Potential multiple disease progression pathways in female patients with Alzheimer's disease inferred from transcriptome and epigenome data of the dorsolateral prefrontal cortex

Short title: Progression pathways of AD

Kousei Honda[1], Akinori Awazu[1,2,*]

[1] Graduate School of Integrated Sciences for Life, Hiroshima University, Higashihiroshima, Hiroshima 739-8526, Japan

[2] Research Center for the Mathematics on Chromatin Live Dynamics, Hiroshima University, Higashihiroshima, Hiroshima 739-8526, Japan

* Send correspondence to: Akinori Awazu; Email: awa@hiroshima-u.ac.jp

Keywords: Alzheimer's disease, transcriptome**,** epigenome

ABSTRACT

Late-onset Alzheimer's disease (AD) is a typical type of dementia for which therapeutic strategies have not yet been established. The database of the Rush Alzheimer's Disease study by the ENCODE consortium contains transcriptome and various epigenome data. Although the Rush AD database may contain a satisfactory amount of data for women, the amount of data for men remains insufficient. Here, based on an analysis of publicly available data from female patients, this study found that AD pathology appears to be nonuniform; several different typical forms of cognitive deterioration were observed. First, cluster analysis was performed on individuals diagnosed with "No Cognitive Impairment (NCI)," "Mild Cognitive Impairment (MCI)," and "Alzheimer's Disease (AD)" stages in clinical trials using gene expression. NCI was found to contain at least two substages, whereas MCI and AD contained four and six substages, respectively. The epigenome data, in particular genome-wide H3k4me3 distribution data, also supported the existence of multiple AD substages. However, *APOE* gene polymorphisms of individuals seemed to not correlate with disease stage. An inference of adjacency networks among substages, evaluated via partition-based graph abstraction using the gene expression profiles of individuals, suggested the possibility of multiple typical disease progression pathways from NCI to different AD substages through various MCI substages. These results will aid future research on the pathological features and mechanisms of AD and may induce changes in the strategic bases of conventional AD prevention and treatment.

INTRODUCTION

Alzheimer's disease (AD) is the most common cause of dementia, accounting for 60–70 % of dementia cases in older adults [1, 2]. AD can be classified into the familial type, which is strongly influenced by genetic factors, and the late-onset (sporadic) type, which is strongly influenced by aging and environmental factors, with the latter accounting for the majority of AD cases [3]. No treatment or preventive methods for AD have yet been established because the fundamental pathogenic mechanism of sporadic AD remains unclear.

Plaques of abnormally aggregated amyloid β (Aβ) are often observed in the brains of patients with AD and are considered to be related to the onset and progression of the disease [4-7]. However, cases have been reported in which AD developed without the accumulation of Aβ [8] and inhibiting Aβ production even worsened symptoms

(Francesca et al., 2023). Factors other than Aβ plaque formation, such as the abnormal aggregation of tau protein [9], the expression of apolipoprotein E variant APOE4 [10], herpes virus and coronaviruses (including SARS-CoV2) infections [11-14], diabetes [15, 16], brain inflammation [17], oxidative damage in cells [18-20], endoplasmic reticulum and ciliary dysplasia [21-24], myelin dysplasia [25, 26], and retrotransposons activities [27-29] have also been reported to influence the onset and progression of AD. Owing to the involvement of the aforementioned factors, the causes and symptoms of AD are thought to vary widely from patient to patient. Further, variations have been previously reported in symptoms and disease progression rates among patients [30-33].

In recent years, transcriptome and genome-wide analysis data of nucleotide polymorphisms have been collected from patients with AD, and various differentially expressed genes and patient-specific SNPs have been identified in healthy individuals and patients with AD [34-36]. Notably, these characteristics are not always seen equally across all studies. One reason for this is the variations in technical factors among studies. However, the diversity in molecular conditions and gene expression patterns in the brain among patients with AD could also be a major contributor [37-41].

Thus, AD progression is expected to exhibit various disease states. Classifying and characterizing these disease states and progression pathways is necessary to enable appropriate state-dependent treatment. In addition to a comparison of the features of the entire transcriptome, a comparison of the entire epigenome status among patients should be useful for such classification and characterization of disease diversity. The database of the Rush Alzheimer's Disease study by the ENCODE consortium (https://www.encodeproject.org/) [42, 43] contains transcriptome and various epigenome data such as histone modifications (H3K27ac, H3K27me3, and H3K4me3) and CCCTC-binding factor (CTCF) binding site distribution from individuals in the "No Cognitive Impairment (NCI)," "Mild Cognitive Impairment (MCI)," and "Alzheimer's Disease (AD)" groups that are available for unlimited public use (termed Rush AD database thereafter). Therefore, analyzing these publicly available data of patients with different cognitive impairment stages is expected to elucidate the diversity in the onset and progression of late-onset AD in more detail.

However, although the Rush AD database seemed to contain a satisfactory amount of data for women, the amount of data for men seemed insufficient, particularly for patients with MCI and AD, where the number of female patients in NCI, MCI, and AD

groups was 25, 29, and 28, respectively, whereas that of male patients was 14, 4, and 7, respectively. Therefore, in this study, the classification of disease status and AD progression pathways were performed using transcriptome and epigenome data of female patients in the Rush AD database as a first step of the analysis of the diversity of patients with AD using multiomics data. Our findings are expected to be useful not only for female patients but also for understanding symptoms in male patients, even though some features such as disease progression rate are different between male and female patients [44, 45].

In this study, patient diversity was considered by assessing the deviation from healthy individuals and diversity of disease progression. For this purpose, we first searched for differentially expressed gene sets between healthy individuals and patients. The gene sets searched here were those which contained the largest number of genes from gene sets that could actually distinguish healthy individuals from patients using cluster analysis. The candidate genes included therein were selected by identifying statistically significant differences in their expression levels. Second, we focused on the variety in the expression patterns of those gene sets between patients to show the existence of diverse conditions and subgroups within patients. The existence of these subgroups was also supported by comparisons of epigenomic data between individuals. Third, the relationship between subgroups and structure of the disease progression pathway was predicted based on the similarity in gene expression between subgroups.

RESULTS

Fundamental features of the analyzed transcriptome and epigenome data

RNA sequencing (RNA-seq) data (read count and TPM) of the dorsolateral prefrontal cortex of female individuals belonging to the NCI, MCI, and AD groups were acquired from the Rush AD database. In this dataset, the number of RNA-seq data of individuals in the NCI, MCI, and AD groups were 25, 29, and 28, respectively (Table 1, Table S1a–S1b).

To examine the quality and bias of analyzed data, we estimated one of the size factors of the RNA-seq data, that is, the median of the ratios (MOR) defined in the recent literature for DESeq [46], using read count data (Table S1a). The ratio [standard deviation of MOR over individuals]/[average MOR over individuals] was 0.1075. Similarly, we evaluated the first and third quartiles of the ratio (1OR and 3OR) between

the average and standard deviation, which were 0.1147 and 0.1006, respectively. These results suggested that the size factor would not considerably input bias among samples or affect the comparison of either the read count data or TPM values.

When examining the significance of the difference in gene expression levels between individuals aged <85 years and those >90 years, we did not detect any genes exhibiting a false discovery rate (FDR) < 0.1 (likelihood ratio test and two-tailed *t*-test) (Table S1a-S1b). This suggested that the variations in gene expression levels among the NCI, MCI, and AD groups was not due to the influence of age but rather due to differences in clinical stages.

In addition to RNA-seq data, the bigwig format data of fold change over control (FC) on genome-wide histone modifications (H3K4me3, H3K27ac, and H3K27me3) and CTCF binding site distribution obtained through chromatin immunoprecipitation sequencing (ChIP-seq) were acquired from the Rush AD database for 39 individuals: 11, 11, and 17 individuals from the NCI, MCI, and AD groups, respectively (Table S1c). In this study, the regions with FC > 5 were defined as the regions of peak intensities of these epigenome marks [47].

Cluster analysis of differentially expressed gene sets in the majority of patients with MCI

To extract a differentially expressed set of genes among patients with NCI, MCI, or AD, we performed the likelihood ratio test, which is one of the most familiar tests for estimating differentially expressed genes (DEGs) from read count data, for each gene using the DESeq2 software [48] (Table S1a). Consequently, we performed hierarchical clustering of individuals using genes with P-values below the threshold value P^{D}_{NMA} = 0.01, 0.02, 0.03, 0.04, and 0.05 (Figs. 1 and S1a–S1e). When $P^{D}_{NMA} \leq 0.05$, a cluster with significantly more patients with MCI (MCI cluster) and a cluster with patients with NCI, AD, and a small number of patients with MCI was formed via the first divergence of the dendrogram ($P < 0.05$, Fisher's exact test; Fig. S1a–S1e). However, even though the gene set for clustering contained only genes with smaller P values, patients with NCI, MCI, and AD were not from their respective cluster; no clusters comprising solely of patients with NCI, MCI, or AD were formed through the first divergence of the dendrogram.

161 We then performed the same analysis using analysis of variance (ANOVA), which is
162 the most familiar test for evaluating differences of characteristic values among three or
163 more groups, to determine the P-values using TPM values (Table S1b), where P^{A}_{NMA}
164 was defined instead of P^{D}_{NMA}. Herein, ANOVA was generally better for detecting
165 differences and was expected to identify more candidate genes in a gene set than the
166 likelihood ratio test. Consequently, an MCI cluster and a cluster with patients with NCI,
167 AD, and a small number of patients with MCI were formed, as in the previous case (Fig.
168 S1f-S1j). Additionally, when the P^{A}_{NMA} was 0.04, a cluster comprising solely patients
169 with MCI was formed through the first divergence of the dendrogram (Fig. 1a). In
170 contrast, we did not observe any clear separation between patients with NCI and those
171 with AD, regardless of the P^{A}_{NMA} value.
172
173 Notably, we did not obtain any gene set when applying the P^{D}_{NMA} and P^{A}_{NMA} values
174 that enabled the complete discrimination of patients with NCI, MCI, and AD. In
175 addition, depending on the P^{D}_{NMA} and P^{A}_{NMA} values, the patients with MCI who
176 deviated from the MCI cluster varied; however, 24 patients with MCI were always
177 included in the MCI cluster (Fig. S1). This suggests that the MCI cluster consisted of 24
178 patients with MCI, named MCI_{Major} individuals, whereas the 5 other patients with MCI
179 (M3, M12, M14, M16, and M28) who tended to deviate from the MCI_{Major} individuals,
180 were named MCI_{NCI_AD} individuals. We assumed that the MCI_{Major} and MCI_{NCI_AD}
181 individuals were classified to the substages named MCI_{Major} and MCI_{NCI_AD},
182 respectively.
183
184 Moreover, the set of genes with P-values of the above ANOVA $\leq P^{A}_{NMA} = 0.04$ was
185 named the "M-NA set" as it significantly differed between the majority of patients with
186 MCI from those with NCI and AD (Fig. 1b, Table S2a). In a 3-dimensional U-map
187 using the M-NA set, we clearly observed the deviation of the MCI_{Major} individuals from
188 those at other substages, with MCI_{NCI_AD} detected in a mixed group of patients with NCI
189 and AD (Fig. 1c).
190
191 The classification result of patients with MCI into MCI_{Major} and MCI_{NCI_AD} individuals
192 seemed to be independent of the age of individuals. For example, the numbers of
193 patients over and under 90 years old in MCI_{Major} were 19 and 5, whereas those in
194 MCI_{NCI_AD} were 3 and 2, respectively (P-value of the Fisher's exact test = 0.5688;
195 Table S1a–S1b).
196

197 **Functional features of genes in the M-NA set**
198 We also considered the functional features of genes in the M-NA set. In this set, we
199 obtained 594 genes (290 upregulated and 304 downregulated genes in MCI_{Major}); of
200 them, 349 genes (164 upregulated and 185 downregulated in MCI_{Major}) were annotated
201 (Fig. 1b) as protein coding or noncoding RNA genes using Entrez [49]. Herein, both
202 sets of upregulated and downregulated genes in MCI_{Major} contained some previously
203 reported AD risk-associated genes; in particular, the upregulated gene set contained
204 *CYP19A1, GSTM1, HSD11B1, IL10, MT-RNR1, PCK1, TRAK2, UBE2D1*, and *XBP1*,
205 whereas the downregulated gene set contained *MT-ATP8* and *NME8* [50].
206
207 In addition, we performed enrichment analysis of all genes including genes with higher
208 and lower expression levels in MCI_{Major} individuals (Table S2b–S2d). We identified that
209 genes with higher expression levels in MCI_{Major} individuals were enriched in GO terms
210 for the functions "response to lipopolysaccharide," "response to molecule of bacterial
211 origin," "Toll-like receptor signaling pathway," "regulation of MHC class II
212 biosynthetic process," "vitamin metabolic process," "water-soluble vitamin metabolic
213 process," "chemotaxis," and "taxis" (P-value $< 10^{-3}$ and FDR < 0.5) (Table S2c).
214 Whereas, we identified that genes with lower expression levels in MCI_{Major} individuals
215 were enriched in GO terms for the functions "mRNA 5′-splice site recognition" and
216 "mRNA cis splicing, via spliceosome" (P-value $< 10^{-3}$ and FDR < 0.5) (Table S2d).
217 Recent studies have suggested that some of these functional features are characteristic
218 of AD pathology; for example, the lipopolysaccharide found in the wall of all Gram-
219 negative bacteria was reported to play a role in causing AD [51, 52], Toll-like receptors
220 were found to participate in neuronal cell death by apoptosis [53, 54], MHC class II-
221 expressing cells were suggested to be involved in the degradation of neurons [55],
222 vitamins were reported to affect the generation and clearance of Aβ [56], and microglia
223 were shown to exhibit chemotaxis for Aβ clearance [57, 58].
224
225 **Differentially expressed gene sets between patients with NCI and AD**
226 To extract a differentially expressed gene set between patients with NCI and those with
227 AD, we performed hierarchical clustering of individuals using the genes satisfying the
228 following criteria: mean expression level (TPM) was higher in patients with AD than in
229 those with NCI, with smaller P-values in the likelihood ratio test than the P^D_{NA}
230 threshold, or mean expression level (TPM) was lower in patients with AD than in those
231 with NCI, with smaller P-values in the likelihood ratio test than the P^D_{AN} threshold.
232 Subsequently, we searched the sets of P^D_{NA} and P^D_{AN} values that divided the patients

with NCI and those with AD into different clusters in the first branch of the dendrogram. Within the range $P^D_{NA} \leq 0.05$ and $P^D_{AN} \leq 0.05$, we identified the finite but narrow area of P^D_{NA} and P^D_{AN}, $P^D_{NA} = 0.01$ and $P^D_{AN} = 0.02$, $P^D_{NA} = 0.01$ and $P^D_{AN} = 0.03$, $P^D_{NA} = 0.02$ and $P^D_{AN} = 0.03$, $P^D_{NA} = 0.03$ and $P^D_{AN} = 0.03$, and $P^D_{NA} = 0.05$ and $P^D_{AN} = 0.02$, where patients with NCI were completely separated from those with AD (Fig. 2a). Thus, we obtained the largest set of genes separating the NCI and AD groups at $P^D_{NA} = 0.05$ and $P^D_{AN} = 0.02$ (Fig. S2a, Table S3a); this gene set was named the N-A_{D1} set. Notably, the number of genes with a lower mean expression level (TPM) in patients with AD than in those with NCI was larger when $P^D_{NA} = 0.03$ and $P^D_{AN} = 0.03$ than when $P^D_{NA} = 0.05$ and $P^D_{AN} = 0.02$ (Fig. S2b, Table S3b). Notably, this gene set contained all genes in other gene sets with P^D_{NA} and P^D_{AN} smaller than 0.03 that could separate the NCI and AD groups. Thus, we focused on the gene set obtained at $P^D_{NA} = 0.03$ and $P^D_{AN} = 0.03$ and named it the "N-A_{D2} set."

