accumulation of tau proteinopathy is closely linked to the onset of cognitive decline in older adults, and our model proposes a key role for the Ast.10 subpopulation in mediating this toxic effect of tau. Along with Ast.10, at this stage of the AD cascade, we observed an increase in multiple subpopulations that are part of the prAD-late community (Ast.10, Oli.7, Ast.9, Mic.11 and others), all upregulating various stress response genes, suggesting coordinated functional changes across multiple cell types leading to cognitive dysfunction. Notably, our model predicts that not only tau but also Mic.13 abundance affects the proportion of Ast.10 independent of tau, suggesting that the proportion of this astrocyte subpopulation may be a point of convergence for different processes leading to cognitive dysfunction.

While Mic.12 offers a good target for the prevention of Aβ accumulation enhancing therapeutic options centred on anti-Aβ antibodies, averting the polarization of microglia and astrocytes into Mic.13 and Ast.10, respectively, opens new routes for interventions into preventing the symptomatic manifestations of AD—cognitive impairment and dementia—that are the key clinical targets. While these cell subpopulations that we prioritize echo the transcriptomic characteristics of glial subpopulations that have been reported previously as associated with AD in case–control analyses[1,2,6,10,14,16], our study refines the distinction between these subpopulations and, importantly, provides specific hypotheses as to which elements of the AD cascade each astrocytic and microglial subpopulation contributes to: our results are actionable as we propose molecularly specific hypotheses. Moreover, we highlight the relative vulnerability of specific SST^{+} neuronal subpopulation as previously indicated[1], and prioritize oligodendroglia, vascular and other cell subpopulations within coordinated cellular communities that will require further investigation.

Methodologically, our research solves critical challenges in the study of molecular processes involved in slowly progressive diseases such as AD. First, we show that rigorous statistical methodologies, such as mediation analysis, enable us to propose likely causal drivers despite the limitations of cross-sectional post-mortem datasets and simplifying assumptions in our modelling (Fig. 3). Furthermore, we were able to predict cellular compositions from bulk RNA-seq (using CelMod[1]) for different cell types, and these inferred cell subpopulation proportions created a robust, independent replication dataset that confirmed our results and enhanced our statistical power when deployed in a meta-analysis combining snRNA-seq and bulk RNA-seq data (Figs. 3 and 5). Finally, the BEYOND strategy facilitated the data-driven reconstruction of disease trajectories, independent of clinicopathologic phenotypes such that unbiased analyses can then relate these trajectories to AD-related traits (Fig. 5); this is an important extension of previous methods that integrate pathologies to infer disease pseudotime[35]. Moreover, BEYOND provided a refined characterization of the cellular communities found in the older brain (Fig. 6), some of which may become targets of therapeutic development. The BEYOND approach can be generalized by integrating data across brain regions and data modalities. Thus, our study provides a conceptual framework applicable to study any dynamic process with coordinated changes in cellular environments.

The relatively large sample size of our study enabled robust and novel discoveries, yet larger sample sizes of genetically and pathologically diverse individuals across brain regions are necessary to identify additional, minor trajectories that might exist and to better characterize the ABA trajectory. Moreover, additional data modalities on a large scale (for example, protein expression, post-transcriptional modifications and spatial information) will help to refine our models. Our study highlights that AD is a disease of the whole brain environment where we may need to modulate cellular communities, not singular cell subpopulations, to restore homeostasis and preserve cognitive function. We have generated precise hypotheses as to where in the sequence of events to AD a particular subpopulation or community may have a role, informing future therapeutic development and clinical trial design. Moreover, we highlight the need for partitioning of individuals in terms of their likelihood of being on the prAD or the ABA trajectory: given that these trajectories are based on diverging cellular environments, an individual's response to a therapeutic intervention could be quite different depending on their cellular trajectory. Furthermore, ignoring this heterogeneity will reduce the statistical power of AD studies, as including individuals not on the prAD trajectory will add noise to a study. We have therefore provided a cellular foundation for a new perspective on AD pathophysiology in which a pathologic cellular community becomes the true target of therapeutic development, and the shared molecular signals of this community provide an important substrate for the development of interventions.

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/s41586-024-07871-6.

1. Cain, A. et al. Multicellular communities are perturbed in the aging human brain and Alzheimer's disease. *Nat. Neurosci.* **26**, 1267–1280 (2023).
2. Mathys, H. et al. Single-cell atlas reveals correlates of high cognitive function, dementia, and resilience to Alzheimer's disease pathology. *Cell* **186**, 4365–4385 (2023).
3. Yang, A. C. et al. A human brain vascular atlas reveals diverse mediators of Alzheimer's risk. *Nature* **603**, 885–892 (2022).
4. Pandey, S. et al. Disease-associated oligodendrocyte responses across neurodegenerative diseases. *Cell Rep.* **40**, 111189 (2022).
5. Luquez, T. et al. Cell type-specific changes identified by single-cell transcriptomics in Alzheimer's disease. *Genome Med.* **14**, 136 (2022).
6. Sadick, J. S. et al. Astrocytes and oligodendrocytes undergo subtype-specific transcriptional changes in Alzheimer's disease. *Neuron* **110**, 1788–1805 (2022).
7. Tuddenham, J. F. et al. A cross-disease human microglial framework identifies disease-enriched subsets and tool compounds for microglial polarization. Preprint at *bioRxiv* https://doi.org/10.1101/2022.06.04.494709 (2022).
8. Gerrits, E. et al. Distinct amyloid-β and tau-associated microglia profiles in Alzheimer's disease. *Acta Neuropathol.* **141**, 681–696 (2021).
9. Habib, N. et al. Disease-associated astrocytes in Alzheimer's disease and aging. *Nat. Neurosci.* **23**, 701–706 (2020).
10. Grubman, A. et al. A single-cell atlas of entorhinal cortex from individuals with Alzheimer's disease reveals cell-type-specific gene expression regulation. *Nat. Neurosci.* **22**, 2087–2097 (2019).
11. Keren-Shaul, H. et al. A unique microglia type associated with restricting development of Alzheimer's disease. *Cell* **169**, 1276–1290.e17 (2017).
12. 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).
13. Zhang, B. et al. Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer's disease. *Cell* **153**, 707–720 (2013).

14. Mathys, H. et al. Single-cell transcriptomic analysis of Alzheimer's disease. *Nature* **570**, 332–337 (2019).
15. Lau, S.-F., Cao, H., Fu, A. K. Y. & Ip, N. Y. Single-nucleus transcriptome analysis reveals dysregulation of angiogenic endothelial cells and neuroprotective glia in Alzheimer's disease. *Proc. Natl Acad. Sci. USA* **117**, 25800–25809 (2020).
16. Zhou, Y. et al. Human and mouse single-nucleus transcriptomics reveal TREM2-dependent and TREM2-independent cellular responses in Alzheimer's disease. *Nat. Med.* **26**, 131–142 (2020).
17. Garcia, F. J. et al. Single-cell dissection of the human cerebrovasculature in health and disease. *Nature* **603**, 893–899 (2022).
18. Dubois, B. et al. Revising the definition of Alzheimer's disease: a new lexicon. *Lancet Neurol.* **9**, 1118–1127 (2010).
19. Bennett, D. A., Schneider, J. A., Arvanitakis, Z. & Wilson, R. S. Overview and findings from the religious orders study. *Curr. Alzheimer Res.* **9**, 628–645 (2012).
20. Bennett, D. A. et al. Overview and findings from the Rush Memory and Aging Project. *Curr. Alzheimer Res.* **9**, 646–663 (2012).
21. Habib, N. et al. Massively parallel single-nucleus RNA-seq with DroNc-seq. *Nat. Methods* **14**, 955–958 (2017).
22. Habib, N. et al. Div-Seq: single nucleus RNA-seq reveals dynamics of rare adult newborn neurons. *Science* **353**, 925–928 (2016).
23. Bennett, D. A. et al. Neuropathology of older persons without cognitive impairment from two community-based studies. *Neurology* **66**, 1837–1844 (2006).
24. The National Institute on Aging, and Reagan Institute Working Group on Diagnostic Criteria for the Neuropathological Assessment of Alzheimer's Disease. Consensus recommendations for the postmortem diagnosis of Alzheimer's disease. *Neurobiol. Aging* **18**, S1–S2 (1997).
25. Kang, H. M. et al. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat. Biotechnol.* **36**, 89–94 (2018).
26. Marques, S. et al. Oligodendrocyte heterogeneity in the mouse juvenile and adult central nervous system. *Science* **352**, 1326–1329 (2016).
27. Waller, R., Mandeya, M., Viney, E., Simpson, J. E. & Wharton, S. B. Histological characterization of interneurons in Alzheimer's disease reveals a loss of somatostatin interneurons in the temporal cortex. *Neuropathology* **40**, 336–346 (2020).
28. Bateman, R. J. et al. Clinical and biomarker changes in dominantly inherited Alzheimer's disease. *N. Engl. J. Med.* **367**, 795–804 (2012).
29. Jack, C. R. et al. NIA-AA Research Framework: toward a biological definition of Alzheimer's disease. *Alzheimers Dement.* **14**, 535–562 (2018).
30. Selkoe, D. J. & Hardy, J. The amyloid hypothesis of Alzheimer's disease at 25 years. *EMBO Mol. Med.* **8**, 595–608 (2016).
31. Carpenter, A. E. et al. CellProfiler: image analysis software for identifying and quantifying cell phenotypes. *Genome Biol.* **7**, R100 (2006).
32. Felsky, D. et al. Neuropathological correlates and genetic architecture of microglial activation in elderly human brain. *Nat. Commun.* **10**, 409 (2019).
33. Setty, M. et al. Characterization of cell fate probabilities in single-cell data with Palantir. *Nat. Biotechnol.* **37**, 451–460 (2019).
34. Stassen, S. V., Yip, G. G. K., Wong, K. K. Y., Ho, J. W. K. & Tsia, K. K. Generalized and scalable trajectory inference in single-cell omics data with VIA. *Nat. Commun.* **12**, 5528 (2021).
35. Iturria-Medina, Y. et al. Unified epigenomic, transcriptomic, proteomic, and metabolomic taxonomy of Alzheimer's disease progression and heterogeneity. *Sci. Adv.* **8**, eabo6764 (2022).

Methods

Experimental design: study participants, AD traits and ethics

Data were derived from individuals enrolled in one of two longitudinal clinical–pathologic cohort studies of ageing and dementia, the Religious Orders Study (ROS)[19] and the Rush Memory and Aging Project (MAP)[20], collectively referred to as ROSMAP. All of the participants are without known dementia at enrolment, have annual clinical evaluations and agree in advance to brain donation at death. At death, the brains undergo a quantitative neuropathologic assessment, and the participant's rate of cognitive decline is calculated from the longitudinal cognitive measures that include up to 25 yearly evaluations[36]. Each study was approved by an Institutional Review Board of Rush University Medical Center. All of the participants signed an informed consent, Anatomic Gift Act and repository consent. For this study, we selected 465 participants, blinded to their neuropathologic and clinical traits, and based on availability of frozen pathologic material from the dorsolateral prefrontal cortex (DLPFC, BA9), including only participants with RNA integrity number (RIN) > 5 and post-mortem interval (PMI) < 24 h, as in our previous studies[1]. After defining our cell subpopulations using data from all participants, we excluded 28 participants without whole-genome sequencing (WGS) data or with shallow sequencing resulting in a low number of nuclei within the library, retaining 437 participants for disease associations and statistical analysis (Extended Data Fig. 1b), each with at least 868 nuclei assigned (Extended Data Fig. 1k). Our study cohort includes diverse individuals across the full range of the pathological and clinical stages of AD. The demographic and clinicopathologic characteristics are described in Fig. 1, Supplementary Table 1 and Extended Data Fig. 1.

Pathological measures were collected as part of the ROSMAP cohorts (previously described[37–39]). We focused our analysis on three quantitative AD-related traits, which have a larger statistical power compared to discrete classifications. (1) Rate of cognitive decline: uniform structured clinical evaluations, including a comprehensive cognitive assessment, are administered annually to the ROS and MAP participants. The ROS and MAP methods of assessing cognition have been extensively summarized in previous publications[40–42]. Scores from 19 cognitive performance tests common in both studies, 17 of which were used to obtain a summary measure for global cognition as well as measures for five cognitive domains of episodic memory, visuospatial ability, perceptual speed, semantic memory and working memory. The summary measure for global cognition is calculated by averaging the standardized scores of the 17 tests, and the summary measure for each domain is calculated similarly by averaging the standardized scores of the tests specific to that domain. To obtain a measurement of cognitive decline, the annual global cognitive scores are modelled longitudinally with a mixed effects model, adjusting for age, sex and education, providing person-specific random slopes of decline (which we refer to as cognitive decline). Further details of the statistical methodology have been previously described[43]. (2) Aβ and tau pathology burden: quantification and estimation of the burden of parenchymal deposition of Aβ and the density of abnormally phosphorylated tau-positive neurofibrillary tangles levels present at death (which we refer to as Aβ and tau pathology, respectively). Tissue was dissected from eight regions of the brain: the hippocampus, entorhinal cortex, anterior cingulate cortex, midfrontal cortex, superior frontal cortex, inferior temporal cortex, angular gyrus and calcarine cortex. Sections (20 μm) from each region were stained with antibodies for the Aβ and tau protein, and quantified using image analysis and stereology. Measurements were summarized to provide a global measure of Aβ and tau burdens. For the trait association, causality prediction and dynamics modelling analyses, we used the measurements of Aβ and tau in the midfrontal cortex (referred to as neocortical Aβ and -tau). The load of these pathologies at the midfrontal cortex are a potentially better proxy for the pathology in the DLPFC brain region compared with the pathology load across the entire brain. Furthermore, we note that we are assessing cellular changes in the fresh-frozen DLPFC samples from one hemisphere of each brain in relation to the measures of cortical Aβ and tau burdens measured in the midfrontal cortex of the opposite, fixed hemisphere in which the standard, structured neuropathologic assessment is conducted. Moreover, we use in our analysis measures quantifying final cognitive diagnosis[44], Braak stage[23,45], CERAD score[23,46] and NIA-Reagan criteria[23,24].

