Deciphering cellular transcriptional alterations in Alzheimer's disease brains

Xue Wang[1*], Mariet Allen[2], Shaoyu Li[3], Zachary S. Quicksall[1], Tulsi A. Patel[2], Troy P. Carnwath[2], Joseph S. Reddy[1], Minerva M. Carrasquillo[2], Sarah J. Lincoln[2], Thuy T. Nguyen[2], Kimberly G. Malphrus[2], Dennis W. Dickson[2], Julia E. Crook[1], Yan W. Asmann[1], Nilüfer Ertekin-Taner[2,4*]

[1] Department of Health Sciences Research, Mayo Clinic Florida, Jacksonville, FL, USA.

[2] Department of Neuroscience, Mayo Clinic Florida, Jacksonville, FL, USA.

[3] Department of Mathematics and Statistics, University of North Carolina at Charlotte, Charlotte, NC, USA.

[4] Department of Neurology, Mayo Clinic Florida, Jacksonville, FL, USA.

* Corresponding Authors

E-mails: Wang.Xue@mayo.edu (XW) and Taner.Nilufer@mayo.edu (NET)

Author e-mails:

Xue Wang, wang.xue@mayo.edu

Mariet Allen, Allen.Mariet@mayo.edu

Shaoyu Li, Shaoyu.Li@uncc.edu

Zachary S. Quicksall, Quicksall.Zachary@mayo.edu

Tulsi A. Patel, Patel.Tulsi@mayo.edu

Troy P. Carnwath, Carnwath.Troy@mayo.edu

Joseph S. Reddy, Reddy.Joseph@mayo.edu

Minerva M. Carrasquillo, Carrasquillo.Minerva@mayo.edu

Sarah J. Lincoln, Lincoln.Sarah@mayo.edu

Thuy T. Nguyen, Nguyen.Thuy2@mayo.edu

Kimberly G. Malphrus, Malphrus.Kimberly@mayo.edu

Dennis W. Dickson, Dickson.Dennis@mayo.edu

Julia E. Crook, Crook.Julia@mayo.edu

Yan W. Asmann, Asmann.Yan@mayo.edu

Nilüfer Ertekin-Taner, taner.nilufer@mayo.edu

Abstract

Large-scale brain bulk-RNAseq studies identified molecular pathways implicated in Alzheimer's disease (AD), however these findings can be confounded by cellular composition changes in bulk-tissue. To identify cell intrinsic gene expression alterations of individual cell types, we designed a bioinformatics pipeline and analyzed three AD and control bulk-RNAseq datasets of temporal and dorsolateral prefrontal cortex from 685 brain samples. We detected cell-proportion changes in AD brains that are robustly replicable across the three independently assessed cohorts. We applied three different algorithms including our in-house algorithm to identify cell intrinsic differentially expressed genes in individual cell types (CI-DEGs). We assessed the performance of all algorithms by comparison to single nucleus RNAseq data. We identified consensus CI-DEGs that are common to multiple brain regions. Despite significant overlap between consensus CI-DEGs and bulk-DEGs, many CI-DEGs were absent from bulk-DEGs. Consensus CI-DEGs and their enriched GO terms include genes and pathways previously implicated in AD or neurodegeneration, as well as novel ones. We demonstrated that the detection of CI-DEGs through computational deconvolution methods is promising and highlight remaining challenges. These findings provide novel insights into cell-intrinsic transcriptional changes of individual cell types in AD and may refine discovery and modeling of molecular targets that drive this complex disease.

Keywords: Alzheimer's disease, gene expression, transcriptome, RNA sequencing, deconvolution, cell-specific gene expression, neurodegeneration, bioinformatics

62 **Background**

63 Alzheimer's disease (AD) is a neurodegenerative disease that affects ~5.7 million
64 patients with annual cost of more than $230 billion in the US[1]. Effective disease-modifying
65 drugs are still elusive despite the urgent need and increasing global burden[2,3]. Pathologically, AD
66 is marked by amyloid-beta plaques and neurofibrillary tangles, along with neuronal loss and
67 gliosis in the affected brain regions. Transcriptome-wide expression levels have been analyzed
68 from bulk brain tissue of hundreds of AD patients and neuropathologically normal controls[4-8] to
69 discover genes and biological pathways that are perturbed in and/or lead to AD. Systems biology
70 and bioinformatics analysis of these data have implicated altered pathways in AD including
71 immune response [6] and myelin metabolism [4,5]. However, a fundamental knowledge gap remains
72 concerning whether disease-associated changes in brain gene expression are due to changes in
73 cellular composition of the AD samples secondary to disease neuropathology, or due to changes
74 in the intrinsic regulation/activity of genes in the central nervous system (CNS) cells. From a
75 clinical perspective, it is difficult to target changes in cellular composition secondary to
76 neuropathology, whereas intrinsic changes in gene expression that may drive AD progression are
77 potentially "druggable".

78 We expect that identifying cell-intrinsic differentially expressed genes (CI-DEGs) of
79 individual CNS cell types will reveal novel insights into the genes and pathways that could
80 potentially identify drug targets and lead to AD therapeutics. This approach circumvents the
81 influence of cell-composition changes that can impact disease associated DEGs obtained from
82 bulk tissue transcriptome analysis. RNA sequencing (RNAseq) studies from single cell, single
83 nucleus or purified adult human CNS cells[9-11] can be used to identify CI-DEGs. Even though
84 such single cell-type RNAseq data can effectively serve as a reference to annotate bulk-tissue

85 transcriptome data[4], such approaches remain costly, cumbersome and limited in sample sizes. On
86 the other hand, there exist large-scale bulk brain RNAseq datasets[5,8,12], which can be leveraged to
87 identify CI-DEGs through analytic deconvolution of bulk tissue expression into signals of
88 individual cell types by using cell proportions or their proxies[13,14].

89 In this study, we describe the design and application of a bioinformatics pipeline that uses
90 cell type marker genes to estimate cell proportion[15,16] to deconvolute bulk tissue transcriptome
91 data using three computational approaches[13,14,17] and to subsequently identify CI-DEGs. We
92 applied our pipeline to the analysis of three post-mortem brain datasets, one from dorsolateral
93 prefrontal cortex (DLPFC)[8] and two from temporal cortex (TCX)[4,12,18] regions, comprised of 685
94 unique samples. Consensus CI-DEGs common to both TCX and DLPFC regions were analyzed
95 for enrichment of gene ontology (GO) terms. We compared the results of consensus CI-DEGs to
96 consensus bulk-DEGs. In addition, for the DLPFC[8] dataset that had both bulk and single nucleus
97 RNAseq[19] (snRNAseq) data, we compared the CI-DEGs from the computational deconvolution
98 to CI-DEGs obtained from snRNAseq[19].

99 To our knowledge, this is the first study to detect consensus CI-DEGs and their enriched
100 gene ontology (GO) terms from multiple brain regions using multiple computational
101 deconvolution algorithms for AD and control RNAseq samples. Our study illustrates the cell
102 proportion landscape of AD and control brain regions assessed in three independent RNAseq
103 studies[4,7,8,12]. We identify consensus CI-DEGs many of which are not observed in bulk-DEG
104 analysis and characterize their cell-type specificity. GO terms that are enriched for CI-DEGs
105 implicate cell intrinsic transcriptional alterations that may influence AD, rather than be a result
106 of cell-proportion changes in this disease. These CI-DEGs and their biological pathways may
107 serve as refined molecular targets for therapeutic discoveries and disease modeling in AD. Our

108 study also demonstrates that detection of CI-DEGs through computational deconvolution
109 methods is promising while some challenges remain.
110

111 Results

112 Cellular composition in three brain cohorts from two brain regions

113 We analyzed three cohorts each consisting of post-mortem brains from AD and control
114 subjects (**Table S1**), namely the Rush Religious Orders Study and Memory and Aging Project
115 dorsolateral prefrontal cortex (DLPFC)[7,8], Mayo Clinic temporal cortex (TCX-Mayo)[4,12], and
116 Mount Sinai VA Medical Center Brain Bank temporal cortex (TCX-MSBB)[18]. We generated the
117 TCX-Mayo RNAseq dataset, and downloaded DLPFC and TCX-MSBB RNAseq datasets from
118 the AMP-AD knowledge portal on Synapse (www.synapse.org).

119 Cell proportions (**Table S2**) were estimated for DLPFC, TCX-Mayo and TCX-MSBB
120 datasets independently using the digital sorting algorithm (DSA) method[16] and the top 100
121 marker genes (**Table S3**) obtained from R package BRETIGEA[15] for each of the following cell
122 types – neuron, oligodendrocyte, microglia, oligodendrocyte progenitor cell (OPC), astrocyte and
123 endothelial cell.

124 An inspection about the pairwise correlation between marker genes (**Fig 1a**) revealed that
125 markers of OPC have poor median pairwise Pearson correlation values of 0.12 in DLPFC, 0.11
126 in TCX-Mayo and 0.06 in TCX-MSBB respectively, whereas among the other five cell types
127 neuronal markers have the highest median correlation (0.68 in DLPFC, 0.78 in TCX-Mayo and
128 0.67 in TCX-MSBB), and microglia markers have the lowest correlation (0.37 in DLPFC, 0.42
129 in TCX-Mayo and 0.44 in TCX-MSBB). In addition, a computer simulation study (**Fig S1**)
130 demonstrated that the estimated proportions of OPC were not robust upon using different

131 selection of marker genes. Therefore, we did not include OPC in downstream analyses in this
132 study.

133 In all three datasets, neuronal cell proportion estimates were significantly lower in AD
134 compared to controls (**Fig 1b**). The magnitude of this decrease was the greatest for TCX-Mayo
135 (AD mean proportion = 28.0%, Control = 35.7%; ratio of AD:control cell proportions=0.78),
136 followed by TCX-MSSM (AD = 42.3%, control = 49.3%; ratio=0.87) and DLPFC (AD = 42.4%,
137 control = 47.4%; ratio=0.89). The estimated proportions of microglia were significantly higher in
138 AD vs. controls for all datasets, with higher magnitude in TCX-Mayo (AD:control ratio=1.19)
139 and TCX-MSBB (AD:control ratio=1.19) than for DLPFC (AD:control ratio=1.06). The
140 estimated proportions of astrocytes and endothelial cells were significantly higher in AD vs.
141 controls for DLPFC and TCX-Mayo datasets, although the magnitude was greater in TCX-Mayo
142 (1.40 and 1.30 respectively) than in DLPFC (1.07 and 1.14 respectively) for both cell types.
143 Oligodendrocyte proportion is significantly higher in AD in DLPFC with AD:control ratio 1.14
144 and TCX-MSBB with AD:control ratio 1.27, although remains unchanged in TCX-Mayo with
145 the ratio 0.94. Collectively, these findings demonstrate that the proportions of CNS cell types are
146 different in post-mortem AD vs. control brains for most cell types. Although these proportional
147 changes with AD are mostly consistent across the different studies, their extent varies across
148 brain regions, with TCX tending towards higher magnitude of neuronal loss and microglia
149 proliferation than DLPFC. It needs to be emphasized that the cell proportion changes estimated
150 here are relative values, rather than absolute cell proportion changes between ADs and controls.