We performed the same analysis using two-tailed *t*-tests to determine the P-values, defining the P^t_{NA} and P^t_{AN} instead of the P^D_{NA} and P^D_{AN}, respectively. Herein, the *t*-test was generally better for detecting differences and was expected to identify more candidate genes in a gene set than the likelihood ratio test. Searching in the ranges $P^t_{NA} \leq 0.05$ and $P^t_{AN} \leq 0.05$, patients with NCI were completely separated from those with AD in 12 of 15 cases of P^t_{NA} and P^t_{AN} set for $P^t_{NA} \leq P^t_{AN}$ (Fig. 2a). When we searched the wider range $P^t_{NA} \leq 0.1$ and $P^t_{AN} \leq 0.1$, patients with NCI were completely separated from those with AD in 37 of 55 cases of P^t_{NA} and P^t_{AN} set for $P^t_{NA} \leq P^t_{AN}$ (Fig. 2a). We thus obtained the largest set of genes separating the NCI and AD groups at $P^t_{NA} = 0.08$, and $P^t_{AN} = 0.1$ (Fig. S2c, Table S3c). This gene set was named the "N-A_t set."

Hierarchical clustering using a combination of the expression levels of all genes in the N-A_t and N-A_{D1} sets also separated patients with NCI from those with AD (Fig. 2b, Table S3d). Conversely, the expression of all genes in the N-A_t and N-A_{D2} sets did not separate patients with NCI from those with AD. However, clustering using a combination of the expression levels of genes in the N-A_t set and a part of the N-A_{D2} set with low P-values completely separated patients with NCI from those with AD (Fig. 2c, Table S3e). Such gene sets are expected to involve various typical differentially expressed genes between patients with NCI and those with AD. In this study, the gene set consisting of the N-A_t and N-A_{D1} sets, which could completely separate patients with NCI from those with AD, was named the "N-A1 set" (Table S3d), whereas the gene set

consisting of the N-A_t set and maximum number of genes in the N-A_{D2} set with low P-values was named the "N-A2 set" (Table S3e).

Patients with AD can be classified into three groups

From the results of the hierarchical clustering for patients with NCI and AD performed using the N-A1 set, we noticed the emergence of three subgroups of patients with AD (Figs. 2b and 3a). Thus, we performed ANOVA on the expression levels of genes that were higher in patients with AD than in those with NCI among the three assumed subgroups (1020 genes). We identified 367 genes whose expression levels differed significantly (FDR < 0.1) among the 3 subgroups. Performing k-means cluster analysis of the expression levels of these genes in patients with AD, we identified two characteristic clusters: a cluster consisting of 96 genes the expression levels of which were low in only one subgroup (Cluster CA1) and another consisting of 271 genes the expression levels of which were high in only one subgroup (Cluster CB1) (Fig. 3a).

Thereafter, we performed a similar analysis for patients with NCI. Based on the shape of the dendrogram, patients with NCI were divided into three or four subgroups (Fig. 2b, Fig. 3a). However, when patients with NCI were divided into three subgroups, ANOVA revealed only nine genes with FDR < 0.1 among the three subgroups (Table S3d). Moreover, when patients with NCI were divided into four subgroups, ANOVA revealed only two genes with FDR < 0.1 among the four subgroups (Table S3d). Based on these values, the number of genes in clusters CA1 and CB1 was sufficiently large.

Collectively, these results suggested that the gene expression patterns in patients with AD, which can be classified into three types, were highly nonuniform. Patients with AD exhibiting high gene expression levels in both clusters, CA1 and CB1, were classified to the AD_{HH} subgroup (named AD_{HH} individuals), those exhibiting low gene expression levels in both CA1 and CB1 were classified in the AD_{LL} subgroup (named AD_{LL} individuals), whereas those exhibiting high gene expression levels in cluster CA1 and low gene expression levels in cluster CB1 were classified in the AD_{HL} subgroup (named AD_{HL} individuals).

Hierarchical clustering of patients with NCI and AD using the N-A2 set also revealed three possible subgroups (Figs. 2c and 3b). By performing the same analysis described above, we identified gene clusters similar to cluster CA1 containing 86 genes (cluster CA2) and cluster CB1 containing 234 genes (cluster CB2). Accordingly, patients with

304 AD were divided into AD_{HH}, AD_{LL}, and AD_{HL} groups based on the expression profiles
305 of genes in clusters CA2 and CB2. The members of the AD_{HH}, AD_{LL}, and AD_{HL} groups
306 were the same in both the N-A1 and N-A2 sets (Fig. S3a).
307
308 Importantly, variations were observed among the AD_{HH}, AD_{HL}, and AD_{LL} individuals
309 when different P^t_{NA} and P^t_{AN} values were employed to determine the genes in the N-At
310 set. However, for small changes in the P^t_{NA} and P^t_{AN} values, less variation in individuals
311 was observed among subgroups; for example, when using $P^t_{NA} = 0.06$ and $P^t_{AN} = 0.1$
312 instead of $P^t_{NA} = 0.08$ and $P^t_{AN} = 0.1$ to determine the N-At. Herein, the gene set named
313 N-A3 was defined as that consisting of genes in the $N\text{-}A_t$ set and maximum number of
314 genes in $N\text{-}A_{D1}$ set with low P-values that could completely separate patients with NCI
315 from those with AD, whereas the gene set named N-A4 was defined as that consisting
316 of genes in the $N\text{-}A_t$ set and maximum number of genes in $N\text{-}A_{D4}$ set with low P-values
317 that could completely separate patients with NCI from those with AD. The classification
318 of patients with AD into the AD_{HH}, AD_{LL}, and AD_{HL} groups was similar to that obtained
319 for N-A1, N-A2, N-A3, and N-A4, where 24 of the 28 individuals belonged to the same
320 subgroups (Fig. S3). This suggested the robustness of our method in classifying the
321 AD_{HH}, AD_{HL}, and AD_{LL} individuals.
322
323 In addition, the classification result of patients with AD into AD_{HH}, AD_{HL}, and AD_{LL}
324 individuals seemed to be independent of the ages of individuals. For example, the
325 numbers of individuals aged over and under 90 years in AD_{HH} were eight and two, those
326 in AD_{LL} were six and four, and those in AD_{HL} were four and four, respectively (P-value
327 = 0.4850).
328
329 **Common functional features of the N-A1 and N-A2 sets**
330 We also considered the common functional features of the N-A1 and N-A2 sets. The N-
331 A1 set contained 1477 genes (1020 upregulated and 457 downregulated, AD), with 933
332 genes (683 upregulated and 250 downregulated) annotated as protein coding or
333 noncoding RNA genes using Entrez [49] (Fig. 3a). The N-A2 set contained 1424 genes
334 (947 upregulated and 477 downregulated, AD), with 889 genes (619 upregulated and
335 270 downregulated) annotated as protein coding or noncoding RNA genes using Entrez
336 [49] (Fig. 3b). The upregulated genes in both the N-A1 and N-A2 set contained *AR*,
337 *CBS*, *CLU*, *CYP19A1*, *DBH*, *GNA11*, *HLA-A*, *IQCK* , *MPO*, *MS4A4E*, *MX1*, *NLRP3*,
338 *SDC2*, *TAP2*, and *ZFHX3* (Table S3d–e), all of which have been reported to be AD
339 risk-associated genes in recent studies [50, 59-61]. Similarly, the downregulated genes

340 in both the N-A1 and N-A2 set contained *COL25A1* and *TARDBP* (Table S3d–e), which
341 have also been reported to be AD risk-associated genes [50].
342
343 Additionally, the clusters CA1, CB1, CA2, and CB2 contained 71, 215, 64, and 181
344 annotated genes, respectively (Fig. 3). We identified *CLU*, *SDC2*, and *ZFHX3* in both
345 the CA1 and CA2 clusters, whereas *DBH*, *GNA11*, *IQCK*, *MPO*, and *NLRP3* were
346 identified in both the CB1 and CB2 clusters (Table S3d–e), thereby suggesting that
347 these genes are responsible for the differences in the features among the three AD
348 subgroups [50, 61].
349
350 Next, we performed enrichment analysis of all genes, including those with higher and
351 lower expression levels in patients with AD for both the N-A1 and N-A2 set (Table
352 S3f–S3o). We identified that genes with higher expression levels in patients with AD in
353 both the N-A1 and N-A2 set were commonly enriched in GO terms for "forebrain
354 dorsal/ventral pattern formation," "mRNA 5′-splice site recognition," and "negative
355 regulation of NF-kappaB transcription factor activity" (P-value < 10^{-3} and FDR < 0.5)
356 (Tables S3g and S3j). Whereas, genes with lower expression levels in patients with AD
357 in both the N-A1 and N-A2 set were enriched in GO terms for "negative regulation of
358 intrinsic apoptotic signaling pathway in response to DNA damage by p53 class
359 mediator," "negative regulation of DNA damage response, signal transduction by p53
360 class mediator," and "coronary vasculature morphogenesis" (P-value < 10^{-3} and FDR <
361 0.5) (Table S3h, S3k). In the nervous system, NF-κB is known as one of the crucial
362 components in the molecular switch that converts short- to long-term memory [62].
363 Thus, the upregulation of genes that suppress NF-κB expression is expected to
364 exacerbate AD pathology. In addition, the downregulation of genes that suppress
365 apoptosis in response to DNA damage would also promote AD progression through the
366 enhancement of cell death in the central nervous system [63, 64].
367
368 We found that genes in both the CA1 and CA2 clusters were enriched in GO terms for
369 "response to xenobiotic stimulus", "cellular response to lipid", "negative regulation of
370 small molecule metabolic process", "lung development", "cellular response to external
371 stimulus", "negative regulation of carbohydrate metabolic process", "neuron migration",
372 and "positive regulation of epithelial cell differentiation" (P-value < 10^{-3} and FDR <
373 0.5) (Tables S3l and S3n). Whereas, genes in both the CB1 and CB2 clusters were
374 enriched in GO terms for "mitotic spindle assembly checkpoint signaling" and "L-alpha-
375 amino acid transmembrane transport" (P-value < 10^{-3} and FDR < 0.5) (Tables S3m and

376 S3o). Notably, recent studies have suggested that the perturbed cerebral carbohydrate
377 metabolism contributes to AD pathogenesis [65, 66], whereas neuron migration ability
378 contributes to neurogenesis in the brain [67, 68]. These results suggested that the three
379 AD subgroups may differ in these functional features.
380
381 We also evaluated the chromosomal distribution of genes belonging to the CA1 or CB1
382 cluster in N-A1 set, and that of those belonging to the CA2 or CB2 clusters in N-A2 set
383 (Fig. 4). Compared with the chromosomal distribution of all human coding and
384 noncoding genes, both gene sets belonging to the CA1 or CB1 clusters and CA2 or CB2
385 clusters seemed to be accumulated at Chr. 1, 7, 9, 16, 19, and 22 (Fig. 4).
386
387 **Epigenome data analysis supported the classification of patients with AD into three**
388 **groups**
389 The data from the dorsolateral prefrontal cortex of the Rush AD database analyzed in
390 the present study also included information on the genome-wide distribution of
391 epigenomic markers such as histone modifications (H3K4me3, H3K27me3, and
392 H3K27ac) and CTCF binding site distribution for approximately half of all individuals
393 (Table S1). Hence, to classify these individuals based on features other than
394 transcriptome, we next evaluated the similarity of such epigenomic features among
395 individuals using Jaccard coefficients of the genome-wide distribution of previously
396 defined peak regions of epigenome status. At the first divergence of the dendrogram
397 according to hierarchical clustering using Jaccard coefficients, individuals were divided
398 into two groups: a group with a typical epigenome status distribution and another one
399 with an untypical distribution for the respective epigenomic markers (Figs. 5–6 and S4).
400 The Jaccard coefficients of peak regions of epigenome status among individuals
401 suggested that more AD_{HL} individuals exhibited untypical patterns compared with those
402 in other patients with AD (Figs. 5–6 and S4, Table 2); in particular, the AD_{HL} group had
403 a significantly larger number of patients with H3K4me3 modifications compared with
404 the AD_{HH} and AD_{HL} groups (Table 2a). Similar features were obtained for H3K27ac,
405 H3K27me3, and CTCF (Table 2b–2d); however, these results were not significant.
406 Moreover, AD_{HL} individuals tended to exhibit low epigenetic peak region density (Figs.
407 5–6 and S4).
408
409 Further clustering revealed that individuals exhibiting typical H3K4me3 status were
410 further divided into two groups, named Typical 1 and Typical 2 (Fig. 5). Based on their
411 specific epigenome features, more AD_{HH} individuals belonged to the Typical 2

412 subgroup compared with other patients with AD (Table 3e). These facts suggest that the
413 AD_{HH}, AD_{HL}, and AD_{LL} individuals exhibited differences in the genome wide
414 distribution of H3K4me3 modifications.
415
416 **No effect of *APOE* gene polymorphisms in classifying individuals**
417 To examine the relationships between the *APOE* genotype and the subgroup to which
418 individuals belong, the *APOE* genotype of each individual was identified using the
419 RNA-seq read data from the dorsolateral prefrontal cortex of individuals in the Rush
420 AD database. Accordingly, no individuals were found to simultaneously possess the ε2
421 and ε4 alleles. Therefore, individuals were classified in three categories: APOE2 carrier,
422 possessing one or two ε2 alleles; APOE3 carrier, possessing two ε3 alleles; and APOE4
423 carrier, possessing one or two ε4 alleles. Notably, the number of APOE2, APOE3, and
424 APOE4 carriers was not correlated with the clinical diagnostic stage of individuals
425 (Tables 3a and S1a–b). In addition, the number of individuals also seemed to be
426 independent of the two MCI (MCI_{Major} and MCI_{NCI_AD}) and three AD (AD_{HH}, AD_{HL},
427 and AD_{LL}) substages (Table 3b–c, Table S1a–b).
428
429 **Adjacency network among substages**
430 The data on the expression levels of genes in the M-NA set revealed differences
431 between patients with typical MCI, NCI, and AD, whereas those in the N-A1 or N-A2
432 sets revealed differences between patients with NCI and AD. To clarify the
433 relationships among disease substages and the pathways of disease progression in these
434 substages, we explored the adjacency networks among substages using M-NA and N-
435 A1 sets, and M-NA and N-A2 sets by performing the partition-based graph abstraction
436 (PAGA) method [69]. Using PAGA, we generated a graph of substages in which the
437 edge weights represented connection confidence. Notably, connection confidence
438 between NCI and AD substages often involved finite values (Table S4); however, the
439 direct connection between NCI and AD substages seemed unnatural in disease
440 progression. Therefore, we assumed that only connections with confidence values
441 sufficiently larger than the maximum confidence values of the connection between NCI
442 and AD substages (MD_{N_A}) were effective for the adjacency network among substages
443 (Fig. S5). Subsequently, we used connections with confidence values larger than twice
444 that of MD_{N_A} to construct the adjacency network (Table S4).
445
446 Changes in gene expression patterns that characterize disease states are considered to
447 occur gradually as the disease progresses. Therefore, we observed that disease appeared

to progress from NCI to AD along the previously obtained adjacency network. Note that the adjacency network obtained using the M-NA and N-A1 sets (Fig S5a) and that constructed using the M-NA and N-A2 sets (Fig. S5b) exhibited a slightly different structure because of differences in the contained genes in these sets. Thus, we focused on the network constructed using common connections obtained from both networks constructed using the M-NA and N-A1 sets (Fig S5a) and M-NA and N-A2 sets (Fig. S5b) (Fig. 7). We did not obtain any connections from NCI to MCI_{Major}, the major group of MCI individuals, in any of the networks (Figs. 7 and S5a–b), which seemed unnatural for AD progression. Thus, we performed more detailed subgroup classifications of patients with NCI, MCI, and AD to reveal the relationship between NCI and MCI_{Major} individuals.

