Possibilities of multiple disease progression pathways in female Alzheimer's disease patients 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. Based on an analysis of publicly available data of women patients, this study found that the pathology of AD was suggested to be nonuniform; however, 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 and epigenome data. NCI was found to contain at least two substages, whereas MCI and AD were found to contain four and six substages, respectively. An inference of adjacency networks among substages, evaluated via partition-based graph abstraction using the gene expression profiles of individuals, suggested the possibilities of multiple typical disease progression pathways from NCI through different MCI substages and then to different AD substages. These results will aid future research on the pathological features and mechanisms of AD and induce changes in the strategic bases of conventional AD prevention and treatment.

INTRODUCTION

Alzheimer's disease (AD) is the most common cause of dementia and accounts for 60–70 % of dementia cases in older adults (Gatz et al., 2006; World Health Organization, 2017). 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 (Suresh et al., 2023). 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 (Hardy and Higgins, 1992; Hardy and Selkoe, 2002; Jin et al., 2011; Selkoe et al., 2016). However, cases have been reported in which AD develops without the accumulation of Aβ (Sekiya et al., 2018) 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 (Arriagada et al., 1992), the expression of apolipoprotein E variant APOE4 (Strittmatter et al., 1993), herpes virus and

coronaviruses (including COVID-19) infections (Itzhaki et al., 1997; Osaka et al., 2022; Readhead et al., 2018; Wang et al., 2022), diabetes (Kandimalla et al., 2017; Ott et al., 1999), brain inflammation (McGeer et al., 1995), oxidative damage in cells (Hirai et al., 2001; Nunomura et al., 2001; Wang et al., 2014), endoplasmic reticulum and ciliary dysplasia (Chakravarthy et al., 2012; Morelli et al., 2017; Plácido et al., 2014; Vorobyeva et al., 2018), myelin dysplasia (Bartzokis, 2004, 2011), and retrotransposons activities (Cheng et al., 2021; Macciardi et al., 2022; Ravel et al., 2021) 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, the variations in symptoms and disease progression rates among patients have been previously reported (Duara et al., 2021; Komarova et al., 2011; Younas et al., 2020, 2024).

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 (Jansen et al., 2019; Wightman et al., 2021; Yuen et al., 2020). It is notable that these characteristics are not always seen equally in all studies. One reason for this is the variations in technical factors among studies. However, the diversity of molecular conditions and gene expression patterns in the brain among patients with AD could also be a major contributor (Bellenguez et al., 2022; Johnson et al., 2022; Neff et al., 2021; Tijms et al., 2020, 2021).

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 various patients should be useful for such classification and characterization of disease diversity. The database of the Rush Alzheimer’s Disease study by ENCODE consortium (https://www.encodeproject.org/) (ENCODE Project Consortium, 2012; Luo et al., 2020) contains transcriptome data and various epigenome data such as histone modifications (H3K27ac, H3K27me3, and H3K4me3) and CCCTC-binding factor (CTCF) binding distribution of individuals in "No Cognitive Impairment (NCI)," "Mild Cognitive Impairment (MCI)," and "Alzheimer’s Disease (AD)" groups that are available for unlimited public use (in the following, call Rush AD database). Therefore, the analyses of these publicly available data of patients with different cognitive

impairment stages are expected to elucidate the diversity in the onset and progression of late-onset AD in more detail.

However, the Rush AD database seemed to contain a satisfactory amount of data for women individuals but the amount of data for men seemed insufficient, particularly for MCI and AD individuals, where the number of women individuals in NCI, MCI, AD groups was 25, 29, and 28, respectively, whereas that of men was 14, 4, and 7, respectively. Therefore, in this study, the classification of disease status and progression pathway of women AD patients were performed using transcriptome and epigenome data of women individuals in the Rush AD database as a first step of the analysis of the diversity of AD patients using multi-omics data. Of course, this study would be 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 were known to be different between male and female patients (Davis et al., 2020; Lopez-Lee et al., 2024).

In this study, patient diversity was considered by assessing the deviation from healthy individuals and the diversity of disease progression. For this purpose, we first searched for gene sets that characterize the differences between healthy individuals and patients. The gene set searched here was that which contained the largest number of genes from gene sets that could actually classify healthy subjects from patients using cluster analysis. The candidate genes included therein were selected through statistical tests of differences in expression levels. Second, we focused on the variety of 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 the structure of the disease progression pathway was predicted based on the similarity of 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 (Anders et al., 2010), 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 facts 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.

From the Rush AD database, in addition to RNA-seq data, the narrow peak data on genome-wide histone modifications (H3K27ac, H3K27me3, and H3K4me3) and CTCF binding distribution obtained by chromatin immunoprecipitation sequencing (ChIP-seq) could be acquired for 39 individuals: 11, 11, and 17 individuals of NCI, MCI, and AD patients, respectively (Table S1c). Here, for ChIP-seq data of H3K27ac, H3K27me3, H3K4me3, and CTCF, the correlation coefficients between the total read number mapped to the human genome and the number of narrow peaks were obtained as 0.193 (P-value = 0.239), 0.146 (P-value = 0.374), 0.027 (P-value = 0.872), and 0.333 (P-value = 0.038), respectively. These results suggested the number of narrow peaks of H3K27ac, H3K27me3, and H3K4me3 seemed to be independent of the total number of mapped reads. For the data of CTCF, the correlation between the total number of mapped reads and the number of narrow peaks was not considered 0 but expected sufficiently small with which the total number of mapped reads did not influence the analytic results of narrow peaks seriously.

Cluster analysis of individuals' provided gene set that characterized the differences in major MCI individuals from others

To extract a set of genes that might characterize the differences among the NCI, MCI, and AD groups, 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 (Love et al., 2014) (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 (Fig. 1, Fig. S1a-S1e). When $P^{D}_{NMA} \leq 0.05$, a cluster with significantly more MCI individuals (MCI cluster) and a cluster with NCI, AD, and a small number of MCI individuals was formed by 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, individuals with NCI, MCI, and AD were not from their respective cluster; no clusters comprising solely of individuals with NCI, MCI, or AD were formed by the first divergence of the dendrogram.

We then performed the same analysis using analysis of variance (ANOVA), which is the most familiar test for evaluating differences of characteristic values among three or more groups, to determine the P-values using TPM values (Table S1b), where P^{A}_{NMA} was defined instead of P^{D}_{NMA}. Here, ANOVA was generally better for detecting differences and was expected to identify more candidate genes in a gene set than the likelihood ratio test. As a result, an MCI cluster and a cluster with NCI, AD, and a small number of MCI individuals were formed, as in the previous case (Fig. S1f-S1j). Additionally, when the P^{A}_{NMA} was 0.04, a cluster comprising solely of individuals with MCI was formed by the first divergence of the dendrogram (Fig. 1a). In contrast, we did not observe any clear separation between participants with NCI and those with AD, regardless of the P^{A}_{NMA} value.

It should be noted that we did not obtain any gene set by the P^{D}_{NMA} and P^{A}_{NMA} values that enabled the complete discrimination of NCI, MCI, and AD individuals. Additionally, depending on the P^{D}_{NMA} and P^{A}_{NMA} values, the individuals with MCI who deviated from the MCI cluster varied; however, 24 MCI individuals were always included in the MCI cluster (Fig. S1). This suggested that the MCI cluster consisted of 24 MCI individuals, named MCI_{Major} individuals, whereas the five other MCI individuals (M3, M12, M14, M16, and M28) who tended to deviate from the MCI_{Major} individuals, were named MCI_{NCI_AD} individuals. We assumed that the MCI_{Major} and

195 MCI_{NCI_AD} individuals were at the substages named MCI_{Major} and MCI_{NCI_AD},
196 respectively.
197
198 Moreover, the set of genes with P-values of the above ANOVA $\leq P^{A}_{NMA} = 0.04$ was
199 named the "M-NA set" as it characterized the differences between majority of MCI
200 individuals from NCI and AD individuals (Fig. 1b, Table S2a). In a 3-dimensional U-
201 map using M-NA set, we clearly observed the deviation of the MCI_{Major} individuals
202 from individuals at other substages, where MCI_{NCI_AD} were found in a mixed group of
203 NCI and AD individuals (Fig. 1c).
204
205 Note that the classification result of MCI individuals into MCI_{Major} and MCI_{NCI_AD}
206 individuals seemed to be independent of the ages of individuals. For example, the
207 numbers of individuals over and under 90 years old in MCI_{Major} were 19 and 5, and
208 those in MCI_{NCI_AD} were 3 and 2, respectively (Table S1a-S1b), by which the P-value of
209 Fisher's exact test was obtained 0.5688.
210
211 **Functional features of genes in the M-NA set**
212 We also considered the functional features of genes in the M-NA set. In this set, we
213 obtained 594 genes (290 upregulated and 304 downregulated genes in MCI_{Major}), of
214 which 349 genes (164 upregulated and 185 downregulated in MCI_{Major}) were annotated
215 (Fig. 1b) as protein coding or noncoding RNA genes using Entrez (Schuler et al., 1996).
216 Herein, both sets of upregulated and downregulated genes in MCI_{Major} contained some
217 previously reported AD risk-associated genes; the upregulated gene set contained
218 *CYP19A1, GSTM1, HSD11B1, IL10, MT-RNR1, PCK1, TRAK2, UBE2D1*, and *XBP1*,
219 whereas the downregulated gene set contained *MT-ATP8* and *NME8* (Hu et al. 2017).
220
221 Additionally, we performed enrichment analysis of all genes including genes with
222 higher and those with lower expression levels in the MCI_{Major} individuals (Table S2b–
223 S2d). We identified that genes with higher expression levels in MCI_{Major} individuals
224 were enriched in GO terms for the functions "response to lipopolysaccharide,"
225 "response to molecule of bacterial origin," "Toll-like receptor signaling pathway,"
226 "regulation of MHC class II biosynthetic process," "vitamin metabolic process," "water-
227 soluble vitamin metabolic process," "chemotaxis," and "taxis" (P-value $< 10^{-3}$ and FDR
228 < 0.5) (Table S2c). Whereas, we identified that genes with lower expression levels in
229 MCI_{Major} individuals were enriched in GO terms for the functions "mRNA 5'-splice site
230 recognition" and "mRNA cis splicing, via spliceosome" (P-value $< 10^{-3}$ and FDR < 0.5)

(Table S2d). Recent studies have suggested that some of these functional features characterize the pathology of AD; for example, the lipopolysaccharide found in the wall of all Gram-negative bacteria was reported to play a role in causing AD (Singh et al., 2022; Zhan et al., 2018), Toll-like receptors participate in neuronal cell death by apoptosis (Adhikarla et al., 2021; Wu et al., 2022), MHC class II-expressing cells may be involved in the degradation of neurons (Parachikova, et al., 2007), vitamins affect the generation and clearance of Aβ (Grimm et al., 2016), and microglia exhibit chemotaxis for the clearance of Aβ (Lau et al., 2023; Rogers et al., 2001).

