Multiple Alzheimer's disease progression pathways inferred from transcriptome 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, dorsolateral prefrontal cortex, transcriptome, disease progression pathways

ABSTRACT

Late-onset Alzheimer's disease (AD) is a typical type of dementia for which therapeutic strategies have not yet been established. The heterogeneity in symptom by the variability of pathogenic factors may account for the difficulties in preventing and treating this disease. In this study, based on an analysis of publicly available data, the pathology of AD was suggested to be non-uniform, but there were several different typical forms of cognitive deterioration. First, a cluster analysis of subjects diagnosed with "No Cognitive Impairment (NCI)," "Mild Cognitive Impairment (MCI)," and "Alzheimer's Disease (AD)" stages from clinical trials was performed using gene expression data. NCI was found to contain at least two substages, and 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 subjects, suggested 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, accounting 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 type (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, and early detection and the improvement of diagnostic accuracy remain important issues to be solved.

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). Recently, a clinical drug that modestly delays the progression of AD symptoms by removing extracellularly accumulated Aβ was approved by the FDA for public use (Van Dyck, et al., 2023). However, this drug does not lead to a complete cure. Additionally, 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 that have been suggested to influence the onset and progression of AD have also been reported. Viral infections such as herpes virus and coronaviruses (including COVID-19) (Itzhaki et al., 1997; Readhead et al., 2018; Osaka et al., 2022; Wang et al., 2022), diabetes (Ott et al., 1999; Kandimalla et al., 2017), the activity of retrotransposons such as Alu and LINE1 (Cheng et al., 2021; Ravel et al., 2021; Macciardi et al., 2022), inflammation in the brain (McGeer et al., 1995), lipid metabolism in neurons (Strittmatter et al., 1993), the abnormal aggregation of tau protein (Arriagada et al., 1992), oxidative damage in cells and mitochondrial dysfunction (Nunomura et al., 2001; Hirai et al., 2001; Wang et al., 2014), endoplasmic reticulum and ciliary dysplasia (Chakravarthy et al., 2012; Plácido et al., 2014; Morelli et al., 2017; Vorobyeva et al., 2018), and myelin dysplasia (Bartzokis, 2004; Bartzokis, 2011) have been reported to influence the onset and progression of AD. Owing to the involvement of these factors, the causes and symptoms of AD are thought to vary widely from patient to patient, and variations in symptoms and disease progression rates have been reported (Komarova et al., 2011; Younas et al., 2020; Duara et al., 2021; Younas et al., 2024). However, little is known about the origins of this diversity.

In recent years, transcriptome and genome-wide analysis data of nucleotide polymorphisms have been collected for patients with AD, and various differentially expressed genes and AD patient-specific SNPs have been identified in healthy subjects and patients with AD (Jansen et al., 2019; Yuen et al., 2020; Wightman et al., 2021). It should be noted that these characteristics are not always seen equally in all studies, and that there were some differences in how each characteristic appeared depending on the sample from which the data were obtained. One reason for this is the variations in technical factors among studies. However, the diversity of molecular conditions in the brain among patients with AD could also be a major contributor (Tijms et al., 2020; Neff et al., 2021; Tijms et al., 2021; Bellenguez et al., 2022; Johnson et al., 2022). Additionally, such diversity in molecular conditions may be related to the diversity in symptoms and clinical manifestations (Komarova et al., 2011; Younas et al., 2020; Duara et al., 2021; Younas et al., 2024).

Thus, because AD might involve various disease states during progression, the administration of appropriate treatments according to each disease state is key for healing and preventing further disease progression. To enable appropriate state-dependent treatment, it is necessary to classify and characterize these disease states and progression pathways. To evaluate this diversity, a comparison of the features of the entire transcriptome data among various patients should be useful. Therefore, this study analyzed publicly available transcriptome data from groups of patients with different cognitive impairment stages to elucidate the diversity in the onset and progression of late-onset AD and its classifications This research provides the basis for developing appropriate treatments tailored to each Alzheimer's disease patient's disease characteristics.

RESULTS

Fundamental features of the analyzed data

Gene expression data (TPM and read count) of the dorsolateral prefrontal cortex obtained from the "No Cognitive Impairment (NCI)," "Mild Cognitive Impairment (MCI)," and "Alzheimer's Disease (AD)" subjects were acquired from ENCODE consortium's Rush Alzheimer's Disease study (https://www.encodeproject.org/) (ENCODE Project Consortium, 2012; Luo et al., 2020). In this dataset, the number of NCI, MCI, and AD subjects who were female were 25, 29, and 28, respectively, while there were 14, 4, and 7 males, respectively. It should be noted here that recent clinical research suggested gender differences in the onset and symptoms of AD (Lopez-Lee et al., 2024). Therefore, to eliminate the influence of sex differences and variations in the number of male subjects, only data from female subjects were analyzed (Table 1, Table S1).