Detailed classification of patients with MCI by comparing between patients with NCI and those with MCI and between patients with MCI and those with AD

To classify patients with MCI more comprehensively, we conducted comparisons between patients with NCI and those with MCI and between patients with MCI and those with AD. First, we performed hierarchical clustering of individuals using the genes satisfying the following criteria: the mean expression level (TPM) was higher in the MCI group than that in the NCI group with smaller P-values in the likelihood ratio test than the P^D_{MN} threshold or the mean expression level (TPM) was lower in the MCI group than that in the NCI group, with smaller P-values in the likelihood ratio test than the P^D_{NM} threshold. We performed cluster analysis using each gene set within the range $P^D_{NM} \leq 0.05$ and $P^D_{MN} \leq 0.05$. However, we did not identify any cases in which patients with NCI were completely separated from those with MCI (Figs. 8a and S6a). However, clusters with significantly more patients with NCI (NCI clusters) and MCI clusters were formed within all cases of P^D_{NM} and P^D_{MN} except in the case of $P^D_{NM} = 0.05$ and $P^D_{MN} = 0.01$ (Fig. 8a).

When we used two-tailed *t*-tests instead of the likelihood ratio test to determine the P-values, our results were qualitatively the same. P^t_{MN} and P^t_{NM} were defined instead of P^D_{MN} and P^D_{NM}, respectively. Analysis using each gene set within the ranges $P^t_{MN} \leq 0.05$ and $P^t_{NM} \leq 0.05$ revealed no cases in which patients with NCI were completely separated from those with MCI (Figs. 8a and S6a). However, NCI and MCI clusters were formed within all cases of P^t_{NM} and P^t_{MN} (Fig. 8a).

Next, we performed the same analyses for patients with MCI and those with AD. In this case, P^D_{MA}, P^D_{AM}, P^t_{MA}, and P^t_{AM} were given as values corresponding to P^D_{NM}, P^D_{MN}, P^t_{NM}, and P^t_{MN}, respectively. Similar to the above results, we did not identify any cases in which patients with MCI were completely separated from those with AD (Fig. 8b and S6b). However, MCI and AD clusters were formed within all cases of P^D_{MA}, P^D_{AM}, P^t_{MA}, and P^t_{AM} except in the case of $P^D_{MA} = 0.05$ and $P^D_{AM} = 0.01$ and that of $P^t_{MA} = 0.05$ and $P^t_{AM} = 0.01$ (Fig. 8b).

In this study, variations were observed in patients with MCI who deviated from the MCI cluster depending on P^D_{NM}, P^D_{MN}, P^t_{NM}, P^t_{MN}, P^D_{MA}, P^D_{AM}, P^t_{MA}, and P^t_{AM} values. For each patient with MCI, we counted the number of times they belonged to the NCI cluster upon changing the P^D_{NM}, P^D_{MN}, P^t_{NM}, and P^t_{MN} values, and those belonging to the AD cluster upon changing the P^D_{MA}, P^D_{AM}, P^t_{MA}, and P^t_{AM} values (Fig. 8c). Consequently, individuals M3, M4, M5, M8, M12, M13, M14, M16, M28, and M29 belonged to the NCI cluster more than once, whereas individuals M3, M5, M9, M12, M14, M16, M23, M27, and M28 belonged to the AD cluster more than once.

Notably, individuals M3, M12, M14, M16, and M28, (same as those in MCI_{NCI_AD}), clearly exhibited a greater number of deviations from patients with NCI and AD than MCI_{Major} individuals (Fig. 8c). To classify MCI_{Major} individuals in more detail, we performed k-means clustering using the number of times each individual belonged to the NCI cluster and that belonging to the AD cluster (Figs. 8c and S6c). As a result, MCI_{Major} individuals were divided into three subtypes: $MCI_{Typical}$ individuals (17 individuals) who rarely or never belonged to other clusters; MCI_{NCI} individuals (4 individuals) who may belong to NCI clusters only; and MCI_{AD} individuals (3 individuals) who may belong to AD clusters only (Table S1a–S1b). Thus, MCI was divided into four substages: $MCI_{Typical}$, MCI_{NCI}, MCI_{AD}, and MCI_{NCI_AD}.

Detailed classification of patients with NCI by comparing them with patients with MCI

Similar analysis showed that individuals N11, N12, N18, N19, N21, and N22 belonged to the MCI cluster more than once when the P^D_{NM}, P^D_{MN}, P^t_{NM}, and P^t_{MN} values were changed (Fig. 9a). After k-means clustering of patients with NCI using the number of times each individual belonged to the MCI cluster, patients with NCI were divided into two clusters: one comprising N11, N21, and N22, and another comprising those who never belonged to the MCI cluster (Figs. 9a and S7a). Therefore, these individuals were

classified as typical patients with NCI belonging to the NCI cluster, named $NCI_{Typical}$ individuals, who were assumed to be in the $NCI_{Typical}$ substage. Individuals N12, N18, and N19 were classified as patients with NCI close to MCI, named NCI_{MCI} individuals, who were assumed to be in the NCI_{MCI} substage (Table S1a–S1b).

Detailed classification of patients with AD by comparing them with patients with MCI

Similarly, 17 patients with AD belonged to the MCI cluster more than once, when the P^{D}_{MA}, P^{D}_{AM}, P^{t}_{MA}, and P^{t}_{AM} values were changed (Fig. 9b). Using k-means clustering of patients with AD based on the number of times each individual belonged to the MCI cluster, patients with AD were divided into three clusters, with individuals A10, A14, A15, A17, A18, A19, A22, and other patients with AD who never belonged to the MCI cluster being included in the same cluster (Figs. 9b and S7b). Therefore, these individuals were expected to be in one of the typical AD substages: AD_{HH}, AD_{HL}, or AD_{LL} (Table S1a–S1b). In contrast, individuals A2, A3, A4, A6, A7, A8, A9, A11, A20, and A21 were assumed to be at one of the AD substages close to MCI and were named AD_{HH_MCI}, AD_{HL_MCI}, or AD_{LL_MCI} (Table S1a–S1b).

Adjacency network among substages obtained after detailed classifications of individuals suggests multiple disease progression pathways from NCI to AD

Detailed classifications of individuals suggested that patients with NCI were classified into $NCI_{Typical}$ and $NCI_{MCI,}$ patients with MCI were classified into $MCI_{Typical}$, MCI_{NCI}, MCI_{AD}, and $MCI_{NCI_AD,}$ and patients with AD were classified into AD_{HH}, AD_{HH_MCI}, AD_{HL}, AD_{HL_MCI}, AD_{LL}, and AD_{LL_MCI}. We constructed adjacency networks among these 12 substages using gene expression data from the M-NA and N-A1 sets (Fig. S8a, Table S5a) as well as from the M-NA and N-A2 sets (Fig. S8b, Table S5b) as previously described. Subsequently, we focused on the adjacency network that consisted of common connections between the network from the M-NA and N-A1 sets, and that from the M-NA and N-A2 sets (Fig. 10). We identified more than four progression paths from NCI to AD that passed through different substages, as follows: I) from NCI ($NCI_{Typical}$ or NCI_{MCI}) to AD_{LL} through MCI_{NCI_AD}, II) from NCI to MCI_{NCI} and from MCI_{NCI} to AD_{LL} through AD_{HH_MCI} and AD_{LL_MCI}, respectively, III) from NCI to MCI_{NCI} and from MCI_{NCI} to AD_{HH} through $MCI_{Typical}$ or AD_{HH_MCI}, and IV) from NCI to AD_{HL} or AD_{LL} via MCI_{NCI}, MCI_{AD}, or AD_{HL_MCI}. This suggested that at least four disease progression pathways are active in the three classified AD substages.

555 **Comparisons with the findings of recent studies**
556 Many recent studies have performed transcriptome analyses and identified various
557 significant differentially expressed genes (significant DEGs) between healthy (control)
558 individuals and patients with AD, which they termed "AD risk genes" [50, 59-61]. The
559 cluster analysis of patients with NCI, MCI, and AD and that of patients with NCI and
560 AD was performed using these AD risk genes (Fig. S9, Table S6). However, these
561 results differed from the individual clustering results in Figures 1, 2, and 3. Notably,
562 most of the AD risk gene data used here were obtained from research on the
563 hippocampus, whereas the data analyzed in the previous sections of this study were
564 obtained from the prefrontal cortex. A recent study showed that gene expression
565 patterns differ between the hippocampus and prefrontal cortex [39]. Hence, the
566 observed inconsistencies were attributed to these differences.
567
568 From the dataset analyzed in the previous sections, no significant DEGs were detected
569 between patients with NCI and those with AD, whereas only nine significant DEGs
570 were identified among patients with NCI, MCI, and AD (Tables S1a–S1b and S7).
571 Herein, significant DEGs were evaluated via the likelihood test using DEseq2 with FDR
572 < 0.01. Notably, the abovementioned study reported that not many significant DEGs
573 were yielded in areas belonging to the prefrontal cortex [39]. These facts supported
574 the consistency between the results of the present and those of recent
575 studies.
576
577
578 **DISCUSSION**
579 In this study, the analysis of data from women individuals in the Rush AD database
580 suggested that the pathological progression pathway of AD is nonuniform, exhibiting
581 different typical forms of deterioration. First, cluster analysis using gene expression data
582 from individuals diagnosed with NCI, MCI, and AD suggested that NCI can be
583 classified in at least two substages, whereas MCI and AD can be classified in four and
584 six substages, respectively. Differences among AD substages were identified based on
585 epigenomic features, in particular the genome-wide distribution of H3K4me3. However,
586 the *APOE* gene polymorphisms of individuals appeared to not correlate with the disease
587 substage. In addition, some of the genes and gene sets that were suggested in recent
588 studies to be AD-associated contributed to the characterization of each AD substage in
589 the present study. Second, an inference of adjacency networks among the different
590 substages suggested the existence of four typical NCI → AD disease progression

591 pathways from different MCI to different AD substages. These findings on disease
592 progression features will facilitate further advancements in AD research. Moreover,
593 they are expected to induce changes in the strategic bases of AD prevention, treatment,
594 and studies for biomarkers [70-72].
595
596 Recent studies using hippocampal and frontal lobe transcriptomes have also suggested
597 the existence of multiple AD subtypes. However, the relationship between multiple
598 states, such as their connection in a series on the same disease progression path or on
599 different paths, has not been clarified. Furthermore, these studies did not focus on the
600 diversity of NCI and MCI. In contrast, in this study, both patients with NCI and MCI
601 were divided into multiple subgroups at different substages based on their respective
602 gene expression characteristics. In addition, estimating the connectivity among
603 substages clearly revealed the existence of multiple pathological pathways of disease
604 progression.
605
606 This study analyzed data from the dorsolateral prefrontal cortex obtained from the Rush
607 AD database, rather than the hippocampus, which has been the focus of many previous
608 studies. The reason for analyzing this dataset in this study was that, it included not only
609 transcriptome data but also data on the genome-wide distribution of histone
610 modifications (H3K4me3, H3K27me3, H3K27ac) and CTCF binding sites (Table S1).
611 Such datasets can potentially reveal the relationship between a wide range of
612 transcriptional regulatory states and disease conditions. Recently, by analyzing such
613 datasets, changes in CTCF binding and transcription activities were shown to be
614 associated with changes in pathological conditions [73]. The analysis of such
615 epigenome data in the present study suggests that the genome-wide distribution of
616 H3K4me3 differed among individuals belonging in three different AD substages.
617 However, the mechanism by which such differences in the epigenetic status are related
618 to gene expression patterns remains unclear and warrants elucidation in future studies.
619
620 Notably, many recent studies have performed transcriptome analyses to identify
621 significant DEGs between healthy individuals and patients with AD. However, by
622 focusing on the expression patterns of more widespread candidate genes rather than the
623 significant DEGs, we identified different characteristics appearing in the dorsolateral
624 prefrontal cortex between healthy individuals and patients with AD and that disease
625 substages occur in response to changes in the disease state.
626

627 In addition, many studies have focused on gene expression patterns in the hippocampus.
628 One such previous research suggested the existence of multiple states in the
629 hippocampus of patients with AD based on differences in the gene expression profiles
630 [39]. However, as previously mentioned, the gene expression patterns differ between the
631 hippocampus and prefrontal cortex, and these differences exhibit further variations
632 between healthy individuals and patients with AD. Hence, different results would be
633 obtained even from the same analysis. Importantly, the hippocampal data analyzed in
634 previous studies can be analyzed in the same way as in this study; connections between
635 substages can be extracted and pathological routes can be estimated, all of which are
636 crucial issues that need to be addressed in the future.
637
638 The present study has some limitations. We only classified female individuals as we
639 used transcriptome and epigenomic data of female patients from the Rush AD database.
640 However, the insights of this study are expected to be useful not only for female
641 patients but also for understanding the rapid progression of symptoms in male patients.
642 To validate this expectation, the dataset from male individuals should also be analyzed.
643 In this dataset, however, the number of female patients in the NCI, MCI, and AD groups
644 was 25, 29, and 28, respectively, whereas that of men was 14, 4, and 7, respectively,
645 that is, three times smaller than that of female individuals (seven and four times smaller
646 than those of female patients with MCI and AD). Moreover, the number of male
647 patients with MCI and AD in the database (four and seven) was smaller than that of
648 individuals in each of the three AD groups obtained for female individuals. With such
649 restricted data, performing similar analyses on male individuals or sex-independent
650 analyses was impossible. Obtaining more data on male patients is one of the essential
651 issues that must be resolved in the future for performing such analyses and eliminating
652 potential sex-related bias.
653
654 The likelihood ratio test using DESeq2 and *t*-test were used to evaluate the differences
655 in RNA-seq data between different groups when searching for gene sets to characterize
656 and classify individuals. The *t*-test is generally better for detecting differences and was
657 expected to identify more candidate genes in a gene set than the likelihood ratio test or
658 other nonparametric testing methods. However, the *t*-test produces more false positives
659 than other testing methods when data deviate from a Gaussian distribution. The gene
660 sets used in this study were expected to contain few false-positive genes because they
661 were chosen under imposing the additional condition that healthy individuals could be
662 clearly distinguished from patients through cluster analysis. However, inappropriate

genes may have still been retained. For more precise consideration, verification is necessary using various nonparametric tests.