Gene sets that differentiate groups of NCI and AD individuals

To extract a gene set that may characterize the difference between the NCI and AD individuals, we performed hierarchical clustering of individuals using the genes satisfying the following criteria: mean expression level (TPM) was higher in AD individuals than in NCI individuals, with smaller P-values in the likelihood ratio test than the P^D_{NA} threshold, or mean expression level (TPM) was lower in AD individuals than in NCI individuals, with smaller P-values in the likelihood ratio test than the P^D_{AN} threshold. Subsequently, we searched the sets of P^D_{NA} and P^D_{NA} values that divided the NCI and AD individuals 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 the NCI individuals and AD individuals completely separated (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. It should also be noted that the number of genes with a lower mean expression level (TPM) in the AD individuals than in the NCI individuals 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). Of course, this gene set contained all genes in other gene sets with smaller P^D_{NA} and P^D_{AN} 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. Here, 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 $P^t_{AN} \leq 0.05$, the NCI individuals and the AD individuals were completely

separated 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 $P^t_{AN} \leq 0.1$, the NCI and AD individuals were completely separated 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 the NCI and AD individuals (Fig. 2b, Table S3d). Conversely, the expression of all genes in the N-A_t and N-A_{D2} sets did not separate the NCI and AD individuals. 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 the NCI and AD individuals (Fig. 2c, Table S3e). Such gene sets are expected to involve various typical genes that characterize the differences between NCI and AD individuals. In this study, the gene set consisting of the N-A_t and N-A_{D1} sets, which could completely separate the NCI and AD individuals, was named the "N-A1 set" (Table S3d), whereas the gene set consisting of the N-A_t set and the maximum number of genes in the N-A_{D2} set with low P-values was named the "N-A2 set" (Table S3e).

AD individuals can be classified into three groups

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

Thereafter, we performed a similar analysis for the NCI individuals. Based on the shape of the dendrogram, the NCI individuals were divided into three or four subgroups (Fig. 2b, Fig. 3a). However, when the NCI individuals were assumed to be divided into three subgroups, ANOVA revealed only nine genes with FDR < 0.1 among the three subgroups (Table S3d). Additionally, when NCI individuals were divided into four

subgroups, ANOVA revealed only two genes with FDR < 0.1 among the three 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 of AD individuals, which can be classified into three types, were highly nonuniform. AD individuals exhibiting high gene expression levels in both clusters CA1 and CB1 were regarded as individuals in the AD_{HH} subgroup (named AD_{HH} individuals), those exhibiting low gene expression levels in both CA1 and CB1 were regarded as individuals in the AD_{LL} subgroup (named AD_{LL} individuals), and those exhibiting high gene expression levels in cluster CA1 and low gene expression levels in cluster CB1 were regarded as individuals in the AD_{HL} subgroup (named AD_{HL} individuals).

Hierarchical clustering of the NCI and AD individuals using the N-A2 set also exhibited three possible subgroups (Fig. 2c, Fig. 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). Similar to a previous method, AD individuals were divided into AD_{HH}, AD_{LL}, and AD_{HL} groups based on the expression profiles of genes in clusters CA2 and CB2. The members of the AD_{HH}, AD_{LL}, and AD_{HL} groups were the same in both the N-A1 and N-A2 sets (Fig. S3a).

Importantly, the AD_{HH}, AD_{HL}, and AD_{LL} individuals generally varied when different P^t_{NA} and P^t_{AN} values were employed to determine the genes in the N-At set. However, for small changes in the P^t_{NA} and P^t_{AN} values, less exchange of individuals was observed among subgroups. For example, considering the case where $P^t_{NA} = 0.06$ and $P^t_{AN} = 0.1$ were employed instead of $P^t_{NA} = 0.08$ and $P^t_{AN} = 0.1$ to determine the N-At. Herein, the gene set named N-A3 was defined as that consisting of genes in the N-A_t set and the maximum number of genes in N-A_{D1} set with low P-values that could completely separate the NCI and AD individuals, whereas the gene set named N-A4 was defined as that consisting of genes in the N-A_t set and the maximum number of genes in N-A_{D4} set with low P-values that could completely separate the NCI and AD individuals. The classification of AD individuals into the AD_{HH}, AD_{LL}, and AD_{HL} groups was similar to the results obtained for N-A1, N-A2, N-A3, and N-A4, where 24 of the 28 individuals belonged to the same subgroups (Fig. S3). This suggested the robustness of our method in classifying the AD_{HH}, AD_{HL}, and AD_{LL} individuals.

339 Additionally, note that the classification result of AD individuals into AD_{HH}, AD_{HL}, and
340 AD_{LL} individuals seemed to be independent of the ages of individuals. For example, the
341 numbers of individuals over and under 90 years old in AD_{HH} were 8 and 2, those in
342 AD_{LL} were 6 and 4, and those in AD_{HL} were 4 and 4, respectively, by which the P-value
343 of Fisher's exact test was obtained 0.4850.
344
345 **Common functional features of the N-A1 and N-A2 sets**
346 We also considered the common functional features of the N-A1 and N-A2 sets. The N-
347 A1 set contained 1477 genes (1020 upregulated and 457 downregulated, AD), and 933
348 genes (683 upregulated and 250 downregulated) were annotated as protein coding or
349 noncoding RNA genes using Entrez (Schuler et al., 1996) (Fig. 3a). The N-A2 set
350 contained 1424 genes (947 upregulated and 477 downregulated, AD), and 889 genes
351 (619 upregulated and 270 downregulated) were annotated as protein coding or
352 noncoding RNA genes using Entrez (Schuler et al., 1996) (Fig. 3b). The upregulated
353 genes in both the N-A1 and N-A2 set contained *AR*, *CBS*, *CLU*, *CYP19A1*, *DBH*,
354 *GNA11*, *HLA-A*, *IQCK* , *MPO*, *MS4A4E*, *MX1*, *NLRP3*, *SDC2*, *TAP2*, and *ZFHX3*
355 (Table S3d-e), all of which were reported to be AD risk-associated genes in recent
356 studies (Hu et al. 2017; Xiang et al. 2018; Rahman et al. 2019; Yang et al. 2022).
357 Similarly, the downregulated genes in both the N-A1 and N-A2 set contained *COL25A1*
358 and *TARDBP* (Table S3d-e), which were also reported to be AD risk-associated genes
359 (Hu et al. 2017).
360
361 Additionally, the clusters CA1, CB1, CA2, and CB2 contained 71, 215, 64, and 181
362 annotated genes, respectively (Fig. 3). We found *CLU*, *SDC2*, and *ZFHX3* in both the
363 CA1 and CA2 clusters, whereas *DBH*, *GNA11*, *IQCK*, *MPO*, and *NLRP3* were found in
364 both the CB1 and CB2 clusters (Table S3d-e), thereby suggesting that these genes are
365 responsible for the differences in the features among the three substages of AD (Hu et
366 al. 2017; Yang et al. 2022).
367
368 Next, we performed enrichment analysis of all genes, including those with higher and
369 those with lower expression levels in the AD individuals for both the N-A1 and N-A2
370 set (Table S3f–S3o). We identified that genes with higher expression levels in AD
371 individuals in both the N-A1 and N-A2 set were commonly enriched in GO terms for
372 "forebrain dorsal/ventral pattern formation," "mRNA 5'-splice site recognition," and
373 "negative regulation of NF-kappaB transcription factor activity" (P-value $< 10^{-3}$ and
374 FDR < 0.5) (Table S3g, S3j). Whereas, genes with lower expression levels in AD

individuals in both the N-A1 and N-A2 set were enriched in GO terms for the functions "negative regulation of intrinsic apoptotic signaling pathway in response to DNA damage by p53 class mediator," "negative regulation of DNA damage response, signal transduction by p53 class mediator," and "coronary vasculature morphogenesis" (P-value < 10^{-3} and FDR < 0.5) (Table S3h, S3k). In the nervous system, NF-κB is known as one of the crucial components in the molecular switch that converts short- to long-term memory (Kaltschmidt et al., 2015). Thus, the upregulation of genes that suppress NF-κB expression is expected to exacerbate the pathology of AD. Additionally, the downregulation of genes that suppress the apoptosis in response to DNA damage would also promote AD progression through the enhancement of cell death in the central nervous system (Lin et al., 2020; Ranganathan et al., 2022).

We found that genes in both the CA1 and CA2 clusters were enriched in GO terms for the functions "response to xenobiotic stimulus", "cellular response to lipid", "negative regulation of small molecule metabolic process", "lung development", "cellular response to external stimulus", "negative regulation of carbohydrate metabolic process", "neuron migration", and "positive regulation of epithelial cell differentiation" (P-value < 10^{-3} and FDR < 0.5) (Table S3l, S3n). Whereas, genes in both the CB1 and CB2 clusters were enriched in GO terms for the functions "mitotic spindle assembly checkpoint signaling" and "L-alpha-amino acid transmembrane transport" (P-value < 10^{-3} and FDR < 0.5) (Table S3m, S3o). Notably, recent studies suggested that the perturbed cerebral carbohydrate metabolism contributes to the pathogenesis of AD (Chen et al., 2013; Cunnane et al., 2010), whereas the neuron migration ability contributes to neurogenesis in brain (Esteve et al., 2022; Young-Pearse et al., 2007). These facts suggested that the three substages of AD may be characterized by differences in these functional features.

Additionally, the chromosomal distribution of genes that belong to CA1 or CB1 cluster in N-A1 set, and that to CA2 or CB2 clusters in N-A2 set, were evaluated (Fig. 4). Compared to the chromosomal distribution of all human coding and noncoding genes, both genes set belonging to CA1 or CB1 clusters and that to CA2 or CB2 clusters seemed to accumulate at Chr. 1, 7, 9, 16, 19, and 22 (Fig. 4).