Nucleus isolation and single-nucleus RNA library preparation

To increase library throughput and reduce batch effects and price, we profiled single nuclei in pooled batches of samples. Each batch of samples for library construction consisted of eight participants, except batch B63, which comprised seven participants. Despite random assignment of samples to batches, they were balanced for clinical and pathological diagnosis and for sex (Extended Data Fig. 1a). DLPFC tissue specimens were received frozen from the Rush Alzheimer's Disease Center. We observed variability in the morphology of these tissue specimens with differing amounts of grey 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 grey matter tissue was transferred into the dounce homogenizer (Sigma-Aldrich, D8938) with 2 ml of NP40 lysis buffer (0.1% NP40, 10 mM Tris pH 8.0, 146 mM NaCl, 1 mM $CaCl_2$, 21 mM $MgCl_2$, 40 U ml^{-1} of RNase inhibitor (Takara, 2313B)). Tissue was gently dounced 25 times with pestle A, followed by 25 times with pestle B while on ice, 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 to a final volume of 5 ml and then immediately centrifuged with a swinging-bucket rotor at 500*g* for 5 min at 4 °C. The samples were processed two at a time, the supernatant was removed and the pellets were left on ice 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 preseparation filters (Miltenyi, 130-101-812) and counted using the Nexcelom Cellometer Vision and a 2.5 μg $μl^{-1}$ API stain at 1:1 dilution with cellometer cell counting chamber (Nexcelom, CHT4-SD100-002). In total, 5,000 nuclei from each of 8 participants were then pooled into one sample, and the 40,000 nuclei in around 15–30 μl volume were loaded into two channels on the 10x scRNA-seq platform using the Chromium Single Cell 3′ Reagent Kits version 3. Libraries were generated according to the manufacturer's protocol. In brief, single nuclei were partitioned into nl-scale gel beads in emulsion (GEMs) in the Chromium controller instrument, where cDNAs from the same cell shares a common 10x barcode from the bead. Amplified cDNA is measured by Qubit HS DNA assay (Thermo Fisher Scientific, Q32851) and quality assessed using the BioAnalyzer (Agilent, 5067-4626). This WTA (whole-transcriptome amplified) material was diluted to <8 ng ml^{-1} and processed through v3 library construction, and resulting libraries were quantified again using the Qubit and BioAnalzyer systems. 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 1 million reads per channel. The same libraries of batches B10–B63 were resequenced at The New York Genome Center using Illumina NovaSeq 6000. Sequencing data from both Broad Institute and New York Genome Center were used for analysis.

Preprocessing and quality control steps for snRNA-seq data

For each of the 127 pooled libraries, we performed the following steps: (1) library alignment and background noise removal; (2) demultiplexing; (3) application of a normalization and clustering pipeline; (4) classification of nuclei for cell types; (5) removal of low-quality nuclei using a cell-type-specific threshold; (6) detection of doublets for removal. Further details of each step are provided below.

Library alignment and background noise removal. Libraries were aligned to the GRCh38 pre-mRNA transcriptome and unique molecular identifier (UMI)-collapsing were inferred using the CellRanger[47] toolkit (v.6.0.0, chemistry V3, 10x Genomics). To control for technical artifacts of background ambient RNA molecules, we ran CellBender[48] remove-background (v.0.2.0) utility over the gene expression matrices generated by CellRanger. In brief, CellBender is an unsupervised method for inferring empty- and cell-containing droplets, learning the background RNA distribution of the empty droplets, and removal of such technical background noise. We therefore retrieve uncontaminated cell-containing droplets (cuda flag set, epochs=300, learning rate=1e-5, z-dim=50). Each of the 127 libraries was processed while setting the number of expected cells according to the number of nuclei estimated by CellRanger.

Demultiplexing. We demultiplexed nuclei and inferred participants of origin in our pooled libraries using available genotype data of the participants (WGS or arrays). On the basis of the participants' polymorphic sites and each nucleus' genotype data obtained from the snRNA-seq reads, we assigned each nucleus back to its original participant using the demuxlet[25] software (v.0.1-beta). From the WGS-based VCF file of 1,196 ROS/MAP individuals, we extracted single-nucleotide polymorphisms (SNPs) that were in transcribed regions, passed a filter of GATK, and at least one of the eight individuals had its alternative allele. The extracted SNP genotype data were fed to demuxlet along with a BAM file generated by CellRanger. For libraries in which not all eight individuals had previously been genotyped, we used freemuxlet (https://github.com/statgen/popscle; v.0.1-beta), which clusters droplets on the basis of 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 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. After demultiplexing, we included data for 465 individuals for the generation of the cell atlas and 437 individuals passing quality thresholds, with sufficient number of nuclei assigned (nuclei per participant: minimum = 868, median = 3,590, maximum = 11,530; Extended Data Fig. 1k) for statistical analyses, including the trait associations, mediation modelling and BEYOND analysis.

Normalization and clustering pipeline. The following pipeline was executed on the RNA count matrix: normalization and scaling (Seurat package (v.4)[49], SCTransform, variable.features.n=2000, conserve.memory=TRUE), dimensionality reduction (Seurat, RunPCA, npcs=30), construction of shared neighbour graph (Seurat, FindNeighbors, dims=1:30) and Louvain community detection clustering (Seurat, FindClusters, Resolution=0.2, algorithm=1).

Automatic classification of cell types. We automatically classified nuclei into one of the following eight major cell types: excitatory neurons, inhibitory neurons, astrocytes, microglia, oligodendrocytes, OPCs, endothelial and pericytes. The automatic annotation of nuclei was done by a weighted ElasticNet-regularized logistic regression classifier, fitted over our previous cell atlas of the human ageing DLPFC from 24 individuals[1] with a total of 182,739 nuclei (Extended Data Fig. 1c). The gene count matrix of the previous atlas[1] was log normalized (Seurat, NormalizeData) and scaled (Seurat, ScaleData, method=vst) over the top 700 variable features (Seurat, FindVariableFeatures, excluding non-coding non-annotated loci with a pattern of ^(AL|AC|LINC)\\d+).

To select the optimal regularization parameter we applied ten-fold cross-validation (glmnet package[50,51], cv.glmnet) over randomly selected 75% of the data. To ensure the capture of rare cell types such as pericytes, we weighed samples as $1/n_k$ for the number of nuclei of the cell type present in the training set. We selected the ElasticNet mixing parameter of $\alpha = 0.25$ (to increase the sparsity of the fitted model) by evaluating test accuracy over the remaining 25% of the data. The fitted model used only 121 features and achieved a test accuracy of 99.95.

Removal of low-quality cells. Low-quality nuclei were identified by the total number of UMIs (#UMIs) and the number of unique genes (#Genes). As different brain cell types have inherently different RNA quantities, we learned cell-type-specific thresholds over these parameters. Thresholds were optimized based on hand annotation of 10 pooled libraries and applied to all 127 libraries to classify low-quality cells, and remove such cells from the downstream analysis (Extended Data Fig. 1d). Clusters of the 10 pooled libraries were manually curated to low- and high-quality clusters based on the #UMIs and #Genes distributions (Seurat, VlnPlot). We then selected the cell-type-specific thresholds as the median of all optimal #UMIs and #Genes parameter pairs, scored using the harmonic mean of the precision and recall. #UMIs and #Genes thresholds were as follows: excitatory neurons, 2,232 and 1,916; inhibitory neurons, 800 and 100; astrocytes, 800 and 616; microglia, 400 and 253; oligodendrocytes, 400 and 253; OPCs, 695 and 253; vascular cells, 400 and 253; and pericyte cells, 400 and 100, respectively (Extended Data Fig. 1e,f). Low-quality clusters were removed as well by a Soft-SVM classifier fitted over the same 10 pooled libraries and using the (1) proportion of nuclei annotated as low quality (by #UMIs and #Genes threshold); (2) average entropy of cell-type prediction, and (3) the proportion of doublets by the demuxlet algorithm.

Doublet detection. Between-sample doublets were identified by the demuxlet algorithm, based on the sample barcodes (Extended Data Fig. 1g). Within-sample doublets were predicted in silico based on their RNA profiles. To predict doublets, we ran DoubletFinder[52] (DoubletFinder_v3, pN=0.5, pK=75/(1.5*(#nuclei in library)), nExp=0, sct=T) over each of the libraries (Extended Data Fig. 1h). Thresholds for DoubletFinder predictions were determined separately for each library, based on the maximal Matthew's correlation coefficient compared to the demuxlet-identified doublets (Extended Data Fig. 1i). Furthermore, as DoubletFinder is not designed to identify doublets of the same cell type, we modified it to simulate doublets from parent nuclei of different cell types, inferred based on the cell-type classification (https://github.com/GreenGilad/DoubletFinder). Using high-resolution clustering of the nuclei (Seurat, FindClusters, Resolution=1.5) we expanded and marked as a doublet any nuclei predicted to be a demultiplexed doublet, a DoubletFinder doublet or belonging to a cluster consisting of more than 70% DoubletFinder doublets.

Unified UMAP space. To compute a UMAP embedding consisting of all nuclei, we created a Seurat object over a subpopulation of nuclei and projected the remaining nuclei onto the UMAP space computed for the subpopulation. As, at the time of writing, Seurat is limited in the counts matrix size, we were unable to create a single 1.65-million-nuclei large object. We randomly sampled 30 out of 127 libraries and created a single Seurat object consisting of 400,000 nuclei. We then followed a similar pipeline of normalizing, scaling, principal component analysis (PCA) and UMAP embedding as described above (using 4,000 variable features and a PCA space of 50 dimensions). These PCA and UMAP spaces are referred to as reference spaces.

The remaining nuclei were then projected as follows. Each library was normalized and scaled using only the variable genes used by the reference (Seurat, SCTransform, specifying residual.features as the reference variable genes). The scaled data were then projected onto the reference PC space using the reference's feature loadings. Now that both sets are embedded in the same PC space, the remaining nuclei were projected onto the reference's UMAP space (Seurat, ProjectUMAP, setting reference and query reductions as 'pca').

Subclustering analysis

On the basis of the cell-type classification, we partitioned nuclei into subsets of the different cell classes, and we performed the subclustering analysis separately per cell type. Endothelial and pericyte cells were analysed together as part of the vascular niche, and OPCs and oligodendrocytes were analysed together as part of the Oligodendroglia. For each cell class, we performed the following analysis steps: (1) removed genes expressed in fewer than 15 nuclei as well as non-coding non-annotated genes (pattern=^(AC\\d+{3}|AL\\d+{3}|AP\\d+{3}|LINC\\d+{3})); (2) removed cells with more than 10% mitochondrial RNA (Seurat, PercentageFeatureSet, pattern='^MT-'); and (3) removed any residual doublet cells that DoubletFinder did not detect. In the case of oligodendrocytes and OPCs, we did not remove nuclei predicted as oligodendrocyte–OPC doublets.

We iteratively ran the normalization and clustering pipeline as described for each library to remove additional low-quality or doublet nuclei, with the following parameters: (1) SCTransform method with variable.features.n = 4000 (for most cells) and n=2000 in the vascular niche subset; (2) RunPCA with npcs=50; (3) FindNeighbors with dims=1:50, FindClusters with Resolution=1.5, algorithm=4 (Leiden), method=igraph; and (4) RunUMAP with dims=1:50, min.dist=0.1. After each such run, we removed clusters with low-quality cells or a high percentage of doublet cells (following the guidelines described above). To validate the doublet annotations, we plotted the expression level of canonical cell-type markers, and small clusters expressing multiple markers of multiple cell types were also removed.

After cleaning the nuclei per cell type, we performed the subclustering analysis per cell type. For each cell type the subclustering was performed over multiple resolutions (Seurat FindClusters, algorithm=4, method=igraph), and the resolution was determined by the (1) differential gene expression per cluster (Seurat, FindAllMarkers, test.use=negbinom); (2) functional annotations of the differential signatures; and (3) the proportions of clusters across individuals. Clusters that did not have differential genes or clusters specific to a single individual were united with the neighbouring cluster with a similar RNA profile (based on Seurat BuildClusterTree). These steps were done separately for the endothelial-, mural- (pericytes and SMCs) and fibroblasts cells making up the vascular niche subset, as well as for the oligodendrocytes and OPC cells making up the oligodendroglia subset.

Owing to Seurat's current counts matrix size limitation, excitatory neurons were split into two major sets (one consisting of upper cortical layer 2–4 pyramidal neurons, *CUX2*$^+$ nuclei and the other consisting of all others deeper cortical layer pyramidal neurons, referred to as *CUX2*$^-$). Clean-up and subclustering analysis were performed for each subset separately. Neuronal clusters at the borders were then reclustered to avoid misclassifications due to the split.

Clustering quality, comparison to published datasets and neuronal subtype classification

Clustering quality was tested by evaluating cluster identity agreement within the shared nearest-neighbour graphs. For each cell type, we conducted a pairwise comparison across subclusters, showing the fraction of shared nearest neighbours, shared within a cluster and between clusters. Moreover, we compared our cell subpopulations with recently published human and mouse annotations by computing gene signature scores (Seurat, AddModuleScore; see Extended Data Fig. 3a–c; the genes used in each signature can be found in the 'Code availability' section). We then evaluated a signature's enrichment in each subpopulation using a Wilcoxon rank-sum test (wilcox.test method, x=within subpopulation signature values, y=between subpopulations signature values, alternative=greater). *P* values were corrected for multiple hypothesis testing using the Benjamini–Hochberg procedure. For neuronal cells, we predicted the class membership of each of our snRNA-seq nuclei profiles using singleR[53] package. We fitted the classifier separately for inhibitory and excitatory neurons using the published dataset of the Brain Cell Types Database from the Allen Brain Atlas, human atlas of M1 cortical region mapped by 10x genomics (https://celltypes.brain-map.org/rnaseq/human_m1_10x). Each nucleus was labelled by the maximum prediction value (using pruned.labels output), and accordingly assigned to cortical layers, markers and neuropeptide expression.

Differential expression and functional annotation

Within a given cell class subpopulation, differentially expressed genes (DEGs) were computed between each cluster and the rest using a negative binomial test and controlling for the post-mortem interval and participants' batch (Seurat FindAllMarkers, test.use=negbinom, latent.vars=batch, pmi). In the case of excitatory neurons, as we were not able to create a single Seurat object to facilitate all cells, we partitioned genes into six groups. We then merged *CUX2*$^+$ and *CUX2*$^-$ cells, taking a single group of genes at a time and running the negative binomial test as above. After merging the results, we corrected for multiple-hypothesis testing in the same manner as is performed by Seurat.

Functional annotation was inferred by (1) identifying statistically significant enriched gene pathways and gene sets in the upregulated and downregulated differential gene signatures and (2) grouping them based on their similarities. We identified upregulated and downregulated pathways using clusterProfiler[54] (compareCluster, formula=id~cluster+direction) using KEGG (fun=enrichKEGG, organism=hsa), Reactome (fun=enrichPathway) and GO (fun=enrichGo, OrgDb=org.Hs.eg.db, ont=BP,MF,CC). We set the background universe gene set to the set of all genes present in the data. Redundancy between GO terms were removed (ClusterProfiler, simplify).

To better understand the functionality captured by the pathways and overcome the issue of redundancy within the databases—where multiple different enriched pathways capture the same sets of differential genes and therefore probably reflect the same enriched function—we clustered the inferred upregulated or downregulated pathways of a given subpopulation. Given a list of pathways, we first computed two one-hot encoding matrices of pathways over genes: an evidence matrix and a prior matrix, both sharing the same rows (pathways). The columns of the evidence matrix are the union of all DEGs found for the particular subpopulation and take a value of 1 at place *i*,*j* if and only if DEG *j* is part of pathway *i*. The columns of the prior matrix are the union of all genes that are part of the pathways of the particular subpopulation, and take a value of 1 at place *i*,*j* if and only if gene *j* is part of pathway *i*. Both matrices were then converted into pathways pairwise similarity matrices (simplifyEnrichment[55], term_similarity, method=kappa). We used the kappa coefficient as the measure of similarity to consider the chance of a given gene to be associated with two different pathways. The partitioning of the pathways was done over the joint adjacency matrix being the element-wise average of the prior and evidence similarity matrices using the binary-cut algorithm (simplifyEnrichment, cluster_terms, method=binary_cut). The partitioning cut-off passed to the algorithm was adjusted manually based on the visual inspection of the joined adjacency matrix.