In addition to the transcriptome of each individual, the database targeted for the present analysis also contained various epigenome data; however, other information on each individual, such as postmortem interval, brain PH, and medical history was not included. Therefore, although this study has yielded some unprecedented findings, the validation and characterization of each substage through comparisons with other physiological data could not be included. However, such analysis may be achieved by combining meta-analysis with databases focusing on other feature groups.

EXPERIMENTAL PROCEDURES

Data processing

Gene expression data (read count and TPM) from the dorsolateral prefrontal cortex of patients with NCI, MCI, and AD were acquired from the Rush AD database. In the present analysis, the data of genes with average expression levels (TPM) in patients with NCI, MCI, and AD were smaller than 0.1, whereas those of spike-in-RNAs were eliminated.

The data provided by the Rush AD study in ENCODE does not publish information on RNA quality such as the postmortem interval of samples and RIN values. However, according to ENCODE standards, no specific alerts regarding the quality of the RNA-seq data used in this study were provided. Thus, the quality of the data was assumed to be guaranteed. The set of genes whose expression changes between patients with NCI and AD and among patients with NCI, MCI, and AD, involved many genes identified as AD risk genes in previous studies. This further supported the lack of data bias that would contradict previous studies.

Statistical analyses

Two-tailed *t*-tests and ANOVA were performed using Python (version 3.9.16). All *t*-tests in this study were Welch's *t*-tests. The false discovery rate (FDR) was estimated through a set of P-values using the Benjamini–Hochberg procedure [74] implemented in MultiPy (https://puolival.github.io/multipy/) [75]. The search for DEGs via the likelihood ratio test using DESeq2 and the calculation of P-values for the Fisher's exact test were performed using R (version 4.1.3).

699 **Cluster analyses**

700 Cluster analyses were performed using the values of $X = \log_2(TPM + 1) - C$, where C
701 was chosen to satisfy the average of X over individuals = 0 for each gene. iDEP.96
702 (http://bioinformatics.sdstate.edu/idep96/) was used for hierarchical clustering [76]. In
703 hierarchical clustering, the group-average method, was used to create a dendrogram,
704 using the correlation coefficient as the distance. If the cluster formed by the first branch
705 of hierarchical clustering satisfied the following criteria for a certain group (NCI, MCI,
706 and AD), then that cluster was regarded as a cluster including significantly more
707 individuals belonging to that group. These criteria were as follows: i) containing more
708 than 2/3 of the individuals belonging to that group, and ii) Fisher exact test resulting in a
709 P-value < 0.05. Further, k-means cluster analysis was performed using the Python
710 package scikit-learn (version 1.0.2) [77]. The number of clusters in each k-means
711 cluster analysis was determined using the elbow method.
712 The construction of UMAP was performed using sc.tl.pca() [77] with default
713 parameters, scanpy.pp.neighbors() with "n_neighbors" = 30 [78], and the
714 scanpy.tl.umap() functions with n_components = 3 [69] of scanpy (version 1.9.3)
715 according to the literature on basic workflows for clustering analysis using Scanpy
716 (https://scanpy-tutorials.readthedocs.io/en/latest/pbmc3k.html). All principal
717 components were used for clustering using the scanpy.pp.neighbors() function. The
718 obtained results were qualitatively independent of the parameters of these applications.
719

720 **Epigenomic data analysis**

721 The Jaccard coefficient of the genome-wide distribution of epigenomic markers among
722 individuals was estimated as follows. The bigwig formatted files of fold change over
723 control of these epigenome markers were acquired from individuals whose ChIP-seq
724 data of H3K4me3, H3K27ac, H3K27me3, and CTCF were provided in the Rush AD
725 database. From each bigwig format data, bedgraph format data was obtained using
726 bigWigToBedGraph [79]. From the bedgraph format data, bed format data of peak
727 regions were obtained following the removal of regions with FC < 5. The Jaccard
728 coefficients between each pair of individuals were estimated using Bedtools (version
729 2.31.0). The density of each epigenomic marker on each chromosome was estimated
730 according to the following equation: [Total length of genome regions with narrow peaks
731 in ChIP-seq data in the chromosome obtained from the bed formatted file (bp)]/[Total
732 length of chromosome obtained from GRCh38 (bp)].
733

734 **Analysis of *APOE* gene polymorphisms**

RNA-seq read data (fastq format data) from the dorsolateral prefrontal cortex of patients with NCI, MCI, and AD were acquired from the Rush AD database. Fastp (https://github.com/OpenGene/fastp) was used to remove low-quality reads and trim adapter sequences. Following fastp processing, the reads were mapped to the CDS sequence of the *APOE* exon 4 of GRCh38 using Hisat2 (https://daehwankimlab.github.io/hisat2/), and the results were obtained as a bam format file. Variant calling was performed using the obtained bam format file and CDS sequence data of the GRCh38 *APOE* exon 4 using bcftools (https://github.com/samtools/bcftools). After filtering out low-quality calls (QUAL<20), genotypes were determined based on the type of mutation and allele frequency.

Enrichment analysis

Enrichment analysis was performed for the M-NA, N-A1, and N-A2 sets using Metascape (v3.5.20240101) (https://metascape.org/gp/index.html#/main/step1) [80]. The biological process terms in Gene Ontology were the main focus of this study.

Construction of an adjacency network among substages

The adjacency network among substages was estimated via PAGA [69] using the gene expression data of each individual. Each group at each substage was regarded as a partition in PAGA. PAGA can generate a graph of substages in which the edge weights represent connection confidence. PAGA was performed using sc.tl.pca() [77] with default parameters, scanpy.pp.neighbors() with “n_neighbors” = 2 [78], and the scanpy.tl.paga() functions [69] of scanpy (version 1.9.3) according to the literature for trajectory inference using transcriptome data (https://scanpy-tutorials.readthedocs.io/en/latest/paga-paul15.html). All principal components were used for clustering using the scanpy.pp.neighbors() function. Additionally, to extract the essential network among substages, the value of the parameter "n_neighbors" of the function scanpy.pp.neighbors() was set as two, which is the minimum value in the range of this parameter. Thus, a network with nonredundant connections among substages was expected to be obtained.

AUTHOR CONTRIBUTIONS

Kousei Honda and Akinori Awazu conceived and designed the study. Kousei Honda performed the analyses. Kousei Honda and Akinori Awazu drafted the manuscript.

ACKNOWLEDGEMENTS

This research was supported by JST Moonshot R&D (Grant Number JPMJMS2021) and a JSPS KAKENHI grant (award number 21K06124, 24K09248). Computations were partially performed on the NIG supercomputer at the ROIS National Institute of Genetics.

REFERENCES

1. Gatz, M., Reynolds, C. A., Fratiglioni, L., Johansson, B., Mortimer, J. A., Berg, S.,... & Pedersen, N. L. (2006). Role of genes and environments for explaining Alzheimer disease. *Archives of General Psychiatry*, *63*, 168–174. https://doi.org/10.1001/archpsyc.63.2.168

2. World Health Organization. (2017). Global action plan on the public health response to dementia 2017–2025.

3. Suresh, S., Singh, A., Rushendran, R., Vellapandian, C., & Prajapati, B. (2023). Alzheimer's disease: the role of extrinsic factors in its development, an investigation of the environmental enigma. *Frontiers in Neurology*, *14*, 1303111. https://doi.org/10.3389/fneur.2023.1303111

4. Hardy, J. A., & Higgins, G. A. (1992). Alzheimer's disease: the amyloid cascade hypothesis. *Science*, *256*, 184–185. https://doi.org/10.1126/science.1566067

5. Hardy, J., & Selkoe, D. J. (2002). The amyloid hypothesis of Alzheimer's disease: progress and problems on the road to therapeutics. *Science*, *297*, 353–356. https://doi.org/10.1126/science.1072994

6. Jin, M., Shepardson, N., Yang, T., Chen, G., Walsh, D., & Selkoe, D. J. (2011). Soluble amyloid β-protein dimers isolated from Alzheimer cortex directly induce Tau hyperphosphorylation and neuritic degeneration. *Proceedings of the National Academy of Sciences*, *108*, 5819–5824. https://doi.org/10.1073/pnas.1017033108

7. Selkoe, D. J., & Hardy, J. (2016). The amyloid hypothesis of Alzheimer's disease at 25 years. *EMBO Molecular Medicine*, *8*(6), 595–608. https://doi.org/10.15252/emmm.201606210

8. Sekiya, M., Wang, M., Fujisaki, N., Sakakibara, Y., Quan, X., Ehrlich, M. E.,... Iijima, K. M. (2018). Integrated biology approach reveals molecular and pathological interactions among Alzheimer's Aβ42, Tau, TREM2, and TYROBP in Drosophila models. *Genome Medicine*, *10*(1), 1–20. https://doi.org/10.1186/s13073-018-0530-9

9. Arriagada, P. V., Growdon, J. H., Hedley-Whyte, E. T., & Hyman, B. T. (1992). Neurofibrillary tangles but not senile plaques parallel duration and severity of Alzheimer's disease. *Neurology*, *42*, 631–631. https://doi.org/10.1212/wnl.42.3.631

10. Strittmatter, W. J., Saunders, A. M., Schmechel, D., Pericak-Vance, M., Enghild, J., Salvesen, G. S., & Roses, A. D. (1993). Apolipoprotein E: high-avidity binding to beta-amyloid and increased frequency of type 4 allele in late-onset familial Alzheimer disease. *Proceedings of the National Academy of Sciences*, *90*, 1977–1981. https://doi.org/10.1073/pnas.90.5.1977

11. Itzhaki, R. F., Lin, W. R., Shang, D., Wilcock, G. K., Faragher, B., & Jamieson, G. A. (1997). Herpes simplex virus type 1 in brain and risk of Alzheimer's disease. *The Lancet*, *349*, 241–244. https://doi.org/10.1016/s0140-6736(96)10149-5

12. Osaka, R., Kobayashi, N., Shimada, K., Ishii, A., Oka, N., & Kondo, K. (2022). VP26, a herpes simplex virus type 1 capsid protein, increases DNA methylation in COASY promoter region. *Brain, Behavior, & Immunity - Health*, *26*, 100545. https://doi.org/10.1016/j.bbih.2022.100545

13. Readhead, B., Haure-Mirande, J. V., Funk, C. C., Richards, M. A., Shannon, P., Haroutunian, V.,... Dudley, J. T. (2018). Multiscale analysis of independent Alzheimer's cohorts finds disruption of molecular, genetic, and clinical networks by human herpesvirus. *Neuron*, *99*, 64–82. https://doi.org/10.1016/j.neuron.2018.05.023

14. Wang, L., Davis, P. B., Volkow, N. D., Berger, N. A., Kaelber, D. C., & Xu, R. (2022). Association of COVID-19 with new-onset Alzheimer's disease. *Journal of Alzheimer's Disease*, *89*(2), 1–4. https://doi.org/10.3233/jad-220717

15. Kandimalla, R., Thirumala, V., & Reddy, P. H. (2017). Is Alzheimer's disease a type 3 diabetes? A critical appraisal. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, *1863*, 1078–1089. https://doi.org/10.1016/j.bbadis.2016.08.018

16. Ott, A., Stolk, R. P., van Harskamp, F., Pols, H. A. P., Hofman, A., & Breteler, M. M. B. (1999). Diabetes mellitus and the risk of dementia: The Rotterdam Study. *Neurology*, *53*, 1937–1937. https://doi.org/10.1212/wnl.53.9.1937

17. McGeer, P. L., & McGeer, E. G. (1995). The inflammatory response system of brain: implications for therapy of Alzheimer and other neurodegenerative diseases. *Brain Research Reviews*, *21*, 195–218. https://doi.org/10.1016/0165-0173(95)00011-9

18. Hirai, K., Aliev, G., Nunomura, A., Fujioka, H., Russell, R. L., Atwood, C. S.,... Smith, M. A. (2001). Mitochondrial abnormalities in Alzheimer's disease. *Journal of Neuroscience*, *21*, 3017–3023. https://doi.org/10.1523/jneurosci.21-09-03017.2001

19. Nunomura, A., Perry, G., Aliev, G., Hirai, K., Takeda, A., Balraj, E. K.,... Smith, M. A. (2001). Oxidative damage is the earliest event in Alzheimer disease. *Journal of Neuropathology & Experimental Neurology*, *60*, 759–767. https://doi.org/10.1093/jnen/60.8.759

20. Wang, X., Wang, W., Li, L., Perry, G., Lee, H. G., & Zhu, X. (2014). Oxidative stress and mitochondrial dysfunction in Alzheimer's disease. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, *1842*, 1240–1247. https://doi.org/10.1016%2Fj.bbadis.2013.10.015

21. Chakravarthy, B., Gaudet, C., Ménard, M., Brown, L., Atkinson, T., LaFerla, F. M.,... & Whitfield, J. (2012). Reduction of the immunostainable length of the hippocampal dentate granule cells' primary cilia in 3xAD-transgenic mice producing human Aβ1-42 and tau. *Biochemical and Biophysical Research Communications*, *427*(1), 218–222. https://doi.org/10.1016/j.bbrc.2012.09.056