Epigenome data analysis supported classification of AD individuals into three groups

The data from the dorsolateral prefrontal cortex of Rush AD analyzed in the present study also included data on genome-wide distributions of epigenomic markers such as

411 histone modification (H3K4me3, H3K27me3, and H3K27ac) and CTCF binding sites
412 distributions for approximately half of all individuals (Table S1). Subsequently, we
413 evaluated the similarity of such epigenomic features among individuals using Jaccard
414 coefficients of the genome-wide distributions of epigenome status to classify them
415 based on other transcriptome features. At the first divergence of the dendrogram
416 according to the hierarchical clustering of individuals using Jaccard coefficients,
417 individuals were divided into two groups: a group with a typical distribution of
418 epigenome status and that with an untypical one, for respective epigenomic markers
419 (Fig. 5-6, Fig S4). The Jaccard coefficients of H3K27ac, H3K27me3, and CTCF among
420 individuals suggested that more AD_{HL} individuals exhibited untypical patterns
421 compared with those in other AD individuals (Fig. 5, Table 2a–2c, Fig. S4).
422 Additionally, a larger number of AD_{HH} exhibited untypical genome-wide distribution
423 patterns of H3K4me3 compared with those in other AD individuals (Fig. 6, Table 2d).
424 These facts suggested that the AD_{HH}, AD_{HL}, and AD_{LL} individuals exhibited different
425 epigenomic features.
426
427 Furthermore, we evaluated the epigenetic differences between individuals using the
428 density of each epigenomic marker on each chromosome. As a result, we observed a
429 large difference in the density of histone modification on chromosomes 17, 19, and X
430 between Typical and Untypical groups, and the density of CTCF binding mainly on
431 chromosomes 17 and 19. This fact suggested that the density of H3K27ac and
432 H3K27me3 on chromosomes 17, 19, and X, and CTCF on chromosomes 17, 19 differed
433 between AD_{HL} individuals and those at other substages. Additionally, H3K4me3 on
434 chromosomes 17, 19, and X also differed between AD_{HH} individuals and those at other
435 substages.
436
437 Notably, chromosomes 17, 19, and X contain the Tau, APOE, and KDM6A genes,
438 respectively, which are deeply involved in the progression of Alzheimer's disease.
439 However, they were found neither M-NA set, N-A1 set, nor N-A2 set.
440
441 **Adjacency network among substages**
442 The data on the expression levels of genes in the M-NA set revealed the differences
443 between typical MCI, NCI, and AD individuals, whereas those in the N-A1 or N-A2
444 sets revealed the differences between NCI and AD individuals. To clarify the
445 relationships among disease substages and the pathways of disease progression through
446 these substages, we described the adjacency networks among substages using these

447 gene sets by performing the partition-based graph abstraction (PAGA) method (Wolf et
448 al., 2019). Using PAGA, we generated a graph of substages in which the edge weights
449 represented confidence in the presence of connections. Notably, the confidence of
450 connections between NCI and substages in AD often involved finite values (Table S4);
451 however, the direct connection between NCI and substages in AD seemed unnatural in
452 disease progression. Therefore, we assumed that only connections with confidence
453 values sufficiently larger than the maximum confidence values of the connection
454 between NCI and substages in AD (MD_{N_A}) were effective for the adjacency network
455 among substages (Fig. S5). Subsequently, we used connections with confidence values
456 larger than twice that of MD_{N_A} to construct the adjacency network (Table S4).
457
458 Changes in gene expression patterns that characterize disease states are considered to
459 occur gradually as the disease progresses. Therefore, the disease appeared to progress
460 from NCI to AD along the previously obtained adjacency network. Note that the
461 adjacency network obtained using the M-NA and N-A1 sets (Fig S5a) and that
462 constructed using the M-NA and N-A2 sets (Fig. S5b) exhibited a little different
463 structure because the contained genes in these sets were a little different. Thus, we
464 focused on the network constructed using common connections obtained from both
465 networks constructed using the M-NA and N-A1 sets (Fig S5a) and the M-NA and N-
466 A2 sets (Fig. S5b) (Fig. 7). We did not obtain any connections from NCI to MCI_{Major},
467 which is the major group of MCI individuals, in any of the networks (Fig. 7, Fig S5a-b),
468 which seemed unnatural for the progression of AD. Thus, we performed more detailed
469 subgroup classifications of NCI, MCI, and AD individuals to reveal the relationship
470 between NCI and MCI_{Major} as described below.
471
472 **Detailed classification of MCI by comparing between NCI and MCI and between**
473 **MCI and AD**
474 To classify MCI individuals more comprehensively, we conducted comparisons
475 between NCI and MCI individuals and between MCI and AD individuals. First, we
476 performed hierarchical clustering of individuals using the genes satisfying the following
477 criteria: mean expression level (TPM) was higher in the MCI group than in the NCI
478 group with smaller P-values in the likelihood ratio test than the P^D_{MN} threshold or the
479 mean expression level (TPM) was lower in the MCI group than in the NCI group, with
480 smaller P-values in the likelihood ratio test than the P^D_{NM} threshold. We performed
481 cluster analysis using each gene set within the range $P^D_{NM} \leq 0.05$ and $P^D_{MN} \leq 0.05$.
482 However, we did not identify any cases in which the NCI and MCI individuals were

483 completely separated (Fig. 8a, Fig. S6a). However, clusters with significantly more NCI
484 individuals (NCI clusters) and MCI clusters were formed within all cases of P^D_{NM} and
485 P^D_{MN} except a case of $P^D_{NM} = 0.05$ and $P^D_{MN} = 0.01$ (Fig. 8a).
486
487 When we used two-tailed *t*-tests instead of the likelihood ratio test to determine the P-
488 values, the results were qualitatively the same. P^t_{MN} and P^t_{NM} were defined instead of
489 P^D_{MN} and P^D_{NM}, respectively. Analysis using each gene set within the ranges $P^t_{MN} \leq 0.05$
490 and $P^t_{NM} \leq 0.05$ revealed no cases in which the NCI and MCI individuals were
491 completely separated (Fig. 8a, Fig. S6a). However, the NCI and MCI clusters were
492 formed within all cases of P^t_{NM} and P^t_{MN} (Fig. 8a).
493
494 Next, we performed the same analyses for the MCI and AD individuals. In this case,
495 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},
496 and P^t_{MN}, respectively. Similar to the above results, we did not identify any cases in
497 which the MCI and AD individuals were completely separated (Fig. 8b, Fig. S6b).
498 However, the MCI and AD clusters were formed within all cases of P^D_{MA}, P^D_{AM}, P^t_{MA},
499 and P^t_{AM} except a 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}
500 $= 0.01$ (Fig. 8b).
501
502 In this study, the MCI individuals who deviated from the MCI cluster varied depending
503 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 individual with
504 MCI, we counted the number of times they belonged to the NCI cluster upon changing
505 the P^D_{NM}, P^D_{MN}, P^t_{NM}, and P^t_{MN} values, and those belonging to the AD cluster upon
506 changing the P^D_{MA}, P^D_{AM}, P^t_{MA}, and P^t_{AM} values (Fig. 8c). Consequently, individuals
507 M3, M4, M5, M8, M12, M13, M14, M16, M28, and M29 belonged to the NCI cluster
508 more than once, whereas individuals M3, M5, M9, M12, M14, M16, M23, M27, and
509 M28 belonged to the AD cluster more than once.
510
511 Notably, individuals M3, M12, M14, M16, and M28, (same as those in MCI_{NCI_AD}),
512 clearly exhibited a larger number of deviations from NCI and AD than those in MCI_{Major}
513 (Fig. 8c). To classify the MCI_{Major} individuals in more detail, we performed k-means
514 clustering using the number of times each individual belonged to the NCI cluster and
515 that belonging to the AD cluster (Fig. 8c, Fig. S6c). As a result, the MCI_{Major} individuals
516 were divided into three subtypes: $MCI_{Typical}$ individuals (17 individuals), who rarely or
517 never belonged to other clusters; MCI_{NCI} individuals (four individuals), who may
518 belong to NCI clusters only; and MCI_{AD} individuals (three individuals), who may

519 belong to AD clusters only (Table S1a-S1b). Thus, the MCI stage was divided into four
520 substages: $MCI_{Typical}$, MCI_{NCI}, MCI_{AD}, and MCI_{NCI_AD}.
521
522 **Detailed classification of NCI by comparing between NCI and MCI individuals**
523 The analysis performed in the previous section also showed that individuals N11, N12,
524 N18, N19, N21, and N22 belonged to the MCI cluster more than once when the P^D_{NM},
525 P^D_{MN}, P^t_{NM}, and P^t_{MN} values were changed (Fig. 9a). After k-means clustering of the
526 NCI individuals using the number of times each individual belonged to the MCI cluster,
527 the NCI individuals were divided into two clusters: one comprising N11, N21, and N22,
528 and the other comprising NCI individuals who never belonged to the MCI cluster (Fig.
529 9a, Fig. S7a). Therefore, these individuals were classified as typical NCI individuals
530 belonging to the MCI cluster and named $NCI_{Typical}$ individuals, and were assumed to be
531 in the $NCI_{Typical}$ substage. Individuals N12, N18, and N19 were classified as NCI
532 individuals close to MCI and named NCI_{MCI} individuals, and were assumed to be in the
533 NCI_{MCI} substage (Table S1a-S1b).
534
535 **Detailed classification of AD by comparing between AD and MCI individuals**
536 Similarly, 17 AD individuals belonged to the MCI cluster more than once, when the
537 P^D_{MA}, P^D_{AM}, P^t_{MA}, and P^t_{AM} values were changed (Fig. 9b). Using k-means clustering of
538 AD individuals based on the number of times each individual belonged to the MCI
539 cluster in the previous section, AD individuals were divided into three clusters, with
540 individuals A10, A14, A15, A17, A18, A19, A22, and other AD individuals who never
541 belonged to the MCI cluster being included in the same cluster (Fig. 9b, Fig S7b).
542 Therefore, these participants were expected to be in one of the typical AD substages:
543 AD_{HH}, AD_{HL}, or AD_{LL} (Table S1a-S1b). In contrast, individuals A2, A3, A4, A6, A7,
544 A8, A9, A11, A20, and A21 were assumed to be at one of the AD substages close to
545 MCI and named AD_{HH_MCI}, AD_{HL_MCI}, or AD_{LL_MCI} (Table S1a-S1b).
546
547 **Adjacency network among substages obtained after detailed classifications of**
548 **individuals suggests multiple disease progression pathways from NCI to AD**
549 Detailed classifications of individuals suggested that NCI was classified into $NCI_{Typical}$
550 and NCI_{MCI}; MCI was classified into $MCI_{Typical}$, MCI_{NCI}, MCI_{AD}, and MCI_{NCI_AD}; and
551 AD was classified into AD_{HH}, AD_{HH_MCI}, AD_{HL}, AD_{HL_MCI}, AD_{LL}, and AD_{LL_MCI}. We
552 constructed adjacency networks among these 12 substages using gene expression data
553 from the M-NA and N-A1 sets (Fig. S8a, Table S5a) and the M-NA and N-A2 sets (Fig.
554 S8b, Table S5b) as previously described. Subsequently, we focused on the adjacency

network that consisted of common connections between the network from the data in the M-NA and N-A1 sets and that from the data in the M-NA and N-A2 sets (Fig. 10). We identified four or more 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}, 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}, and AD_{HL_MCI}. This suggested that at least four disease progression pathways are responsible for the progression of the three classified AD substages.