Computation of subpopulation proportions

We divided the subpopulations according to the following seven major cell type groups: excitatory neurons, inhibitory neurons, astrocytes, microglia (which included monocytes and macrophages), oligodendrocytes, OPCs and vascular niche (which included endothelial cells, pericytes, SMC and fibroblasts). Rare subpopulations of erythrocytes, CD8$^+$ T cells, NK cells and neutrophils, for which we have low total abundances, were not included. This resulted in a final set of 91 subpopulations used for the statistical analysis. We then calculated for every participant the subpopulation proportion within the relevant

cell type group, to create the subpopulation proportion matrix whose rows represent participants and columns represent subpopulations:

$$[P]_{ds} = \frac{n_{ds}}{\Sigma_{k \in C(s)} n_{ds}}$$

for d a participant, s a subpopulation, $C(s)$ the set of all subpopulations in the cell type of s and n_{ds} the number of nuclei of subpopulation s in participant d. We refer to the row $[P]_d$ as the cellular environment of participant d. A column $[P]_{.,s}$ is the vector of subpopulation proportions for subpopulation s across all participants. The rows and columns of P were used for association and causality prediction analyses as well as the BEYOND analysis (see sections below).

Statistical analysis associating subpopulations to AD-related traits

Statistical associations between traits and subpopulations were tested by regressing traits on the square-root proportions of the subpopulation: trait = $\sqrt{[P]_{.s}}$ for subpopulation s and P the proportions matrix (defined in the section above). To remove potential confounding effects, we adjusted for (1) age at death, sex and post-mortem interval, and (2) snRNA-seq library quality reflected by the number of cells captured and total genes detected for each participant. The results were corrected for multiple-hypothesis testing by calculating the FDR (p.adjust, method=BH) within each tested trait. We applied a similar approach when associating CelMod predicted subpopulation proportion (see the section below) and AD-related traits, using RIN scores as the measure of library quality. To integrate the statistical associations calculated over the discovery snRNA-seq and the replication bulk-predicted datasets, we applied a meta-analysis approach using a weighted z-statistic[56]:

$$z = \frac{\sqrt{n_{\text{sn}}}\, t_{\text{sn}} + \sqrt{n_{\text{bulk}}}\, t_{\text{bulk}}}{\sqrt{n_{\text{sn}} + \sqrt{n_{\text{bulk}}}}}$$

Meta-analysis results were then corrected for multiple-hypothesis testing (p.adjust, method=BH).

Inferring cell-state proportions from bulk RNA-seq. We used bulk RNA-seq data generated from DLPFC of 1,092 samples. As described previously[57], RNA-seq data were generated in three sessions: (1) 10 batches of 739 participants were sequenced at the Broad Institute, using Qiagen's miRNeasy Mini Kit and RNase-free DNase set. RNA was quantified using Nanodrop and RNA quality was evaluated using the Agilent Bioanalyzer. Samples with RIN score > 5 and quantity threshold > 5 µg were submitted for library construction. Sequencing was conducted using the Illumina HiSeq2000 system with 2 × 101 bp reads for a targeted coverage of 50 million paired reads. (2) Two batches of 124 participants were sequenced at the New York Genome Center, using the KAPA Stranded RNA-seq kit with RiboErase. Sequencing was performed on the Illumina NovaSeq 6000 system using 2 × 100 bp reads targeting 30 million paired reads per sample. (3) One batch of 229 samples was sequenced at the Rush Alzheimer's Disease Center, using the Chemagic RNA tissue kit, with RNA quality number calculated using the Fragment Analyzer. A total of 500 ng total RNA was used to generate the RNA-seq library, after rRNA depletion with RIbogold. Libraries were sequenced on the Illumina NovaSeq 6000 system with 2 × 150 bp reads at 40–50 million reads per sample.

RNA-seq reads were aligned to the GRCh38 human genome and gene-level counts were normalized to transcripts per million (TPM). Outlier samples were removed based on MDS plots and quantiles of expression profiles, and genes with median TPM < 10 were filtered out to reduce technical noise. $\log_2$-transformed TPM values in the expression matrix were included in a linear regression model with the following covariates: age at death, sex, batch, library size, percentage of coding bases, percentage of aligned reads, percentage of ribosomal bases, percentage of UTR bases, median 5′ to 3′ bias, median CV coverage, post-mortem interval and study index (ROS or MAP).

We used the CelMod[1] package (https://github.com/MenonLab/Celmod) to infer cell subpopulation proportions from the residuals of the bulk RNA-seq data, for participants without snRNA-seq profiles. As described previously, CelMod uses a consensus regression model trained on matched snRNA-seq proportions and bulk RNA-seq to extrapolate cell subpopulation proportions to samples with only bulk RNA-seq available. In brief, the approach builds a regression model for each bulk RNA-seq gene using the cell type proportions as the predictor variable. Genes are then ranked by goodness-of-fit for each cell type, and the mean of the top genes is used as the prediction for the proportion of a given cell type in samples having only bulk data. The only free parameter—the number of genes used in the consensus—is determined by a fivefold cross-validation of the training set. Notably, CelMod is not run simultaneously for all cell subpopulations but, rather, for the subpopulations within each broad cell class separately to reduce the effects of noise that may result when trying to estimate the proportion of many cell subsets at the same time.

With a total of 419 overlapping participants between the snRNA-seq and bulk RNA-seq samples, we fitted a CelMod model using 315 (75%) participants. The remaining 104 participants were kept as the hold-out set to assess the reliability of the predictions. Similar to the trait associations above, we applied CelMod to the proportions square-root. To reduce the sensitivity of to the specific selection of training samples, we trained an ensemble of 100 CelMod models, each fitted over a different random train–test split (setting seeds 1 to 100). We averaged a participant's predicted subpopulation proportion, obtaining a single estimated subpopulation proportion for each participant. After running CelMod on each major cell type group separately, we calculated the Spearman correlation of the (averaged) predicted proportion versus the snRNA-seq proportion on the holdout set. This per-cell type Spearman correlation provides an estimate of the reliability of the prediction for each subpopulation. Only subpopulations for which correlation was positive and significant (FDR < 0.01) were considered to be reliable and used in downstream analysis. The same ensemble strategy was used to obtain subpopulation predicted proportions of the 673 replication sample participants (participants with bulk RNA-seq but no snRNA-seq data).

Causal modelling

In this section, we focused on the most robust subpopulation–AD trait associations by selecting those with the association FDR < 0.01 with at least two tested AD traits: Mic.12, Mic.13, Ast10 and Oli.7. We first assessed the association of *APOE* ε4, the strongest genetic risk factor of AD, with the identified key cell states. Subsequently, we tested whether AD endophenotypes (starting with Aβ) mediate the observed *APOE* ε4–cell state association. This step uses *APOE* ε4 as the genetic anchor that guides the direction of effect, leveraging the fact that the genetic variation is not subject to reverse causation[58]. We then built on the widely accepted causal chain of AD pathogenesis that starts with Aβ, leading to tau accumulation and cognitive decline[28–30], and performed a series of causal mediation analyses to determine the most plausible location of each cell state in the core sequence of AD (Aβ→tau→cognitive decline). In mediation models, we used AD endophenotypes and square-rooted cell state proportions as continuous variables and performed causal mediation analyses using a nonparametric bootstrap option with 10,000 simulations in the R mediation package[59]. We then constructed a SEM based on the mediation analyses using the R lavaan package[60]. We first calculated the residuals of each variable by regressing age, sex, PMI, the estimated number of cells (per individual) and the number of detected genes (per individual). An exception was cognitive decline, as this variable has already been residualized against demographic covariates (random slope from

linear mixed effect model, adjusting for baseline age, sex and years of education)[43]. We did not allow cognitive decline to be the parent (causal) node, as cognitive decline is a symptom rather than the cause of AD pathophysiology and cellular environment changes. To replicate our findings in an independent set of individuals, we repeated the same procedure, using the CelMod-predicted cell subpopulation proportions data from ROSMAP participants who did not contribute to the quality-controlled snRNA data (replication sample).

BEYOND analysis

The BEYOND strategy is composed of four major steps: (1) learning the cellular landscape manifold; (2) identifying axes of cellular change; (3) fitting subpopulation and trait dynamics; and (4) grouping subpopulations into cellular communities. To implement this strategy, we used the subpopulation proportion matrix (see above). We represent a participant's cellular environment by its compositional profile of subpopulations (that is, matrix rows). The cellular landscape is thus represented by the vector space spanned by the columns of the matrix. For convenience, we stored the proportion matrix as an AnnData[61] object.

To perform steps 1 and 2, we also clustered the participants based on their cellular environments. We computed their *k*-neighbourhood (scanpy, scanpy.pp.neighbours, n_neighbors=10, use_rep="X", metric="cosine") and then clustered using the Leiden community detection algorithm (scanpy, scanpy.tl.leiden, Resolution=0.25). We further refined cluster #0 (scanpy, scanpy.tl.leiden, Resolution=0.75, restrict_to="0"). This provided the participant's clustering assignment seen in Extended Data Fig. 8a.

Step 1: learning the cellular landscape manifold. We embedded cellular environments using three manifold learning algorithms: (1) PHATE (scanpy, external.tl.phate, k=10, n_components=3, a=40, knn_dist=euclidean, mds_dist=euclidean, mds_solver=smacof); UMAP[62] (scanpy, *tl.umap*, maxiter=3000, spread=3); and *t*-SNE[63] (scanpy, *tl.tsne*, n_pcs=0, use_rep="X", learning_rate=100). To better visualize the main area of the landscape, we also computed a 2D PHATE embedding excluding participants of clusters 9 and 10 (scanpy, external.tl.phate, k=15, n_components=2, a=100, knn_dist=euclidean, mds_dist=correlation, mds_solver=smacof).

Step 2: identifying axes of cellular change. To robustly identify and model axes of cellular change, we performed a pseudotime analysis using two algorithms: Palantir[33] and VIA[34]. In brief, Palantir projects the input data onto a multidimensional diffusion space and constructs a neighbouring graph. It then iteratively refines the shortest paths from a user-defined starting point to each given datapoint, defining this relative distance as the pseudotime. Lastly, it constructs a Markov chain using the neighbouring graph and inferring directionality by the pseudotime. Palantir defines trajectories as the terminal states of the Markov chain and trajectory probabilities as the probability of a given datapoint to reach each of the terminal states. VIA begins by representing the input data as a cluster graph. It computes the pseudotime using a two-step procedure: first as the expected hitting time for a lazy-teleporting random walk along the undirected cluster graph starting from a user-defined starting point, and then refines it by running Markov chain Monte Carlo (MCMC) simulations. VIA defines trajectories by consensus voting on terminal states and trajectory probabilities as the visitation frequencies under the lazy-teleporting MCMC simulations.

We ran Palantir over the cellular landscape excluding participant clusters 9 and 10 (Palantir, run_diffusion_maps and determine_multiscale_ space, n_components=5, knn=50; Palantir, run_palantir, knn=30, max_iterations=100, scale_components=F, n_jobs=1, use_early_cell_as_start=F). As VIA was designed to model trajectories in disconnected graphs, we ran it over the full cellular landscape of all 437 participants (VIA, knn = 20, too_big_factor=0.2 and 0.075, small_pop=5, preserve_disconnected=True). For both algorithms we provided the same starting point, which was calculated as the medoid participant of participant clusters 1 and 3. Participants of these clusters exhibit high proportions of homeostatic glial subpopulations.

Step 3: fitting and plotting dynamics. Dynamics were computed by regressing the feature values over the pseudotime in a specific trajectory using a generalized additive model, similarly to the strategy used by both Palantir[33] and VIA[34]. Features used were participants' traits, subpopulation proportions or community proportions. We then used the fitted model to predict the final dynamics over equidistant pseudotime values. In detail, we spline-fitted a generalized additive model for feature *y* in trajectory *j*:

$$y_i \sim s(\mathrm{ps}_i), \text{ weighted by } T_{ij} \text{ and where } \mathrm{ps}_i \le \mathrm{ps}(j)$$

where *i* is a participant, *j* is a trajectory, ps_i is the pseudotime of participant *i*, ps(*j*) is the terminal pseudotime of trajectory *j* and T_{ij} the trajectory probability matrix of participant i in trajectory *j* (mgcv[64], gam, formula = $y \sim$ s(ps), weights = $T_{.j}$). The final dynamics were predicted over equidistant pseudotime values in the range [0, ps(*j*)] (mgcv predict.gam, se.fit=TRUE). When plotting the dynamics we presented the predicted feature values over the equidistant pseudotime values, as well as a confidence interval area of predicted value ±2 × s.e., as retrieved from predict.gam.

Step 4: constructing cellular communities. Partitioning of subpopulations into cellular communities was done using two sorts of measures: (1) similarity in subpopulation dynamics and (2) co-occurrences of subpopulations across participants. Similarities of dynamics were calculated using a weighted adaptive RBF kernel over the *z*-scored dynamics matrix, computed as follows: we represented each subpopulation *s* by its dynamics along both trajectories $x_s \in \mathbb{R}^{t_1+t_2}$ for t_1 and t_2 the number of equidistant pseudotime values used for the predicted dynamics (see section above). We obtained $\tilde{x}_s$ by centring and standardizing x_s: $\tilde{x}_s = (x_s - \mathrm{mean}(x_s))/\mathrm{s.d.}(x_s)$. We then computed *M*, the Mahalanobis distance between every two subpopulations, as $M_{sk} = (x_s - x_k)^\top W (x_s - x_k)$ for *W*, a diagonal matrix of which the first t_1 diagonal values are $1/t_1$ and the rest $1/t_2$ (that is, equally weighing both trajectories). Lastly, we calculated the adjacency matrix as $A_{\mathrm{dyn}} = \exp(-M/\sigma\sigma^\top)$, where the division is performed element-wise and clipped to zero values smaller than 10^{-4}. $\boldsymbol{\sigma}$ is the vector of local densities at each subpopulation calculated as $\boldsymbol{\sigma}_s = \sqrt{\Sigma_{k \in \mathcal{N}(s)} M_{sk}}$ and $\mathcal{N}(s)$ is the set of 5-nearest neighbours of *s* according to *M*. The co-occurrence matrix A_{co} was calculated as the pairwise spearman correlation matrix of subpopulation proportions across all participants (stats, cor, use=pairwise.complete.obs, method=spearman). That is, $[A_{\mathrm{co}}]_{sk} = \mathrm{cor}([P]_{.s}, [P]_{.k})$ for *P* the subpopulation proportions matrix defined in the sections above.