22. Morelli, A., Sarchielli, E., Guarnieri, G., Coppi, E., Pantano, D., Comeglio, P.,... Vannelli, G. B. (2017). Young human cholinergic neurons respond to physiological regulators and improve cognitive symptoms in an animal model of Alzheimer's disease. *Frontiers in Cellular Neuroscience*, *11*, 339. https://doi.org/10.3389/fncel.2017.00339

23. Plácido, A. I., Pereira, C. M. F., Duarte, A. I., Candeias, E., Correia, S. C., Santos, R. X.,... Moreira, P. I. (2014). The role of endoplasmic reticulum in amyloid precursor protein processing and trafficking: implications for Alzheimer's disease. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, *1842*, 1444–1453. https://doi.org/10.1016/j.bbadis.2014.05.003

24. Vorobyeva, A. G., & Saunders, A. J. (2018). Amyloid-β interrupts canonical Sonic hedgehog signaling by distorting primary cilia structure. *Cilia*, *7*(1), 1–11. https://doi.org/10.1186/s13630-018-0059-y

25. Bartzokis, G. (2004). Age-related myelin breakdown: a developmental model of cognitive decline and Alzheimer's disease. *Neurobiology of Aging*, *25*, 5–18. https://doi.org/10.1016/j.neurobiolaging.2003.03.001

26. Bartzokis, G. (2011). Alzheimer's disease as homeostatic responses to age-related myelin breakdown. *Neurobiology of Aging*, *32*, 1341–1371. https://doi.org/10.1016/j.neurobiolaging.2009.08.007

27. Cheng, Y., Saville, L., Gollen, B., Veronesi, A. A., Mohajerani, M., Joseph, J. T., & Zovoilis, A. (2021). Increased Alu RNA processing in Alzheimer brains is linked to gene expression changes. *EMBO Reports*, *22*, e52255. https://doi.org/10.15252/embr.202052255

28. Macciardi, F., Giulia Bacalini, M., Miramontes, R., Boattini, A., Taccioli, C., Modenini, G.,... Mapstone, M. (2022). A retrotransposon storm marks clinical phenoconversion to late-onset Alzheimer's disease. *GeroScience*, *44*, 1525–1550. https://doi.org/10.1007/s11357-022-00580-w

29. Ravel-Godreuil, C., Znaidi, R., Bonnifet, T., Joshi, R. L., & Fuchs, J. (2021). Transposable elements as new players in neurodegenerative diseases. *FEBS Letters*, *595*, 2733–2755. https://doi.org/10.1002/1873-3468.14205

30. Duara, R., & Barker, W. (2021). Heterogeneity in Alzheimer's disease diagnosis and progression rates: implications for therapeutic trials. *Neurotherapeutics*, *19*, 8–25. https://doi.org/10.1007/s13311-022-01185-z

31. Komarova, N. L., & Thalhauser, C. J. (2011). High degree of heterogeneity in Alzheimer's disease progression patterns. *PLoS Computational Biology*, *7*, e1002251. https://doi.org/10.1371/journal.pcbi.1002251

32. Younas, N., Zafar, S., Shafiq, M., Noor, A., Siegert, A., Arora, A. S., … Zerr, I. (2020). SFPQ and Tau: Critical factors contributing to rapid progression of Alzheimer's disease. *Acta Neuropathologica*, *140*, 317–339. https://doi.org/10.1007/s00401-020-02178-y

33. Younas, A., Younas, N., Iqbal, M. J., Ferrer, I., & Zerr, I. (2024). Comparative interactome mapping of Tau-protein in classical and rapidly progressive Alzheimer's disease identifies subtype-specific pathways. *Neuropathology and Applied Neurobiology*, *50*, e12964. https://doi.org/10.1111/nan.12964

34. Jansen, I. E., Savage, J. E., Watanabe, K., Bryois, J., Williams, D. M., Steinberg, S.,... Posthuma, D. (2019). Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. *Nature Genetics*, *51*, 404–413. https://doi.org/10.1038/s41588-018-0311-9

35. Wightman, D. P., Jansen, I. E., Savage, J. E., Shadrin, A. A., Bahrami, S., Holland, D.,... Posthuma, D. (2021). A genome-wide association study with 1,126,563 individuals identifies new risk loci for Alzheimer's disease. *Nature Genetics*, *53*, 1276–1282. https://doi.org/10.1038%2Fs41588-021-00921-z

36. Yuen, S. C., Zhu, H., & Leung, S. W. (2020). A systematic bioinformatics workflow with meta-analytics identified potential pathogenic factors of Alzheimer's disease. *Frontiers in Neuroscience*, *14*, 209. https://doi.org/10.3389/fnins.2020.00209

37. Bellenguez, C., Küçükali, F., Jansen, I. E., Kleineidam, L., Amin, N., Naj, A. C., ... Lambert, J. (2022). New insights into the genetic etiology of Alzheimer's disease and related dementias. *Nature Genetics*, *54*, 412–436. https://doi.org/10.1038/s41588-022-01024-z

38. Johnson, E. C., Carter, E. K., Dammer, E. B., Duong, D. M., Gerasimov, E. S., Liu, Y.,... Seyfried, N. T. (2022). Large-scale deep multi-layer analysis of Alzheimer's disease brain reveals strong proteomic disease-related changes not observed at the RNA level. *Nature Neuroscience*, *25*, 213–225. https://doi.org/10.1038/s41593-021-00999-y

39. Neff, R. A., Wang, M., Vatansever, S., Guo, L., Ming, C., Wang, Q.,... Zhang, B. (2021). Molecular subtyping of Alzheimer's disease using RNA sequencing data reveals novel mechanisms and targets. *Science Advances*, *7*, eabb5398. https://doi.org/10.1126/sciadv.abb5398

40. Tijms, B. M., Gobom, J., Reus, L., Jansen, I., Hong, S., Dobricic, V.,... Visser, P. J. (2020). Pathophysiological subtypes of Alzheimer's disease based on cerebrospinal fluid proteomics. *Brain*, *143*, 3776–3792. https://doi.org/10.1093/brain/awaa325

41. Tijms, B. M., Gobom, J., Teunissen, C., Dobricic, V., Tsolaki, M., Verhey, F.,... Visser, P. J. (2021). CSF proteomic Alzheimer's disease-predictive subtypes in cognitively intact amyloid negative individuals. *Proteomes*, *9*, 36. https://doi.org/10.3390/proteomes9030036

42. ENCODE Project Consortium. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, *489*(7414), 57. https://doi.org/10.1038/nature11247

43. Luo, Y., Hitz, B. C., Gabdank, I., Hilton, J. A., Kagda, M. S., Lam, B.,... Cherry, J. M. (2020). New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. *Nucleic Acids Research*, 48, D882–D889. https://doi.org/10.1093/nar/gkz1062

44. Davis, E. J., Broestl, L., Abdulai-Saiku, S., Worden, K., Bonham, L. W., Miñones-Moyano, E., ... & Dubal, D. B. (2020). A second X chromosome contributes to

982 resilience in a mouse model of Alzheimer's disease. *Science translational*
983 *medicine*, *12*(558), eaaz5677. https://doi.org/aaz5677
984
985 45. Lopez-Lee, C., Torres, E. R. S., Carling, G., & Gan, L. (2024). Mechanisms of sex
986 differences in Alzheimer's disease. *Neuron*.
987 https://doi.org/10.1016/j.neuron.2024.01.024
988
989 46. Anders, S., & Huber, W. (2010). Differential expression analysis for sequence count
990 data. *Genome Biology*, *11*, R106. https://doi.org/10.1186/gb-2010-11-10-r106
991
992 47. Shin, H., Liu, T., Duan, X., Zhang, Y., & Liu, X. S. (2013). Computational
993 methodology for ChIP-seq analysis. *Quantitative Biology*, *1*(1), 54-
994 70. https://doi.org/10.1007/s40484-013-0006-2
995
996 48. Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change
997 and dispersion for RNA-seq data with DESeq2. *Genome biology*, *15*, 1-21.
998 https://doi.org/10.1186/s13059-014-0550-8
999
1000 49. Schuler, G. D., Epstein, J. A., Ohkawa, H., & Kans, J. A. (1996). Entrez: Molecular
1001 biology database and retrieval system. In Methods in enzymology (Vol. 266, pp.
1002 141-162). Academic Press. https://doi.org/10.1016/S0076-6879(96)66012-1
1003
1004 50. Hu, Y. S., Xin, J., Hu, Y., Zhang, L., & Wang, J. (2017). Analyzing the genes
1005 related to Alzheimer's disease via a network and pathway-based
1006 approach. *Alzheimer's research & therapy*, *9*, 1-15. https://doi.org/10.1186/s13195-
1007 017-0252-z
1008
1009 51. Singh, S., Sahu, K., Singh, C., & Singh, A. (2022). Lipopolysaccharide induced
1010 altered signaling pathways in various neurological disorders. *Naunyn-*
1011 *Schmiedeberg's Archives of Pharmacology*, *395*(3), 285-294.
1012 https://doi.org/10.1007/s00210-021-02198-9
1013
1014 52. Zhan, X., Stamova, B., & Sharp, F. R. (2018). Lipopolysaccharide associates with
1015 amyloid plaques, neurons and oligodendrocytes in Alzheimer's disease brain: a
1016 review. *Frontiers in aging neuroscience*, *10*, 42.
1017 https://doi.org/10.3389/fnagi.2018.00042

53. Adhikarla, S. V., Jha, N. K., Goswami, V. K., Sharma, A., Bhardwaj, A., Dey, A., ... & Jha, S. K. (2021). TLR-mediated signal transduction and neurodegenerative disorders. *Brain sciences*, *11*(11), 1373. https://doi.org/10.3390/brainsci11111373

54. Wu, L., Xian, X., Xu, G., Tan, Z., Dong, F., Zhang, M., & Zhang, F. (2022). Toll-like receptor 4: a promising therapeutic target for alzheimer's disease. *Mediators of inflammation*, *2022*. https://doi.org/10.1155/2022/7924199

55. Parachikova, A., Agadjanyan, M. G., Cribbs, D. H., Blurton-Jones, M., Perreau, V., Rogers, J., ... & Cotman, C. W. (2007). Inflammatory changes parallel the early stages of Alzheimer disease. *Neurobiology of aging*, *28*(12), 1821-1833. https://doi.org/10.1016/j.neurobiolaging.2006.08.014

56. Grimm, M. O., Mett, J., & Hartmann, T. (2016). The impact of vitamin E and other fat-soluble vitamins on Alzheimer s disease. *International journal of molecular sciences*, *17*(11), 1785. https://doi.org/10.3390/ijms17111785

57. Lau, S. F., Wu, W., Wong, H. Y., Ouyang, L., Qiao, Y., Xu, J., ... & Ip, N. Y. (2023). The VCAM1–ApoE pathway directs microglial chemotaxis and alleviates Alzheimer's disease pathology. *Nature Aging*, *3*(10), 1219-1236. https://doi.org/10.1038/s43587-023-00491-1

58. Rogers, J., & Lue, L. F. (2001). Microglial chemotaxis, activation, and phagocytosis of amyloid β-peptide as linked phenomena in Alzheimer's disease. *Neurochemistry international*, *39*(5-6), 333-340. https://doi.org/10.1016/S0197-0186(01)00040-7

59. Xiang, S., Huang, Z., Wang, T., Han, Z., Yu, C. Y., Ni, D., ... & Zhang, J. (2018). Condition-specific gene co-expression network mining identifies key pathways and regulators in the brain tissue of Alzheimer's disease patients. *BMC medical genomics*, *11*, 39-51. https://doi.org/10.1186/s12920-018-0431-1

60. Rahman, M. R., Islam, T., Turanli, B., Zaman, T., Faruquee, H. M., Rahman, M. M., ... & Moni, M. A. (2019). Network-based approach to identify molecular signatures and therapeutic agents in Alzheimer's disease. *Computational biology and chemistry*, *78*, 431-439. https://doi.org/10.1016/j.compbiolchem.2018.12.011

61. Yang, A. C., Vest, R. T., Kern, F., Lee, D. P., Agam, M., Maat, C. A., ... & Wyss-Coray, T. (2022). A human brain vascular atlas reveals diverse mediators of Alzheimer's risk. *Nature*, *603*(7903), 885-892. https://doi.org/10.1038/s41586-021-04369-3

62. Kaltschmidt, B., & Kaltschmidt, C. (2015). NF-KappaB in long-term memory and structural plasticity in the adult mammalian brain. *Frontiers in molecular neuroscience*, *8*, 69. https://doi.org/10.3389/fnmol.2015.00069

63. Lin, X., Kapoor, A., Gu, Y., Chow, M. J., Peng, J., Zhao, K., & Tang, D. (2020). Contributions of DNA damage to Alzheimer's disease. *International Journal of Molecular Sciences*, *21*(5), 1666. https://doi.org/10.3390/ijms21051666

64. Ranganathan, R., Sapozhnikov, G., Ni, W., Li, S., & Song, Y. (2023). Recent developments in the role of DNA damage response and understanding its implications for new therapeutic approaches in Alzheimer's disease. *Translational Medicine of Aging*. https://doi.org/10.1016/j.tma.2023.06.004

65. Chen, Z., & Zhong, C. (2013). Decoding Alzheimer's disease from perturbed cerebral glucose metabolism: implications for diagnostic and therapeutic strategies. *Progress in neurobiology*, *108*, 21-43. https://doi.org/10.1016/j.pneurobio.2013.06.004

66. Cunnane, S., Nugent, S., Roy, M., Courchesne-Loyer, A., Croteau, E., Tremblay, S., ... & Rapoport, S. I. (2011). Brain fuel metabolism, aging, and Alzheimer's disease. *Nutrition*, *27*(1), 3-20. https://doi.org/10.1016/j.nut.2010.07.021

67. Esteve, D., Molina-Navarro, M. M., Giraldo, E., Martínez-Varea, N., Blanco-Gandia, M. C., Rodríguez-Arias, M., ... & Lloret, A. (2022). Adult neural stem cell migration is impaired in a mouse model of Alzheimer's disease. *Molecular Neurobiology*, 1-15. https://doi.org/10.1007/s12035-021-02620-6

68. Young-Pearse, T. L., Bai, J., Chang, R., Zheng, J. B., LoTurco, J. J., & Selkoe, D. J. (2007). A critical function for β-amyloid precursor protein in neuronal migration

revealed by in utero RNA interference. *Journal of Neuroscience*, *27*(52), 14459-14469. https://doi.org/10.1523/JNEUROSCI.4701-07.2007

69. Wolf, F. A., Hamey, F. K., Plass, M., Solana, J., Dahlin, J. S., Göttgens, B.,... Theis, F. J. (2019). PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biology*, *20*, 1–9. https://doi.org/10.1186%2Fs13059-019-1663-x