Results for comparisons with recent studies

Many recent studies performed transcriptome analyses and identified various significant differentially expressed genes (significant DEGs) between healthy (control) individuals and patients with AD as "AD risk genes" (Hu et al. 2017; Xiang et al. 2018; Rahman et al. 2019; Yang et al. 2022). The cluster analysis of NCI, MCI, and AD individuals and that of NCI and AD individuals was performed using these AD risk genes (Fig. S9, Table S6). The results differed from the individual clustering results in Fig. 1, 2, and 3. It is noteworthy that most of the data of AD risk genes used here were obtained from research on the hippocampus, while the data analyzed in the previous sections of this study were obtained from the prefrontal cortex. One of the recent studies has shown that gene expression patterns differ between the hippocampus and the prefrontal cortex (Nef et al., 2021). These facts were expected to be the reason for such inconsistency.

From the data set analyzed in the previous sections, no significant DEGs between NCI and AD individuals and only 9 significant DEGs among NCI, MCI, and AD individuals were obtained (Table S1a-S1b, Table S7). Here, significant DEGs were evaluated by the likelihood test using DEseq2 with FDR < 0.01. Notably, the abovementioned study reported that not many significant DEGs were yielded in areas belonging to the prefrontal cortex (Nef et al., 2021). These facts suggested to support the consistency between the results of recent and present studies.

DISCUSSION

In this study, by the analysis of the data from women individuals in Rush AD database, the pathological progression pathway of AD was suggested to be nonuniform, with different typical forms of deterioration. First, cluster analysis of individuals using gene

expression data from those diagnosed with NCI, MCI, and AD through clinical trials suggested that NCI contained at least two substages, whereas MCI and AD contained four and six substages, respectively. The differences among AD substages were characterized based on epigenomic features such as the genome-wide distributions of histone modifications and CTCF binding. Additionally, some of the genes and gene sets with respective GO terms that were suggested as AD-associated in recent studies seemed to contribute to the characterization of each AD substage. Second, an inference of adjacency networks among the different substages suggested the existence of four typical NCI → AD disease progressions from different MCI to different AD substages. These typical pathological progression pathways were unknown, so these findings on disease progression features will contribute to the development of AD research. Additionally, they are expected to induce changes in the strategic bases of AD prevention and treatment.

Recent studies using hippocampal and frontal lobe transcriptomes have also suggested the existence of multiple AD subtypes. However, the relationship between multiple states, such as their connection in series on the same disease progression path or on different paths, has not been clarified. Furthermore, these studies did not focus on the details of the diversity of NCI and MCI. In contrast, in this study, both NCI and MCI individuals were divided into multiple subgroups at different substages based on their respective gene expression characteristics. Additionally, estimating the connectivity among substages clearly revealed multiple pathological pathways for disease progression.

This study analyzed data from the dorsolateral prefrontal cortex of Rush AD database, rather than the hippocampus, which has been the focus of many previous studies. The reason for the analysis of this dataset in this study was that, for some individuals, this dataset included not only transcriptome data but also data of genome-wide distributions of histone modification (H3K4me3, H3K27me3, H3K27ac) and CTCF binding sites (Table S1). Such datasets can potentially reveal the relationship between a wide range of transcriptional regulatory states and disease conditions. Recently, a relationship between changes in CTCF binding and transcription activities associated with changes in pathological conditions was found by analyzing such datasets (Patel et al., 2023). The analysis of these epigenome data in the present study suggested that individuals in three different AD substages exhibited different respective epigenomic features, particularly in chromosomes 17, 19, and X.

It is also notable that these three subgroups of AD were classified according to the difference in the expression level of genes that tended to be enriched on Chr. 1, 7, 9, 16, 19, and 22 (Fig. 4). Here, the differences in expression levels of genes on chromosome 19 among three AD groups may be directly related to the different epigenetic status of chromosome 19 among three AD groups. However, the expression of genes on other chromosomes may be indirectly influenced by the differences in the epigenetic status of chromosomes 17, 19, and X. For these facts, we proposed the following hypothesis regarding the possible relationships between the differences in epigenome status and those in expression levels of genes among three AD subgroups.

First, changes in gene expression are strongly influenced by histone modifications in the promoters and enhancers of each gene but are not significantly influenced by modifications in other regions (even gene regions) in general. Therefore, it is not unnatural that differences in the distribution of histone modifications across the genome of chromosomes 17, 19, and X. Moreover, the distribution of CTCF binding that controls them did not directly lead to significant differences in the expression of genes on those chromosomes.

In contrast, changes in the distribution of histone modifications throughout the genome change the distribution of heterochromatin along the sequence of each genome, which changes the physical structure and properties of the genome (Komoto et al., 2022; Fujishiro et al., 2022). Such changes in the structure and properties of each chromosome are also known to dramatically change the intranuclear location of each genomic region (Finan et al., 2011; Komoto et al., 2022; Fujishiro et al., 2022). Furthermore, it is known that such differences in the intranuclear location of chromosomes and local genomic regions change the gene expression of each chromosome through the formation and deformation of transcription factories (Iborra et al., 1996; Osborne et al., 2004). Based on these findings, the differences in the genome-wide distribution of histone modifications in chromosomes 17, 19, and X among the three AD subgroups induced different tendencies in the arrangement of each chromosome within the nucleus among the three AD subgroups. These differences may result in differences in the expression of genes contained in chromosomes 1, 7, 9, 16, 19, and 22.

To verify this hypothesis, it is necessary to carry out fluorescent imaging of intranuclear chromosomes in brain cells of various AD patients in the future.

Additionally, the following hypothesis can also be proposed. The histone modifications and CTCF bindings influence transcript splicing and transcriptional start sites, thus, contributing to various isoform formations (Alharbi et al., 2021; Awazu et al., 2023; Podlaha et al., 2014; Zhou et al., 2013). Each isoform may have different functions or may have opposite effects on the regulation of the expression of other genes (Yamamoto et al., 2016; Stevens et al., 2019; Awazu et al., 2023). Therefore, differences in the epigenetic status of chromosomes 17, 19, and X may alter the proportions of each isoform produced by the genes on these chromosomes, thereby influencing gene expression on other chromosomes.

However, the transcriptome data analyzed in the present study were obtained from short read sequencing, so that the estimation of the transcription levels of each isoform had various difficulties. In the future, long read sequencing should be performed for the precise estimation of the transcription levels of each isoform.

We also expected that, it may be possible to clarify the relationship between the differences in gene expression and the differences in the epigenomic state in chromosomes, such as 17, 19 and X, by examining the differences in each gene expression and the differences in the epigenetic state of each promoter and enhancer through a more detailed analysis of the epigenomic profile and gene expression profile. We consider that such considerations are also an important future challenge.

It is noted that many recent studies performed transcriptome analyses to identify significant DEGs between healthy and AD individuals. However, by focusing on the expression patterns of more widespread candidate genes rather than the significant DEGs, we found that different characteristics appeared in the dorsolateral prefrontal cortex between healthy and AD individuals and that disease substages occur in response to changes in the disease state.

Additionally, many studies focused on gene expression patterns in the hippocampus. One such previous research suggested that multiple states exist in the hippocampus of patients with AD based on differences in the gene expression profiles (Neff et al., 2021). Between the hippocampus and prefrontal cortex, as mentioned previously, gene expression patterns and their variations between healthy individuals and AD patients differ, and then, different results would be obtained even from the same analysis.

Importantly, the hippocampal data analyzed in previous studies can be analyzed in the same way as in this study; connections between substages can be extracted and pathological routes can be estimated, which are crucial issues that need to be addressed in the future.

However, the present study also had some limitations. We classified female individuals through the analyses of transcriptome and epigenomic data of women in the Rush AD database. The insights are expected to be useful not only for female patients but also for understanding the rapid progression of symptoms in male patients. To validate this expectation, the dataset from male individuals should be analyzed. In this dataset, however, the number of women patients in the NCI, MCI, and AD groups was 25, 29, and 28, respectively, whereas that of men was 14, 4, and 7, respectively, which was 3 times smaller than that of female individuals in total (7 times and 4 times smaller than those on female individuals for MCI and AD). Additionally, the number of men individuals with MCI and AD in the database (4 and 7) was fewer than that of individuals in each of the three AD groups obtained for women individuals. With such little data on male individuals, it was impossible to perform similar analyses on male individuals or sex-independent analyses. To perform such analyses, obtaining more data on male patients is one of the essential future issues that must be resolved for eliminating the potential sex-related bias.

To evaluate the differences in RNA-seq data between different groups when searching for gene sets for characterizing and classifying individuals, the likelihood ratio test using DESeq2, which is one of the typical methods, and *t*-test were used. The *t*-test is generally better for detecting differences and was expected to identify more candidate genes in a gene set than the likelihood ratio test or other nonparametric testing methods. However, the *t*-test produces more false positives than other testing methods when the fluctuations in the data deviate from a Gaussian distribution. The gene sets used in this study were expected to contain few false-positive genes because they were chosen under imposing the additional condition that the healthy individuals and patients could be clearly distinguished through cluster analysis. However, the possibility that inappropriate genes were retained is a valid case. 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 unprecedented knowledge, the validation and characterization of each substage through comparisons with other physiological data could not be presented. It is possible that such analysis could be processed using meta-analysis combined with databases focusing on other feature groups.

EXPERIMENTAL PROCEDURES

Data processing

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

Note that 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, any specific alerts regarding data quality according to ENCODE standards were not given to the RNA-seq data used in this study. Thus, the quality of the data was assumed to be guaranteed. In fact, the set of genes whose expression changes between NCI and AD and among NCI, MCI, and AD, involved many genes obtained from previous studies as AD risk genes. This also supports the lack of data bias that would contradict previous reports.

Statistical analyses

Two-tailed *t*-tests and ANOVA were performed using Python (Ver.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 (Benjamini et al., 1995) implemented in MultiPy (https://puolival.github.io/multipy/) (Puoliväli et al., 2020). 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 (Ver. 4.1.3).