When examining the significance of the difference in gene expression levels between subjects under 85 years old and subjects over 90 years old, no genes exhibiting a false discovery rate (FDR) < 0.1 (two-tailed t-test and DESeq2) were found (Table S1). This suggests that the variations in gene expression levels among the NCI, MCI, and AD subjects was not due to the influence of age, but rather due to the differences in clinical stages.

Cluster analysis of the subjects

To extract a set of genes that might characterize the differences among the NCI, MCI, and AD groups, analysis of variance (ANOVA) was performed for each gene among the three groups. Consequently, hierarchical clustering of subjects using genes with ANOVA P-values below the threshold P^{A}_{NMA} = 0.01, 0.02, …, and 0.05, was performed (Fig. 1, Fig. S1a). When $P^{A}_{NMA} \leq 0.05$, a cluster with significantly more MCI subjects (MCI cluster) and a cluster with NCI, AD, and a small number of MCI subjects was formed by the first divergence of the dendrogram ((P < 0.05, Fisher's exact test; Fig. S1a). When the P^{A}_{NMA} was 0.04, a cluster comprising solely of subjects with MCI was formed by the first divergence of the dendrogram (Fig. 1a). In contrast, no clear separation was observed between participants with NCI and those with AD, regardless of the P^{A}_{NMA} value.

Furthermore, the same analysis was performed using DESeq2 instead of ANOVA to determine the P-values, where P^{D}_{NMA} was defined instead of P^{A}_{NMA}. As a result, an MCI cluster and a cluster with NCI, AD, and a small number of MCI subjects were formed, as in the case of ANOVA. However, no clusters with only MCI subjects were formed (Fig. S1b).

It should be noted that depending on the P^{A}_{NMA} and P^{D}_{NMA} values, the subjects with MCI who deviated from the MCI cluster varied; however, 24 MCI subjects were always included in the MCI cluster (Fig. S1a, S1b). This suggests that the MCI cluster consisted of 24 MCI subjects, named MCI_{Major} subjects, while the five other MCI subjects (M3, M12, M14, M16, and M28) who tended to deviate from the MCI_{Major} subjects, were named MCI_{NCI_AD} subjects. According to our data, the MCI_{Major} subjects were assumed to be at the substage named MCI_{Major}, and the MCI_{NCI_AD} subjects were assumed to be at the substage named MCI_{NCI_AD}.

Moreover, the set of genes with P-values of the above ANOVA $\leq P^{A}_{NMA} = 0.04$ was named the "M-NA set" since it characterized the differences between typical MCI subjects from NCI and AD subjects (Fig. 1b, Table S2a).

Functional features of genes in the M-NA set

The functional features of genes in the M-NA set were also considered. In this set, 594 genes were obtained (290 upregulated and 304 downregulated genes, MCI_{Major}), of which 349 genes (164 upregulated and 185 downregulated) were annotated (Fig. 1b). Enrichment analysis revealed that the function "mRNA 5'-splice site recognition" was found as a significantly enriched GO term for genes with lower expression levels in MCI_{Major} subjects compared to others (Fig. 1b, Table S2b–S2d).

Gene sets that differentiate groups of NCI and AD subjects

In order to extract a gene set that may characterize the difference between the NCI and AD subjects, hierarchical clustering of subjects was performed using the genes satisfying the following criteria: mean expression level (TPM) was higher in AD subjects than in NCI subjects, with P-values from two-tailed t-tests smaller than the threshold P^{t}_{NA}, or mean expression level (TPM) was lower in AD subjects than in NCI subjects, with P-values in two-tailed t-tests smaller than threshold P^{t}_{AN}. The sets of P^{t}_{NA} and P^{t}_{NA} values that divided the NCI and AD patients into different clusters in the first branch of the dendrogram were then searched. Searching in the ranges $P^{t}_{NA} \leq 0.05$ and $P^{t}_{AN} \leq 0.05$, the NCI subjects and the AD subjects were completely separated (Fig. 2a). When the wider range $P^{t}_{NA} \leq 0.1$ and $P^{t}_{AN} \leq 0.1$ was searched, the largest set of genes separating the NCI and AD groups was obtained at $P^{t}_{NA} = 0.08$, and $P^{t}_{AN} = 0.1$ (Fig. S2a, Table S3a). This gene set was named the "N-A_t set."