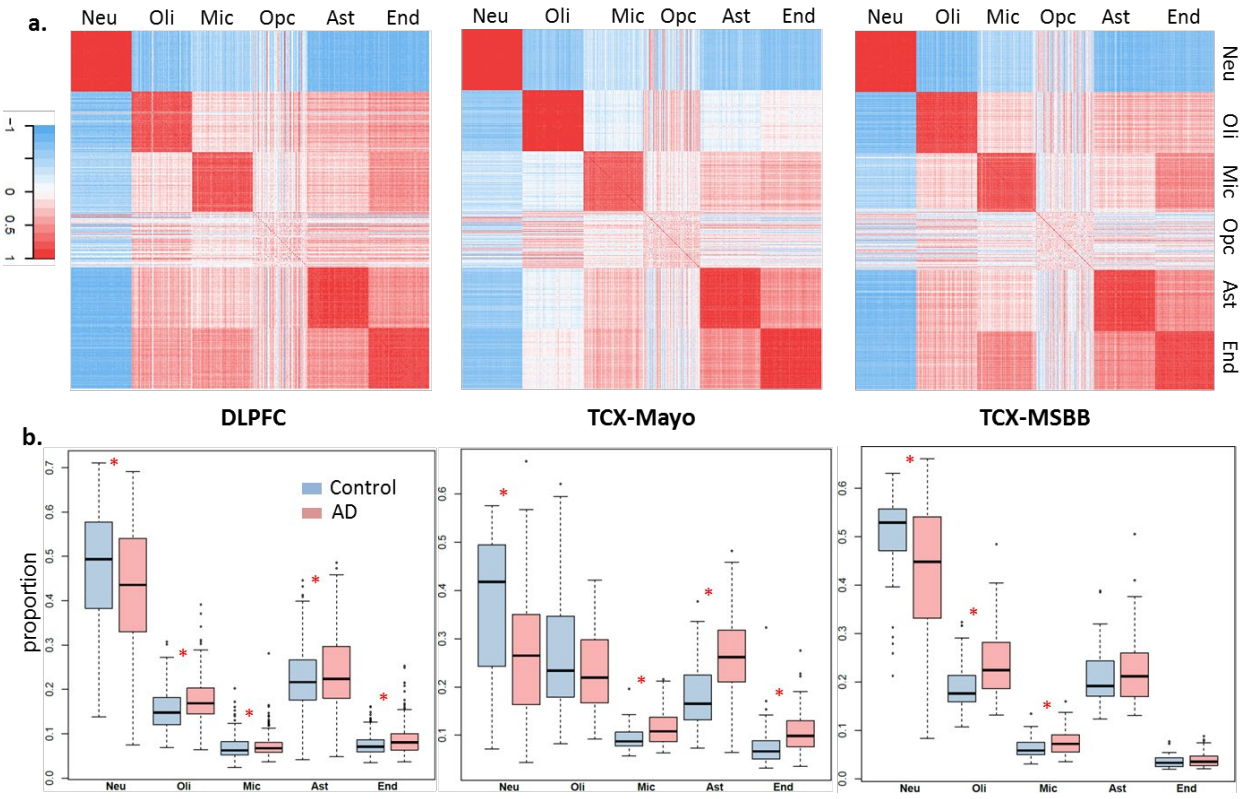

Fig.1: **a)** Pearson correlation between marker gene expressions in six cell types. Marker genes are from literature. **b)** Estimated cell proportions in DLFPC, TCX-Mayo and TCX-MSBB datasets in five cell types. Red asterisk indicates differences between cell proportions in AD and control groups at nominal p value 0.05 from Wilcoxon rank sum test.

Differential Expression Analyses

In this study, three computational approaches were applied to identify cell intrinsic differential expression in individual cell types (CI-DEGs, **Table S4-S6**), namely CellCODE[14], PSEA[13] and our method WLC. Differentially expressed genes from bulk brain tissue (bulk-DEGs) were identified through linear regression without adjusting for cellular composition (**Table S7**). For the DLPFC, TCX-Mayo and TCX-MSBB datasets, we obtained bulk-DEGs and

164 CI-DEGs from the three computer algorithms for neuronal, oligodendrocytic, microglial,
165 astrocytic and endothelial cell types respectively.

166 We compared bulk-DEGs across the three datasets (**Fig 2a**, top panel). Similarly, CI-
167 DEGs from CellCODE, PSEA and WLC are compared across datasets (**Fig 2a**, lower panels),
168 such that CI-DEGs shared between datasets are required to be consistent in the designated cell
169 type. All DEGs are identified at nominal p-value cutoff 0.05 and shared CI-DEGs have the same
170 direction of change in the compared datasets. The ratio of overlap between any two datasets over
171 all DEGs, i.e. the number in overlapping areas of the Venn diagram over the total number (**Fig**
172 **2a**, top panel), is 30.0% or 1711/5697 in up-regulated bulk-DEGs, and 34.8% or 2214/6371 in
173 down-regulated bulk-DEGs. This ratio of overlap in bulk-DEGs is much higher than that in CI-
174 DEGs (2.7%, 4.7% and 10.0% in up-regulated genes from CellCODE, PSEA and WLC
175 respectively; 3.1%, 6.8% and 9.3% in down-regulated genes from CellCODE, PSEA and WLC
176 respectively).
177

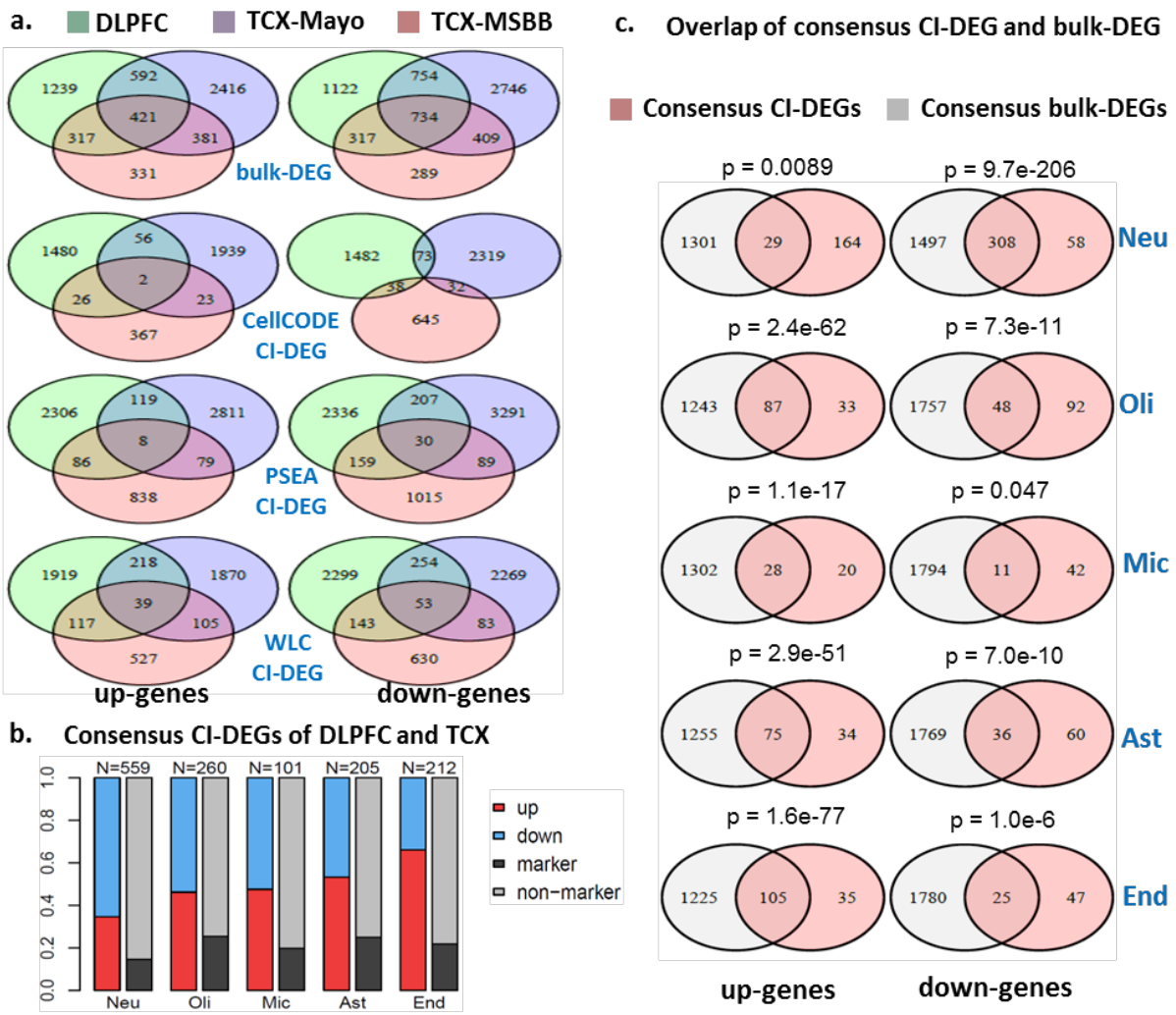

Fig.2: **a)** Overlap across three independent RNAseq datasets of bulk-DEGs (upper panel) and CI-DEGs (lower panels) from three computational approaches. **b)** Consensus CI-DEGs between DLPFC and TCX brain regions, which consist of consensus CI-DEGs between DLPFC and TCX-Mayo, or between DLPFC and TCX-MSBB. c) Overlap between consensus CI-DEGs and consensus bulk-DEGs, per cell type. The p-values of overlap are from hypergeometric tests.