Cluster analyses

Cluster analyses were performed using the values of $X = \log_2(TPM + 1) - C$, where C was chosen to satisfy the average of X over individuals = 0 for each gene. iDEP.96 (http://bioinformatics.sdstate.edu/idep96/) was used for hierarchical clustering (GE et

al., 2018). For hierarchical clustering, the group-average method, using the correlation coefficient as the distance, was used to create a dendrogram. If the cluster formed by the first branch of hierarchical clustering satisfied the following criteria for a certain group (NCI, MCI, and AD), then that cluster was regarded as a cluster in which there were significantly more individuals belonging to that group. These criteria were as follows: i) containing more than 2/3 of the individuals belonging to that group , and ii) Fisher exact test resulting in a P-value < 0.05. K-means cluster analysis was performed using the Python package scikit-learn (Ver.1.0.2) (Pedregosa et al., 2011). The number of clusters in each k-means cluster analysis was determined using the elbow method.
The construction of UMAP was performed using sc.tl.pca() (Pedregosa et al., 2011) with default parameters, scanpy.pp.neighbors() with "n_neighbors" = 30 (McInnes et al., 2018), and the scanpy.tl.umap() functions with n_components = 3 (Wolf et al., 2019) of scanpy (Ver.1.9.3) according to the literature on basic workflows for clustering analysis using Scanpy (https://scanpy-tutorials.readthedocs.io/en/latest/pbmc3k.html). All principal components were used for clustering using the scanpy.pp.neighbors() function. The obtained results were qualitatively independent of the parameters of these applications.

Epigenomic data analysis

The Jaccard coefficient of the genome-wide distributions of epigenomic markers among individuals was estimated as follows. The bed formatted files of the narrow peaks of these epigenome markers and their bam formatted files of the read mapping results were acquired for individuals whose ChIP-seq data of H3K27ac, H3K27me3, H3K4me3, and CTCF exist in Rush AD database. The total number of mapped reads was counted from the bam formatted file using samtools (Ver. 1.20). The correlation between the number of narrow peaks and the total number of mapped reads for each ChIP-seq data was evaluated by Pearson correlation coefficient. The Jaccard coefficients between each pair of individuals were estimated using Bedtools (Ver. 2.31.0). The density of each epigenomic marker on each chromosome was estimated according to the following equation: [Total length of genome regions with narrow peak of ChIP-seq data in chromosome obtained by bed formatted file (bp)]/[Total length of chromosome obtained by GRCh38 (bp)].

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) (Zhou et

al., 2019). 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 the substages was estimated via PAGA (Wolf et al., 2019) 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 confidence in the presence of connections. PAGA was performed using sc.tl.pca() (Pedregosa et al., 2011) with default parameters, scanpy.pp.neighbors() with “n_neighbors” = 2 (McInnes et al., 2018), and the scanpy.tl.paga() functions (Wolf et al., 2019) of scanpy (Ver.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 parameter "n_neighbors" of the function scanpy.pp.neighbors() was chosen as 2, which is the minimum value in the range of this parameter. Thus, a network with nonredundant connections among the 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

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

Alharbi, A. B., Schmitz, U., Bailey, C. G., & Rasko, J. E. (2021). CTCF as a regulator of alternative splicing: new tricks for an old player. *Nucleic acids research*, *49*(14), 7825-7838. https://doi.org/10.1093/nar/gkab520

Alves, F., Kalinowski, P., & Ayton, S. (2023). Accelerated brain volume loss caused by anti-β-amyloid drugs: a systematic review and meta-analysis. *Neurology*, *100*, e2114–e2124. https://doi.org/10.1212/wnl.0000000000207156

Anders, S., & Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, *11*, R106. https://doi.org/10.1186/gb-2010-11-10-r106

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

Awazu, A., Takemoto, D., Watanabe, K., & Sakamoto, N. (2023). Possibilities of skin coat color-dependent risks and risk factors of squamous cell carcinoma and deafness of domestic cats inferred via RNA-seq data. *Genes to Cells*, *28*(12), 893-905. https://doi.org/10.1111/gtc.13076

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

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

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

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

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

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

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

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

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 resilience in a mouse model of Alzheimer's disease. *Science translational medicine*, *12*(558), eaaz5677. https://doi.org/aaz5677

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

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

Finan, K., Cook, P. R., & Marenduzzo, D. (2011). Non-specific (entropic) forces as major determinants of the structure of mammalian chromosomes. *Chromosome research*, *19*, 53-61. https://doi.org/10.1007/s10577-010-9150-y

Fujishiro, S., & Sasai, M. (2022). Generation of dynamic three-dimensional genome structure through phase separation of chromatin. *Proceedings of the National Academy of Sciences*, *119*(22), e2109838119. https://doi.org/10.1073/pnas.2109838119

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

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

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

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

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

Higginbotham, L., Ping, L., Dammer, E. B., Duong, D. M., Zhou, M., Gearing, M., … Seyfried, N. T. (2020). Integrated proteomics reveals brain-based cerebrospinal fluid biomarkers in asymptomatic and symptomatic Alzheimer's disease. *Science Advances*, *6*, eaaz9360. https://doi.org/10.1126/sciadv.aaz9360

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

Iborra, F. J., Pombo, A., Jackson, D. A., & Cook, P. R. (1996). Active RNA polymerases are localized within discrete transcription 'factories' in human nuclei. *Journal of cell science*, *109*(6), 1427-1436. https://doi.org/10.1242/jcs.109.6.1427

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

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

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

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

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

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

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

Komoto, T., Fujii, M., & Awazu, A. (2022). Epigenetic-structural changes in X chromosomes promote Xic pairing during early differentiation of mouse embryonic stem cells. *Biophysics and Physicobiology*, *19*, e190018. https://doi.org/10.2142/biophysico.bppb-v19.0018

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

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

Lopez-Lee, C., Torres, E. R. S., Carling, G., & Gan, L. (2024). Mechanisms of sex differences in Alzheimer's disease. *Neuron*. https://doi.org/10.1016/j.neuron.2024.01.024

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology*, *15*, 1-21. https://doi.org/10.1186/s13059-014-0550-8

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

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

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

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

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

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

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

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

Osborne, C. S., Chakalova, L., Brown, K. E., Carter, D., Horton, A., Debrand, E., ... & Fraser, P. (2004). Active genes dynamically colocalize to shared sites of ongoing transcription. *Nature genetics*, *36*(10), 1065-1071. https://doi.org/10.1038/ng1423

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

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

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

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.

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

Podlaha, O., De, S., Gonen, M., & Michor, F. (2014). Histone modifications are associated with transcript isoform diversity in normal and cancer cells. *PLoS computational biology*, *10*(6), e1003611. https://doi.org/10.1371/journal.pcbi.1003611

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

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

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

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

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

Schuler, G. D., Epstein, J. A., Ohkawa, H., & Kans, J. A. (1996). Entrez: Molecular biology database and retrieval system. In Methods in enzymology (Vol. 266, pp. 141-162). Academic Press. https://doi.org/10.1016/S0076-6879(96)66012-1

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

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

Singh, S., Sahu, K., Singh, C., & Singh, A. (2022). Lipopolysaccharide induced altered signaling pathways in various neurological disorders. *Naunyn-Schmiedeberg's Archives of Pharmacology*, *395*(3), 285-294. https://doi.org/10.1007/s00210-021-02198-9

Stevens, M., & Oltean, S. (2019). Modulation of the apoptosis gene Bcl-x function through alternative splicing. *Frontiers in Genetics*, *10*, 804. https://doi.org/10.3389/fgene.2019.00804

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

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

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

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

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

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

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

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

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

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

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

Yamamoto, H., Rundqvist, H., Branco, C., & Johnson, R. S. (2016). Autocrine VEGF isoforms differentially regulate endothelial cell behavior. *Frontiers in cell and developmental biology*, *4*, 99. https://doi.org/10.3389/fcell.2016.00099

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

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

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

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

Zhan, X., Stamova, B., & Sharp, F. R. (2018). Lipopolysaccharide associates with amyloid plaques, neurons and oligodendrocytes in Alzheimer's disease brain: a review. *Frontiers in aging neuroscience*, *10*, 42. https://doi.org/10.3389/fnagi.2018.00042

Zhou, H. L., Luo, G., Wise, J. A., & Lou, H. (2013). Regulation of alternative splicing by local histone modifications: potential roles for RNA-guided mechanisms. *Nucleic acids research*, *42*(2), 701-713. https://doi.org/10.1093/nar/gkt875

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

1206 **Table 1**. Summary of the individuals included in this study (see Table S1a and S1b for
1207 more details).

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

1208

1209 **Table 2.** Number of individuals exhibiting typical genome-wide distributions of each
1210 epigenetic state, E_T, and that exhibiting the untypical one, E_U, in the three AD substages
1211 (AD_{HH}, AD_{LL}, and AD_{HL}); H3K27ac (a), H3K27me3 (b), CTCF (c), and H3K4me3 (d).
1212 The E_T and E_U in each substage were obtained from Figures 5, 6, and S4. P-values
1213 obtained using the Fisher's exact test for each table were 0.04328 (a), 0.04299 (b),
1214 0.04328 (c), and 0.1829 (d).
1215
1216 (a)

H3K27ac	AD_{HH}	AD_{LL}	AD_{HL}
# of individuals with typical (E_T)	6	4	3
# of individuals with untypical (E_U)	0	0	4

1217
1218 (b)

H3K27me3	AD_{HH}	AD_{LL}	AD_{HL}
# of individuals with typical (E_T)	5	4	2
# of individuals with untypical (E_U)	1	0	5

1219
1220 (c)

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

1221
1222 (d)

H3K4me3	AD_{HH}	AD_{LL}	AD_{HL}
# of individuals with typical (E_T)	2	2	6
# of individuals with untypical (E_U)	4	2	1