Furthermore, the same analysis was performed using DESeq2 instead of two-tailed t-tests to determine the P-values, defining the P^{D}_{NA} and P^{D}_{AN} instead of the P^{t}_{NA} and P^{t}_{AN}, respectively. As a result of searching within the range $P^{D}_{NA} \leq 0.05$ and $P^{D}_{AN} \leq 0.05$, the finite but narrow area of P^{D}_{NA} and P^{D}_{AN} was found where the NCI subjects and AD subjects completely separated (Fig. 2a). Here, the largest set of genes separating the NCI and AD groups was obtained at $P^{D}_{NA} = 0.05$ and $P^{D}_{AN} =$

0.02 (Fig. S2b, Table S3b). 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 subjects than in the NCI subjects 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. S2c, Table S3c). Thus, the gene set obtained at P^D_{NA} = 0.03 and P^D_{AN} = 0.03 was named the "N-A_{D2} 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 subjects (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 subjects. 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 subjects (Fig. 2c, Table S3e). Such gene sets are expected to involve various typical genes that characterize the differences between NCI and AD subjects. In this study, the gene set consisting of the N-A_t and N-A_{D1} sets was named the "N-A1 set" (Table S3d), and that consisting of the N-A_t set and the appropriate part of the N-A_{D2} set was named the "N-A2 set" (Table S3e).

AD subjects can be classified into three groups

From the results of the hierarchical clustering for NCI and AD subjects performed using the N-A1 set, there appeared to be three subgroups of AD subjects (Fig. 2b, 2d). Thus, assuming that AD subjects could be classified into these three subgroups, ANOVA of the gene expression levels among the three assumed subgroups was performed for genes whose expression levels were higher in AD subjects than in NCI subjects (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 subjects with AD, two characteristic clusters were identified: a cluster consisting of 271 genes whose expression levels were high in only one subgroup (Cluster CA1) and another cluster consisting of 96 genes whose expression levels were low in only another subgroup (Cluster CB1) (Fig. 2d).

A similar analysis was performed for the NCI subjects. Based on the shape of the dendrogram, the NCI subjects were divided into three or four subgroups (Fig. 2b, 2d). However, when the NCI subjects 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 subjects were divided into four subgroups, ANOVA revealed only two genes with FDR < 0.1 among the three subgroups (Table S3d). Compared to 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 subjects were highly non-uniform, which can be classified into three types. AD subjects exhibiting high gene expression levels in both clusters CA1 and CB1 were regarded as subjects in the AD_{HH} subgroup (named AD_{HH} subjects), those exhibiting low gene expression levels in both CA1 and CB1 were regarded as subjects in the AD_{LL} subgroup (named AD_{LL} subjects), and those exhibiting low gene expression levels in cluster CA1 and high gene expression levels in cluster CB1 were regarded as subjects in the AD_{HL} subgroup (named AD_{HL} subjects).

Hierarchical clustering of the NCI and AD subjects using the N-A2 set also exhibited three possible subgroups (Fig. 2c, 2e). By performing the same analysis described above, gene clusters similar to cluster CA1 containing 234 genes (cluster CA2) and cluster CB1 containing 86 genes (cluster CB2) were identified. Similar to a previous method, subjects with AD 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).

It is important to note that the AD_{HH}, AD_{HL}, and AD_{LL} subjects 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, there was less exchange of subjects among subgroups. For example, considering the case where $P^t_{NA} = 0.06$ and $P^t_{AN} = 0.1$, which were employed to determine the N-At, N-A3, and N-A4 sets, N-A3 consists of genes in the N-At set and appropriately chosen genes in $N\text{-}A_{D1}$ and N-A4 consists of genes in the N-At set and appropriately chosen genes in the $N\text{-}A_{D2}$ set. The classification of AD subjects 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 subjects belonged to the same subgroups (Fig. S3). This suggests the robustness of our method in classifying the AD_{HH}, AD_{HL}, and AD_{LL} subjects.

Functional features of the N-A1 and N-A2 sets

The functional features of the N-A1 and N-A2 sets were also considered. The N-A1 set contained 1477 genes (1020 upregulated and 457 downregulated, AD), and 933 genes (683 upregulated and 250 downregulated) were annotated (Fig. 2d). The function “forebrain dorsal/ventral pattern formation” was an enriched GO term for genes with higher expression levels in AD subjects than others after enrichment analysis (Fig. 2d, Table S3f).