We used the Leiden community detection algorithm over the multiplexed 3-layered graph induced by the matrices A_{dyn}, A_{co} after zeroing negative correlations and $-A_{\mathrm{co}}$ after zeroing positive correlations (optimise_partition_multiplex method, leidenalg Python package[60,61]), with layers weighted as 1, 1, −1, respectively. We used the RBERVertexPartition partitioning model[62,63], specifying the resolution parameter for each layer as the value maximizing the modularity (resolution_profile method, leidenalg Python package, partition_type=RBERVertexPartition, resolution_range=[10^{-2}, 10], number_iterations=−1). We further refined communities based on the dendrogram calculated within each community based on the dendrogram calculated within each community using $A_{\mathrm{dyn}} + A_{\mathrm{co}}$.

Once subpopulations were partitioned into cellular communities, we assigned participants with a community proportion by averaging the normalized subpopulation proportions, and further normalization such that community proportions for every participant sum to one.

Community dynamics along trajectories were calculated using the same procedure as used for traits or subpopulation dynamics.

smFISH quantification and analysis

RNAscope experiment. The RNAscope Multiplex Fluorescent Reagent Kit v2 (ACD, 323100) was used to perform the RNAscope experiments on 15 individuals from the New York Brain Bank (6 cognitively unimpaired, no AD; 1 mild cognitive impairment, MCI; 8 AD). 6 µM paraffin-embedded tissue sections were deparaffinized with CitriSolv Clearing Agent (Decon Laboratories, 1601) for 20 min at room temperature, followed by an ethanol series (100%, 100%, 70%) for 30 s each. The slides were then put in distilled water for 1 min at room temperature. The slides were then incubated with hydrogen peroxide for 10 min at room temperature, then washed with distilled water twice to stop the hydrogen peroxide reaction. Antigen retrieval was performed with pH 6.0 citrate (Sigma-Aldrich, C9999) and heating with a microwave for 25 min at 400 W. After 5 min in tap water, the slides were immersed in 100% ethanol for 1 min. Once the slides had fully dried at room temperature, the Super Pap Pen Liquid Blocker (Newcomer Supply, 6505) was used for drawing a hydrophobic barrier around the tissue section. The slides were blocked for 30 min at room temperature using the RNAscope Co-Detection Antibody Diluent.

All of the slides were stained with IBA1, a microglial marker (Wako, 01127991) diluted in RNAscope Co-Detection Antibody Diluent at 1:50 and incubated for 2 h at room temperature. The slides were then washed with PBS-T (PBS + 0.1% Tween-20) and then submerged in 10% neutral buffered formalin (Sigma-Aldrich, HT5011) for 1 h at room temperature. After washing with PBS-T, the slides were treated with RNAscope protease plus and incubated for 40 min at 40 °C in the RNAscope HybEz II oven. After washing with distilled water, RNAscope probe mix was added to the slides and incubated for 2 h at 40 °C. Finally, the slides were washed with the RNAscope wash buffer and incubated with 5× SSC (Sigma-Aldrich, S6639-1L) overnight at room temperature. The next day, the slides were washed with RNAscope wash buffer and incubated with AMP1 for 30 min at 40 °C. The slides were washed twice with wash buffer and AMP2 was added and incubated for 30 min at 40 °C. After washing, AMP3 was added and incubated for 15 min at 40 °C. After 15 min of incubation at 40 °C with HRP-C1, the slides were washed and Opal 570, diluted in TSA diluent buffer (1:700) and incubated for 30 min at 40 °C. The slides were then washed and, finally, an HRP blocker was added for 15 min. This HRP/Opal/block process was repeated using HRP-C2, HRP-C3 and HRP-C4 depending on the channel of the original probes, and Opal 690) dye. TSA-DIG reagent was used after HRP-C4 and incubated for 30 min at room temperature then blocked with HRP blocker and washed with PBS-T. The secondary antibody (Thermo Fisher Scientific, A11055) for the IBA1 staining was diluted in co-detection antibody diluent at 1:500 and then incubated for 30 min at room temperature. After washing with PBS-T, the slides were incubated with Opal Polaris 780 dye prepared in the antibody diluent/block at 1:700 for 30 min at room temperature. The slides were washed with PBS-T and the lipofuscin autofluorescence was quenched using Trueblack (Biotium, 23007) for 2 min at room temperature and DAPI was added to stain the nucleus. All of the slides were mounted using the Prolong Gold (Thermo Fisher Scientific, P36934). The RNAscope probes used in this experiment were as follows: TPRG1-C1 (ACD, 1047171-C1), MRC1-C3 (ACD, 583921-C3) and CPM-C2 (444811-C2).

Immunohistochemistry. 6 µM paraffin-embedded tissue sections were deparaffinized using CitriSolv Clearing Agent (Decon Laboratories, 1601) for 20 min at room temperature, followed by an ethanol series (100%, 100%, 70%) for 30 s to 1 min each. The slides were then put in distilled water for 1 min at room temperature. Antigen retrieval was performed with pH 6.0 citrate (Sigma-Aldrich, C9999) and heated with a microwave for 25 min at 400 W. After a 5 min immersion in tap water, the slides were blocked using 3% BSA for 30 min and incubated with pTau AT8 primary antibody (Invitrogen, MN1020) overnight at 4 °C. The secondary antibody (Thermo Fisher Scientific, A32787) for the pTau staining was diluted in PBS at 1:500 and then incubated for 1 h at room temperature. After washing with PBS, the lipofuscin autofluorescence was quenched using Trueblack (Biotium, 23007) for 2 min at room temperature. The slides were mounted using the Prolong Gold with Dapi (Thermo Fisher Scientific, P36941). This staining was done on a sister section for each section on which RNAscope data were generated.

Image signal quantification. For all slides, images were acquired using the Nikon Eclipse Ni-E immunofluorescence microscope at a magnification of ×40, and approximately 30 pictures were acquired per individual. The captured images were analysed using CellProfiler[31] software. An extensive pipeline has been developed to automatically segment the microglia and detect transcripts (*CPM*, *MRC1* and *TPRG1*) expressed by IBA1$^+$ cells[12]. DAPI was defined as the primary object using the IdentifyPrimaryObjects module. For IBA1$^+$ cell segmentation, the module EnhanceOrSuppressFeatures was used to enhance the IBA1 signal and to detect the ramification. IBA1$^+$ cells were also defined as the primary object and, using the RelateObjects module, only IBA1$^+$DAPI$^+$ cells were selected. To segment the transcript signals (dots), we used the EnhanceOrSuppressFeatures module with Speckles as the feature type, and the RelateObjects module was used to relate the transcripts signals (*CPM*, *TPRG1* and *MRC1*) to IBA1$^+$DAPI$^+$ cells. The morphology (eccentricity and compactness) of IBA1$^+$DAPI$^+$ cells was measured using the MeasureObjectSizeShape module and the intensity of each transcript as well as IBA1 was measured using the MeasureObjectIntensity module. The data were exported in Excel format using ExportToSpreadsh. For the quantification of pTau (AT8), 35–40 images per individual were captured at ×20 magnification, in a zigzag pattern to cover all six cortical layers. Images were analysed with the Cellprofiler[31] software, where tangles and pTau plaques were size-filtered. The area occupied measurement was used for the analysis.

We assigned microglial subpopulations to the smFISH data based on the expression of the marker genes *CPM*, *TPRG1* and *MRC1*. A cell was classified as a myeloid cell if it was IBA1$^+$DAPI$^+$. None myeloid cells were excluded from further analysis. Cells were then assigned to one of five categories: Mic.13 (*TPRG1*>5 and *MRC1*<10), Mic.12 (*TPRG1*<2, *CPM*>5 and *MRC1*<5), macrophage (*MRC1*≥ 5), other myeloid (*TPRG1*<2, *CPM*<2 and *MRC1*<3) or none if none of the above criteria were met. Cells assigned as 'none' represent insufficient evidence of classifying or not classifying as Mic.12, Mic.13 or macrophages. Once classes were assigned, subpopulation proportions were calculated for each of the RNAscope participants across all five categories.

Spatial transcriptomics analysis

We profiled ten samples of ROSMAP participants using spatial transcriptomics (Visium), including 4 samples predicted to be in the prAD trajectory and 2 samples predicted to be in the ABA trajectory, and 4 predicted to be in the early pseudotime (unassigned to prAD nor ABA).

Visium data generation. For spatial transcriptomics analysis with Visium combined with immunofluorescence, the cerebral cortex and the underlying white matter of fresh frozen brain tissue were obtained from the ROS and MAP cohorts. Tissues were dissected on dry ice to be prepared into a tissue size of 10 × 10 mm^2 in cryoembedding matrix (OCT) and sectioned at 10 µm thickness in duplicate onto a spatial transcriptomics slide with capture probes. The sections were selected based on an RIN greater than 6. The sections were fixed with cold 100% ethanol for 30 min and stained with thioflavin S for 8 min at room temperature (Sigma-Aldrich, T1892). After serial washing with the wash buffer, the slides were incubated in TrueBlack for 2 min to quench the autofluorescence from the lipofuscin. The sections were scanned using the Nikon ECLIPSE immunofluorescence microscope at magnification ×20. The tissue sections were then permeabilized to induce cDNA

synthesis, and a cDNA library was generated. Subsequently, libraries were sequenced and aligned with the bright-field images using the Space Ranger software.

Visium spot-level data processing and cortical layer annotations. For each sample, we loaded the CellRanger output h5 file using Seurat (v.4.1.4) for a per-spot gene quantification and visualization. For quality control, we removed spots with less than 500 expressed genes or with more than 30% mitochondrial RNA, and normalized the data (Seurat, SCTransform). Cortical layer annotation was obtained by clustering the spatial transcriptomics spots into 7 clusters. First, 35 PCs were calculated using the intersection of the top 10,000 highly variable genes and 2,500 cortical layer marker genes from a recent publication[65]. Next, Harmony[66] was used to normalize the PCA embeddings across tissue sections. The BayesSpace[67] method was then used to cluster spots, with the smoothing parameter gamma set to 2. After layer annotations, we removed spots classified as white matter from the rest of the analysis.

Markov affinity-based graph imputation of Visium spots. To overcome data sparsity, we used MAGIC[68] (Markov affinity-based graph Imputation of cells) to denoise the data. We first normalized the raw count data such that the sum of expression values for each spot sums to 1 (phateR, library.size.normalize). Using the square-root transformation of the normalized matrix as an input, we then ran MAGIC on all genes expressed in at least 10 spots (parameters, knn=5, t=3).

Subpopulation spatial density estimation on Visium-profiled tissues. We first defined the glial subpopulations of interests using marker genes as proxy for each of their signatures: Ast.10 (*SMTN*$^+$*SLC38A2*$^+$), Mic.13 (*TREM2*$^+$*GPNMB*$^+$) and Ast.5 (*SERPINA3N*$^+$*OSMR*$^+$). To visualize the spatial distribution for spots enriched in each of these signatures, we calculated and visualized the joint density distribution for each set for markers by kernel density estimation over the spatial coordinates and weighted by the scaled gene expression (Nebulosa v.1.8.0). For each markers pair, Nebulosa calculates the joint density by multiplying their densities.

Spatial transcriptomics statistical associations. To score the abundance of a given glial subpopulation at the spot level, we summed the MAGIC-imputed expression of each marker pair in each spot. To assess the spatial coordination of Mic.13 and Ast.10, we calculated the Pearson correlation between the abundance scores of Ast.10 and Mic.13 within each of the participants, tested for the significance of the correlation being strictly positive (cor.test, method=pearson, alternative="greater") and corrected for multiple hypothesis testing (FDR, p.adjust, method=BH). Comparison between the prAD and ABA participant groups was performed using one-sided *t*-tests, testing that the difference in means is greater than zero (t.test, x=prAD participant correlations, y=ABA participant correlations, alternative="greater"). To evaluate the shift between Ast.10 and Ast.5 enrichment within prAD- and ABA participants we compared within individual expression levels. For prAD participants we tested whether Ast.10 levels were strictly greater than Ast.5 levels (one-sided *t*-test, x=Ast.10 levels, y=Ast.5 levels, alternative="greater"), while, for ABA participants, we tested whether Ast.5 levels were strictly greater than Ast.10 levels (one-sided *t*-test, x=Ast.5 levels, y=Ast.10 levels, alternative="greater"). Results were corrected for multiple hypothesis testing (FDR, p.adjust, method=BH). Comparison between the prAD and ABA participant groups was done over the log-ratio of Ast.10 to Ast.5, and comparing the mean between the groups using a one-sided *t*-test (x=prAD participants' log(Ast.10/Ast.5), y=ABA participants' log(Ast.10/Ast.5), alternative="greater").

Reporting summary

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

Data availability

snRNA-seq data and analysis output are available at the AD Knowledge Portal (https://adknowledgeportal.org). The AD Knowledge Portal is a platform for accessing data, analyses and tools generated by the Accelerating Medicines Partnership (AMP-AD) Target Discovery Program and other National Institute on Aging (NIA)-supported programs to enable open-science practices and accelerate translational learning. The data, analyses and tools are shared early in the research cycle without a publication embargo on secondary use. Data are available for general research use according to the following requirements for data access and data attribution (https://adknowledgeportal.org/DataAccess/Instructions). Access to the content described in this Article is available online: the Synapse database for raw and processed snRNA-seq data (https://www.synapse.org/#!Synapse:syn31512863); the online portal for data browsing (https://github.com/naomihabiblab/BEYOND_DLPFC); the Synapse database for bulk RNA-seq dataset (https://www.synapse.org/#!Synapse:syn3388564); and the Synapse database for the spatial transcriptomics (Visium) dataset (https://www.synapse.org/#!Synapse:syn62110225). Other ROSMAP resources can be requested at the RADC Resource Sharing Hub (https://www.radc.rush.edu). Source data are provided with this paper.

Code availability

The complete code base used in this study is available at GitHub (https://github.com/naomihabiblab/BEYOND_DLPFC) and includes library-level snRNA-seq analysis to remove background RNA, quality control steps, cell type classification and doublet annotation; cell-type-level snRNA-seq analysis to remove low-quality data, subclustering analysis and functional annotations; BEYOND algorithmic approach of visualizing the cellular landscape, fitting subpopulation and trait dynamics, and defining cellular communities; additional analyses presented in this study, such as CelMod fitting and trait associations, causal modelling and validations (smFISH, pTau and morphology analysis, as well as spatial transcriptomics subpopulation co-localization analysis); code for generating the Figures and Extended Data Figures; a guide to reproduce the graphs presented in the study from the Supplementary Tables.