1223
1224

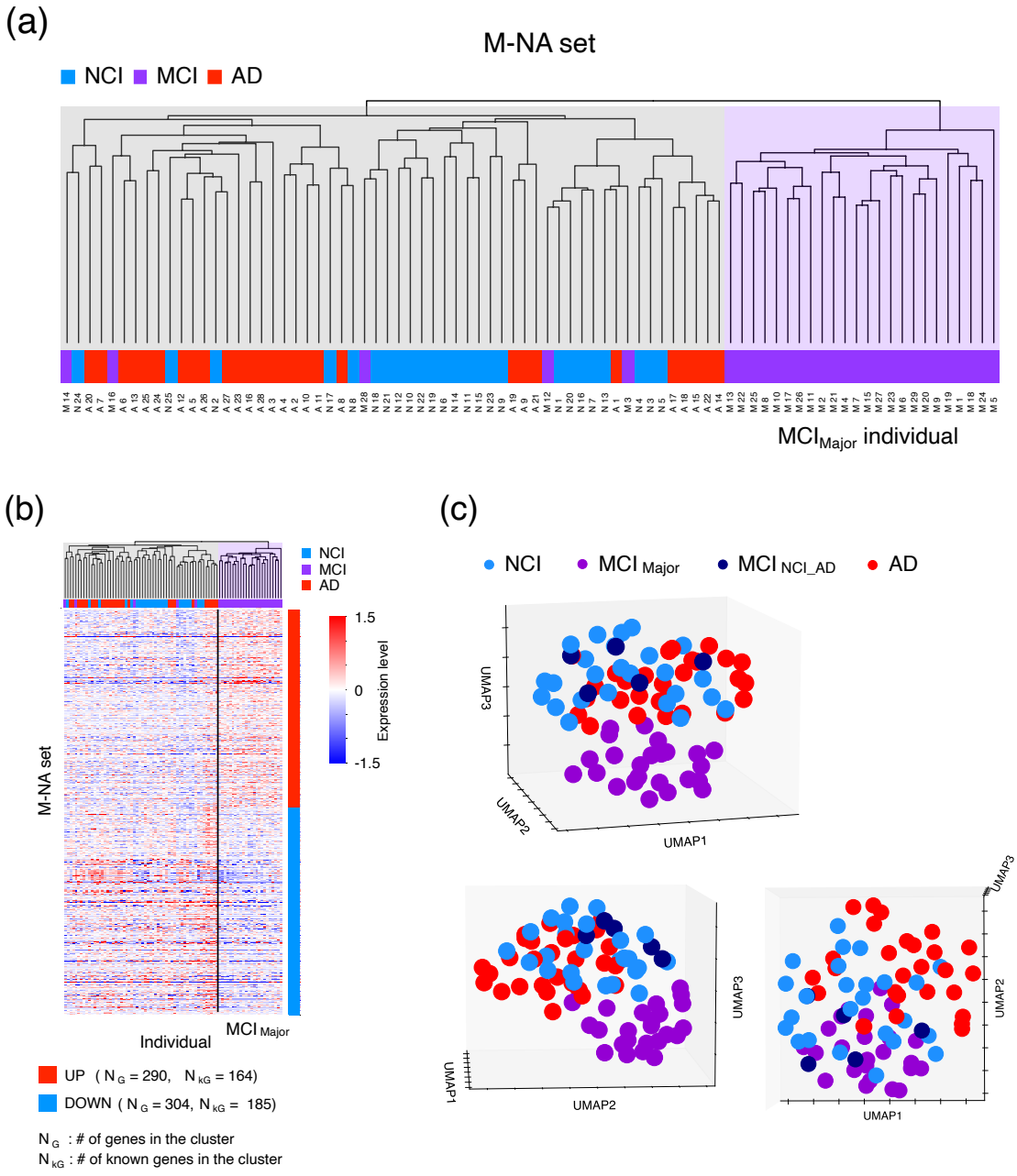
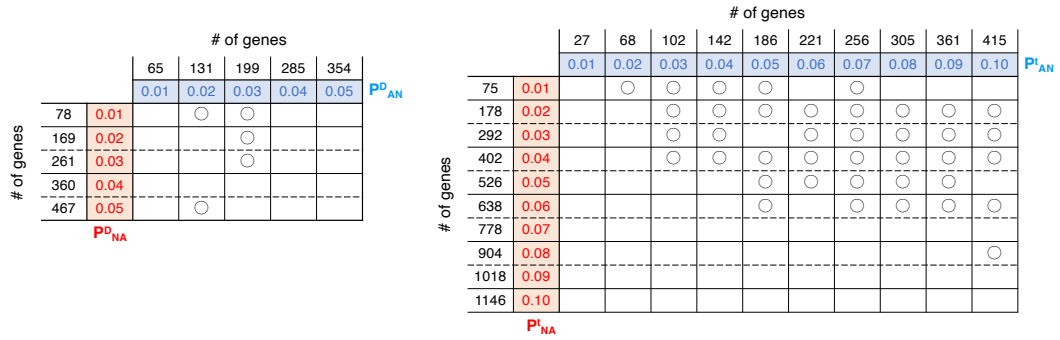


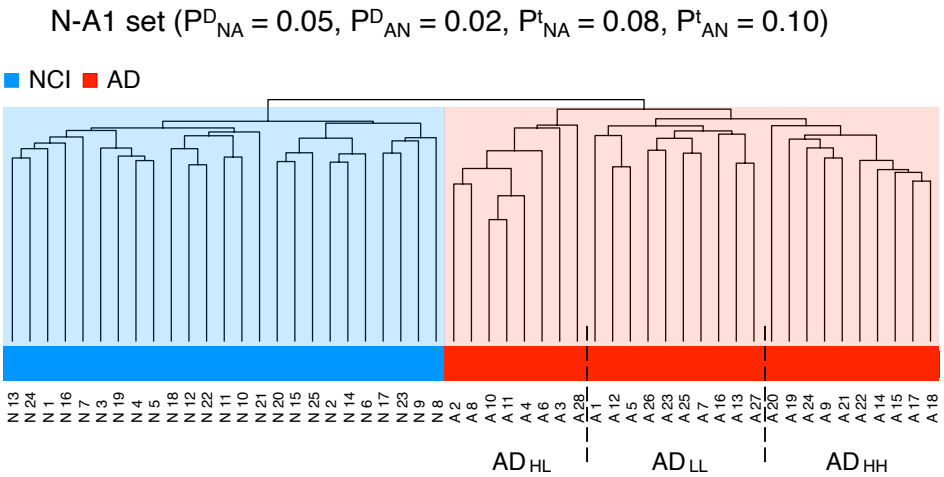
Fig. 1. Results of the hierarchical clustering of individuals, and result of UMAP, obtained by the M-NA set. Result of the hierarchical clustering of individuals in the case of the M-NA set ($P^{A}_{NMA} = 0.04$) (a). The expression level of each gene in the M-NA set of each individual was plotted (b) (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. The MCI_{Major} individuals occupy the right region in the dendrogram. N_G and N_{kG} were 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 the MCI_{Major}

individuals, respectively. Scatter plot of individuals in 3-dimensional space constructed using U-MAP 1, 2, and 3 axes shown from various directions (c).

(a)



(b)



(c)

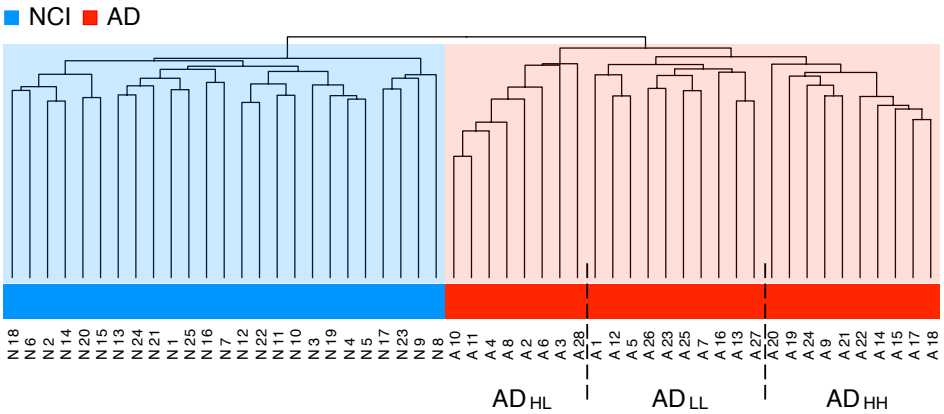


Fig. 2 Phase diagrams of the hierarchical clustering of NCI and AD individuals, and the results of the hierarchical clustering of individuals of the N-A1 and N-A2 sets. Phase diagram of the results of hierarchical clustering using genes where $p \leq P^D_{NA}$ and $p \leq P^D_{AN}$ (a, left), and that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

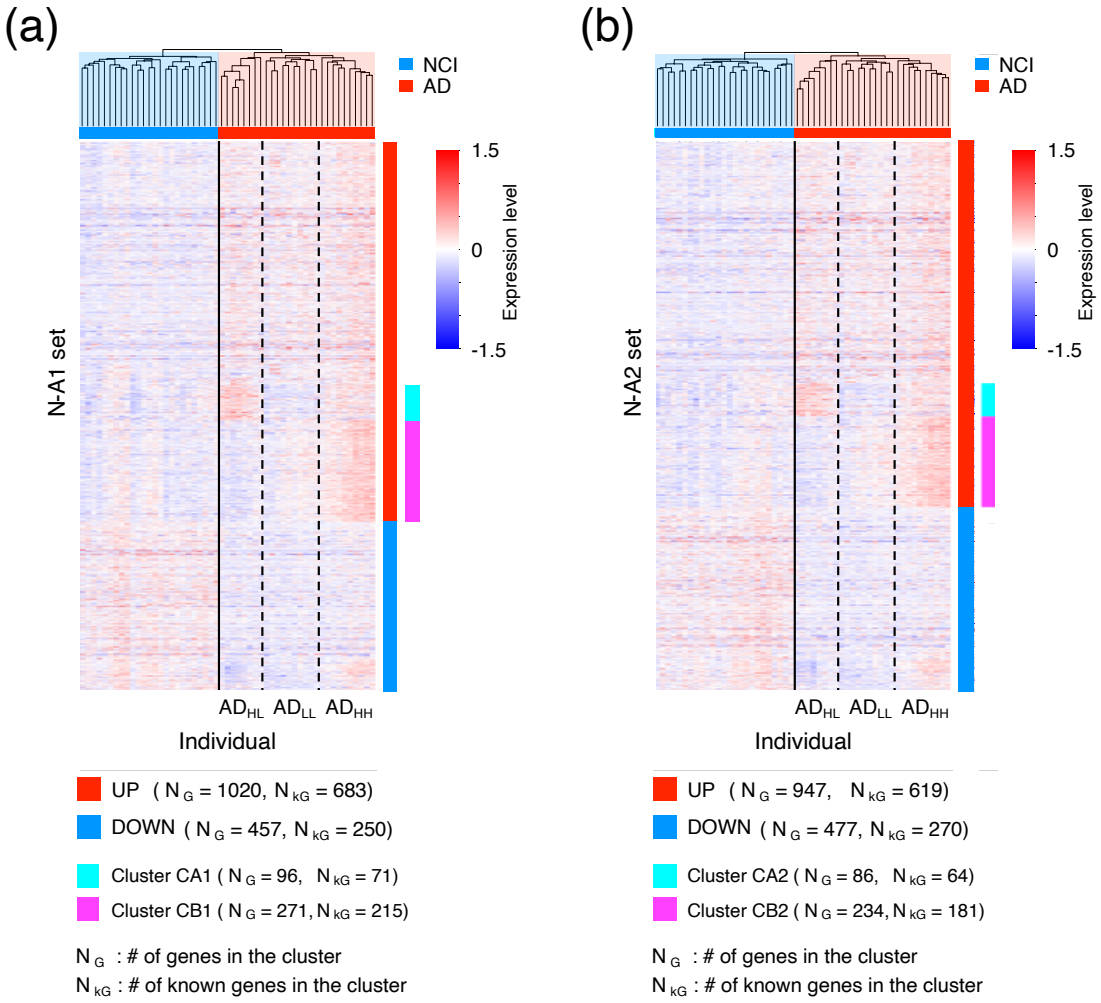


Fig. 3 Results of the hierarchical clustering of individuals of the N-A1 and N-A2 sets and expression level of each gene. The dendrograms and expression level of each gene in the N-A1 set of each individual (a) (see Table S3d for more details) and that in the N-A2 set of each individual was plotted (b) (See Table S3e 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. Broken lines indicate the boundary between the AD_{HL}, AD_{LL}, and AD_{HH}.