The N-A2 set contained 1424 genes (947 upregulated and 477 downregulated, AD), and 889 genes (619 upregulated and 270 downregulated) were annotated (Fig. 2e). The functions “negative regulation of intrinsic apoptotic signaling pathway by p53 class mediator” was found as an enriched

GO term for genes with lower expression levels in AD subjects than others after enrichment analysis (Fig. 2e, Table S3g).

Clusters CA1, CB1, CA2, and CB2 contained 215, 71, 181, and 64 annotated genes, respectively (Fig. 2d-2e). The functions “mitotic spindle assembly checkpoint signaling,” “response to xenobiotic stimulus,” and “negative regulation of small molecule metabolic process” were found to be enriched GO terms to distinguish among AD substages, which were obtained using both the N-A1 and N-A2 sets (Fig. 2d–2e, Tables S3h–S3k). Additionally, functions related to the structural and physiological features of the brain, such as “cerebral cortex development,” “neuron migration,” “forebrain cell migration,” and “astrocyte differentiation,” were found to be enriched GO terms to distinguish the three AD substages, which were obtained in cluster CB1 (Fig. 2d, Table S3h–S3i).

Adjacency network among substages

The data on the expression levels of genes in the M-NA set characterized the differences between typical MCI, NCI, and AD subjects and those in the N-A1 or N-A2 sets characterized the differences between NCI and AD subjects. To clarify the relationships among disease substages and the pathways of disease progression through these substages, the adjacency networks among substages were described using these gene sets based on the partition-based graph abstraction (PAGA) method (Wolf et al., 2019). PAGA generates a graph of substages in which the edge weights represent confidence in the presence of connections. It should be noted that the confidence of connections between NCI and substages in AD often involves finite values (Table S4); however, the direct connection between NCI and substages in AD seems unnatural if the disease progresses. Therefore, only connections with confidence values sufficiently larger than the maximum confidence values of the connection between NCI and substages in AD (MD_{N_A}) were assumed to be effective for the adjacency network among substages (Fig. 3). Subsequently, connections with confidence values larger than twice that of MD_{N_A} were used to construct the adjacency network (Table S4).

It is naturally considered that changes in gene expression patterns that characterize disease states occur gradually as the disease progresses. Therefore, the disease appeared to progress from NCI to AD along the previously obtained adjacency network. In both cases presented in Fig. 3, no connections from NCI to MCI_{Major}, which is the major group of MCI subjects, were obtained (Fig. 3), which seemed unnatural for the progression of AD. Thus, more detailed subgroup classifications of NCI, MCI, and AD subjects were performed to reveal the relationship between NCI and MCI_{Major}, as described below.

Detailed classification of MCI by comparing between NCI and MCI and between MCI and AD

To classify MCI subjects more comprehensively, comparisons between NCI and MCI subjects and between MCI and AD subjects were made. First, hierarchical clustering of subjects was performed using the genes satisfying the following criteria: mean expression level (TPM) was higher in the MCI group than in the NCI group with two-tailed t-test P-values smaller than the threshold P^t_{MN} or the mean expression level (TPM) was lower in the MCI group than in the NCI group, with two-tailed t-test P-values smaller than the threshold P^t_{NM}. Such cluster analysis was performed using each gene set within the range $P^t_{NM} \leq 0.05$ and $P^t_{MN} \leq 0.05$. However, no cases were found where the NCI and MCI subjects were completely separated (Fig. 4a, Fig. S4a). However, within a wide range of P^t_{NM} and P^t_{MN} values, clusters with significantly more NCI subjects (NCI clusters) and MCI clusters were formed (Fig. 4a).

When DESeq2 was used instead of two-tailed t-tests to determine the P-values, the results were qualitatively the same. Here, P^D_{MN} and P^D_{NM} were defined instead of P^t_{MN} and P^t_{NM}, respectively. Analysis using each gene set within the ranges $P^D_{MN} \leq 0.05$ and $P^D_{NM} \leq 0.05$ revealed no cases where the NCI and MCI subjects were completely separated (Fig. 4a). However, the NCI and MCI clusters were formed within a wide range of P^D_{NM} and P^D_{MN} values.

Next, the same analyses were performed for the MCI and AD subjects. Here, P^t_{MA}, P^t_{AM}, P^D_{MA}, and P^D_{AM} were given as values corresponding to P^t_{NM}, P^t_{MN}, P^D_{NM}, and P^D_{MN}, respectively. Similar to the above results, no cases were found in which the MCI and AD subjects were completely separated (Fig. 4b, Fig. S4b). However, the MCI and AD clusters were formed within a wide range of P^t_{MA}, P^t_{AM}, P^D_{MA}, and P^D_{AM} values.