Consensus CI-DEGs between DLPFC and TCX

To obtain the consensus CI-DEGs that are shared between DLPFC and TCX brain regions, we selected those CI-DEGs that are detected in "DLPFC and TCX-Mayo" or in

188 "DLPFC and TCX-MSBB" under any of the three algorithms (**Fig S2**). We combined all such
189 genes, which collectively comprised the consensus CI-DEGs for each cell type (**Fig 2b**).
190 Similarly, consensus bulk-DEGs were the combined set of bulk-DEGs shared between "DLPFC
191 and TCX-Mayo" or "DLPFC and TCX-MSBB".

192 Most consensus CI-DEGs are from neuronal cells (N=559), followed by
193 oligodendrocytes (N=260), whereas microglia contributed the least number (N=101). The
194 majority (65.5% or 366/559) of neuronal CI-DEGs is down-regulated in AD, and the majority
195 (66.0% or 140/212) of endothelial CI-DEGs is up-regulated in AD, with other cell types lying in
196 between. Some of these CI-DEGs are also among the 1000 marker genes of the corresponding
197 cell type from BRETIGEA[15]; 14.7% or 82/559 of neuronal CI-DEGs are also neuronal markers,
198 25.4% or 66/260 of oligodendrocyte CI-DEGs are also oligodendrocyte markers, and other cell
199 types lie in between.

200 With regards to consensus bulk-DEGs (**Fig S3**), 28.2% of them (885/3135) are cell type
201 markers; 10.4% neuronal markers, 5.6% oligodendrocyte, 3.4% microglia, 4.8% astrocyte and
202 4.0% endothelial markers. The above observations indicate that computational deconvolution
203 algorithms could identify CI-DEGs for both marker genes and non-marker genes. Importantly,
204 the proportion of non-marker CI-DEGs is greater than that in bulk-DEGs. This suggests that
205 compared to bulk-DEGs, CI-DEGs may be capturing a greater proportion of expression changes
206 that are not due to mere cell population changes.

207 We also compared the consensus bulk-DEGs with consensus CI-DEGs (**Fig S4**). We
208 determined that only a small fraction (15.0% or 29/193) of the up-regulated neuronal CI-DEGs
209 was also present in up bulk-DEGs although the overlap is still significant (**Fig 2c**). In
210 comparison, most of the up-regulated CI-DEGs of the other four cell types were included in up

211 bulk-DEGs. On the other hand, most (84.2% or 308/366) of the down-regulated neuronal CI-
212 DEGs were also present in down bulk-DEGs, whereas most of the down-regulated CI-DEGs of
213 the other four cell types were absent from this group. Since bulk-DEGs did not adjust for
214 neuronal loss and gliosis in AD (**Fig 1b**), its ability to identify up-regulated neuronal genes and
215 down-regulated glial genes is likely to be compromised. For the same reason, bulk-DEGs may
216 have a false inflation of detecting down-regulated neuronal and up-regulated glial genes.
217

218 **Enriched GO terms of consensus CI-DEGs between DLPFC and TCX**

219 To identify pathways implicated by CI-DEGs that are robust across brain regions, we
220 performed Gene Ontology (GO) enrichment analysis[20,21] for the consensus CI-DEGs, assessing
221 separately those that are up vs. down in AD subjects (**Table S8-S17**). **Fig 3** illustrates the top
222 two enriched GO terms by enrichment p-values, after filtering out terms that encompass less than
223 four CI-DEGs or are cellular compartments.

224 Consensus CI-DEGs revealed biological pathways that are perturbed in AD in specific
225 brain cell types. Some of these pathways have previously been implicated in AD and others are
226 novel. Down-regulated neuronal CI-DEGs were enriched in neuropeptide hormone activity
227 (GO:0005184) and hormone activity (GO:0005179) pathways, which include *VGF* (a.k.a.
228 neuroendocrine regulatory peptide 1)[22] and corticotropin releasing hormone (*CRH*)[23] (**Table**
229 **S13**). Consensus up-regulated neuronal CI-DEGs were significantly enriched in potassium
230 channel activity (GO:0005267) and regulation of ion transport (GO:0043269) pathways (**Table**
231 **S8**). The latter GO term encompasses most of the genes from the former, and also includes other
232 genes involved in neuronal functions such as the glutamate ionotropic receptor NMDA type
233 subunit 1, *GRIN1*[24] and *SYT1*, which encodes the synaptic vesicle protein, synaptotagmin[25].

234 Many of the most significant GO terms are related to key functions of the respective cell
235 types for the glial CI-DEGs, as well. The top enriched pathway of down-regulated CI-DEGs in
236 oligodendrocytes is myelination (GO:0042552), including myelin basic protein (*MBP*)[4],
237 plasmolipin (*PLLP*)[4,5], myelin and lymphocyte protein (*MAL*), and myelin-associated
238 glycoprotein (*MAG*)[26] (**Table S14**). Up-regulated CI-DEGs of oligodendrocytes are enriched in
239 ceramide biosynthetic process (GO:0046513) including ceramide synthase 4 (*CERS4*) and UDP
240 glycosyltransferase 8 (*UGT8*)[5] (**Table S9**). Ceramide is a constituent of sphingomyelin, a
241 sphingolipid which is particularly found in the myelin sheath; and also a multi-functional
242 signaling molecule[27,28]. Hence, both the down-regulated and the up-regulated oligodendroglial
243 consensus CI-DEGs highlight different components of the myelin biology that are perturbed in
244 AD.

245 Similarly, microglial, astrocytic and endothelial CI-DEGs also highlight processes
246 pertinent to the functions of these cell types. Microglial up-regulated CI-DEGs are enriched in
247 inflammatory response (GO:000695) and leukocyte activation (GO:0002696), which includes
248 complement C3a receptor 1 (*C3AR1*)[29], interleukin 18 (*IL18*)[30,31] and CCAAT enhancer binding
249 protein alpha (*CEBPA*)[32] genes (**Table S10**).

250 Astrocytes, a cell type that plays a critical role in maintaining brain energy dynamics[33]
251 and metabolism[34], show enrichment of oxidoreductase activity (GO:0016491) and drug
252 metabolic process (GO:0017144) in down-regulated CI-DEGs which includes genes glutathione
253 S-transferase mu 2 (*GSTM2*)[35] and thioredoxin2 (*TXN2*)[36] (**Table S16**). Astrocytic up-regulated
254 consensus CI-DEGs are enriched for cell-cell junction assembly (GO:0007043) process (**Table**
255 **S11**), including the astrocytic gap junction protein connexin43 (*GJA1*)[37], which was identified as
256 a key regulator associated with AD related outcomes. The other top GO process for astrocytic

up-regulated consensus CI-DEGs is adenylate cyclase-inhibiting G protein-coupled receptor signaling pathway (GO:0007193), which harbors adenylate cyclase 8 (*ADCY8*), involved in memory functions[38].

Finally, endothelial cells, which are crucial in maintaining blood-brain barrier integrity[39,40], show enrichment of up-regulated DEGs in cytoskeleton organization (GO:0007010) and actin filament-based process (GO:0030029) (**Table S12**).

Importantly, some CI-DEGs highlight protein translation as a top perturbed biological pathway. Down-regulated microglial consensus CI-DEGs show enrichment in processes involved in protein translation (GO:0006614 and GO:0006613), which include ribosomal protein encoding genes[41-43] (**Table S15**). Similarly, down-regulated endothelial consensus CI-DEGs also harbor ribosomal protein encoding genes, with enrichment in protein translation related GO processes (GO:0006413 and GO:0006613) (**Table S17**).

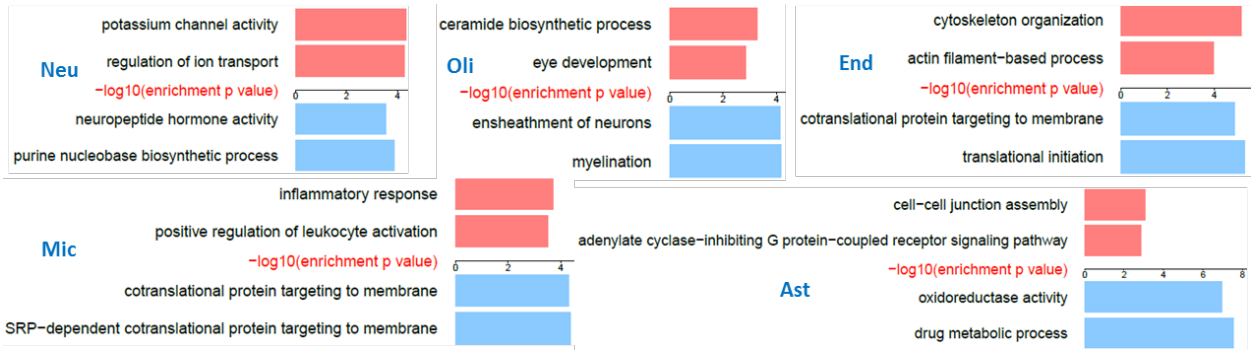

Fig 3: Top two enriched GO terms in up (red) or down-regulated (blue) consensus CI-DEGs between DLPFC and TCX regions, per cell type.

Comparison of CI-DEGs from computational deconvolution vs. snRNAseq

276 We determined the extent to which each of the three computational deconvolution
277 algorithms could detect CI-DEGs from bulk tissue by comparison of their results with those
278 obtained in a published snRNAseq study[19]. The ROSMAP dataset utilized in our study has both
279 bulk RNAseq from DLPFC (bulk-DLPFC) as well as snRNAseq (snDLPFC) in a subset of its
280 participants[19]. We compared the bulk-DLPFC data deconvoluted with three different algorithms
281 with the published snDLPFC[19] data. Endothelial CI-DEGs were not available from the
282 snRNAseq study, therefore overlap of results could be assessed only for four cell types.

283 We tested the overlap between the top CI-DEGs for each cell type obtained from
284 deconvoluted bulk-DLPFC and those from snDLPFC ranked by their p values (**Fig 4a**). We
285 evaluated the overlap for a range of top CI-DEGs up to top 1,000 genes. Overlap for CI-DEGs
286 that are either up (**Fig 4a, upper panel**) or down (**Fig 4a, lower panel**) in AD were assessed
287 separately. Hence, overlapping genes had both similar ranks and direction of effect in both
288 deconvoluted bulk-DLPFC and snDLPFC analyses. We established the significance of overlap
289 using simulations for a range of top ranked genes (N=200, 600 and 1,000) (**Table S18**).