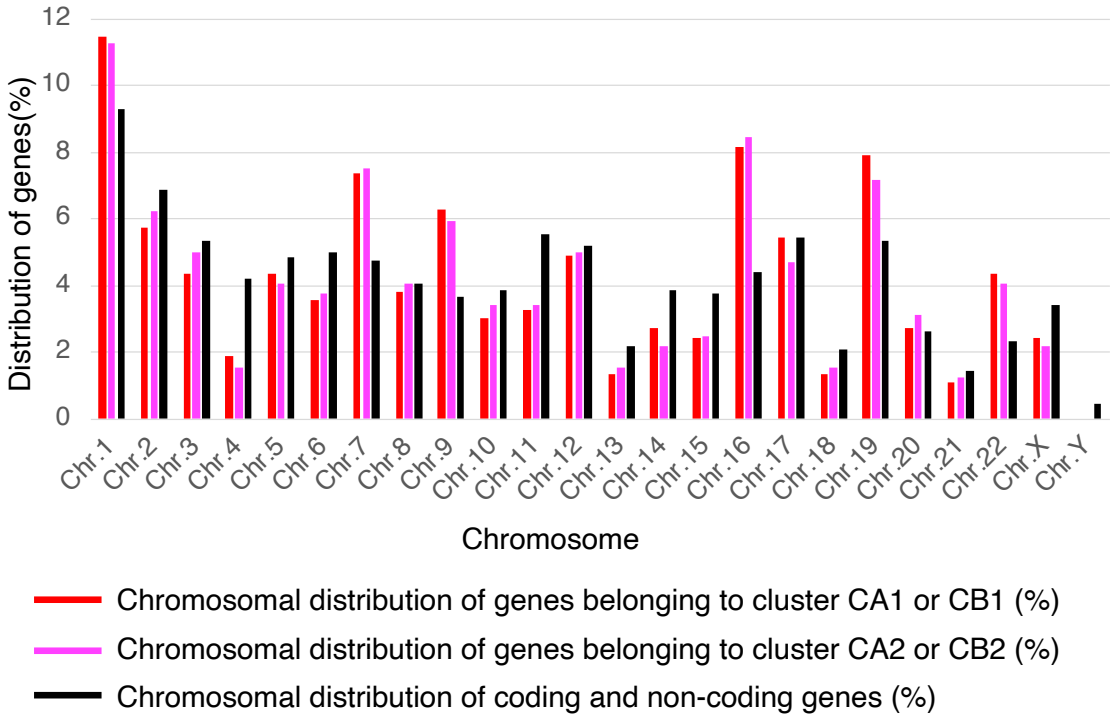


Fig. 4 Chromosomal distribution of genes belonging to the cluster CA1 or CB1 (Fig. 3a), and that to the cluster CA2 or CB2 (Fig. 3b). Chromosomal distribution of coding and non-coding genes was estimated by human genome data GRCh38.

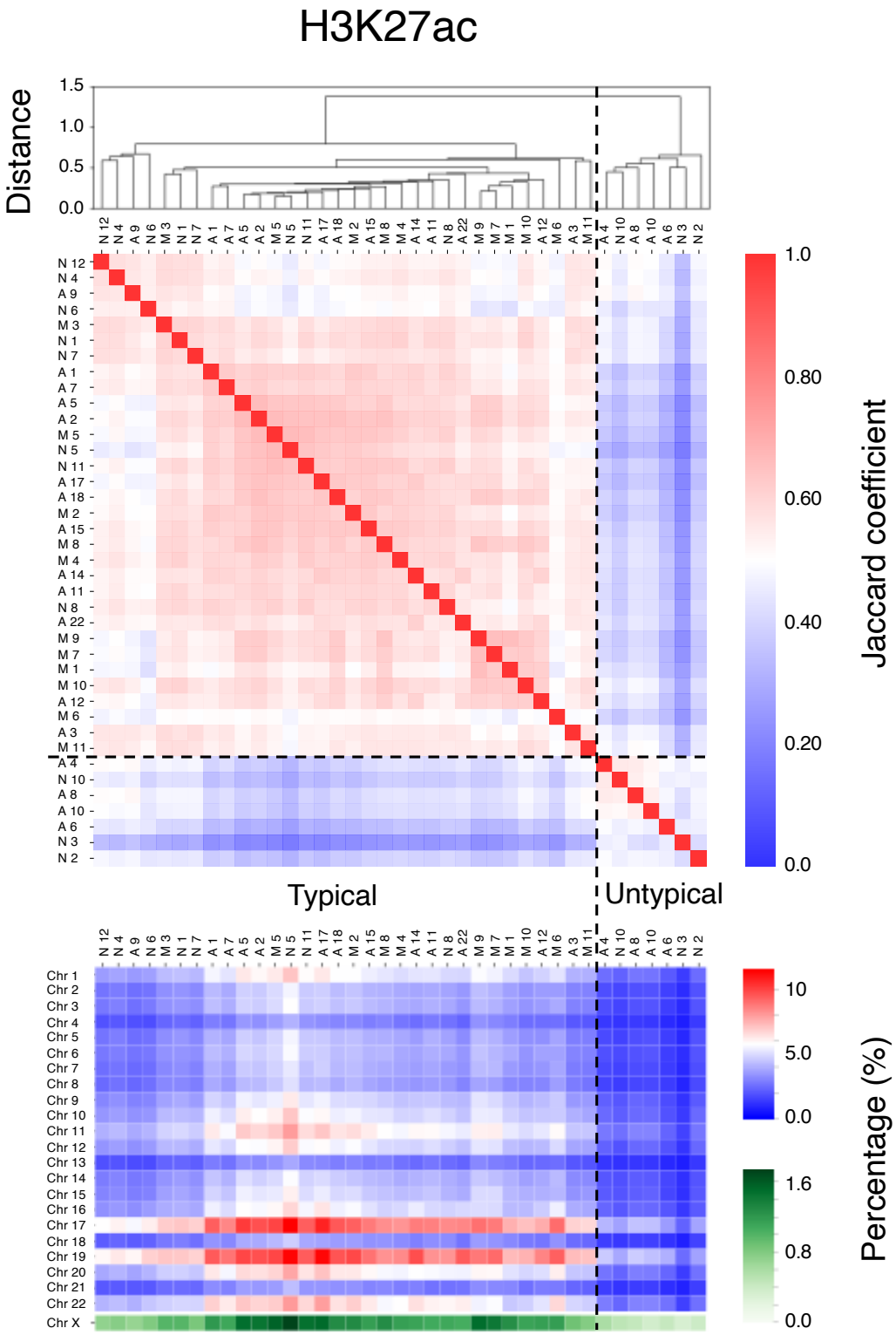


Fig. 5 Jaccard coefficient of the genome-wide distributions of H3K27ac among individuals. Individuals were divided into two groups at the first divergence of the

dendrogram. The epigenetic state exhibited by individuals belonging to the group containing a larger number of individuals was regarded as the "typical" epigenetic state, whereas that containing fewer individuals was regarded as the "untypical" epigenetic state. Dashed lines indicate the boundaries between groups.

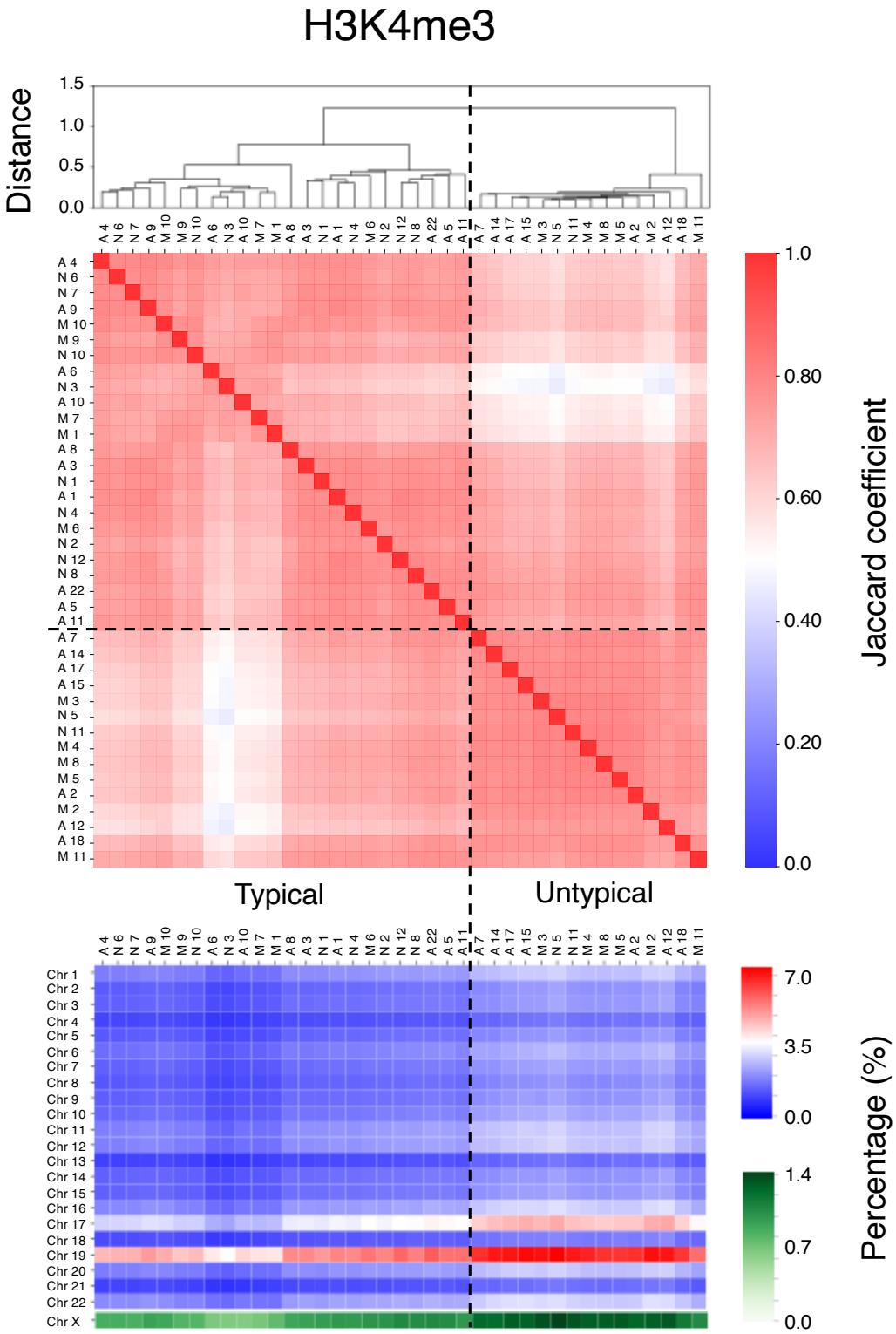


Fig. 6 Jaccard coefficient of the genome-wide distributions of H3K4me3 among individuals. Each object in this figure have the same meaning as that in Fig. 5.