In this study, the MCI subjects who deviated from the MCI cluster varied depending on P^t_{NM}, P^t_{MN}, P^D_{NM}, P^D_{MN}, P^t_{MA}, P^t_{AM}, P^D_{MA}, and P^D_{AM} values. For each subject with MCI, the number of times they belonged to the NCI cluster upon changing the P^t_{NM}, P^t_{MN}, P^D_{NM}, and P^D_{MN} values, and those belonging to the AD cluster upon changing the P^t_{MA}, P^t_{AM}, P^D_{MA}, and P^D_{AM} values were counted (Fig. 4c). Consequently, subjects M3, M4, M5, M8, M12, M13, M14, M16, M28, and M29 belonged to the NCI cluster more than once, whereas subjects M3, M5, M9, M12, M14, M16, M23, M27, and M28 belonged to the AD cluster more than once.

It should be noted that subjects M3, M12, M14, M16, and M28, who were the same subjects in MCI_{NCI_AD}, clearly exhibited a larger number of deviations from NCI and AD than those in MCI_{Major} (Fig. 4c). To classify the MCI_{Major} subjects in more detail, k-means clustering was performed using the number of times each subject belonged to the NCI cluster and the number of times they belonged to the AD cluster (Fig. 4c, Fig. S4b). As a result, the MCI_{Major} subjects were divided into three

subtypes: $MCI_{Typical}$ subjects (17 subjects), who rarely or never belonged to other clusters, MCI_{NCI} subjects (four subjects), who may belong to NCI clusters only, and MCI_{AD} subjects (three subjects), who may belong to AD clusters only (Table S1). Thus, the MCI stage was divided into four substages: $MCI_{Typical}$, MCI_{NCI}, MCI_{AD}, and MCI_{NCI_AD}.

Detailed classification of NCI by comparing between NCI and MCI subjects

The analysis performed in the previous section also showed that subjects N11, N12, N18, N19, N21, and N22 belonged to the MCI cluster more than once when the P^t_{NM}, P^t_{MN}, P^D_{NM}, and P^D_{MN} values were changed (Fig. 5a). After k-means clustering of the NCI subjects using the number of times each subject belonged to the MCI cluster, the NCI subjects were divided into two clusters: one comprising N11, N21, and N22, and the other comprising NCI subjects who never belonged to the MCI cluster (Fig. 5a, Fig. S5a). Therefore, these subjects were classified as typical NCI subjects who belonged to the MCI cluster and named $NCI_{Typical}$ subjects, who were assumed to be in the $NCI_{Typical}$ substage. Subjects N12, N18, and N19 were classified as NCI subjects close to MCI and named NCI_{MCI} subjects, who were assumed to be in the NCI_{MCI} substage (Table S1).

Similarly, 17 subjects with AD belonged to the MCI cluster more than once, with changes in the P^t_{MA}, P^t_{AM}, P^D_{MA}, and P^D_{AM} values (Fig. 5b). Using k-means clustering of AD subjects based on the number of times each subject belonged to the MCI cluster in the previous section, AD subjects were divided into three clusters, where subjects A10, A14, A15, A17, A18, A19, A22, and other AD subjects who never belonged to the MCI cluster were included in the same cluster (Fig. 5b, Fig S5b). Therefore, these participants were expected to be in one of the typical AD substages: AD_{HH}, AD_{HL}, or AD_{LL} (Table S1). In contrast, subjects A2, A3, A4, A6, A7, A8, A9, A11, A20, and A21 were assumed to be at one of the AD substages close to MCI and named AD_{HH_MCI}, AD_{HL_MCI}, or AD_{LL_MCI} (Table S1).

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

Detailed classifications of subjects suggested that NCI was classified into $NCI_{Typical}$ and NCI_{MCI}; MCI was classified into $MCI_{Typical}$, MCI_{NCI}, MCI_{AD}, and MCI_{NCI_AD}; and AD was classified into AD_{HH}, AD_{HH_MCI}, AD_{HL}, AD_{HL_MCI}, AD_{LL}, and AD_{LL_MCI}. Adjacency networks among these 12 substages were constructed as previously performed using gene expression data in the M-NA and N-A1 sets (Fig. 6a, Table S5a) and in the M-NA and N-A2 sets (Fig. 6b, Table S5b). The adjacency network obtained from the data in the M-NA and N-A1 sets exhibited four progression paths from NCI to AD that passed through different substages, as follows (Fig. 6a): 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}. Almost the same network was obtained when data from the M-NA and N-A2 sets were used, except for the presence of a connection between AD_{HH} and AD_{LL} (Fig. 6b). These results suggest that there are at least four disease progression pathways toward at least three classified AD substages.