290 Neuronal CI-DEGs retained their significance of overlap across all comparisons and for
291 all algorithms, except for the top 1,000 up-regulated neuronal CI-DEGs deconvoluted with
292 PSEA. Microglial CI-DEGs had the least numbers of significant overlap for their top ranked
293 genes. Astrocytic and oligodendrocytic top ranked CI-DEGs had significance of overlap between
294 the neuronal and microglial results (**Table S18**). These findings are reflective of the abundance
295 of these cell types, with the most abundant neurons having the most overlap for the top ranked
296 CI-DEGs between deconvoluted bulk-DLPFC and snDLPFC.

297 Amongst these comparisons, we determined that the significance for overlap was best for
298 all algorithms for the top ranked 600 genes. Using WLC deconvoluted results, the overlap for the

299 top 600 CI-DEGs from bulk-DLPFC and snDNPFC are statistically significant for all eight
300 comparisons (**Fig 4a**). For the top 600 genes, overlap with CellCODE results is significant for all
301 except down-regulated oligodendrocyte and up-regulated astrocyte CI-DEGs. For PSEA, none of
302 the microglia CI-DEGs had significant overlap. PSEA results for the top 600 genes were
303 otherwise significant for all but up-regulated oligodendrocyte and down-regulated astrocyte
304 genes.

305 We also performed a comparison of CI-DEGs identified at nominal significance (p-value
306 <0.05) with each algorithm from bulk-DLPFC to nominally significant snDLPFC results (**Fig**
307 **4b, Table S19**). As with the above comparison, genes that are either up or down in both
308 deconvoluted bulk-DLPFC and snDLPFC data were analyzed separately for each cell type.

309 Not surprisingly, down-regulated neuronal CI-DEGs have the greatest overlap (537/3732
310 or 14.4% for WLC, 292/3213 or 9.1% for CellCODE, 415/3516 or 11.8% for PSEA). These
311 overlaps are significant for all three algorithms (**Table S19**). Down-regulated CI-DEGs in
312 microglia show the least proportion of overlap (9/723 or 1.2% for WLC, 4/609 or 0.66% for
313 CellCODE, 16/820 or 2.0% for PSEA) (empirical p-value > 0.05). Significant overlap detected
314 with WLC (all but down-regulated microglia) and PSEA (all but microglial results and up-
315 regulated oligodendrocytes) were similar, whereas CellCODE results had significant overlaps
316 only for the neuronal CI-DEGs (**Table S19**).

317

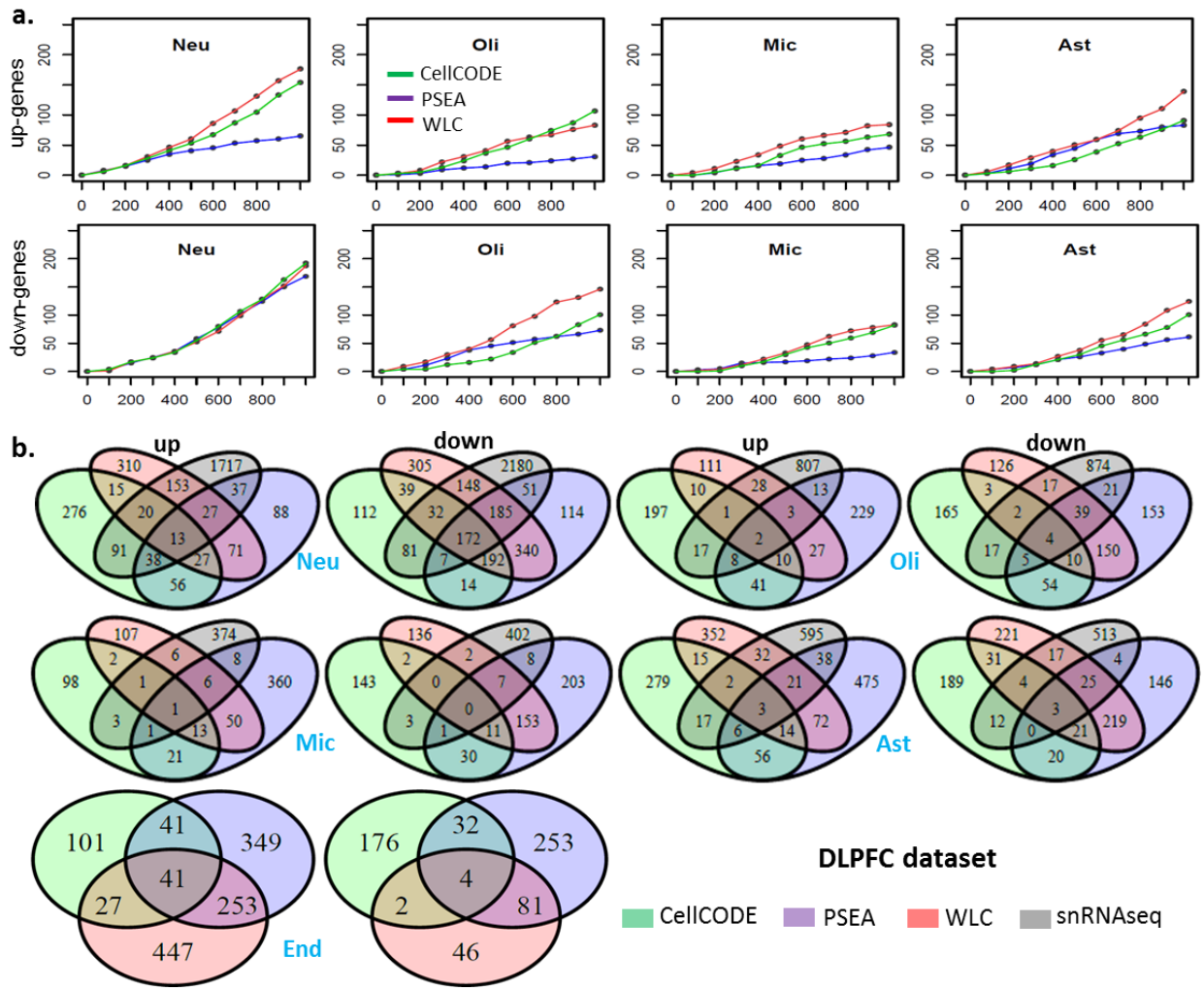

Fig 4: Comparison of CI-DEGs from computational deconvolution with CI-DEGs from snRNAseq on DLPFC dataset. **a.** Upper panel: number of overlapping genes (y-axis) between the top *N* (x-axis) up-regulated genes in snDLPFC and top *N* up-regulated genes from bulk-DLPFC deconvoluted with PSEA, WLC and CellCODE, respectively. Lower panel: number of overlapping genes (y-axis) between the top *N* (x-axis) down-regulated genes in snDLPFC and top *N* down-regulated genes from deconvoluted bulk-DLPFC. **b.** Venn diagram of CI-DEGs from computational deconvolution methods and those from snRNAseq. Overlap is evaluated for both bulk-DLPFC and snDLPFC CI-DEGs detected at nominal p value ≤0.05.

Discussion

There is an increasing number of large scale RNAseq-based transcriptome datasets generated in bulk tissue for many diseases, including brain tissue from patients with AD, other neurodegenerative diseases and controls[5-8,12]. Comparison of such transcriptome data from patient and control individuals has been instrumental in the identification of genes and co-expression networks that are altered in and may therefore be risk factors for these diseases[4-7,44]. The discovery that many of these transcriptional networks harbor genes with disease risk variants provides support for the utility of this bulk-transcriptome approach in deciphering molecules and pathways that are risk factors for these conditions. Nevertheless, there is clear evidence for abundant transcriptome alterations in bulk tissue from affected organs of patients with disease, which appears to be due to changes in cellular composition of that tissue as a consequence of the disease processes[4,45]. Given this, there is a strong need to detect cell-intrinsic transcriptional alterations to be able to distinguish gene expression changes that are upstream of and may therefore be causative factors for disease from those that are merely a result of cell proportion changes that occur due to disease pathology. The discovery of molecules and pathways that are upstream of and risk factors for disease pathology is of paramount importance for development of targeted therapies. This information can also aid in the identification of refined disease biomarkers reflective of disease-causative expression alterations in these conditions. Detection of cell-specific transcriptional changes can also help develop more accurate disease models harboring these cellular alterations. Further, discovery of cell-specific transcriptional alterations in disease may uncover expression changes, particularly in less abundant cell types, which may be missed by the analysis of bulk transcriptome. Thus, there is a growing effort to identify cell-specific expression alterations in human diseases[9-11,14,15,17,46-48].

351 There are two general approaches to decipher cell-specific transcriptional changes in AD.
352 One approach is to conduct single nucleus (snRNAseq), single cell (scRNAseq) or purified cell
353 RNAseq experimentally, followed by data analyses. The alternative approach is to design
354 relatively complex bioinformatics pipelines to decipher cell-intrinsic information of individual
355 cell types from bulk tissue microarray or RNAseq data. The first approach is limited due to
356 significantly higher cost and experimental challenges. Additionally, these approaches may have
357 the drawback that the procedures of dissociating cells break cell-to-cell communication and thus
358 may not reflect the true expression signal *in vivo*. The alternative bioinformatics approach to
359 decipher cell-specific transcriptome alterations from bulk tissue has the potential to avoid the
360 above weaknesses of the experimental approach. Furthermore, a bioinformatics approach can
361 make use of the large amount of existing RNAseq data[4-8,12] from fresh or frozen bulk brain
362 tissues with minimum added cost, and may better reflect the true situation in brain where
363 different cells are in communication.

364 In this study, applying three different algorithms[13,14] including our own novel approach,
365 we estimated cell-intrinsic gene expression for deconvoluted cell types from three large bulk
366 RNAseq datasets[4,8,12,18] from two brain regions. We identified consensus cell-intrinsic
367 transcriptional alterations (CI-DEGs) in AD, which are conserved across cohorts and brain
368 regions. We also performed an in-depth comparison of these CI-DEGs with bulk brain RNAseq
369 data obtained from the same datasets, collectively comprised of 685 unique brain samples. To
370 our knowledge, this is the first study to detect CI-DEGs and their enriched gene ontology (GO)
371 terms in computationally deconvoluted large-scale RNAseq data from AD and control brain
372 samples. Additionally, we conducted a detailed comparison of CI-DEGs deconvoluted from

373 bulk-DLPFC data with the three algorithms and those obtained from snRNAseq[19] (snDLPFC) of
374 a subset of the samples from the same cohort[7,8].