70. Karikari, T. K., Pascoal, T. A., Ashton, N. J., Janelidze, S., Benedet, A. L., Rodriguez, J. L., Chamoun, M., Savard, M., Kang, M. S., Therriault, J., Schöll, M., Massarweh, G., Soucy, J., Höglund, K., Brinkmalm, G., Mattsson, N., Palmqvist, S., Gauthier, S., Stomrud, E., . . . Blennow, K. (2020). Blood phosphorylated tau 181 as a biomarker for Alzheimer's disease: A diagnostic performance and prediction modelling study using data from four prospective cohorts. *The Lancet Neurology*, *19*(5), 422-433. https://doi.org/10.1016/S1474-4422(20)30071-5

71. Therriault, J., Pascoal, T. A., Lussier, F. Z., Tissot, C., Chamoun, M., Bezgin, G., Servaes, S., Benedet, A. L., Ashton, N. J., Karikari, T. K., Kunach, P., Wang, Y., Massarweh, G., Vitali, P., Soucy, J., Blennow, K., Zetterberg, H., & Gauthier, S. (2022). Biomarker modeling of Alzheimer's disease using PET-based Braak staging. *Nature Aging*, *2*(6), 526-535. https://doi.org/10.1038/s43587-022-00204-0

72. Yada, Y., & Naoki, H. (2023). Few-shot prediction of amyloid β accumulation from mainly unpaired data on biomarker candidates. *Npj Systems Biology and Applications*, *9*(1), 1-11. https://doi.org/10.1038/s41540-023-00321-5

73. Patel, P. J., Ren, Y., & Yan, Z. (2023). Epigenomic analysis of Alzheimer's disease brains reveals diminished CTCF binding on genes involved in synaptic organization. *Neurobiology of Disease*, 184, 106192. https://doi.org/10.1016/j.nbd.2023.106192

74. Benjamini, Y., Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistics Society Series B*, *57*, 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

75. Puoliväli, T., Palva, S., & Palva, J.M. (2020). Influence of multiple hypothesis testing on reproducibility in neuroimaging research: A simulation study and Python-

based software. *Journal of Neuroscience Methods*, *337*, 108654. https://doi.org/10.1016/j.jneumeth.2020.108654

76. Ge, S.X., Son, E.W., & Yao, R. (2018). iDEP: an integrated web application for differential expression and pathway analysis of RNA-Seq data. *BMC Bioinformatics*, *19*, 534. https://doi.org/10.1186/s12859-018-2486-6

77. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, É. (2011). Scikit-learn: machine learning in Python. *JMLR*, *12*, 2825–2830.

78. McInnes, L., Healy, J., & Melville, J. UMAP: Uniform manifold approximation and projection for dimension reduction. arXiv:1802.03426. https://doi.org/10.48550/arXiv.1802.03426

79. Kent, W. J., Zweig, A. S., Barber, G., Hinrichs, A. S., & Karolchik, D. (2010). BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics*, *26*(17), 2204-2207. https://doi.org/10.1093/bioinformatics/btq351

80. Zhou, Y., Zhou, B., Pache, L., Chang, M., Khodabakhshi, A. H., Tanaseichuk, O., … Chanda, S. K. (2019). Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nature Communications*, *10*(1),1523. https://doi.org/10.1038/s41467-019-09234-6

1150 **Table 1**. Summary of the individuals included in this study (see Table S1a and S1b for
1151 more details).

Clinical stage	No Cognitive Impairment (NCI)	Mild Cognitive Impairment (MCI)	Alzheimer's Disease (AD)
# of individuals	25	29	28
Individual index	N1 – N25	M1 – M29	A1 – A28

1152

Table 2. Number of individuals exhibiting the typical genome-wide distributions of each epigenetic state, E_T, and of those exhibiting untypical distributions, E_U, among the three AD substages (AD_{HH}, AD_{LL}, and AD_{HL}); H3K4me3 (a), H3K27ac (b), H3K27me3 (c), and CTCF (d). The E_T and E_U in each substage were obtained from Figures 5, 6, and S4. P-values obtained using the Fisher's exact test were 0.023843 (a), 0.193439 (b), 0.193439 (c), and 0.285294 (d). Number of individuals exhibiting E_T were divided into E_{T1} (Typical 1) and E_{T2} (Typical 2) among the three AD substages by further clustering for H3K4me3 obtained from Figure 5 (e). The P-value obtained using the Fisher's exact test was 0.0420316 (e).

(a)

H3K4me3	AD_{HH}	AD_{LL}	AD_{HL}
# of individuals with typical (E_T)	6	4	3
# of individuals with untypical (E_U)	0	0	4

(b)

H3K27ac	AD_{HH}	AD_{LL}	AD_{HL}
# of individuals with typical (E_T)	5	4	3
# of individuals with untypical (E_U)	1	0	4

(c)

H3K27me3	AD_{HH}	AD_{LL}	AD_{HL}
# of individuals with typical (E_T)	5	4	3
# of individuals with untypical (E_U)	1	0	4

(d)

CTCF	AD_{HH}	AD_{LL}	AD_{HL}
# of individuals with typical (E_T)	6	3	4
# of individuals with untypical (E_U)	0	1	3

(e)

H3K4me3	AD_{HH}	AD_{LL}	AD_{HL}
# of individuals with Typical 1 (E_{T1})	1	2	2
# of individuals with Typical 2 (E_{T2})	5	2	1

# of individuals with untypical (E_U)	0	0	4

1174 **Table 3.** Number of APOE2, APOE3, and APOE4 carriers in each clinical diagnostic
1175 stage (a), in the two MCI substages (b), and in the three AD substages (c). The genotype
1176 of each individual is given in Table S1a and S1b. The P-values obtained using the
1177 Fisher's exact test were 0.8841 (a), 0.08148 (b), and 0.4379 (c).
1178
1179 (a)

Clinical Diagnostic Stage	NCI	MCI	AD
# of APOE2 carriers	5	3	4
# of APOE3 carriers	16	20	18
# of APOE4 carriers	4	6	6

1180
1181 (b)

MCI substages	MCI_{Major}	MCI_{NCI_AD}
# of APOE2 carriers	1	2
# of APOE3 carriers	18	2
# of APOE4 carriers	5	1

1182
1183 (c)

AD substages	AD_{HH}	AD_{LL}	AD_{HL}
# of APOE2 carriers	3	1	0
# of APOE3 carriers	6	7	5
# of APOE4 carriers	1	2	3

1184
1185

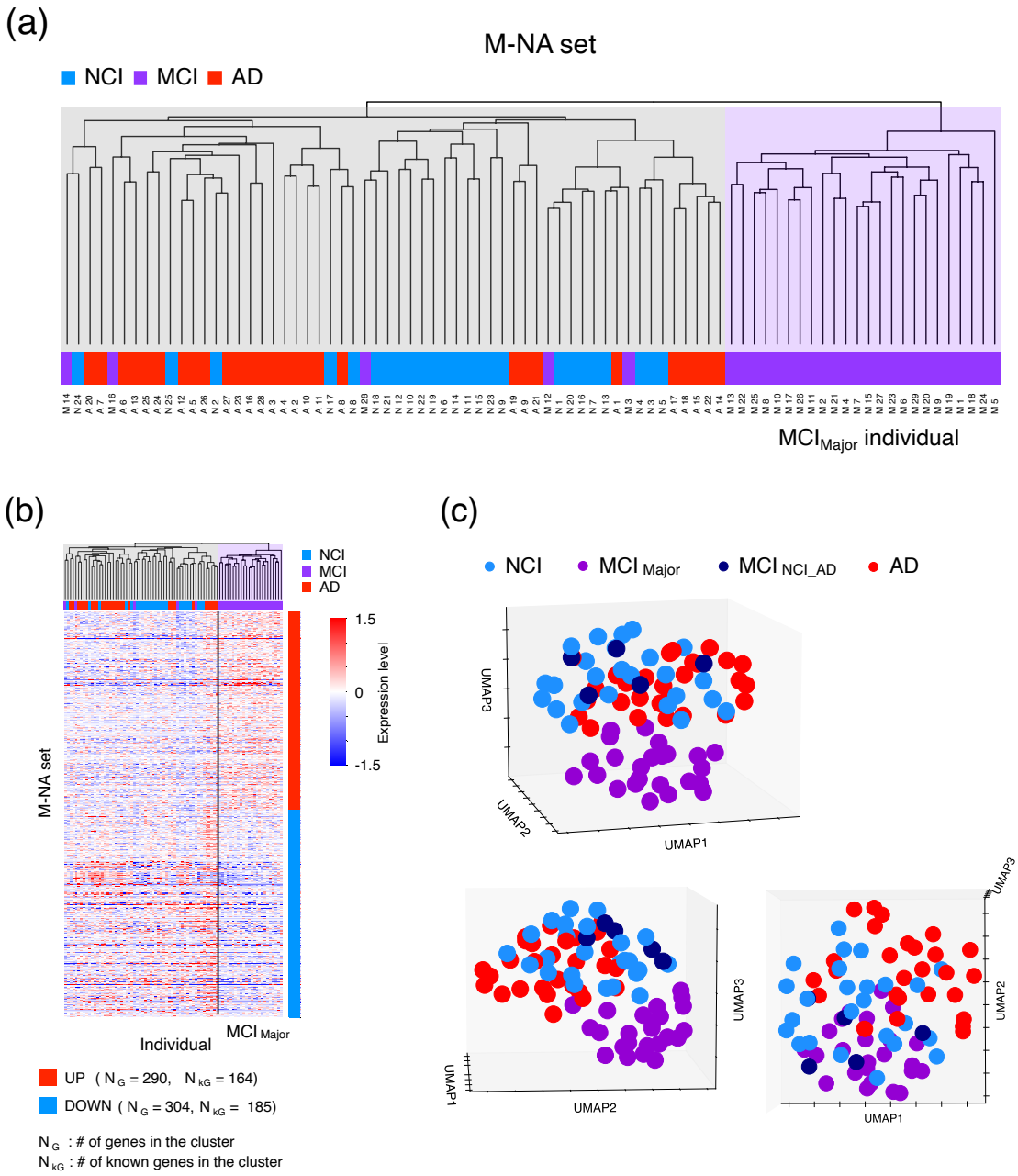
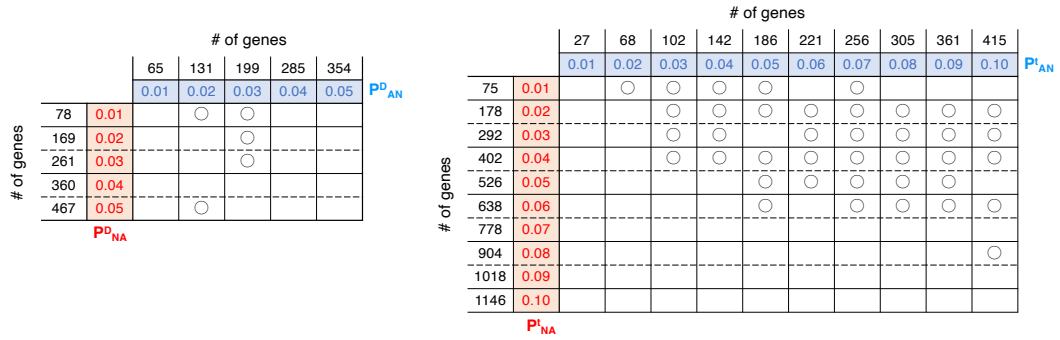


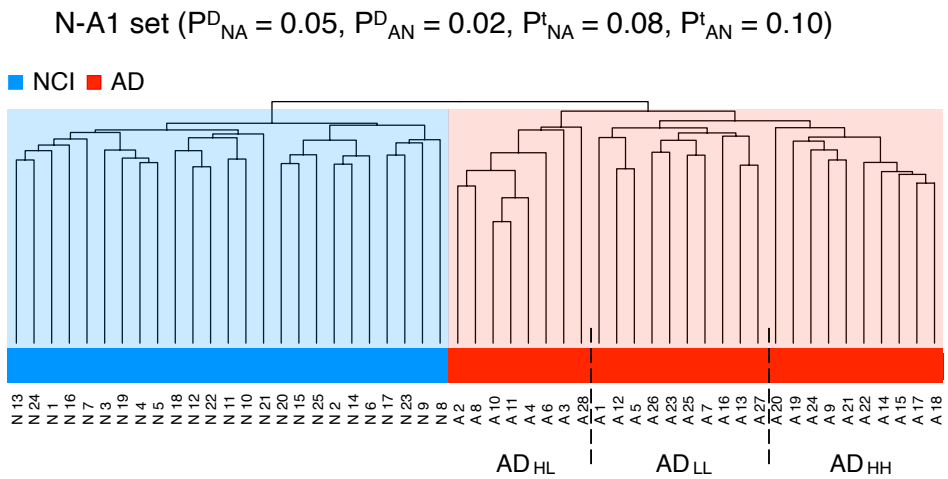
Fig. 1. Results of the hierarchical clustering of individuals and of UMAP obtained using the M-NA set. Result of the hierarchical clustering of individuals when using the M-NA set (P^{A}_{NMA} = 0.04) (a). The expression level of each gene in the M-NA set for each individual was plotted (b) (see Table S2a for more details). The heatmap of the expression levels of each gene was described using $X = \log_2(TPM + 1) - C$, where C was chosen to satisfy the average of X over individuals = 0 for each gene. MCI_{Major} individuals occupy the right region in the dendrogram. N_G and N_{kG} indicate the total number of genes and number of genes annotated as protein coding or noncoding RNA

genes with higher (UP) and lower (DOWN) expression levels in MCI_{Major} individuals, respectively. Scatter plot of individuals in 3-dimensional space constructed using U-MAP with 1, 2, and 3 axes shown from various directions (c).

(a)



(b)



(c)

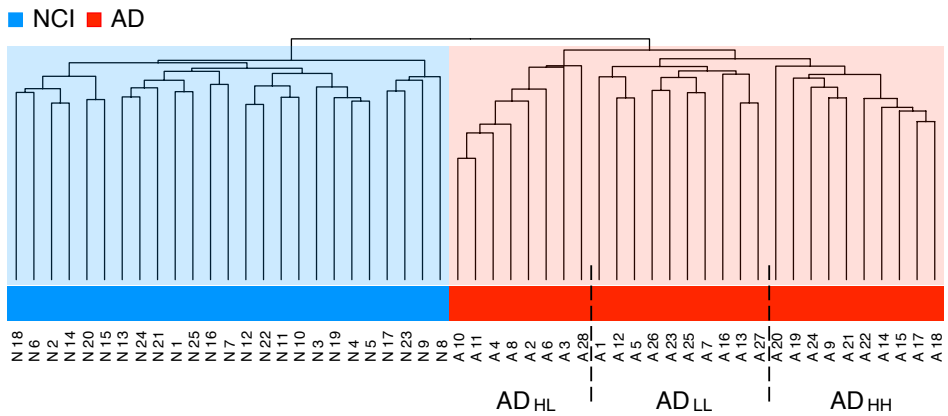


Fig. 2 Phase diagrams of the hierarchical clustering of patients with NCI and AD, and the results of the hierarchical clustering of individuals in the N-A1 and N-A2 sets. Phase diagram of the results of hierarchical clustering when using genes where $p \leq P^D_{NA}$ and $p \leq P^D_{AN}$ (a, left), and that when using genes where $p \leq P^t_{NA}$ and $p \leq P^t_{AN}$ (a, right). Circles indicate the complete separation of patients with NCI from those with

AD into different clusters at the first branch of the dendrogram. The dendrograms obtained using the N-A1 (b) and N-A2 sets (c) are shown. Broken lines indicate the boundary between the AD_{HL}, AD_{LL}, and AD_{HH} groups.