36. Bennett, D. A. et al. Religious Orders Study and Rush Memory and Aging Project. *J. Alzheimers Dis.* **64**, S161–S189 (2018).
37. Boyle, P. A. et al. Attributable risk of Alzheimer's dementia attributed to age-related neuropathologies. *Ann. Neurol.* **85**, 114–124 (2019).
38. Boyle, P. A. et al. To what degree is late life cognitive decline driven by age-related neuropathologies? *Brain* **144**, 2166–2175 (2021).
39. Yu, L. et al. Neuropathologic correlates of human cortical proteins in Alzheimer disease and related dementias. *Neurology* **98**, e1031–e1039 (2022).
40. Wilson, R. S. et al. Education and cognitive reserve in old age. *Neurology* **92**, e1041–e1050 (2019).
41. Wilson, R. S., Wang, T., Yu, L., Bennett, D. A. & Boyle, P. A. Normative cognitive decline in old age. *Ann. Neurol.* **87**, 816–829 (2020).
42. Grodstein, F., Wang, T., Leurgans, S. E., Wilson, R. S. & Bennett, D. A. Modifiable psychosocial risk factors and delayed onset of dementia in older populations: analysis of two prospective US cohorts. *BMJ Open* **12**, e059317 (2022).
43. De Jager, P. L. et al. A genome-wide scan for common variants affecting the rate of age-related cognitive decline. *Neurobiol. Aging* **33**, 1017.e1–1017.e15 (2012).
44. Schneider, J. A., Arvanitakis, Z., Bang, W. & Bennett, D. A. Mixed brain pathologies account for most dementia cases in community-dwelling older persons. *Neurology* **69**, 2197–2204 (2007).
45. Braak, H. & Braak, E. Neuropathological stageing of Alzheimer-related changes. *Acta Neuropathol.* **82**, 239–259 (1991).
46. Mirra, S. S. et al. The Consortium to Establish a Registry for Alzheimer's Disease (CERAD): part II. Standardization of the neuropathologic assessment of Alzheimer's disease. *Neurology* **41**, 479–479 (1991).
47. Zheng, G. X. Y. et al. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).
48. 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).
49. Hao, Y. et al. Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–3587 (2021).
50. Simon, N., Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for Cox's proportional hazards model via coordinate descent. *J. Stat. Softw.* **39**, 1–13 (2011).
51. Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33**, 1–22 (2010).

52. McGinnis, C. S., Murrow, L. M. & Gartner, Z. J. DoubletFinder: doublet detection in single-cell RNA sequencing data using artificial nearest neighbors. *Cell Syst.* **8**, 329–337 (2019).
53. Aran, D. et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat. Immunol.* **20**, 163–172 (2019).
54. Wu, T. et al. clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. *Innovation* **2**, 100141 (2021).
55. Gu, Z. & Hübschmann, D. SimplifyEnrichment: a Bioconductor package for clustering and visualizing functional enrichment results. *Genom. Proteom. Bioinform.* **21**, 190–202 (2023).
56. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
57. Lee, A. J. et al. Multi-region brain transcriptomes uncover two subtypes of aging individuals with differences in Alzheimer risk and the impact of APOEe4. Preprint at *bioRxiv* https://doi.org/10.1101/2023.01.25.524961 (2023).
58. Lawlor, D. A., Harbord, R. M., Sterne, J. A. C., Timpson, N. & Davey Smith, G. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Stat. Med.* **27**, 1133–1163 (2008).
59. Tingley, D., Yamamoto, T., Hirose, K., Keele, L. & Imai, K. mediation: R package for causal mediation analysis. *J. Stat. Soft.* **59**, 1–38 (2014).
60. Rosseel, Y. lavaan: An R package for structural equation modeling. *J. Stat. Softw.* **48**, 1–36 (2012).
61. Virshup, I., Rybakov, S., Theis, F. J., Angerer, P. & Wolf, F. A. anndata: annotated data. Preprint at *bioRxiv* https://doi.org/10.1101/2021.12.16.473007 (2021).
62. McInnes, L. et al. UMAP: uniform manifold approximation and projection. *JOSS* **3**, 861 (2018).
63. van der Maaten, L. & Hinton, G. Visualizing data using *t*-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
64. Wood, S. N. Stable and efficient multiple smoothing parameter estimation for generalized additive models. *J. Am. Stat. Assoc.* **467**, 673-686 (2004).
65. Maynard, K. R. et al. Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nat. Neurosci.* **24**, 425–436 (2021).
66. Korsunsky, I. et al. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).
67. Zhao, E. et al. Spatial transcriptomics at subspot resolution with BayesSpace. *Nat. Biotechnol.* **39**, 1375–1384 (2021).
68. van Dijk, D. et al. Recovering gene interactions from single-cell data using data diffusion. *Cell* **174**, 716–729 (2018).

Acknowledgements We thank the individuals who donated their brain to research through the Rush University Alzheimer's Disease Center. The work was supported by NIH RF1 AG057473 (P.L.D.J. and D.A.B.), U01 AG061356 (P.L.D.J. and D.A.B.), U01 AG046152 (P.L.D.J. and D.A.B.), R01 AG070438 (P.L.D.J.), R01 AG015819 (D.A.B.), U01 AG072572 (P.L.D.J.), R01AG066831 (to V.M.); NIH K23 AG062750 (to H.-S.Y.); NIH NIA K23AG062750 (H.-S.Y.); NIH NIA R01AG080667 (H.-S.Y.); CS-02018-191971; the Israel Science Foundation (ISF) research grant no. 1709/19, the European Research Council grant 853409, the MOST-IL-China research grant no. 3-15687 and the Myers Foundation (N.H.); An Alzheimer's Association Grant through the AD Strategic Fund (ADSF-21-816675) (M.T. and H.-U.K.). A.K.S. was supported by a Minerva Fellowship of the Minerva Stiftung Gesellschaft fuer die Forschung mbH. Diagrams in Figs. 1, 3 and 5 and Extended Data Fig. 6 were created using BioRender.

Author contributions P.L.D.J. and N.H. designed the study. C.M. prepared the single-nucleus libraries and performed sequencing. M.F. performed the sequence alignment and demultiplexing analysis. G.S.G. performed the computational and statistical analyses, with guidance of N.H., H.-S.Y., V.M., P.L.D.J. and with the help of A.C., A.K.S., C.C.W. and L.Z.; H.-S.Y. performed the causality modelling analysis. M.T. performed smFISH, immunohistochemistry experiments and image analysis. M.T., A.S. and H.-U.K. generated the ST data. N.C.-L., H.-U.K. and G.S.G. analysed the ST data. Y.W. analysed the bulk RNA-seq data. G.S.G., H.-S.Y., V.M., N.H. and P.L.D.J. wrote the manuscript, and all of the co-authors provided critical comments. D.A.B. is principal investigator of the parent ROS and MAP studies and obtained funding and performed study supervision. P.L.D.J., A.R. and D.A.B. obtained funding for the project.

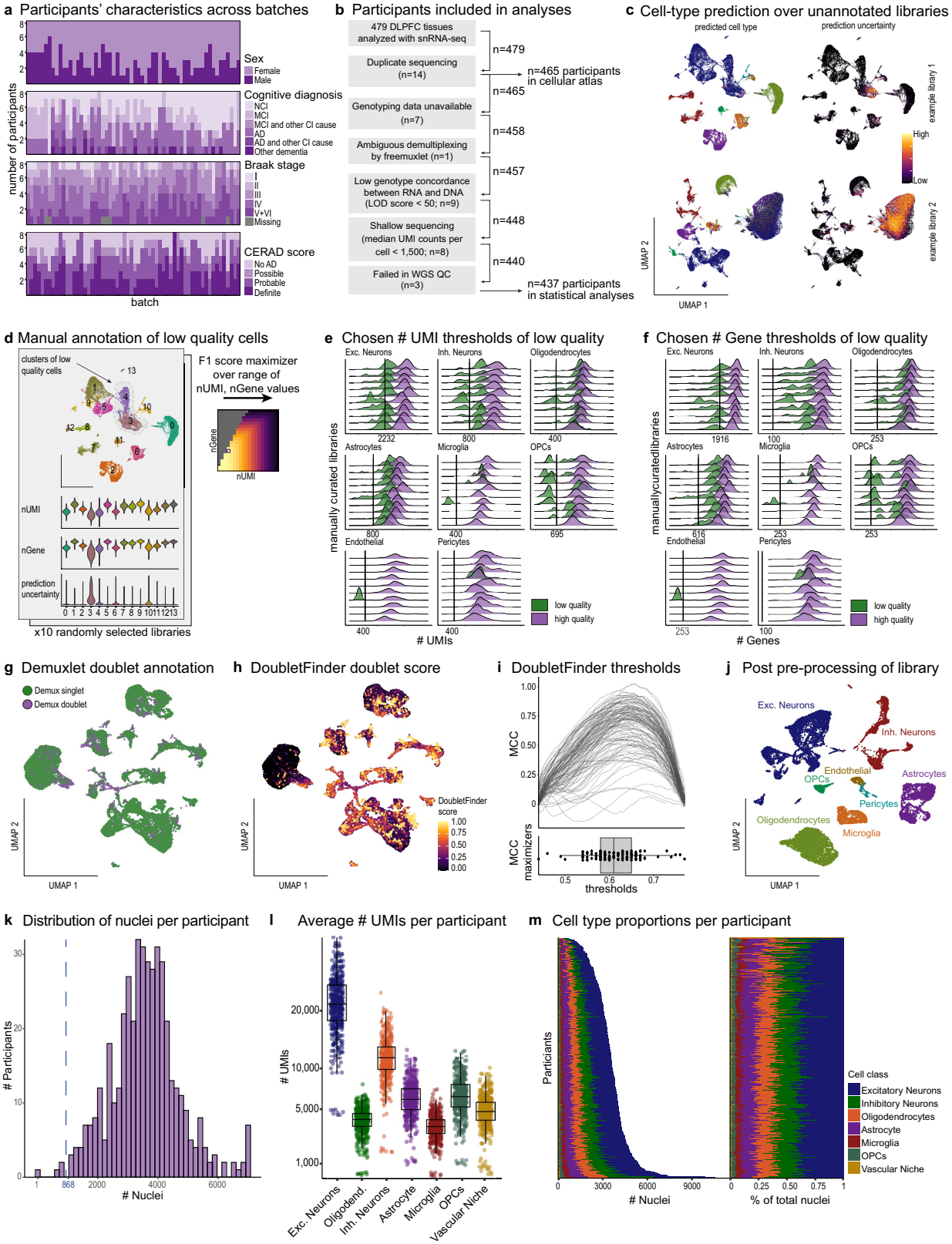
Competing interests A.R. is a founder and equity holder of Celsius Therapeutics, an equity holder in Immunitas Therapeutics and, until 31 August 2020, was an scientific advisory board member of Syros Pharmaceuticals, Neogene Therapeutics, Asimov and Thermo Fisher Scientific. From 1 August 2020, A.R. is an employee of Genentech, a member of the Roche Group. The other authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-024-07871-6.

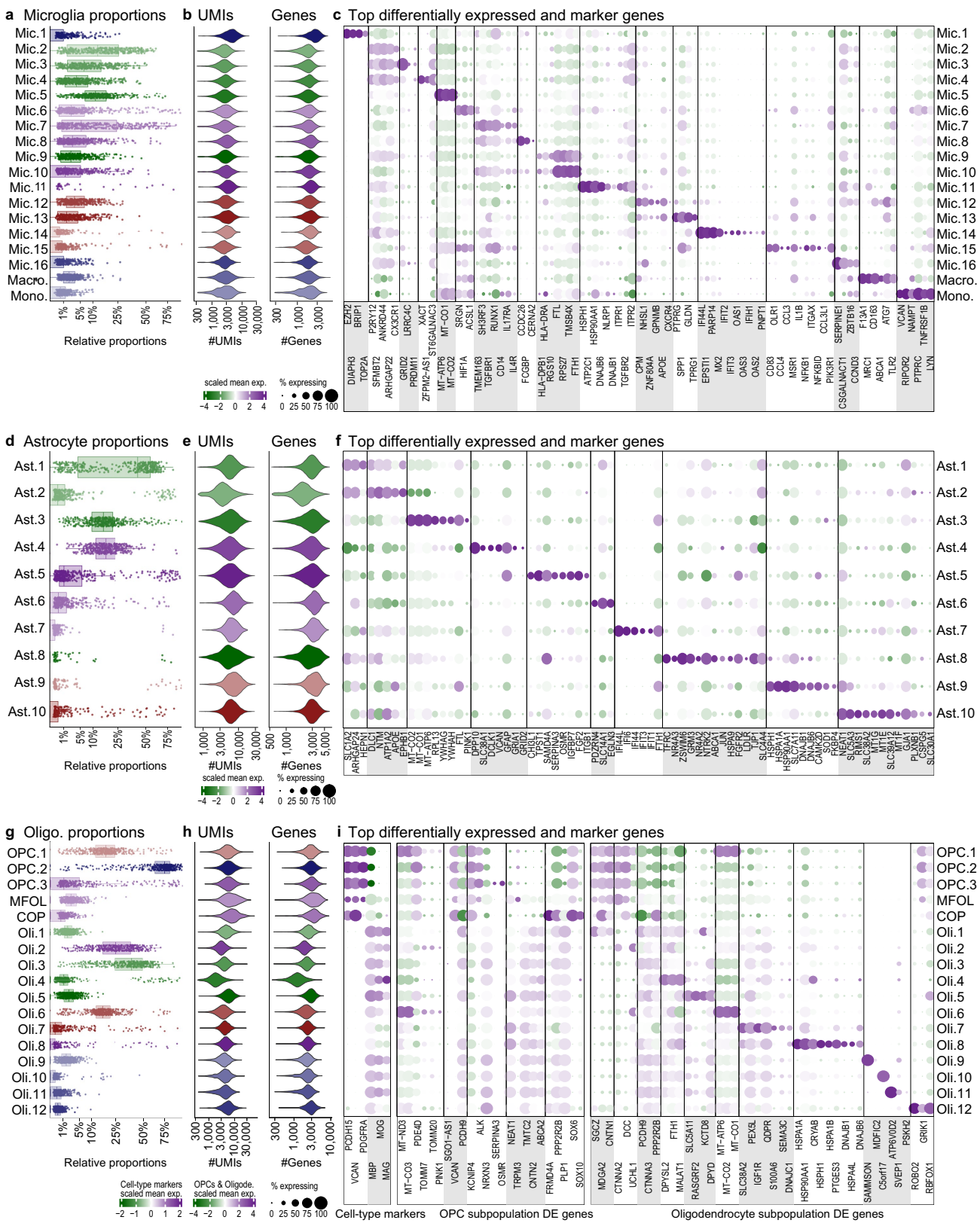
Correspondence and requests for materials should be addressed to Vilas Menon, Naomi Habib or Philip L. De Jager.