375 The main findings from our study can be summarized as follows: **1)** The direction of
376 change in cellular proportions in AD is consistent across two brain regions and three datasets for
377 most cell types, although the magnitude of change seems to vary. Our findings revealed greater
378 neuronal loss and microgliosis in TCX compared to DLPFC. **2)** We identified CI-DEGs and bulk
379 tissue DEGs (bulk-DEGs) independently in two TCX and one DLPFC datasets. The overlap in
380 bulk-DEGs across datasets is greater than that for CI-DEGs. **3)** We performed an in-depth
381 comparison of the consensus CI-DEGs, common to both TCX and DLPFC against the consensus
382 bulk-DEGs detected in these same datasets. We identified significant overlap between consensus
383 CI-DEGs and consensus bulk-DEGs. The extent of overlap between consensus bulk-DEGs and
384 consensus CI-DEGs was greatest for *down-regulated neuronal* genes (p=9.7E-206). This was
385 followed by *up-regulated non-neuronal* genes (p ranging from 1.6E-77 for endothelia to 1.1E-17
386 for microglia). **4)** Despite the statistically significant overlap between consensus bulk vs. CI-
387 DEGs, the majority of the consensus CI-DEGs for *up-regulated neuronal* and *down-regulated*
388 *non-neuronal* genes were not detected in bulk tissue. This finding highlights the potential ability
389 of computational deconvolution approach to identify CI-DEGs that may be missed in bulk-DEGs
390 especially for genes that are not moving in the direction of cell proportion changes. **5**) We
391 identified GO-terms enriched for consensus CI-DEGs, and detected processes that have
392 previously been implicated in AD as well as novel ones. **6**) Using an snRNAseq[19] dataset as
393 comparison, we assessed the performance of our CI-DEG detection algorithm (WLC), and the
394 published CellCODE[14] and PSEA[13] approaches. We determined that WLC had comparable or

395 superior performance in the detection CI-DEGs that had significant overlap with snDLPFC
396 results.

397 Our findings highlight the consistency and reproducibility of our findings across two
398 different brain regions from three different studies conducted separately. We identified similar
399 directions of change in AD:Control cell proportions in TCX and DLPFC. As expected from
400 known AD neuropathology, neuronal populations are significantly lower, and microglial
401 populations are significantly higher in AD vs. control brains in all datasets. Consistent with this
402 pattern of reproducibility, we also found significant overlap of consensus CI-DEGs detected in
403 TCX and DLPFC for all cell types and for both directions of change, i.e. up or down in AD, with
404 consensus bulk-DEGs.

405 Using consensus CI-DEGs, we identified GO terms, which include processes and genes
406 that have previously been implicated in AD, thereby providing further validation of our
407 approach. Detailed discussion of all of the pathways identified with the consensus CI-DEGs is
408 beyond the scope of this study. Instead, we herein highlight some of the CI-DEG enriched
409 pathways.

410 Down-regulated neuronal CI-DEGs were enriched in neuropeptide hormone activity
411 (GO:0005184) pathway. These terms include *VGF* (a.k.a. neuroendocrine regulatory peptide 1),
412 which is selectively expressed in some neurons and shown to be reduced in AD and Parkinson's
413 disease[22]. Corticotropin releasing hormone (CRH), which is a neuronally expressed peptide that
414 mediates stress in the hypothalamic-pituitary-adrenal axis, is also a member of the same GO
415 term. CRH has been implicated in both adverse outcomes related to AD pathology in model
416 systems and epidemiology studies, as well as having an important role in learning and memory[23].
417 Neuronal reduction of *CRH* may either be a potentially neuro-protective response in AD brains

418 or lead to further negative impact on memory. Although the biological implications of our
419 finding remain to be uncovered, our results are aligned with the potential importance of the
420 neuroendocrine system in AD.

421 Interestingly, CI-DEGs also implicate biological processes that are enriched for neuronal-
422 DEGs that are *up* in AD, despite reductions in neuronal cell populations in this condition for both
423 TCX and DLPFC. This suggests that our cell type specific transcriptome deconvolution
424 successfully captures transcript changes that occur in the direction opposite to that of cell-
425 population changes, and that may therefore be missed in bulk-DEG approaches.

426 Indeed, the significant GO terms potassium channel activity (GO:0005267) and
427 regulation of ion transport (GO:0043269) harbor many potassium channels, which are up in AD
428 for neuronal CI-DEGs in both TCX and DLPFC, but do not have consistent results in bulk-DEGs
429 from the same cohorts. These findings highlight the potential utility of cell-specific transcript
430 deconvolution approaches in reducing noise from cell population changes, thereby resulting in
431 consistent detection of true signal. Notably, potassium channels have been reported to be up in
432 AD brains and mouse models of AD[49-51], leading to their suggestion as a potential therapeutic
433 avenue in this condition.

434 Notably, the computational approach we describe can be extended to cellular sub-types.
435 To exemplify this, we utilized published snRNAseq[19] data to identify marker genes of excitatory
436 and inhibitory neurons. Using those and the previously applied markers for the glial cells, we ran
437 DSA algorithm [3] to estimate cell proportion (**Fig S6**). We applied PSEA, WLC and CellCODE
438 to identify consensus CI-DEGs in the same fashion as described (**Table S20**, **Fig S7**). As
439 expected, both excitatory and inhibitory neuron proportions are both significantly lower in AD
440 samples. Interestingly, the up-regulated excitatory neuronal consensus CI-DEGs are enriched in

441 cation channel activity, which is not observed in inhibitory neurons (**Tables S21-S23**). These
442 results highlight the ability of this approach to refine transcriptional alterations from bulk brain
443 data to cellular sub-types.

444 Some consensus CI-DEGs point to AD-related perturbations of key cellular functions for
445 the specific cell type. An example of this is consensus oligodendrocyte CI-DEGs. The *down-*
446 *regulated* oligodendrocyte CI-DEGs are enriched for the myelination GO term (GO:0042552)
447 and those that are *up* in this cell type are enriched in ceramide biosynthetic process
448 (GO:0046513).

449 Down-regulated oligodendrocyte CI-DEGs include myelin basic protein (*MBP*)[4],
450 plasmolipin (*PLLP*)[4,5], myelin and lymphocyte protein (*MAL*), and myelin-associated
451 glycoprotein (*MAG*)[26], even though bulk-DEGs for these genes did not show consistent changes.
452 We[4] and others[5] demonstrated lower levels in AD of co-expression networks of genes implicated
453 in myelination, which is consistent with the present findings from oligodendrocyte CI-DEGs.

454 Ceramide dysregulation has been implicated in AD[52,53]. Increased ceramide species were
455 observed in AD and other neuropathological disorders compared to controls[53], and the activation
456 of the neutral sphingomyelinase–ceramide pathway induces oligodendrocyte death[54]. Our present
457 observation from oligodendrocyte CI-DEGs are consistent with these findings.

458 Another potential utility of the computational deconvolution approach is its
459 complementarity to snRNAseq data. While the latter is able to provide transcriptional data at a
460 cellular level, it can miss information from low abundance cell populations. Indeed, in a recent
461 snRNAseq[19] study of AD and control brain samples, only 121 endothelial cells were identified
462 out of a total of 70,634 cells; and consequently no endothelial DEGs were reported. In contrast,
463 our computational approach identified 140 genes that are up-regulated in endothelial cells in AD

464 and 72 genes that are down-regulated. We further validated the endothelial expression of these
465 genes by confirming their expression in sorted endothelial cells[9]. The significantly enriched GO
466 terms include cytoskeleton organization and translational initiation, for these genes up- and
467 down-regulated, respectively, in endothelia (**Fig 3, Tables S12, S17**). These findings highlight
468 the ability of the computational deconvolution approach to provide biological insights that could
469 be missed by snRNAseq approaches.

470 Despite the biological insights gained from computationally deconvoluted CI-DEGs, they
471 also have some shortcomings. Compared to bulk-DEGs, CI-DEGs between different datasets
472 show less degree of overlap, regardless of the deconvolution algorithm utilized. This highlights
473 the challenge in the field for the ultimate goal of minimizing detection of transcriptional
474 perturbations due to cell proportion changes while maximizing discovery of those that lead to
475 disease. Put differently, CI-DEGs may enhance discovery of true cell-specific transcriptional
476 changes but this may come at the expense of increased false negative findings. In contrast, bulk-
477 DEGs may capture a larger number of perturbed transcripts, but some may be merely due to cell
478 population differences between diseased and healthy tissue. Ultimately, findings from both
479 approaches may provide the greatest utility in detecting true positives.

480 Another analytic caveat is that cell proportion estimation approaches, including those
481 used in this study yield relative values rather than the absolute levels of cell proportions.
482 Detection of absolute levels in a typical bulk RNAseq experiment will be challenging if not
483 impossible due to different mRNA amounts per cell type and the library preparation protocols
484 (e.g. TruSeq® RNA Sample Preparation v2 Guide) which require similar amounts of cDNA from
485 each sample to be sent to the sequencer. This leads to a normalization which would preclude the
486 detection of the absolute cell proportions, as illustrated in an example in **Fig. S8**. Despite these

487 challenges, our findings reveal that analytic deconvolution of bulk RNAseq can detect cell-
488 specific transcriptional changes.

489 Comparison of CI-DEGs deconvoluted from bulk-DLPFC and those detected using an
490 orthogonal approach of snRNAseq[19] from the same cohort (ROSMAP) demonstrates the ability
491 of these computational approaches to identify true cell-specific expression changes in AD. Using
492 our in-house WLC algorithm, there was significant overlap with the results from snRNAseq and
493 CI-DEGs of most cell types (**Fig 4, Tables S18-S19**). CellCODE and PSEA also identified
494 significant overlaps, but to a lesser extent, especially for rarer cell types such as microglia.
495 Hence, our WLC algorithm demonstrates at least comparable performance to these two
496 algorithms[13,14]. This is also supported by the higher degree of overlap among different cohorts
497 for CI-DEGs detected by WLC than the other two algorithms (**Fig 2a**). Due to the challenges in
498 deconvoluting noisy data from human series, different computational approaches may be utilized
499 and combined, and that calls for a more devoted effort in developing such algorithms.

500 In summary, using three distinct computational approaches, we deconvoluted brain bulk-
501 RNAseq data from three large and independent cohorts[8,12,18]. We detected cell population
502 changes that are observed consistently across cohorts, and congruent with the known disease
503 pathology. Although there is significant overlap between consensus CI-DEGs and consensus
504 bulk-DEGs, there are more unique CI-DEGs that change in the direction opposite to that of cell
505 population changes. This suggests that CI-DEGs may have utility in detecting disease-related
506 transcriptional changes above and beyond those due to cell proportion changes. Consensus CI-
507 DEGs identify GO terms, including those for hormone activity, myelin biology and channel
508 activity. The enriched CI-DEGs include genes previously implicated in AD or
509 neurodegeneration, such as *VGF, CRH, MOBP* and *MBP*, and other novel genes.