DISCUSSION

In this study, the pathological progression pathway of AD was suggested to be non-uniform, with different typical forms of deterioration. First, cluster analysis of subjects using gene expression data from those diagnosed with NCI, MCI, and AD through clinical trials suggested that NCI contained at least two substages, and MCI and AD contained four and six substages, respectively. Second, an inference of adjacency networks among the different substages suggests the existence of four typical NCI → AD disease progressions from different MCI substages to different AD substages. These typical pathological progression pathways are unknown, so the findings of such disease progression features will contribute to the development of AD research, which 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 whether they are connected 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 subjects 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, rather than the hippocampus, which has been the focus of many previous studies. The reason why this dataset was analyzed in this study was that, for some subjects, this dataset included not only transcriptome data but also data on epigenomic status such as CCCTC-binding factor (CTCF) binding sites and histone modification (H3K4me3, H3K27me3, H3K27ac) distributions (Table S1). Hence, this dataset can potentially consider the relationship between a wide range of transcriptional regulatory states and disease conditions. In fact, the matrix of Jaccard coefficients of the genome-wide distributions of CTCF, H3K27ac, and H3K27me3 among subjects included in this study suggested that some NCI, $AD_{HL,}$ and AD_{HL_MCI} subjects exhibited atypical genome-wide distribution patterns of protein

binding and histone modifications that deviated from the typical pattern observed in other subjects (Fig. 7a, Tables S6a–S6c, Fig. S6a–S6c, Fig. S7a–S7b). Moreover, more AD_{HL} and AD_{HL_MCI} subjects exhibited atypical patterns compared to other AD subjects. Additionally, some NCI, AD_{HH} and AD_{HH_MCI} subjects appeared to exhibit atypical genome-wide distribution patterns of H3K4me3 that deviated from the typical pattern observed in other subjects (Fig. 7b, Table S6d, Fig. S6d); a larger number of AD_{HH} and AD_{HH_MCI} subjects exhibited atypical patterns compared to other AD subjects. These observations suggest that the specificity of histone modifications in each subject may influence the pathological progression of the disease. Furthermore, a relationship between changes in CTCF binding and transcription activities associated with changes in pathological conditions was recently found by analyzing this dataset (Patel et al., 2023). In the future, the characteristics of different progression pathways and the mechanisms by which each pathway is selected should be comprehensively considered, including the epigenetic state.

Many recent studies performed transcriptome analyses to identify differentially expressed genes (DEGs) between healthy (control) subjects and patients with AD. Such studies did not yield many DEGs in areas belonging to the prefrontal cortex investigated in this study, and no pathological changes were found (Nef et al., 2021). Even in the data analyzed in this study, few DEGs were obtained using the standard method used in conventional studies searching for DEGs, so it was not possible to accurately distinguish between healthy and AD subjects via cluster analysis using only DEGs. However, by focusing on the expression patterns of more widespread candidate genes rather than the DEGs, we found that different characteristics appeared in the dorsolateral prefrontal cortex between healthy and AD subjects and that there are disease substages that occur in response to changes in the disease state.

Previous research focusing on DEGs between patients with AD and healthy individuals suggested that multiple states exist in the hippocampus of patients with AD based on differences in the gene expression profiles (Neff et al., 2021). Based on this knowledge, 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.

EXPERIMENTAL PROCEDURES

Data processing

In this study, gene expression data (TPM and read count) from the dorsolateral prefrontal cortex obtained from the “No Cognitive Impairment (NCI),” “Mild Cognitive Impairment (MCI),” and

"Alzheimer's Disease (AD)" subjects were acquired from the ENCODE consortium's Rush Alzheimer's Disease study (https://www.encodeproject.org/) (ENCODE Project Consortium, 2012; Luo et al., 2020). In the present analysis, the data of genes with average expression levels (TPM) in NCI, MCI, and AD subjects were smaller than 0.1, and those of spike-in-RNAs were eliminated.

As an example of the size factors of the RNA sequencing (RNA-seq) data, the median of the ratios (MOR) defined in the recent literature for DESeq (Anders et al., 2010) was estimated using read count data (Table S1). The ratio [standard deviation of MOR over subjects]/[average MOR over subjects] was 0.10753485. Based on these results, we expected that the size factor would not considerably affect the comparison of the TPM values.