Peer review information *Nature* thanks the anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

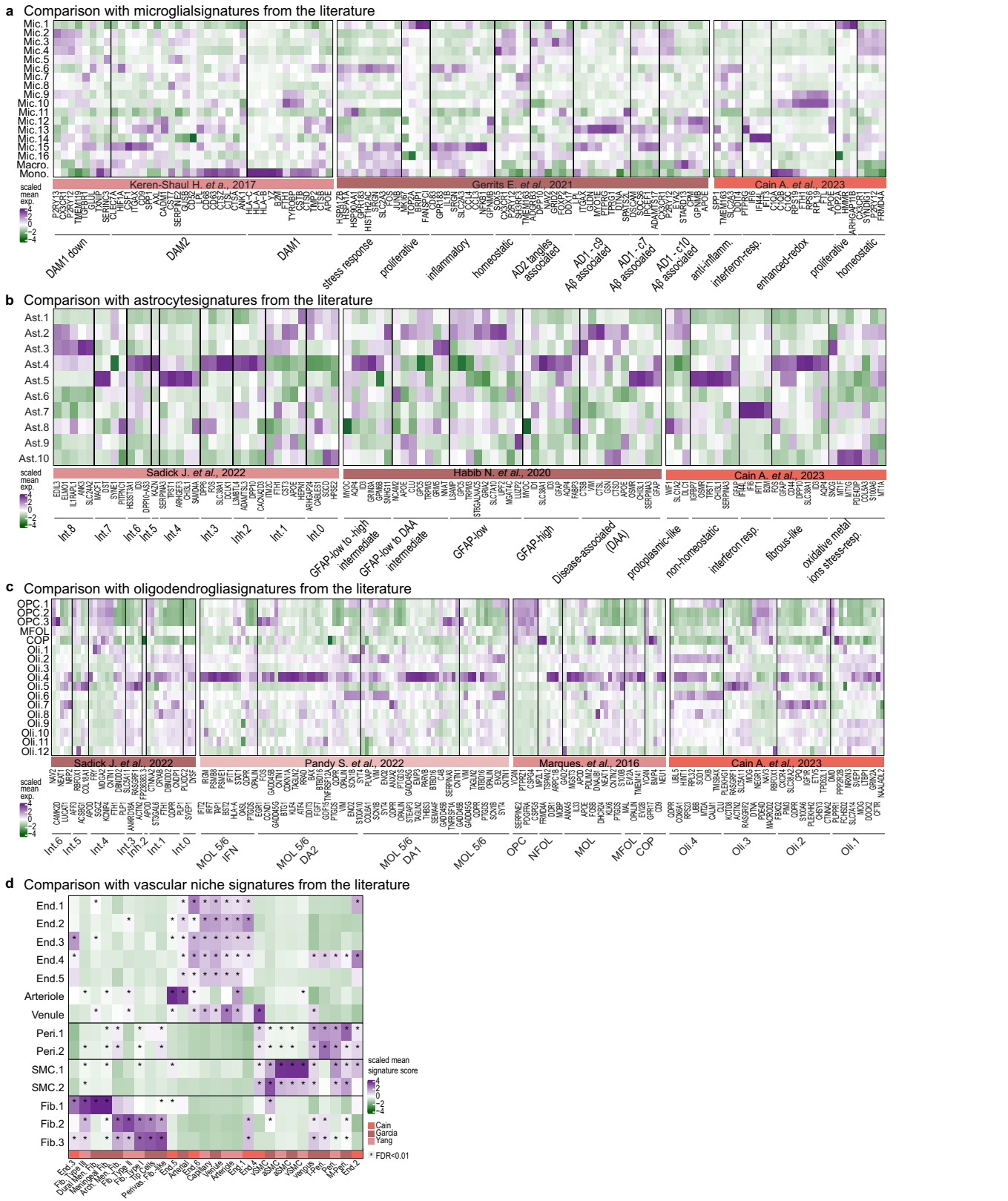


Extended Data Fig. 1 | See next page for caption.

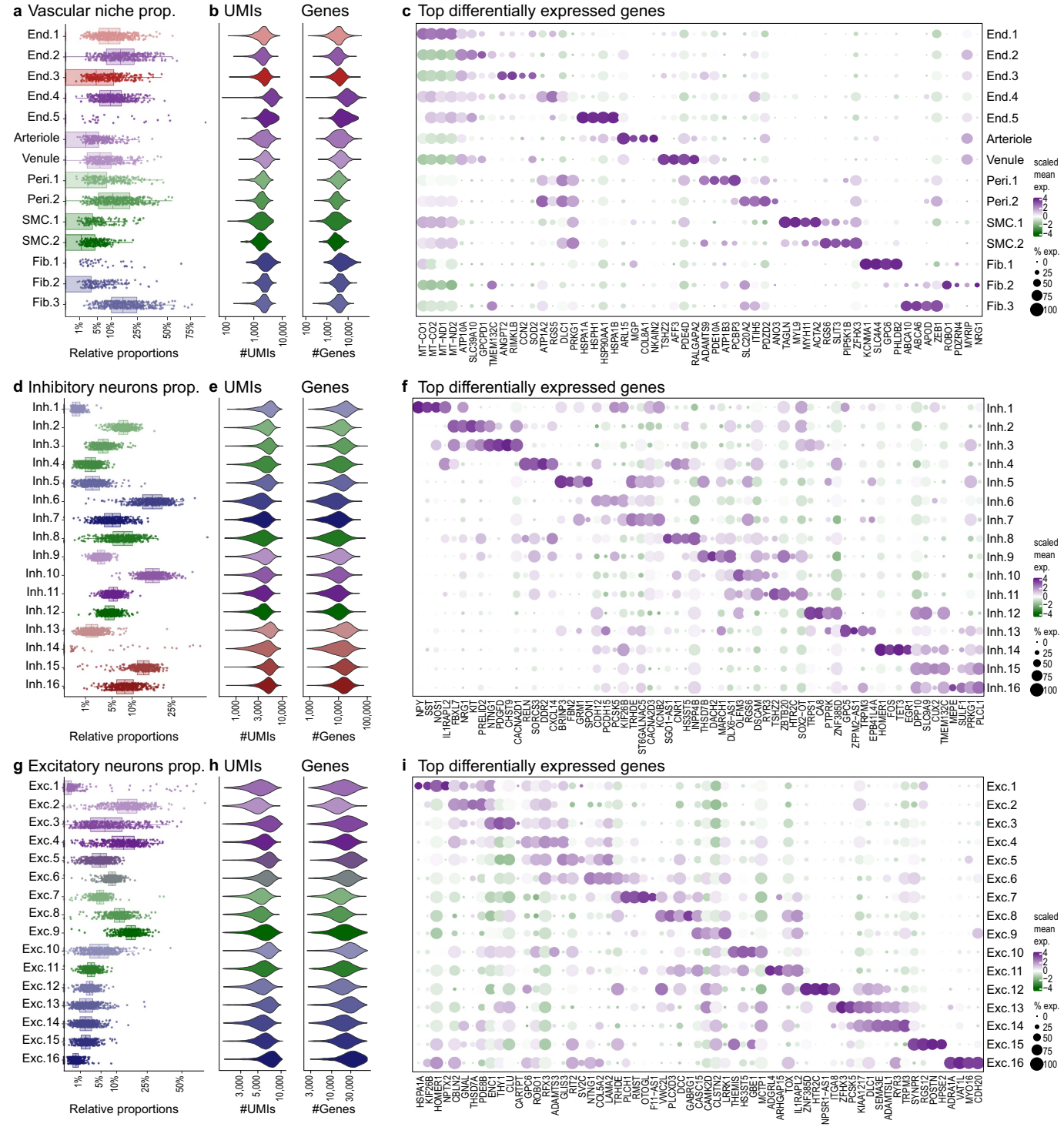
Extended Data Fig. 1 | snRNA-seq libraries preprocessing and quality controls (QCs). (**a**) Distributions across batches of: sex, cognitive diagnosis, Braak stage and CERAD score. (**b**) QC pipeline selecting 465 participants for cell atlas and 437 participants for downstream analysis including snRNA-seq libraries with a sufficient number of robustly assigned nuclei. (**c**) UMAP embeddings of two example snRNA-seq libraries prior to any QCs. Dot colour: predicted cell type (left) and prediction uncertainty (Shannon entropy, right). (**d**) Example snRNA-seq library manually curated for cells of low-quality libraries. Based on the manual curation of 10 such libraries, cell-type-specific low-quality thresholds over the number of UMIs (#UMI) and the number of unique genes (#Genes) were chosen. (**e**,**f**) Distributions of (**e**) # UMI and (**f**) #Genes threshold for low quality nuclei per cell type in the manually curated libraries, indicating selected thresholds. (**g-i**) Detection of doublet cells. UMAP embedding of example library annotated by (**g**) demultiplexing doublet annotation (demuxlet algorithm), or (**h**) DoubletFinder doublet-likelihood scores. (**i**) Distribution of the Matthews correlation coefficient (MCC) scores, reflecting prediction sensitivity, specificity and precision, for a range of thresholds over the DoubletFinder scores, per library (separate line, top), and the maximizing threshold chosen per library (dots, bottom). (**j**) UMAP embedding of the example library (in c) post-QCs, coloured by cell-type prediction. (**k**) Distribution of number of nuclei per participant. Dash line = minimum number of nuclei for a participant in the *Discovery sample*. (**l**) The average number of UMIs per cell type in each participant; Dots: individual participants (n = 465 participants per cell type). (**m**) Number of nuclei per cell type and participant: absolute (left) and proportions (right). Coloured by cell classes. Box: 1st and 3rd quartile, line: median, whiskers extend from box to the highest and lowest values within 1.5 times the distance between the quartiles.



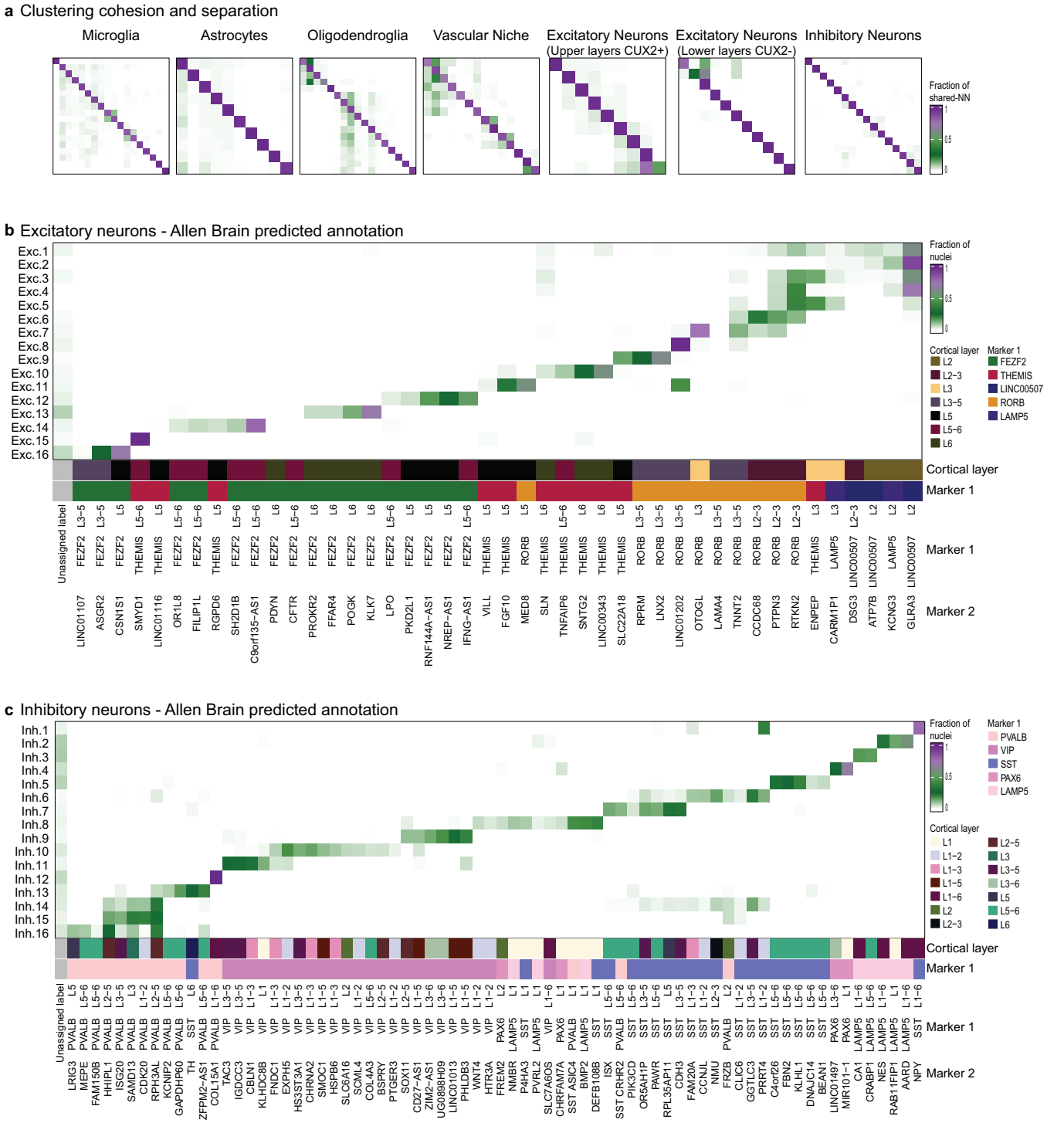
Extended Data Fig. 2 | Glial subpopulation diversity. (**a**,**d**,**g**) Subpopulation proportions across participants for (**a**) microglia, (**d**) astrocytes and (**g**) oligodendroglia cells. Dots: 465 individual participants. Box: 1st and 3rd quartiles; line: median, whiskers extend to the highest and lowest values within 1.5 times the distance between the quartiles. (**b**,**e**,**h**) QC measures. Distributions of number of UMIs and number of Genes detected for each subpopulation of (**b**) microglia, (**e**) astrocytes and (**h**) oligodendroglia. (**c**,**f**,**i**) Selected markers and top differentially expressed genes between subpopulations. Gene expression (columns) across subpopulations (rows) of (**c**) microglia, (**f**) astrocytes and (**i**) oligodendroglia. Dot colour: mean expression in expressing cells. Dot size: percent of cells expressing the gene.