510

511 # Methods

512 **Analysis datasets**

513 We generated the TCX-Mayo data, which consists of temporal cortex RNAseq
514 measurement of 80 AD patients and 28 controls diagnosed according to neuropathologic
515 criteria[4,12]. RNAseq data were processed and quality control (QC) was conducted as
516 described[4,12]. ROSMAP DLPFC[7,8] and TCX-MSBB[18] datasets were downloaded from the AMP-
517 AD Knowledge Portal on Synapse (syn8691134 and syn8691099). We further filtered out non-
518 Caucasian samples and those that had incongruent sex based on provided covariate vs.
519 transcriptome data. All samples were classified as AD or control based on neuropathological
520 data. All TCX-Mayo AD samples had Braak neurofibrillary tangle (NFT) score ≥4. TCX-Mayo
521 controls had Braak score ≤3 and were without any neurodegenerative disease diagnoses. TCX-
522 MSBB AD samples had Braak ≥4 and CERAD ≥2; and controls had Braak ≤3 and CERAD ≤1.
523 DLPFC AD samples form ROSMAP had Braak score ≥4 and CERAD neuritic plaque score ≤2.
524 ROSMAP control samples had Braak ≤3 and CERAD neuritic plaque score ≥3.

525 It should be noted that the CERAD[55] neuritic plaque score as applied by the ROSMAP
526 study is defined such that high CERAD indicates lower neuritic plaque burden and decreased
527 probability of AD. In the MSBB study, higher CERAD indicates higher plaque burden.

528 Raw RNA read counts were normalized using conditional quantile normalization (CQN)
529 method[56] implemented in R cqn package, as previously described[4]. This normalization takes into
530 consideration library size, gene length and GC content. It also performs a log2 transformation so
531 that the resulting distribution for each gene is Gaussian-like. We also determined covariates that
532 contributed significantly to the variation of gene expression in these RNAseq cohorts (**Fig S5**)
533 for adjustment in the analyses.

534

535 **Cell proportion estimation**

536 Digital sorting algorithm (DSA)[16] was applied to estimate cell proportions through R
537 DSA package, function DSA. For each cell type, DSA first computes the average marker gene
538 expression in the analysis dataset, the purpose of which is to construct a variable that better
539 reflects cell proportion variation among subjects. To reduce the effect of outlier expression that
540 is occasionally seen in RNAseq data, we modified the original DSA so that the median instead of
541 mean expression was computed.

542

543 **CI-DEG analysis for individual cell types**

544 In this study we identified CI-DEGs from deconvoluted bulk RNAseq data using three
545 different algorithms, namely PSEA[13], CellCODE[14] and our in-house algorithm WLC. All
546 analyses adjusted for the following variables: Sex, RIN, age at death and batch for DLPFC and
547 TCX-MSBB datasets, and sex, RIN and age at death for TCX-Mayo dataset (**Fig S5**).

548 PSEA[13] applies model selection procedures to select cell type(s) that should be included
549 in baseline (control) or AD condition, and then estimate differential expression in specific cell
550 types (CI-DEGs). We used the R package PSEA to obtain CI-DEG results of PSEA approach,
551 through functions em_quantvg (to generate candidate models) and lmfitst (to fit all candidate
552 models and pick the best one). Expression values used in PSEA are in linear scale (non log-
553 transformed).

554 CellCODE[14] assesses overall gene expression difference between AD and control groups
555 with adjustment of relative cell proportion, followed by assigning the cell type that correlates
556 best with the difference (CI-DEGs). R package CellCODE was used to obtain DEG results of

CellCODE approach, through functions getAllSPVs (to construct surrogate variable using marker genes through singular value decomposition) and lm.coef (to estimate difference between AD and control groups). Strictly speaking, CellCODE measures the overall differential expression rather than CI-DEGs but identifies the cell type that is most correlated with the observed difference between AD and control using cellPopT function. However, for simplicity, we refer to the DEGs from CellCODE as CI-DEGs in this study. Expression values used in CellCODE are in log scale.

Our in-house method WLC, described in the method section, applies weighted linear regression with constraints and model selection procedures, which also estimates differential expression in specific cell types (CI-DEGs). It guarantees the fitted relative gene expression is non-negative. By weighing the expression, it smooths out the extreme data points. The procedures of this algorithm is illustrated by the following high level pseudo code.

Assume cell type 1,2,3,4,5 are neu, oli, mic, ast and end respectively

Fit Eq.(1) to identify a set of cell type T in which the gene is significantly expressed

If the set T is not empty:

    Fit Eq.(2) to identify a set of cell type $\Phi \subseteq T$ *in which the gene is differentiallly expressed*

    Let $\Omega = \{each\ cell\ type\ in\ T, \Phi, T\}$

    For each element $\Theta \in \Omega$

        Fit Eq.(3) with adjustment for other covariates $C_k (1 \leq k \leq m)$

        Keep Akaike information criterion (AIC) of this model fitting

    Identify Θ_{best} *that gives the best AIC.*

580 *Use the estimated values from the best model and obtain p-value from F-test*

581

$$y_i \sim \alpha_0 + \sum_{t\in\{1,2,3,4,5\}} \alpha_t x_{i,t} + \epsilon, \qquad \text{s.t. } \alpha_t \geq 0 \ (1 \leq t \leq 5) \tag{1}$$

$$y_i \sim \beta_0 + \sum_{t\in T} \beta_t x_{i,t} + d_i \sum_{t\in T} \beta_t^{\Delta} x_{i,t} + \epsilon, \quad \text{s.t. } \beta_t \geq 0, \beta_t + \beta_t^{\Delta} \geq 0 \ (t \in T) \tag{2}$$

$$y_i \sim \gamma_0 + \sum_{t\in T} \gamma_t x_{i,t} + d_i \sum_{t\in \Theta} \gamma_t^{\Delta} x_{i,t} + \sum_{k=1}^{m} \lambda_k C_{i,k} + \epsilon, \text{s.t. } \gamma_t \geq 0 (t \in T), \ \gamma_t + \gamma_t^{\Delta} \geq 0 \ (t \in \Theta) \tag{3}$$

582

583 In Eqs.1-3, y_i is the observed expression of a gene in subject i; $x_{i,t}$ is the median marker gene
584 expression of cell type t in subject i; $C_{i,k}$ is covarite k in subject i. In Eq.1, α_t is the overall
585 relative expression in cell type t. In Eq.2, β_t is relative expression at the baseline condition in cell
586 type t; $d_i = 0$ if subject i is in control group, and $d_i = 1$ if subject i is in AD group; therefore,
587 β_t^{Δ} is the difference of relative gene expression between baseline condition and AD condition in
588 cell type t. Of note, due to the biological meaning these coefficents, they must satisfy constraints
589 such that α_t, β_t, $\beta_t + \beta_t^{\Delta}$ be non-negative. In addition, y_i is in linear scale rather than in log
590 scale[57], and these non-log-transformed expression values tend to have extreme data points that
591 need to be weighted down. Based on the above considerations, Eq. 1 was fitted by weighted least
592 square linear regression with contraints, which is implemented in lsei function in R package
593 limSolve. The weight of each observation (w_i) is determined by formulae Eq.3 and Eq.4 below.
594 Intuitively, if the expression of a gene in a sample is extremely distant from the median
595 expression of all samples in the same diagnosis group, the weight of that sample is smaller than 1
596 for that gene.

$$w_i = 1/(1 + \frac{|y_i - \text{median}(y_{AD})|}{1 + \text{median}(y_{AD})^2}) \qquad \text{sample } i \text{ is AD} \tag{4}$$

$$w_i = 1/(1 + \frac{|y_i - \text{median}(y_{control})|}{1 + \text{median}(y_{control})^2}) \qquad \text{sample } i \text{ is control} \tag{5}$$

GO enrichment analysis

Using genes included in the CI-DEG analysis as background genes, p-value for GO term enrichment with consensus CI-DEGs was calculated by "enrichmentAnalysis" function from WGCNA R package[35].

Comparison of CI-DEGs from computational deconvolution vs. snRNAseq:

To determine if computational bulk-tissue RNAseq could reveal true CI-DEGs, we downloaded and utilized a published snRNAseq study[19] from frozen DLPFC tissues (snDLPFC) which compared gene expression from 24 individuals with AD-pathology to that from 24 individuals without AD-pathology in six cell types - excitatory neurons, inhibitory neurons, oligodendrocyte, microglia, oligodendrocyte precursor cells and astrocytes. These snDLPFC samples are from the ROSMAP project, of which we analyzed 474 bulk-DLPFC RNAseq data in our current study. Twenty-four (9 AD cases, 15 controls) of the 48 individuals in snDLPFC study are also included in the bulk-DLPFC dataset. Hence both the snDLPFC and bulk-DLPFC are from the same cohort with significant overlap. Three deconvolution methods were included in this comparison – CellCODE[14], PSEA[13] and WLC, our in-house method.

Two types of comparisons were made between the deconvoluted bulk-DLPFC and snDLPFC results. In the first comparison, we ranked the genes by their p-values of differential expression between AD and control subjects, per cell type. We compared the top *N* up- or down-genes from the snDLPFC study to those identified by each deconvolution algorithm, per cell

618 type. Genes common to both the deconvoluted bulk-DLPFC and snDLPFC were counted and
619 plotted.

620 In the second set of comparisons, CI-DEGs identified from deconvolution methods at
621 nominal significance (p-value <0.05) were compared to those identified in the snRNAseq data
622 (p-value <0.05).

623 To assess if the observed overlap from each set of analyses is significant with regards to
624 overlap between random selections, we obtained empirical p values from computer simulations
625 described below.

626

627 **Empirical p-value for the number of overlapping genes:**

628 The empirical p-values for the number of overlapping genes between snDLPFC and bulk-
629 DLPFC was obtained using a computer simulation as follows. (**A**) Let S_{sn} stand for all genes in
630 snDLFC and S_{bulk} for all genes in bulk-DLFC; (**B**) randomly assign up-regulation on each S_{sn}
631 gene at probability 0.5, and on each S_{bulk} gene at probability 0.5; (**C**) randomly pick *N* genes
632 from up-genes of S_{sn} , randomly pick *N* genes from up-genes of S_{bulk}, and count the number of
633 overlapping genes; (**D**) Steps B-C were repeated 10000 times, and the numbers of overlaps
634 ($M_1, M_2, \ldots, M_{10000}$) were obtained; (**E**) Let M be the number of observed overlapping genes, and
635 the empirical p-value is (1+number of occurrences that M_i >= M)/10001.