Statistical analyses

Two-tailed t-tests and ANOVA were performed using Python (Ver.3.9.16). 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). R (Ver.4.1.3) was used to determine the DEGs using DESeq2 and Fisher's exact test.

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 subjects = 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 subject group (NCI, MCI, AD), then that cluster was regarded as a cluster in which there were significantly more subjects belonging to that group. These criteria were: i) more than 2/3 of the subjects belonging to that group were contained, and ii) Fisher exact test resulted 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.

Enrichiment analysis

Using metascape (v3.5.20240101) (https://metascape.org/gp/index.html#/main/step1) (Zhou et al., 2019), enrichment analysis was performed for the N-MA, N-A1, and N-A2 sets. 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 subject. Each subject 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). Here, 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, the minimum value in the range of this parameter. Thus, a network with nonredundant connections among the substages was expected to be obtained.

ACKNOWLEDGEMENTS

This work was supported by a JSPS KAKENHI grant (award number 21K06124).

REFERENCES

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Van Dyck, C. H., Swanson, C. J., Aisen, P., Bateman, R. J., Chen, C., Gee, M.,... Iwatsubo, T. (2023). Lecanemab in early Alzheimer's disease. *New England Journal of Medicine*, *388*, 9–21. https://doi.org/10.1056/nejmoa2212948

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

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.

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

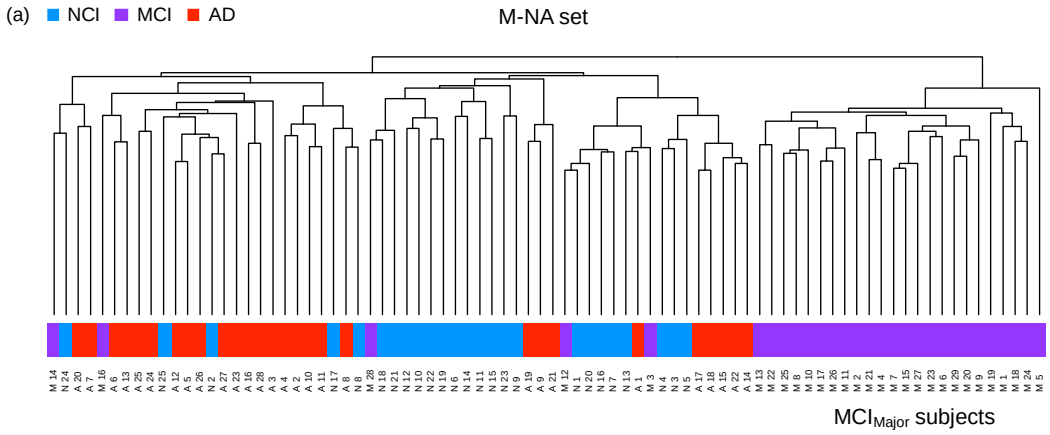
Yuen, S. C., Zhu, H., & Leung, S. W. (2020). A systematic bioinformatics workflow with meta-analytics identified potential pathogenic factors of Alzheimer's disease. *Frontiers in Neuroscience*, *14*, 209. https://doi.org/10.3389/fnins.2020.00209

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

Table 1. Features of the subjects included in this study. The name, ID in the ENCODE database, age, and clinical diagnostic stage of each subject were analyzed (see Table S1 for more details.). Age 90+ indicates that the subject was 90 years of age or older.