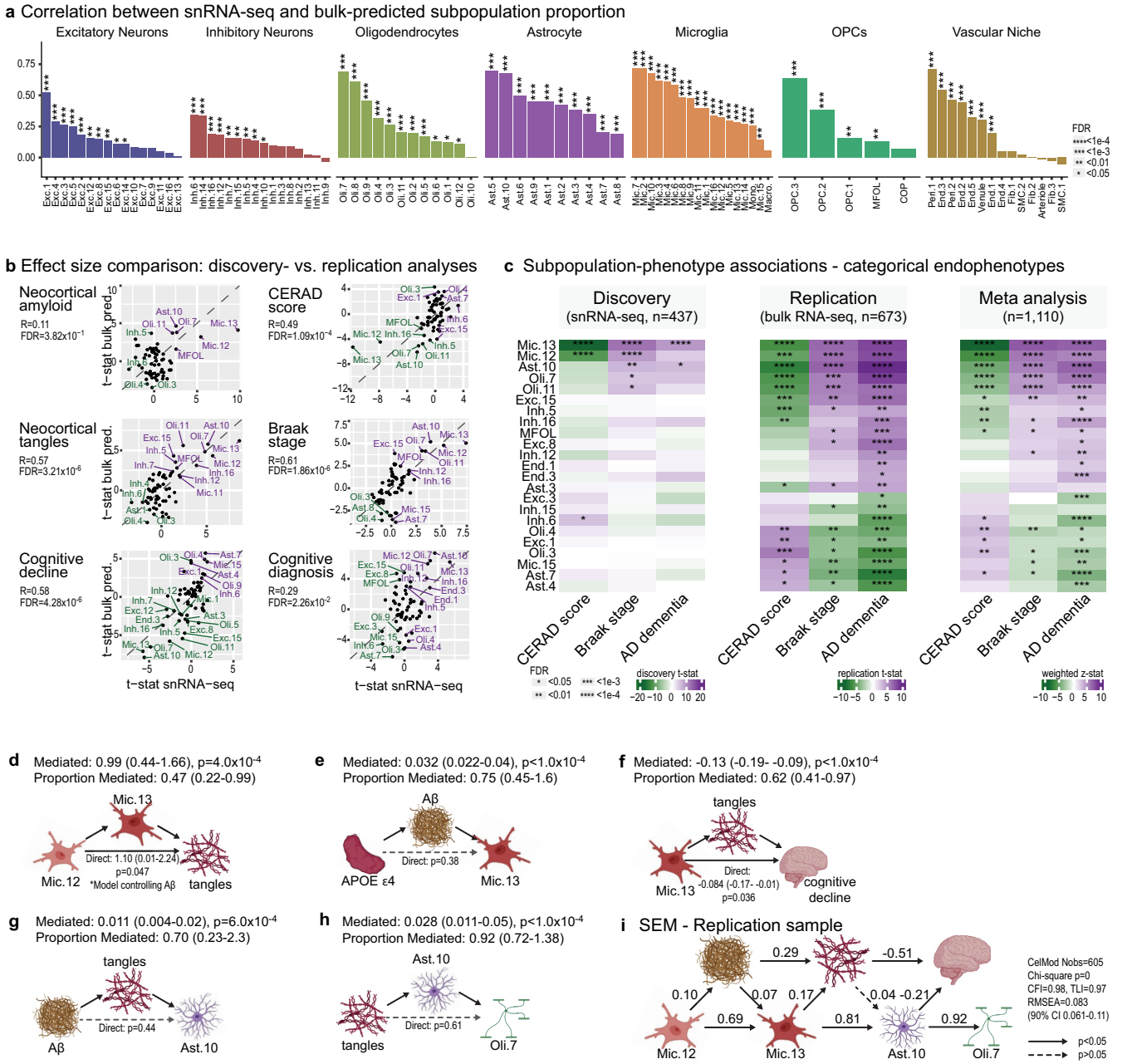
Extended Data Fig. 3 | Comparison of glial subpopulation to previously defined gene signatures. Heatmap of previous signature genes (rows) for cell subpopulations (columns) of (**a**) microglia, (**b**) astrocytes and (**c**) oligodendroglia, separated by signature and split by reference source. Colour-scale: row scaled expression out of expressing cells. Genes defined in multiple signatures appear multiple times. (**d**) Comparison of previous gene signatures to subpopulations of the vascular niche. Scaled mean signature score of published gene signatures (columns) within each subpopulation (rows). (*) = Significantly enriched signatures (U-test, FDR<0.01). Published signatures from[1,3,17].



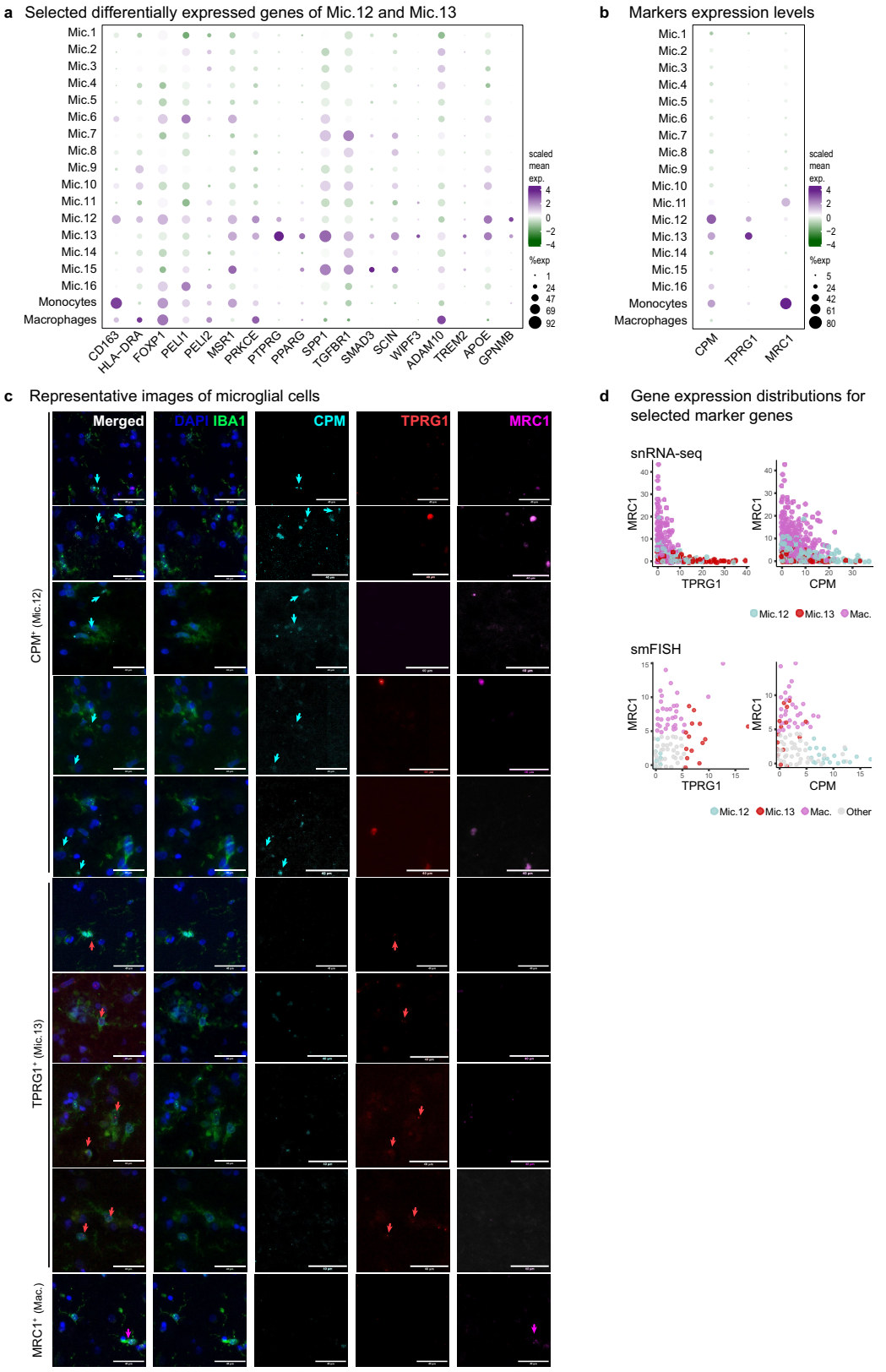
Extended Data Fig. 4 | Vascular niche- and neuronal subpopulation diversity. (**a**,**d**,**g**) Subpopulation proportions across participants for (**a**) vascular niche cells, (**d**) inhibitory neurons and (**g**) excitatory neurons. Dots: 465 individual participants, box: 1st and 3rd quartiles, line: median, whiskers extend to the highest and lowest values within 1.5 times the distance between the quartiles. (**b**,**e**,**h**) QC measures. Distributions of number of UMIs and number of Genes detected for each subpopulation of (**b**) vascular niche, (**e**) inhibitory neurons and (**h**) excitatory neurons. (**c**,**f**,**i**) Selected markers and top differentially expressed genes between subpopulations. Gene expression (columns) across subpopulations (rows) of (**c**) vascular niche, (**f**) inhibitory neurons and (**i**) excitatory neurons. Dot colour: mean expression in expressing cells. Dot size: percent of cells expressing the gene.



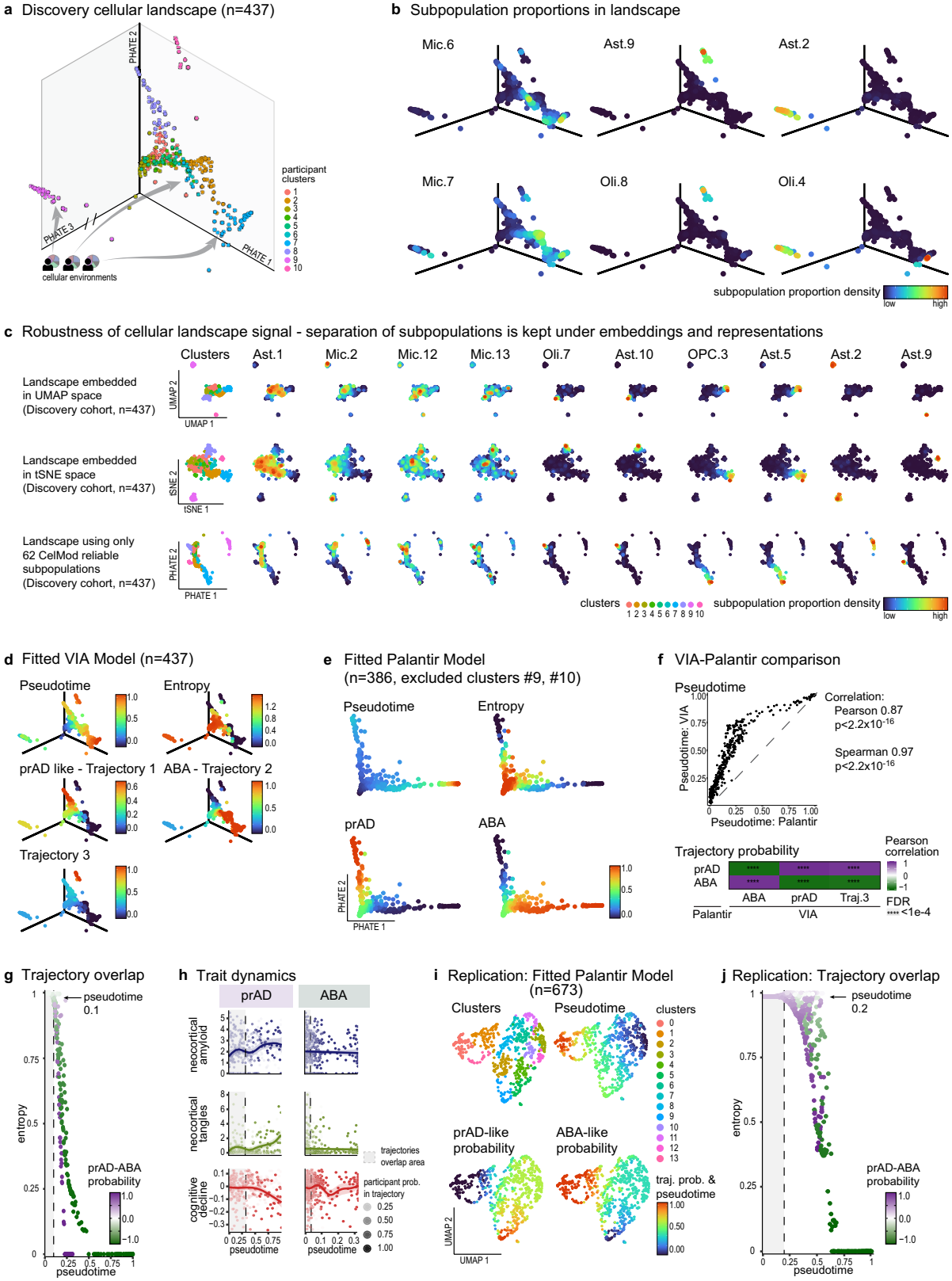
Extended Data Fig. 5 | Clustering quality and neuronal subtypes annotations. (**a**) Quality measures of clustering showing the cohesion within each cluster and separation between clusters. For each cluster (row) showing the distribution of shared nearest neighbours assigned to other clusters (columns) For each cell class. Coloured by the fraction of shared nearest neighbours. Rows and columns are sorted by the cluster number. (**b**-**c**) Classification of (**b**) excitatory and (**c**) inhibitory neuronal cells, by previous annotations provided by the Allen Brain Map. Heatmaps of the percentage of cells from each of our clusters assigned to each of the provided annotations (**Methods**) linking neuronal sub-types to marker genes of inhibitory and excitatory neurons and to cortical layers (colour bar).



Extended Data Fig. 6 | CelMod predictions and endophenotype associations. (**a**) Evaluation of CelMod prediction of subpopulation proportions in bulk RNA samples with matching snRNA-seq measurements in the same participant (n = 419 samples). Spearman correlation between snRNA-seq (actual) and CelMod bulk-predicted proportions, over the held-out set (test set) of participants (Methods). * = FDR corrected p-value. (**b**) Comparison of the estimated effect sizes regressing endophenotypes on subpopulation proportions, for the snRNA-seq (Discovery cohort, x-axis) and the bulk predictions (Replication cohort, y-axis). n = 419. The Spearman correlation between the effect sizes and FDR corrected p-value are shown for each comparison. (**c**) Associating subpopulation proportions to endophenotypes: CERAD score, Braak stage and AD dementia (linear regression controlled for cofounders, FDR<0.05, Methods), showing subpopulations significantly associated with at least one of the tested traits in one on the cohorts: *Discovery* (left, n = 437), *Replication* (centre, n = 673) and the meta-analysis of both cohorts (right, n = 1,110). Colour scale: t-stat. (**d-h**) Causal mediation models which together with Fig. 3f–i position Mic.12, Mic.13, Ast.10 and Oli.7 within the Aβ→tau→cognitive decline AD cascade, indicating direct and mediated effects, as well as proportion of effect mediated. Number of participants: (**d**,**e**,**g**) n = 432, (**f**) n = 413, (**h**) n = 433. (**i**) Validation of the structural equation model (SEM; as in Fig. 3j) in the CelMod predicted subpopulation proportions of the Replication cohort. Arrows show association directionality and relative strength.