636

637 **Identification of excitatory and inhibitory neuronal markers:**

638 Using a published human brain snRNAseq dataset [19], we identified excitatory and
639 inhibitory neuronal markers using Seurat R package FindMarker function. Excitatory neuronal
640 markers are those that **a)** are detected in >= 70% of excitatory neurons, **b)** have average

641 normalized expression in excitatory neurons that are >= 4.5X of that in each of the other cell
642 types (i.e. inhibitory neurons, oligodendrocytes, microglia, astrocytes and OPC), and **c)** have
643 rank sum test p-value < 0.05 in the comparison of expression levels in excitatory neurons and
644 each of the other cell types. Inhibitory neuronal markers were similarly identified except that we
645 required a less stringent detection limit of >= 50% of inhibitory neurons as the >= 70% threshold
646 yielded too few genes. Excitatory neuronal markers were further refined by requiring their
647 presence in the neuronal markers from BRETIGEA [15]. A total of 53 excitatory and 62 inhibitory
648 neuronal markers were identified (**Table S20**). Excitatory and inhibitory CI-DEGs were
649 identified as described above.
650

651 ## Conclusions

652 This study demonstrates the utility of our analytic approach in deciphering cell-specific
653 transcriptional alterations using bulk tissue in a complex disease, provides a comprehensive
654 comparison of our pipeline to existing ones, identifies patterns of cell proportions in AD and
655 control samples across brain regions, discovers novel CI-DEGs with replication across
656 independent cohorts and highlights biological processes with cell-specific expression changes in
657 AD. These findings are expected to refine discovery of molecular therapeutic targets, biomarkers
658 that reflect cellular transcriptional alterations in AD and accelerate generation of more accurate
659 disease models.

660

661 **Abbreviations:**

662 AD: Alzheimer's disease

663 ADCY8: Adenylate cyclase 8

664 AMP-AD: Accelerating Medicines Partnership Alzheimer's Disease

665 Bulk-DEGs: Differentially expressed genes from bulk brain tissue

666 Bulk-DLPFC: Bulk RNA sequencing from dorsolateral prefrontal cortex

667 C3AR1: Complement C3a receptor 1

668 CEBPA: CCAAT enhancer binding protein alpha

669 CERS4: Ceramide synthase 4

670 CI-DEGs: Cell-intrinsic differentially expressed gene(s)

671 CNS: Central nervous system

672 CQN: Conditional quantile normalization

673 CRH: corticotropin releasing hormone

674 DEGs: Differentially expressed genes

675 DLPFC: Dorsolateral prefrontal cortex

676 DSA: Digital sorting algorithm

677 GJA1: Connexin43

678 GO: Gene Ontology

679 GRIN1: Glutamate ionotropic receptor NMDA type subunit 1

680 GSTM2: Glutathione S-transferase mu 2

681 IL18: Interleukin 18

682 MAG: Myelin-associated glycoprotein

683 MAL: Myelin and lymphocyte protein

684 MBP: Myelin basic protein

685 NFT: Neurofibrillary tangle

686 OPC: Oligodendrocyte progenitor cell

687 PLLP: Plasmolipin

688 QC: Quality control

689 RIN: RNA integrity number

690 RNAseq: RNA sequencing

691 ROSMAP: Rush Religious Orders Study and Memory and Aging Project

692 scRNAseq: Single cell RNA sequencing

693 snDLPFC: Single nucleus RNA sequencing from dorsolateral prefrontal cortex

694 snRNAseq: Single nucleus RNA sequencing

695 SYT1: Synaptotagmin

696 TCX: Temporal cortex

697 TCX-Mayo: Mayo Clinic temporal cortex

698 TCX-MSBB: Mount Sinai VA Medical Center Brain Bank temporal cortex

699 TXN2: Thioredoxin2

700 UGT8: UDP glycosyltransferase 8

701 VGF: VGF Nerve Growth Factor Inducible, a.k.a neuroendocrine regulatory peptide 1

702 WGCNA: Weighted gene co-expression network analysis

703

704 # Declarations:

705 **Ethical Approval and Consent to participate:**

706 All data were generated from deceased individual's autopsied brain tissue. This study
707 was approved by Mayo Clinic Institutional Review Board.

708

709 **Consent for publication:**

710 All authors reviewed and approved the final manuscript.

711

712 **Availability of supporting data:**

713 The data used in this manuscript is available to the research community through the
714 AMP-AD knowledge portal on Sage Synapse as follows:

Dataset	Data Type	Description	SynapseID	DoD
Mayo RNAseq TCX	RNASeq Expression	MAPRseq processed, CQN normalized counts	*	na
Mayo RNAseq TCX	Metadata	Individual human and RNAseq	syn3817650	na
Mayo RNAseq TCX	Metadata	Quality Control	syn6126114	na
ROSMAP	RNASeq Expression	Consensus processed RNASeq raw counts	syn8691134	10/2/2019
ROSMAP	Metadata	ID Key	syn3382527	10/2/2019
ROSMAP	Metadata	Individual human	syn3191087	10/2/2019

ROSMAP	Metadata	Assay RNAseq	syn21088596	1/2/2020
MSBB	RNASeq Expression	Consensus processed RNASeq raw counts	syn8691099	10/2/2019
MSBB	Metadata	Individual human	syn6101474	11/22/2019
MSBB	Metadata	Assay RNAseq	syn6100548	10/2/2019
snRNAseqPFC_BA10	Gene Expression	Count matrix	syn18686381	11/1/2019
snRNAseqPFC_BA10	Metadata	Column metadata	syn18686372	11/1/2019
snRNAseqPFC_BA10	Metadata	Gene row names	syn18686382	11/1/2019
snRNAseqPFC_BA10	Metadata	Biospecimen metadata	syn18642936	11/1/2910

DoD = Date of download, “na” indicates data that was generated by study authors and shared within the AMP-AD knowledge portal. *Data not yet available on the AMP-AD knowledge portal, but will be uploaded and shared upon acceptance for publication. SynapseIDs can be searched directly within the portal only (https://adknowledgeportal.synapse.org/). Scripts to perform the analysis and results reported here will be made available upon acceptance of the manuscript for publication.

Competing interests:

The authors declare that they have no competing interests.

Funding

This study was funded by the following studies: NIH/NIA U01 AG046139, RF1 AG051504 and R01 AG061796 (N.E-T.); P50 AG016574 Pilot grant (XW).

Author contributions:

X.W., S.L., M.A. and N.E-T conceived the idea. X.W., M.A. and N.E-T developed and designed the study, and wrote the manuscript. X.W. and S.L. developed the mathematical theory and X.W. performed the computations.D.D. provided tissue from the Mayo Clinic Brain Bank and neuropathologically characterized TCX-Mayo brain samples. T.N., K.M., S.L., M.A. and M.C. performed sample selection and library preparation of the TCX-Mayo dataset. M.A., Z.Q., T.C., T.P. and J.R. performed quality control of DLPFC and TCX-MSBB datasets. J.C. and Y.A. supervised the analytical aspects of the project. N.E-T. provided funding, supervision and direction for the whole study. All authors read and approved the final manuscript.

Acknowledgements:

We thank the patients and their families for their participation, without them these studies would not have been possible. The Mayo RNAseq study data was led by Dr. Nilüfer Ertekin-Taner, Mayo Clinic, Jacksonville, FL as part of the multi-PI U01 AG046139 (MPIs Golde, Ertekin-Taner, Younkin, Price) using samples from The Mayo Clinic Brain Bank. Data collection was supported through funding by NIA grants P50 AG016574, R01 AG032990, U01 AG046139, R01 AG018023, U01 AG006576, U01 AG006786, R01 AG025711, R01 AG017216, R01 AG003949, NINDS grant R01 NS080820, CurePSP Foundation, and support from Mayo Foundation. This study was in part funded by NIH RF1 AG051504 and R01

749 AG061796 (NET). We thank the Mayo Clinic Advanced Genomic Technology Center staff and
750 bioinformatics core facility for all gene expression measurements. XW thanks Dr. E. Aubrey
751 Thompson for text editing and Jeremy Burgess for provide insightful inputs. XW was funded in
752 part by a pilot grant from the Mayo Clinic ADRC (P50 AG016574).

753 The results published here are in whole or in part based on data obtained from the AMP-
754 AD Knowledge Portal (https://adknowledgeportal.synapse.org). Study data were provided by the
755 Rush Alzheimer's Disease Center, Rush University Medical Center, Chicago. Data collection
756 was supported through funding by NIA grants P30AG10161 (ROS), R01AG15819 (ROSMAP;
757 genomics and RNAseq), R01AG17917 (MAP), R01AG30146, R01AG36042 (5hC methylation,
758 ATACseq), RC2AG036547 (H3K9Ac), R01AG36836 (RNAseq), R01AG48015 (monocyte
759 RNAseq) RF1AG57473 (single nucleus RNAseq), U01AG32984 (genomic and whole exome
760 sequencing), U01AG46152 (ROSMAP AMP-AD, targeted proteomics), U01AG46161(TMT
761 proteomics), U01AG61356 (whole genome sequencing, targeted proteomics, ROSMAP AMP-
762 AD), the Illinois Department of Public Health (ROSMAP), and the Translational Genomics
763 Research Institute (genomic).

764 The results published here are in whole or in part based on data obtained from the AMP-
765 AD Knowledge Portal (https://adknowledgeportal.synapse.org/). These data were generated from
766 postmortem brain tissue collected through the Mount Sinai VA Medical Center Brain Bank and
767 were provided by Dr. Eric Schadt from Mount Sinai School of Medicine.