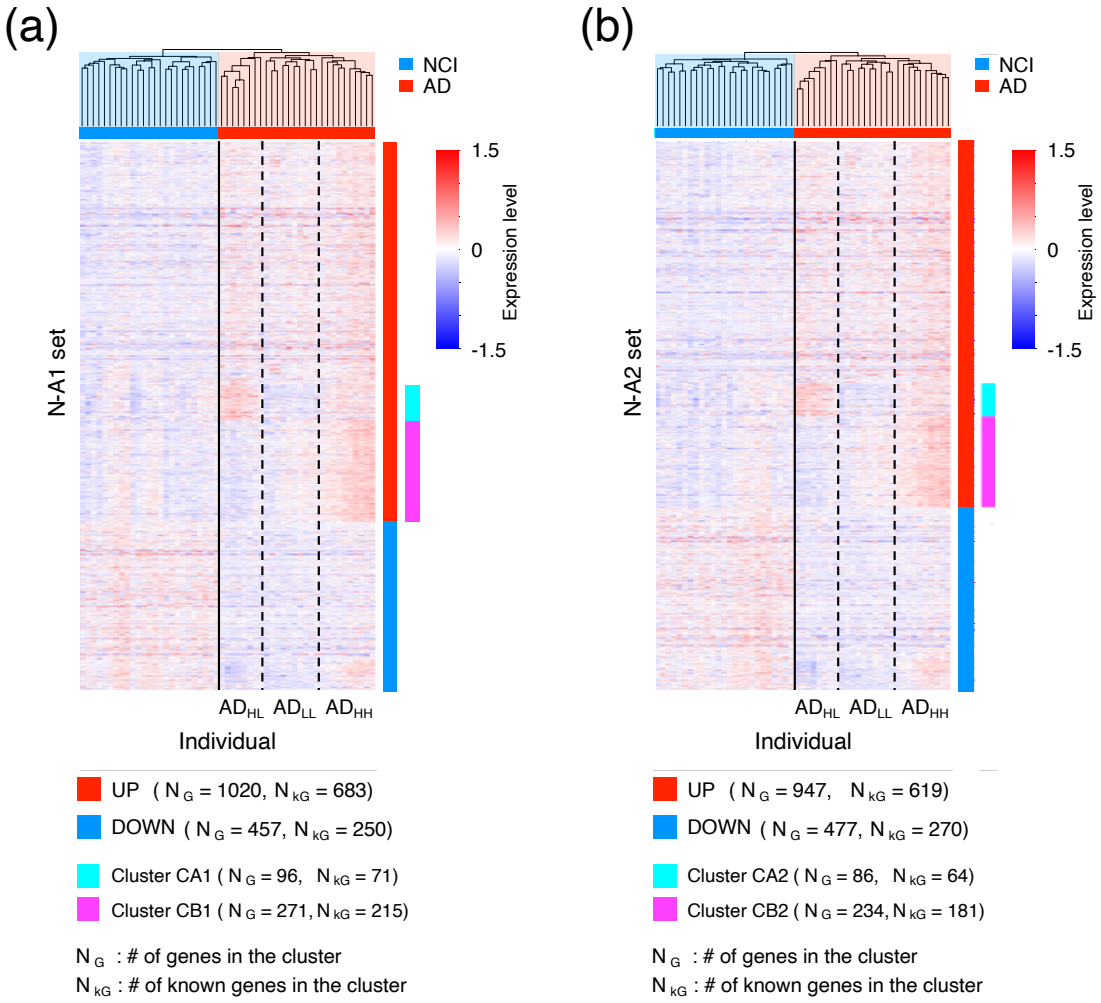


Fig. 3 Results of the hierarchical clustering of individuals in the N-A1 and N-A2 sets and expression level of each gene. The dendrograms and expression level of each gene in the N-A1 set for each individual (a) (see Table S3d for more details) and those in the N-A2 set for each individual (b) (See Table S3e for more details) were plotted. The heatmap of the expression levels of each gene was described using $X = \log_2(TPM + 1) - C$, where C was chosen to satisfy the average of X over individuals = 0 for each gene. Broken lines indicate the boundary between the AD_{HL}, AD_{LL}, and AD_{HH} groups.

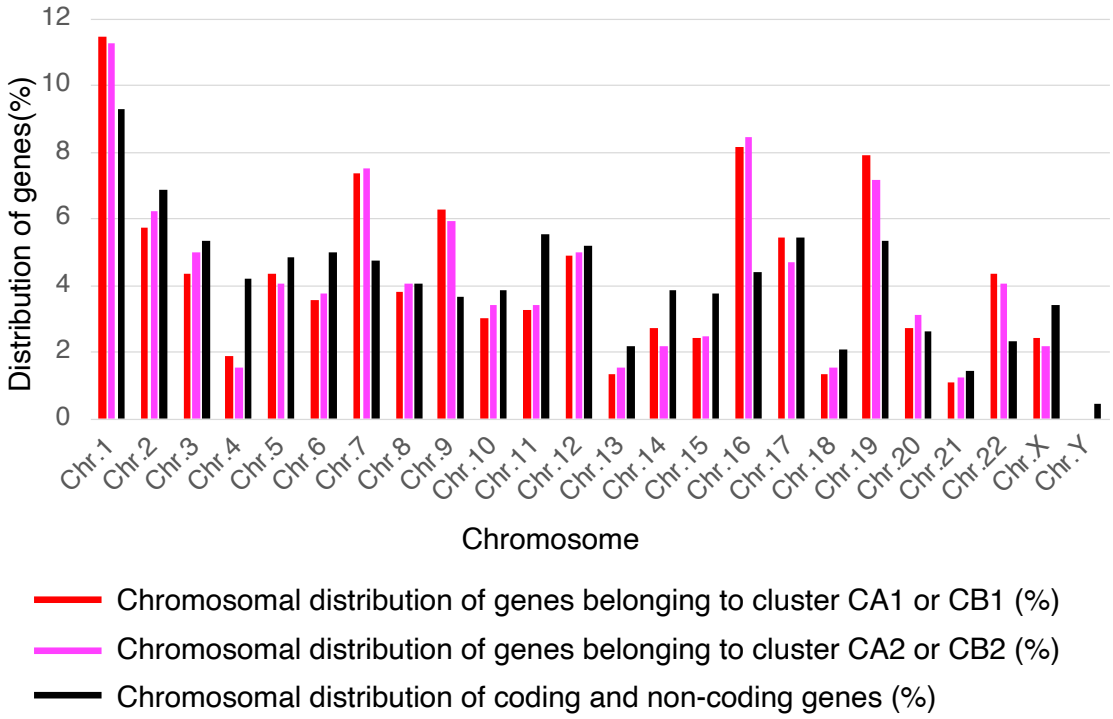
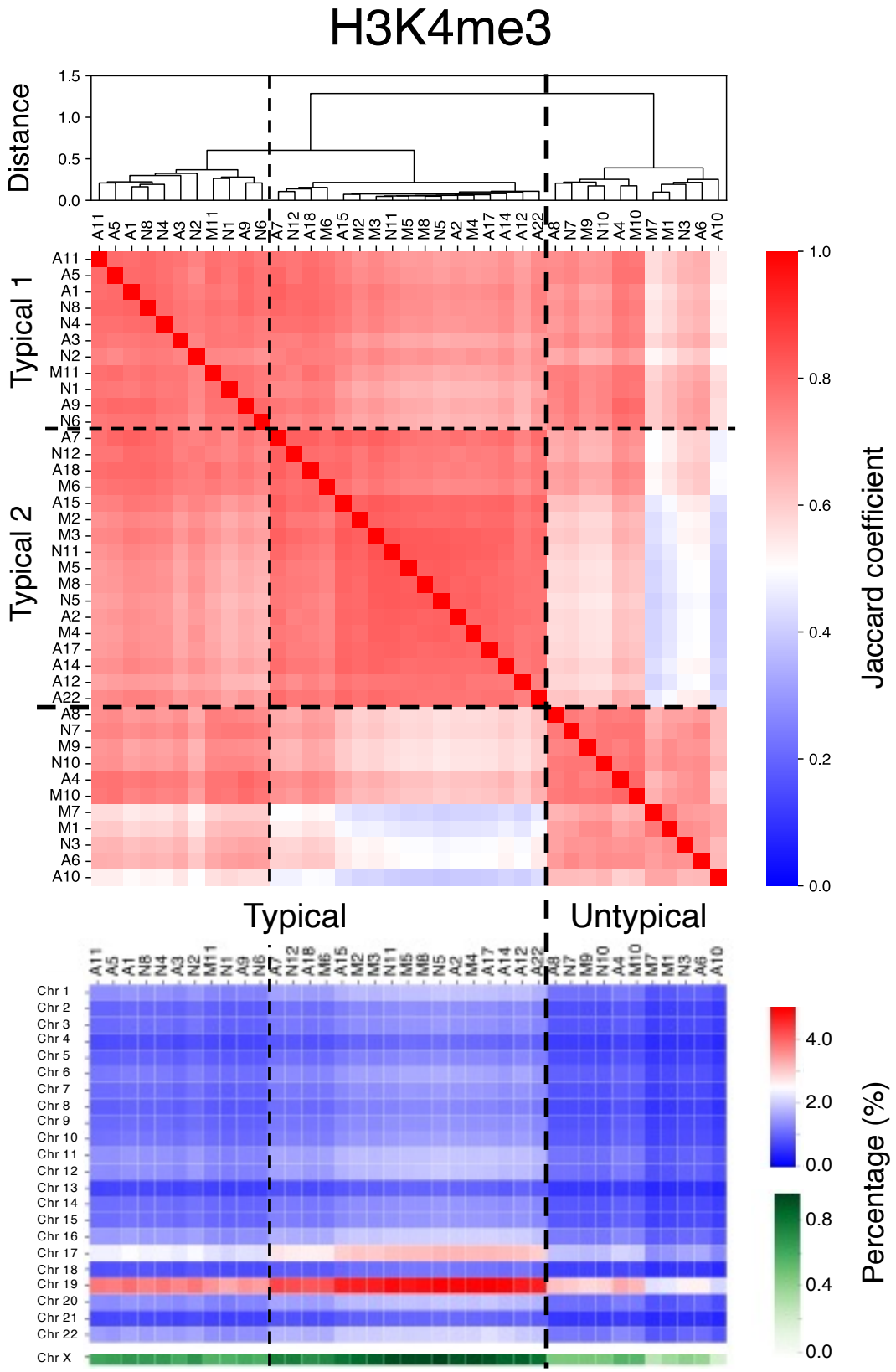
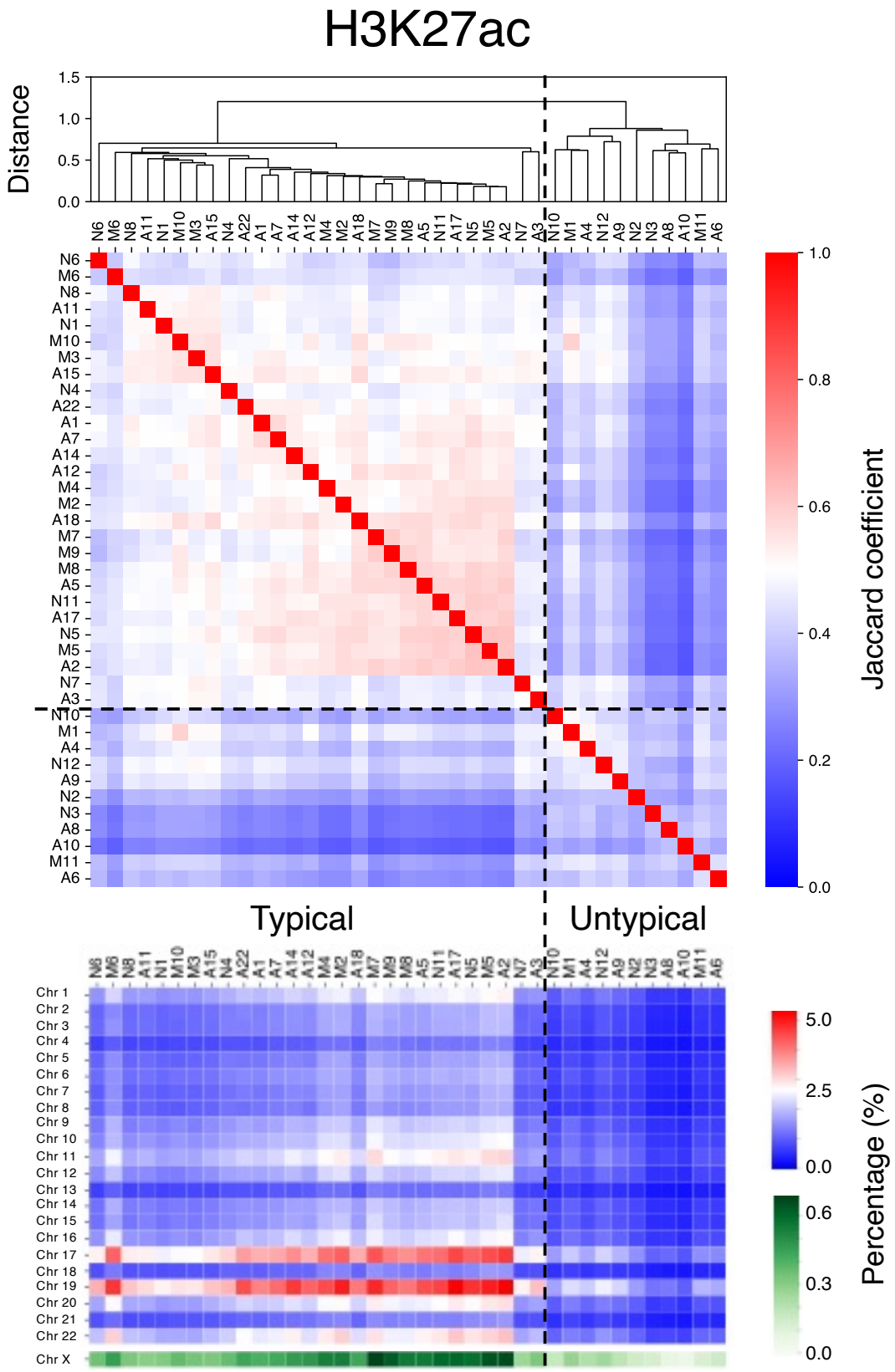


Fig. 4 Chromosomal distribution of genes belonging to the CA1 or CB1 cluster (Fig. 3a), and that of genes belonging to the CA2 or CB2 cluster (Fig. 3b). The chromosomal distribution of coding and non-coding genes was estimated using the human genome data GRCh38.