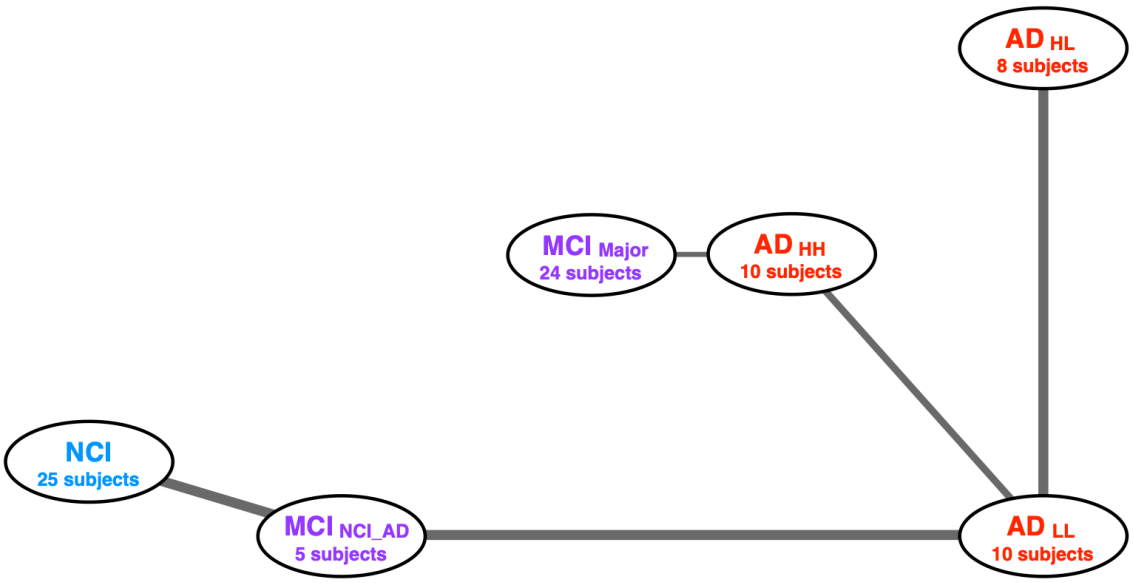


Fig. 7 Adjacency network among substages. The network constructed using common connections obtained from both the network constructed using the M-NA and N-A1 sets (Fig S5a) and that constructed using the M-NA and N-A2 sets (Fig. S5b). The width of each edge indicates the confidence in the presence of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

(a)

# of genes	P^D_{NM} \ P^D_{MN}	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

# of genes	P^t_{NM} \ P^t_{MN}	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

(b)

# of genes	P^D_{MA} \ P^D_{AM}	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

# of genes	P^t_{MA} \ 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

(c)

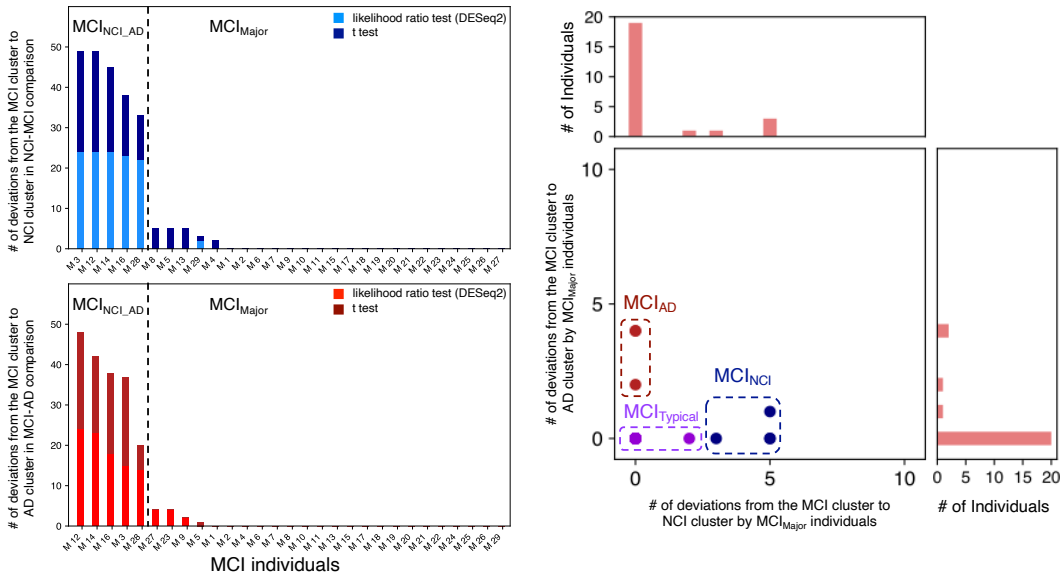


Fig. 8 Detailed classification of the MCI individuals. Number of NCI individuals who belong to the MCI cluster (left value) and that of MCI individuals who belong to the NCI cluster (right value) as a function of P^D_{NM} and P^D_{MN} (a, left), and as a function of P^t_{NM} and P^t_{MN} (a, right). Number of MCI individuals who belong to the AD cluster (left value) and that of AD individuals 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 MCI individual from the MCI to the NCI (c, upper left) and to AD (c, lower left) cluster. Broken lines indicate boundaries between the MCI_{NCI_AD} and MCI_{Major} individuals. Scatter plot of the 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), the MCI_{Major} individuals were divided into $MCI_{Typical}$, MCI_{NCI}, and MCI_{AD}.

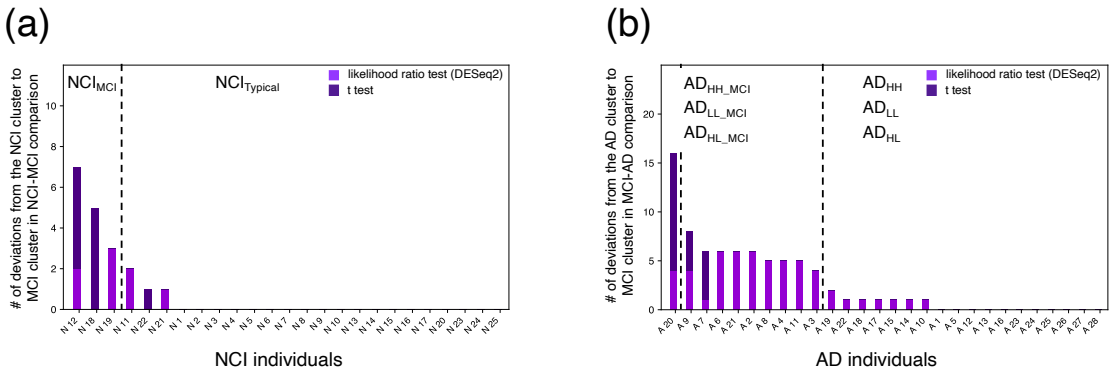


Fig. 9 Detailed classification of the NCI individuals. Number of deviations of each NCI individual from the NCI to the MCI cluster (a), and that of each AD individual from the AD to the MCI cluster (b). Broken lines indicate boundaries among clusters of individuals obtained via k-means clustering (Fig.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

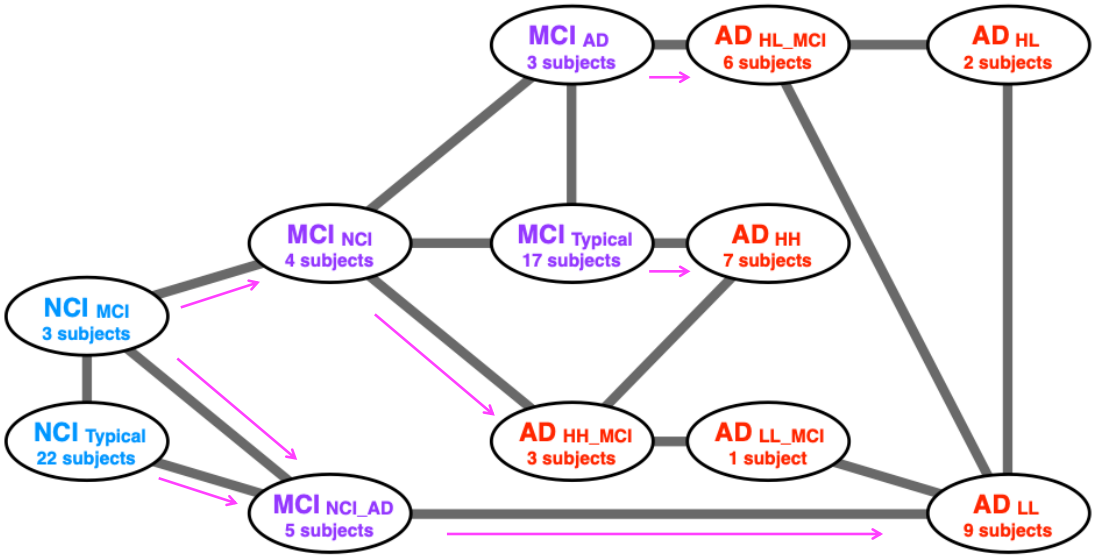


Fig. 10 Adjacency networks among substages. A network containing common connections from both networks obtained using the data in the M-NA and N-A1 sets (Fig. S8a) and that in the M-NA and N-A2 sets (Fig. S8b). The width of each edge indicates the confidence in the presence of a connection between each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