768

769 **Author's information:**

770 Not applicable.

771

References:

1. 2018 Alzheimer's disease facts and figures. *Alzheimer's & Dementia* **14**, 367-429 (2018).
2. Mehta, D., Jackson, R., Paul, G., Shi, J. & Sabbagh, M. Why do trials for Alzheimer's disease drugs keep failing? A discontinued drug perspective for 2010-2015. *Expert opinion on investigational drugs* **26**, 735-739 (2017).
3. Anderson, R.M., Hadjichrysanthou, C., Evans, S. & Wong, M.M. Why do so many clinical trials of therapies for Alzheimer's disease fail? *The Lancet* **390**, 2327-2329 (2017).
4. Allen, M. *et al.* Conserved brain myelination networks are altered in Alzheimer's and other neurodegenerative diseases. *Alzheimer's & Dementia* (2017).
5. McKenzie, A.T. *et al.* Multiscale network modeling of oligodendrocytes reveals molecular components of myelin dysregulation in Alzheimer's disease. *Molecular Neurodegeneration* **12**, 82 (2017).
6. Zhang, B. *et al.* Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer's disease. *Cell* **153**, 707-720 (2013).
7. Mostafavi, S. *et al.* A molecular network of the aging human brain provides insights into the pathology and cognitive decline of Alzheimer's disease. *Nature Neuroscience* **21**, 811-819 (2018).
8. De Jager, P.L. *et al.* A multi-omic atlas of the human frontal cortex for aging and Alzheimer's disease research. *Scientific Data* **5**, 180142 (2018).
9. Zhang, Y. *et al.* Purification and Characterization of Progenitor and Mature Human Astrocytes Reveals Transcriptional and Functional Differences with Mouse. *Neuron* **89**, 37-53 (2016).
10. Darmanis, S. *et al.* A survey of human brain transcriptome diversity at the single cell level. *Proceedings of the National Academy of Sciences* **112**, 7285-7290 (2015).
11. Lake, B.B. *et al.* Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. *Nature biotechnology* **36**, 70-80 (2018).
12. Allen, M. *et al.* Human whole genome genotype and transcriptome data for Alzheimer's and other neurodegenerative diseases. *Scientific Data* **3**, 160089 (2016).
13. Kuhn, A., Thu, D., Waldvogel, H.J., Faull, R.L.M. & Luthi-Carter, R. Population-specific expression analysis (PSEA) reveals molecular changes in diseased brain. *Nat Meth* **8**, 945-947 (2011).
14. Chikina, M., Zaslavsky, E. & Sealfon, S.C. CellCODE: a robust latent variable approach to differential expression analysis for heterogeneous cell populations. *Bioinformatics* **31**, 1584-1591 (2015).
15. McKenzie, A.T. *et al.* Brain Cell Type Specific Gene Expression and Co-expression Network Architectures. *Scientific Reports* **8**, 8868 (2018).
16. Zhong, Y., Wan, Y.-W., Pang, K., Chow, L.M. & Liu, Z. Digital sorting of complex tissues for cell type-specific gene expression profiles. *BMC Bioinformatics* **14**, 89 (2013).
17. Kuhn, A. *et al.* Cell population-specific expression analysis of human cerebellum. *BMC genomics* **13**, 610 (2012).
18. Wang, M. *et al.* The Mount Sinai cohort of large-scale genomic, transcriptomic and proteomic data in Alzheimer's disease. *Sci Data* **5**, 180185 (2018).
19. Mathys, H. *et al.* Single-cell transcriptomic analysis of Alzheimer's disease. *Nature* (2019).
20. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics* **25**, 25-29 (2000).
21. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).

817 22. Cocco, C. *et al.* Distribution of VGF peptides in the human cortex and their selective changes in
818 Parkinson's and Alzheimer's diseases. *Journal of anatomy* **217**, 683-93 (2010).
819 23. Futch, H.S., Croft, C.L., Truong, V.Q., Krause, E.G. & Golde, T.E. Targeting psychologic stress
820 signaling pathways in Alzheimer's disease. *Molecular neurodegeneration* **12**, 49 (2017).
821 24. Wang, R. & Reddy, P.H. Role of Glutamate and NMDA Receptors in Alzheimer's Disease. *J*
822 *Alzheimers Dis* **57**, 1041-1048 (2017).
823 25. Berchtold, N.C. *et al.* Brain gene expression patterns differentiate mild cognitive impairment
824 from normal aged and Alzheimer's disease. *Neurobiol Aging* **35**, 1961-72 (2014).
825 26. McAleese, K.E. *et al.* Parietal white matter lesions in Alzheimer's disease are associated with
826 cortical neurodegenerative pathology, but not with small vessel disease. *Acta neuropathologica*
827 **134**, 459-473 (2017).
828 27. Crivelli, S.M. *et al.* Sphingolipids in Alzheimer's disease, how can we target them? *Adv Drug Deliv*
829 *Rev* (2020).
830 28. Olsen, A.S.B. & Faergeman, N.J. Sphingolipids: membrane microdomains in brain development,
831 function and neurological diseases. *Open Biol* **7**(2017).
832 29. El Gaamouch, F. *et al.* VGF-derived peptide TLQP-21 modulates microglial function through
833 C3aR1 signaling pathways and reduces neuropathology in 5xFAD mice. *Molecular*
834 *neurodegeneration* **15**, 4-4 (2020).
835 30. Tzeng, T.-C. *et al.* Inflammasome-derived cytokine IL18 suppresses amyloid-induced seizures in
836 Alzheimer-prone mice. *Proceedings of the National Academy of Sciences of the United States of*
837 *America* **115**, 9002-9007 (2018).
838 31. White, C.S., Lawrence, C.B., Brough, D. & Rivers-Auty, J. Inflammasomes as therapeutic targets
839 for Alzheimer's disease. *Brain Pathol* **27**, 223-234 (2017).
840 32. Gao, T. *et al.* Transcriptional regulation of homeostatic and disease-associated-microglial genes
841 by IRF1, LXRβ, and CEBPα. *Glia* **67**, 1958-1975 (2019).
842 33. Deitmer, J.W., Theparambil, S.M., Ruminot, I., Noor, S.I. & Becker, H.M. Energy Dynamics in the
843 Brain: Contributions of Astrocytes to Metabolism and pH Homeostasis. *Frontiers in Neuroscience*
844 **13**(2019).
845 34. Bélanger, M., Allaman, I. & Magistretti, Pierre J. Brain Energy Metabolism: Focus on Astrocyte-
846 Neuron Metabolic Cooperation. *Cell Metabolism* **14**, 724-738 (2011).
847 35. Zhang, Q. *et al.* Integrated proteomics and network analysis identifies protein hubs and network
848 alterations in Alzheimer's disease. *Acta neuropathologica communications* **6**, 19-19 (2018).
849 36. Lovell, M.A., Xie, C., Gabbita, S.P. & Markesbery, W.R. Decreased thioredoxin and increased
850 thioredoxin reductase levels in alzheimer’s disease brain. *Free Radical Biology and Medicine* **28**,
851 418-427 (2000).
852 37. Kajiwara, Y. *et al.* GJA1 (connexin43) is a key regulator of Alzheimer's disease pathogenesis. *Acta*
853 *Neuropathol Commun* **6**, 144 (2018).
854 38. de Quervain, D.J. & Papassotiropoulos, A. Identification of a genetic cluster influencing memory
855 performance and hippocampal activity in humans. *Proc Natl Acad Sci U S A* **103**, 4270-4 (2006).
856 39. Shi, Y. *et al.* Rapid endothelial cytoskeletal reorganization enables early blood–brain barrier
857 disruption and long-term ischaemic reperfusion brain injury. *Nature Communications* **7**, 10523
858 (2016).
859 40. Stamatovic, S.M., Keep, R.F. & Andjelkovic, A.V. Brain endothelial cell-cell junctions: how to
860 "open" the blood brain barrier. *Current neuropharmacology* **6**, 179-192 (2008).
861 41. Evans, H.T., Benetatos, J., van Roijen, M., Bodea, L.G. & Gotz, J. Decreased synthesis of
862 ribosomal proteins in tauopathy revealed by non-canonical amino acid labelling. *EMBO J* **38**,
863 e101174 (2019).

42. Garcia-Esparcia, P. *et al.* Altered machinery of protein synthesis is region- and stage-dependent and is associated with alpha-synuclein oligomers in Parkinson's disease. *Acta Neuropathol Commun* **3**, 76 (2015).
43. Koren, S.A. *et al.* Tau drives translational selectivity by interacting with ribosomal proteins. *Acta Neuropathol* **137**, 571-583 (2019).
44. Conway, O.J. *et al.* ABI3 and PLCG2 missense variants as risk factors for neurodegenerative diseases in Caucasians and African Americans. *Molecular neurodegeneration* **13**, 53 (2018).
45. Srinivasan, K. *et al.* Untangling the brain's neuroinflammatory and neurodegenerative transcriptional responses. *Nature communications* **7**, 11295 (2016).
46. Lin, M.-Y. *et al.* Releasing Syntaphilin Removes Stressed Mitochondria from Axons Independent of Mitophagy under Pathophysiological Conditions. *Neuron* **94**, 595-610.e6 (2017).
47. Newberg, L.A., Chen, X., Kodira, C.D. & Zavodszky, M.I. Computational de novo discovery of distinguishing genes for biological processes and cell types in complex tissues. *PloS one* **13**, e0193067 (2018).
48. Li, Z. *et al.* Genetic variants associated with Alzheimer's disease confer different cerebral cortex cell-type population structure. *Genome medicine* **10**, 43 (2018).
49. Angulo, E. *et al.* Up-regulation of the Kv3.4 potassium channel subunit in early stages of Alzheimer's disease. *J Neurochem* **91**, 547-57 (2004).
50. Maezawa, I. *et al.* Kv1.3 inhibition as a potential microglia-targeted therapy for Alzheimer's disease: preclinical proof of concept. *Brain* **141**, 596-612 (2018).
51. Yi, M. *et al.* KCa3.1 constitutes a pharmacological target for astrogliosis associated with Alzheimer's disease. *Mol Cell Neurosci* **76**, 21-32 (2016).
52. Czubowicz, K., Jęśko, H., Wencel, P., Lukiw, W.J. & Strosznajder, R.P. The Role of Ceramide and Sphingosine-1-Phosphate in Alzheimer's Disease and Other Neurodegenerative Disorders. *Molecular neurobiology* **56**, 5436-5455 (2019).
53. Filippov, V. *et al.* Increased ceramide in brains with Alzheimer's and other neurodegenerative diseases. *Journal of Alzheimer's disease : JAD* **29**, 537-547 (2012).
54. Lee, J.-T. *et al.* Amyloid-beta peptide induces oligodendrocyte death by activating the neutral sphingomyelinase-ceramide pathway. *The Journal of cell biology* **164**, 123-131 (2004).
55. 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**, 479-86 (1991).
56. Hansen, K.D., Irizarry, R.A. & Wu, Z. Removing technical variability in RNA-seq data using conditional quantile normalization. *Biostatistics* **13**, 204-216 (2012).
57. Zhong, Y. & Liu, Z. Gene expression deconvolution in linear space. *Nat Meth* **9**, 8-9 (2012).