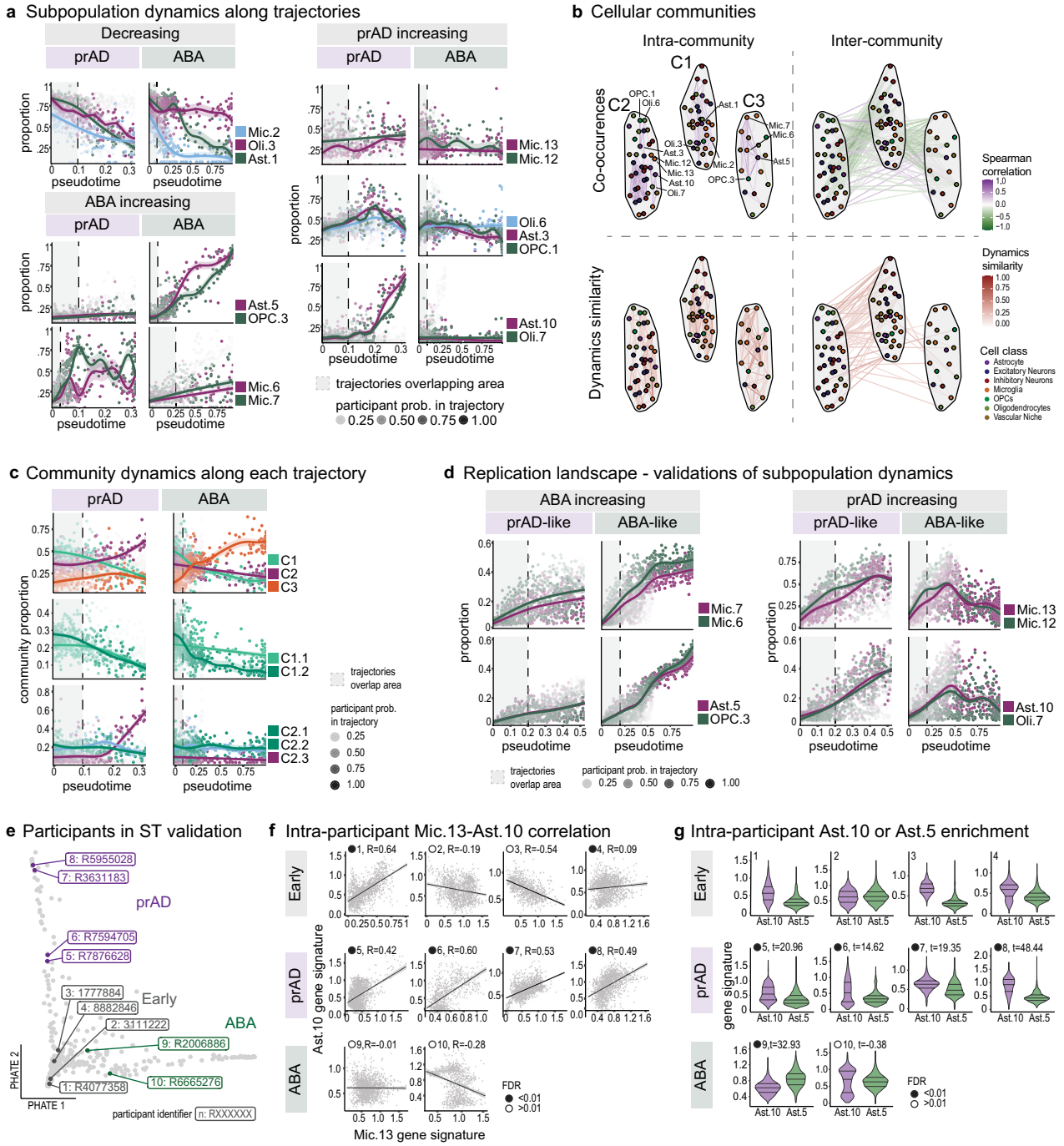


Extended Data Fig. 7 | smFISH quantification analysis. (**a**) Gene expression (columns) across microglial subpopulations (rows) for selected differential genes and known markers. Dot colour: column-scaled mean expression of expressing cells. Dot size: percentage of expressing cells. (**b**) Gene expression of markers used for smFISH across microglial subpopulations. Dot colour and size as in (a). (**c**) A gallery of representative RNAscope smFISH data showing split channels by marker: *CPM*^{high} Mic.12, *TPRG1*^{high} Mic.13, and a MRC1^{high} macrophage, together with IBA1 (green) and DAPI (blue) staining. (**d**) Bivariate expression distributions in snRNA-seq (top) and smFISH (bottom), coloured by assigned snRNA-seq subpopulations (top) predicted assignment in smFISH data (bottom, Methods).



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

Extended Data Fig. 8 | Robustness of cellular landscape modelling by BEYOND. (**a**) 3D PHATE embedding of all 437 snRNA-seq participants, coloured by clustering of participants based on their cellular environments (Methods). (**b**) Distinct patterns of subpopulations along the cellular landscape manifold, showing additional subpopulations to those of Fig. 5c. Participants (dots) are coloured by the locally smoothed proportion of each subpopulation. (**c**) Robustness of the cellular landscape to the embedding method and set of subpopulations used in BEYOND. Participants are coloured by the locally smoothed subpopulation proportion. (**d-e**) Visualizing fitted pseudotime, trajectories and Shannon entropy of trajectory probabilities outputted by: (**d**) VIA (n = 437 participants), and (**e**) Palantir algorithms (n = 386, excluding participant-clusters #9 and #10 in a).(**f**) Robustness of trajectories and pseudotime predictions using different algorithms (over the overlapping n = 386 participants). (Top) Pseudotime assigned for each individual by Palantir compared to VIA. (Bottom) Trajectory probabilities Pearson correlations. Corrected for multiple hypothesis testing (BH). (**g**) Participants' trajectory probabilities entropy drop along pseudotime, in the Palantir model. Dots are coloured by prAD minus ABA trajectory probabilities. The grey area indicates a pseudotime range (0, 0.11) in which the two trajectories are not well separated. (**h**) Trait-dynamics of AD-related traits along the pseudotime in each of the inferred trajectories, showing the datapoints used in fitted curves and error bands showing 0.95 CI (Methods). n = 386 participants. As in Fig. 5f. (**i-j**) Validation of cellular landscapes and trajectories using the *Replication cohort* (n = 673 non-overlapping participants, with the 62 reliable CelMod bulk-predicted subpopulations proportions to represent cellular environments). (**i**) The Palantir model over replication landscape as in (e). (**j**) Trajectory probabilities entropy drop as in (g) but over the replication landscape.



Extended Data Fig. 9 | Cellular dynamics and communities. (**a**) Distinct patterns of subpopulation dynamics, as Fig. 6a with datapoints used to fit the curves. Dots are participants (n = 386). Error bands: 0.95 CI. (**b**) Graph of multi-cellular communities. Nodes: Subpopulations. Edges: co-occurrences (Spearman correlation, green-purple scale) or dynamics similarity (white-red scale). Excluding edges of low similarity (correlation (−0.2, +0.4), dynamics (−0.2, +0.2)) for visualization. See Fig. 6b. (**c**) Distinct dynamic patterns for cellular communities across trajectories. Whole-community dynamics of community proportion along pseudotime in each trajectories. Dots are individual participants (n = 386). See Fig. 6c. Error bands: 0.95 CI. (**d**) Validation in the Replication cohort of coordinated subpopulation dynamics along trajectories. Presented as in Fig. 6a, but over the replication landscape. Error bands: 0.95 CI. (**e-g**) Spatial transcriptomics (ST, Visum, Methods) validations of cellular communities. Dots are individual participants (n = 637). (**e**) Discovery cohort participants included in ST validations (n = 10), and their assignment to trajectories (annotated as: Early, prAD or ABA). 2D PHATE embedding as in Fig. 5b. (**f**) Pearson correlation between Mic.13 and Ast.10 gene signature expression across the Visium spots for each of the participants, grouped by their trajectory assignment. Dots: signature expression in visium spots. Line: regression line. Error bands: 0.95 CI of regression line. P-value: one-sided T-test (positive association), FDR corrected for multiple hypothesis, dot fill colour indicates significance of correlation. (**g**) Divergent association of Ast.10 and Ast.5 between the two trajectories. Distribution of Ast.10 and Ast.5 gene signature expression for each participant (n = 10), grouped by their trajectory assignment. Per participant, significance of the differences of means were tested by one-sided t-test: (1) for prAD participants: Ast.10 levels being strictly higher than Ast.5 levels, and (2) for ABA participants: Ast.5 levels being significantly higher than Ast.10 levels (Methods). Significance level is shown by the dot fill colour.

nature portfolio

Corresponding author(s): Philip De Jager, Naomi Habib

Last updated by author(s): 20.6.2024

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.

Please do not complete any field with "not applicable" or n/a. Refer to the help text for what text to use if an item is not relevant to your study.
For final submission: please carefully check your responses for accuracy; you will not be able to make changes later.

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	Immunofluorescence images were captured with Nikon Eclipse Ni-E. cDNA libraries were assesed and sequenced with the provided software of Agilent BioAnalyzer, 10X Genomics (chemistry V3, Visium), Illumina HiSeqX and Illumina NovaSeq 6000.
Data analysis	General: R (v4.2.3) with publicly available packages, python (v3.9.16) snRNA-seq analysis: CellRanger (v6.0.0.0), CellBender (v0.2.0, download date 01-March-2021); Demuxlet (v0.1-beta), Freemuxlet (v0.1-beta), Seurat (v4.1.0), DoubletFinder (https://github.com/GreenGilad/DoubletFinder/tree/scalable_DoubletFinder), BEYOND analysis: scanpy (v1.9.3), anndata (python, v0.9.2), AnnData (r, v0.7.5.6), Palantir (v1.3.0), VIA (pyvia, v0.1.88, in an independent conda environment with python v3.7) RNAscope image analysis: Images were analyzed using our automated image processing pipeline, which is based on CellProfiler and CellProfiler Analyst, and includes algorithms for cell identification and segmentation, intensity measurement and morphologic feature extraction. Spatial transcriptomics analysis: Seurat (v4.1.4), Rmagic (v2.0.3), SpaceRanger (v1) Other analyses: CelMod (v0.0.0.9), mgcv (v1.8.42), clusterProfiler (v4.6.2) Visualization: ggplot2 (v3.4.2), ComplexHeatmap (v2.14.0), ggnewscale (v0.4.9), gg3D (v0.0.0.9), Nebulosa (v1.14.0) GITHUB for all code: https://github.com/naomihabiblab/BEYOND_DLPFC

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

Synapse database for snRNA-seq data (raw sequenced libraries, processed cell-type Seurat objects, cell-subpopulation mapping) https://www.synapse.org/#!Synapse:syn31512863.
Synapse database for bulk RNA-seq dataset https://www.synapse.org/#!Synapse:syn3388564.
Other ROSMAP resources can be requested at the RADC Resource Sharing Hub at: https://www.radc.rush.edu.

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	We report the sex of each participant as detailed in Supplementary Table 1. Participants' sex was considered in analysis - controlled for when evaluating association of Alzheimer's Disease endophenotypes.
Population characteristics	For this study, we selected 465 participants, blind to their neuropathologic and clinical traits, age and sex, and based on availability of frozen pathologic material from the Dorsolateral Prefrontal Cortex (DLPFC), including only participants with RIN>5 and post mortem interval (PMI) <24 hours. Our study cohort includes diverse individuals across the full range of the pathological and clinical stages of AD. The demographic and clinicopathologic characteristics are described in Supplementary Table 1.
Recruitment	Data were derived from subjects enrolled in one of two longitudinal clinical-pathologic cohort studies of aging and dementia, the Religious Orders Study (ROS) and the Rush Memory and Aging Project (MAP), collectively referred to as ROSMAP. All participants are without know dementia at enrollment and have annual clinical evaluations and agree in advance to brain donation at death. All participants signed an informed consent, Anatomic Gift Act, and repository consent. Selection of ROSMAP participants for this study was done blind to any clinical-pathologic characterization, in an effort to obtain a representative sample of the aged human diversity (as represented within the ROSMAP studies)
Ethics oversight	Each study was approved by an Institutional Review Board of Rush University Medical Center.

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	Single nucleus RNA-seq sample size of 465 ROSMAP participants (Discovery sample) to capture a wide range of clinical-pathological diversity and enable sufficient statistical power for associations with disease traits. Bulk RNA-seq sample size of 1,092 ROSMAP participants was based on available data. Immunohistochemistry/smFISH sample size 15 NY brain bank samples (6 no-AD controls, 1 MCI, 8 AD). Spatial transcriptomics (visium) - 10 samples out of the Discovery sample.
Data exclusions	We excluded 28 of the 465 participants from statistical and association analyses since they had missing genome sequencing data for the assignment of nuclei or had insufficient number of assigned nuclei.
Replication	Each sample in our Discovery cohort was profiled by two nuclei libraries (technical replicates) that were combined for the analysis. Our findings were replicated in an independent set of 673 participants (biological replicates) profiled by bulk RNA-seq (not included in our Discovery cohort profiled by snRNA-seq). These additional samples were used to validate the association of cell subpopulations with disease traits, mediation analysis and causality modeling, as well as the BEYOND landscape and cellular trajectories. We reported the results that were successfully replicated in both the discovery snRNA-seq cohort and the replication cohort. The immunohistochemistry/smFISH was conducted on independent 15 samples from a different brain bank (not profiled by snRNA or bulk RNA).
Randomization	Each batch of samples for library construction consisted of 8 participants, that were randomly selected from the cohort. The batches were

Randomization	balanced for clinical and pathological diagnosis as well as participant sex, and did not have significant technical biases, as seen in Extended Data Fig. 1a.
Blinding	Investigators were blind to any clinical- pathological- and demographical information of the participants during the snRNA-seq, spatial-transcriptomics and immunohistochemistry experimental phase. The data processing and basic atlas analysis were done blind to all participant information and pathologies, including: snRNA-seq quality control steps and sub-clustering, immunohistochemistry quantification of marker genes and anti-phospho Tau antibody AT8, as well as cell subpopulation quantification on spatial transcriptomics data.

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

Methods

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

Antibodies

Antibodies used	Iba1 antibody (Wako, catalog # 011-27991, lot #LEG4278). Iba1 was used at 1/100 dilution. RNAscope probes: Hs-TPRG1-C1 (ACD, catalog # 1047171-C1, Lot #210548), Hs-CPM-C2 (ACD, catalog # 444811-C2, Lot#22108A), Hs-MRC1-C3 (ACD, catalog # 583921-C3, Lot#22108A). pTau AT8 primary antibody (Invitrogen, catalog MN1020). pTau AT8 at 1/300 dilution. The secondary antibodies used for Iba1 and pTau were Alexa Fluor 488 Donkey anti-Goat IgG(H+L) (Dilution 1/500, Ref: A11055, Lot 2747580) and Alexa Fluor 647 Donkey anti-mouse IgG(H+L) (Dilution 1/500, Ref: A31571, Lot 2555690) from Invitrogen.
Validation	Iba1 and pTau AT8 antibody specificity has been confirmed on human post-mortem brain tissues by co-staining with other microglia markers and on healthy brain tissue. Furthermore, Iba1 antibody was cited in more than 1250 papers in 2021. The specificity of the RNAscope probes was validated by ACD. pTau AT8 (Invitrogen, MN1020) has been cited by 905 papers in total, including 469 papers for immunohistochemistry. The purity of the antibody is >95% as determined by SDS-PAGE, and its specificity was also validated in our laboratory by co-staining with Thioflavin S and other anti-pTau antibodies such as the PHF1 antibody. The antibody stained clear neurofibrillary tangles as well as neuropil threads. The dilution of this antibody has been determined by serial dilution ranging from 1/50 to 1/1000. The specificity of the secondary antibodies has been determined by staining secondary antibodies alone, and we did not detect any signals.