H3K4me3
Distance
1.5
1.0
0.5
0.0
A11 A5 A1 N8 N4 A3 N2 M11 N1 A9 N6 A7 N12 A18 M6 A15 M2 M3 N11 M5 M8 N5 A2 M4 A17 A14 A12 A22 A8 N7 M9 N10 A4 M10 M7 M1 N3 A6 A10
Typical 1
Typical 2
Jaccard coefficient
1.0
0.8
0.6
0.4
0.2
0.0
Typical
Untypical
Chr 1
Chr 2
Chr 3
Chr 4
Chr 5
Chr 6
Chr 7
Chr 8
Chr 9
Chr 10
Chr 11
Chr 12
Chr 13
Chr 14
Chr 15
Chr 16
Chr 17
Chr 18
Chr 19
Chr 20
Chr 21
Chr 22
Chr X
Percentage (%)
4.0
2.0
0.0
0.8
0.4
0.0

1229 **Fig. 5** Jaccard coefficient of the genome-wide distribution of H3K4me3 among
1230 individuals. Individuals were divided into two groups at the first divergence of the
1231 dendrogram. The epigenetic state exhibited by individuals belonging to the group
1232 containing the larger number of individuals was regarded as the "typical" epigenetic
1233 state, whereas that exhibited by individuals belonging to the group containing the fewer
1234 individuals was regarded as the "untypical" epigenetic state. Further, individuals
1235 belonging to the "typical" group were divided into two groups, named Typical 1 and
1236 Typical 2, at the second divergence of the dendrogram. Dashed lines indicate the
1237 boundaries between groups.


H3K27ac
Distance
1.5
1.0
0.5
0.0
N6 M6 N8 A11 N1 M10 M3 A15 N4 A22 A1 A7 A14 A12 M4 M2 A18 M7 M9 M8 A5 N11 A17 N5 M5 A2 N7 A3 N10 M1 A4 N12 A9 N2 N3 A8 A10 M11 A6
Jaccard coefficient
1.0
0.8
0.6
0.4
0.2
0.0
Typical
Untypical
Chr 1
Chr 2
Chr 3
Chr 4
Chr 5
Chr 6
Chr 7
Chr 8
Chr 9
Chr 10
Chr 11
Chr 12
Chr 13
Chr 14
Chr 15
Chr 16
Chr 17
Chr 18
Chr 19
Chr 20
Chr 21
Chr 22
Chr X
Percentage (%)
5.0
2.5
0.0
0.6
0.3
0.0

1239 **Fig. 6** Jaccard coefficient of the genome-wide distribution of H3K27ac among
1240 individuals. Analysis was conducted similar to that described in Fig. 5.
1241

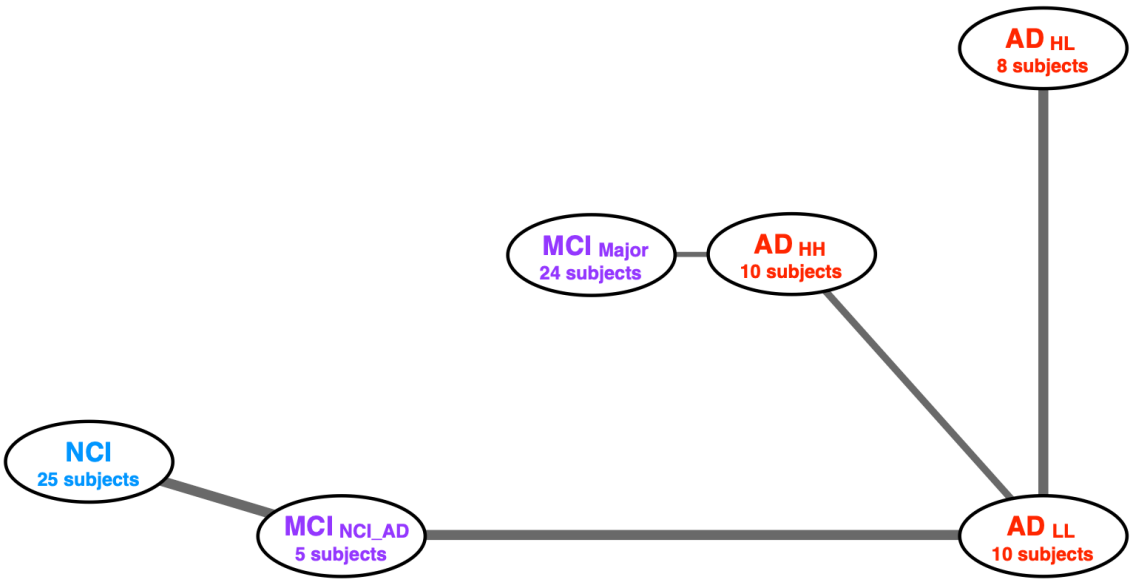


Fig. 7 Adjacency network among substages. The network constructed using common connections obtained from both the network constructed using the M-NA and N-A1 sets (Fig S5a) and that constructed using the M-NA and N-A2 sets (Fig. S5b). The width of each edge indicates the connection confidence between each pair of connected nodes.

(a)

# of genes	P^D_{NM}	77	145	211	286	360
		0.01	0.02	0.03	0.04	0.05
70	0.01	0, 5	0, 5	0, 5	0, 5	0, 5
147	0.02	1, 5	0, 5	0, 5	0, 5	2, 5
224	0.03	0, 5	0, 3	0, 5	0, 5	0, 5
317	0.04	4, 5	1, 5	0, 5	0, 5	0, 5
414	0.05	×	0, 5	0, 5	0, 5	0, 5

P^D_{MN}

# of genes	P^t_{NM}	49	95	158	219	286
		0.01	0.02	0.03	0.04	0.05
78	0.01	1, 3	0, 3	0, 3	1, 4	0, 5
164	0.02	1, 4	2, 2	0, 2	1, 5	0, 5
271	0.03	1, 4	0, 6	1, 2	0, 6	2, 3
392	0.04	1, 4	0, 5	0, 4	0, 8	0, 6
504	0.05	0, 8	0, 7	0, 6	0, 5	0, 5

P^t_{MN}

(b)

# of genes	P^D_{MA}	131	256	388	512	638
		0.01	0.02	0.03	0.04	0.05
138	0.01	3, 5	4, 3	5, 0	2, 9	3, 7
247	0.02	5, 0	5, 0	4, 1	6, 0	2, 5
372	0.03	5, 0	5, 0	6, 0	7, 0	3, 6
478	0.04	5, 9	4, 0	5, 0	5, 0	5, 2
593	0.05	×	4, 0	2, 7	4, 0	5, 0

P^D_{MA}

# of genes	P^t_{AM}	65	140	216	307	389
		0.01	0.02	0.03	0.04	0.05
92	0.01	4, 0	3, 3	4, 2	4, 2	5, 2
183	0.02	5, 1	4, 1	5, 0	5, 0	4, 1
302	0.03	3, 2	3, 2	4, 0	3, 0	4, 0
419	0.04	2, 2	3, 2	4, 0	3, 0	3, 0
526	0.05	×	3, 1	5, 0	5, 0	4, 0

P^t_{MA}

(c)

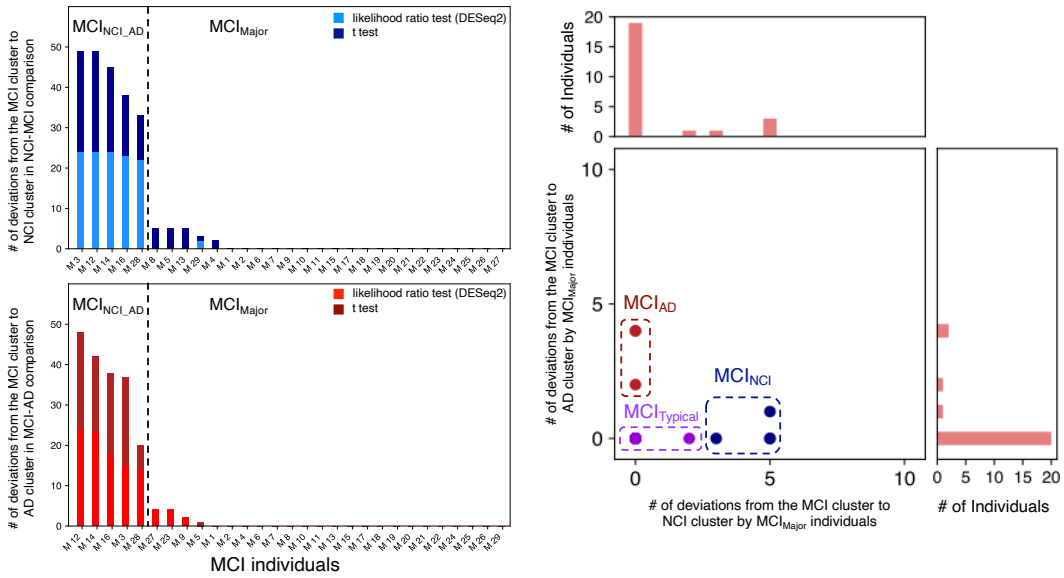


Fig. 8 Detailed classification of patients with MCI. Number of patients with NCI who belong to the MCI cluster (left value) and that of patients with MCI who belong to the NCI cluster (right value) as a function of P^D_{NM} and P^D_{MN} (a, left), and as a function of P^t_{NM} and P^t_{MN} (a, right). Number of patients with MCI who belong to the AD cluster (left value) and that of patients with AD who belong to the MCI cluster (right value) as

a function of P^{D}_{MA} and P^{D}_{AM} (b, left), and as a function of P^{t}_{MA} and P^{t}_{AM} (b, right). Number of deviations of each patient with MCI from the MCI to the NCI (c, upper left) and AD (c, lower left) cluster. Broken lines indicate boundaries between MCI_{NCI_AD} and MCI_{Major} individuals. Scatter plot of MCI_{Major} individuals as a function of the number of times that each individual belonged to the NCI and AD cluster (c, right). Using k-means clustering (Fig. S6c), MCI_{Major} individuals were divided into $MCI_{Typical}$, MCI_{NCI}, and MCI_{AD} groups.

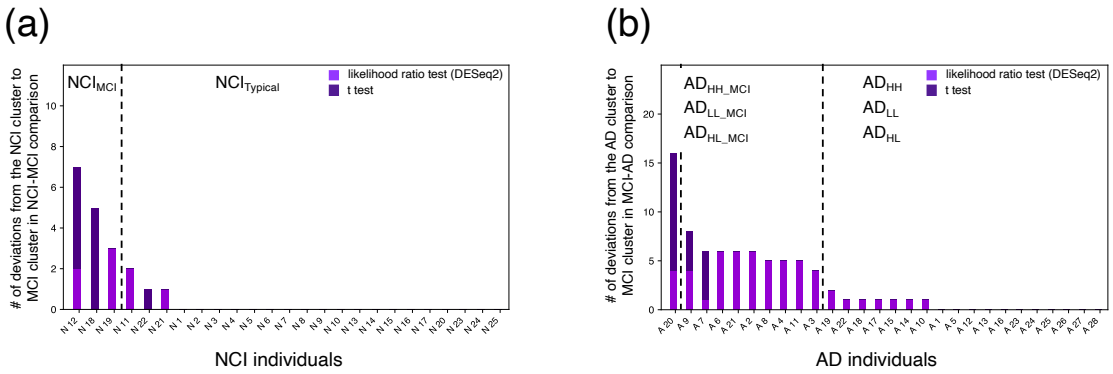


Fig. 9 Detailed classification of patients with NCI. Number of deviations of each patient with NCI from the NCI to the MCI cluster (a), and that of each patient with AD from the AD to the MCI cluster (b). Broken lines indicate boundaries among clusters of individuals obtained via k-means clustering (Fig. S7a and S7b).

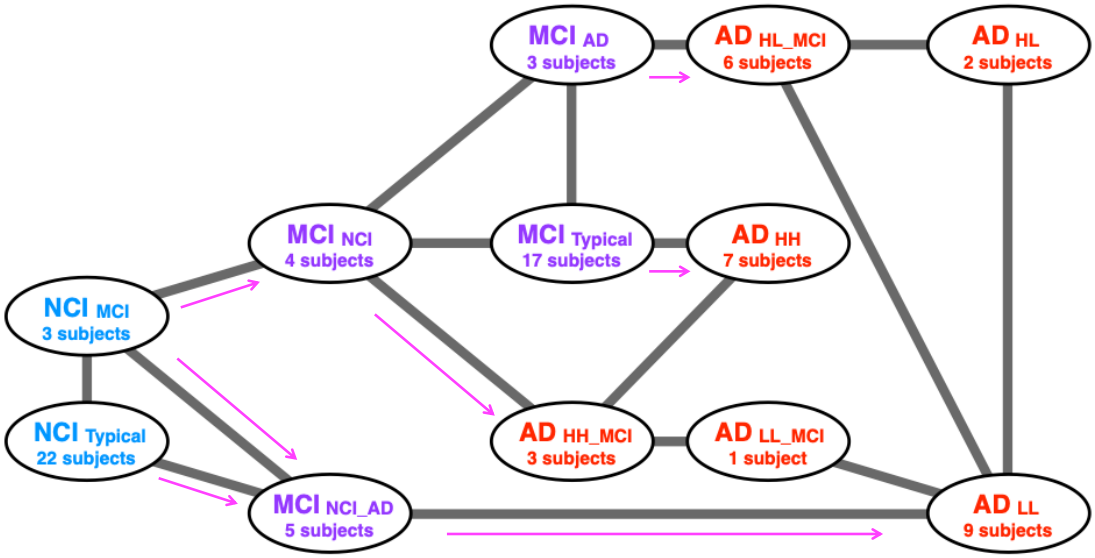


Fig. 10 Adjacency networks among substages. A network containing common connections from both networks obtained using the data in the M-NA and N-A1 sets (Fig. S8a) and that in the M-NA and N-A2 sets (Fig. S8b). The width of each edge indicates the connection confidence between each pair of connected nodes. Arrows indicated the inferred direction of disease progression.