Subject	ID (ENCODE)	Age	Clinical Diagnostic Stage	Subject	ID (ENCODE)	Age	Clinical Diagnostic Stage
N1	ENCDO250PFZ	90+	No Cognitive Impairment	M17	ENCDO170MSG	90+	Mild Cognitive Impairment
N2	ENCDO218FFZ	90+	No Cognitive Impairment	M18	ENCDO183VJV	90+	Mild Cognitive Impairment
N3	ENCDO006DAA	90+	No Cognitive Impairment	M19	ENCDO203GSG	78	Mild Cognitive Impairment
N4	ENCDO609ZOG	79	No Cognitive Impairment	M20	ENCDO343SFL	90+	Mild Cognitive Impairment
N5	ENCDO669IVL	88	No Cognitive Impairment	M21	ENCDO539ZHT	90+	Mild Cognitive Impairment
N6	ENCDO290OPS	82	No Cognitive Impairment	M22	ENCDO853KFT	90+	Mild Cognitive Impairment
N7	ENCDO101GPB	90+	No Cognitive Impairment	M23	ENCDO880FAU	90+	Mild Cognitive Impairment
N8	ENCDO236YSH	78	No Cognitive Impairment	M24	ENCDO893GME	90+	Mild Cognitive Impairment
N9	ENCDO423GGP	87	No Cognitive Impairment	M25	ENCDO896WHW	89	Mild Cognitive Impairment
N10	ENCDO448ZXP	83	No Cognitive Impairment	M26	ENCDO063RHR	87	Mild Cognitive Impairment
N11	ENCDO461DJY	84	No Cognitive Impairment	M27	ENCDO075KHF	90+	Mild Cognitive Impairment
N12	ENCDO640RUC	87	No Cognitive Impairment	M28	ENCDO418RWN	86	Mild Cognitive Impairment
N13	ENCDO081AHA	90+	No Cognitive Impairment	M29	ENCDO943KXE	86	Mild Cognitive Impairment
N14	ENCDO377QXX	82	No Cognitive Impairment	A1	ENCDO997SGX	86	Alzheimer's Disease
N15	ENCDO165PNJ	82	No Cognitive Impairment	A2	ENCDO201EUI	90+	Alzheimer's Disease
N16	ENCDO242JQO	90+	No Cognitive Impairment	A3	ENCDO853VGZ	89	Alzheimer's Disease
N17	ENCDO628ASW	77	No Cognitive Impairment	A4	ENCDO865EMM	90+	Alzheimer's Disease
N18	ENCDO741DAO	85	No Cognitive Impairment	A5	ENCDO907CMO	74	Alzheimer's Disease
N19	ENCDO948PMW	90+	No Cognitive Impairment	A6	ENCDO070VNS	88	Alzheimer's Disease
N20	ENCDO982NCL	85	No Cognitive Impairment	A7	ENCDO097MEH	81	Alzheimer's Disease
N21	ENCDO989QRD	75	No Cognitive Impairment	A8	ENCDO637GUS	90+	Alzheimer's Disease
N22	ENCDO394FZZ	83	No Cognitive Impairment	A9	ENCDO830KFO	89	Alzheimer's Disease
N23	ENCDO921LXT	90+	No Cognitive Impairment	A10	ENCDO080EZF	85	Alzheimer's Disease
N24	ENCDO980GMJ	90+	No Cognitive Impairment	A11	ENCDO847KYQ	90+	Alzheimer's Disease
N25	ENCDO888KBY	86	No Cognitive Impairment	A12	ENCDO730UYW	90+	Alzheimer's Disease
M1	ENCDO471EKG	90+	Mild Cognitive Impairment	A13	ENCDO150EBZ	90+	Alzheimer's Disease
M2	ENCDO672KST	90+	Mild Cognitive Impairment	A14	ENCDO258GJF	90+	Alzheimer's Disease
M3	ENCDO832DBZ	83	Mild Cognitive Impairment	A15	ENCDO597SRH	90+	Alzheimer's Disease
M4	ENCDO915WZE	90+	Mild Cognitive Impairment	A16	ENCDO093CLS	90+	Alzheimer's Disease
M5	ENCDO033BMB	90+	Mild Cognitive Impairment	A17	ENCDO223NTQ	90+	Alzheimer's Disease
M6	ENCDO151OJB	90+	Mild Cognitive Impairment	A18	ENCDO470GFA	90+	Alzheimer's Disease
M7	ENCDO153NUY	90+	Mild Cognitive Impairment	A19	ENCDO473TSU	90+	Alzheimer's Disease

M8	ENCDO354SJE	87	Mild Cognitive Impairment	A20	ENCDO658JXQ	90+	Alzheimer's Disease
M9	ENCDO634UMA	90+	Mild Cognitive Impairment	A21	ENCDO680UUE	90+	Alzheimer's Disease
M10	ENCDO660TGP	90+	Mild Cognitive Impairment	A22	ENCDO843TIL	89	Alzheimer's Disease
M11	ENCDO739EFE	90+	Mild Cognitive Impairment	A23	ENCDO020BIL	90+	Alzheimer's Disease
M12	ENCDO394ZHU	90+	Mild Cognitive Impairment	A24	ENCDO020MJV	90+	Alzheimer's Disease
M13	ENCDO204VNL	90+	Mild Cognitive Impairment	A25	ENCDO427VQK	90+	Alzheimer's Disease
M14	ENCDO305GYI	90+	Mild Cognitive Impairment	A26	ENCDO430XKI	88	Alzheimer's Disease
M15	ENCDO557QXE	90+	Mild Cognitive Impairment	A27	ENCDO651VQQ	90+	Alzheimer's Disease
M16	ENCDO746DTM	90+	Mild Cognitive Impairment	A28	ENCDO801BII	88	Alzheimer's Disease



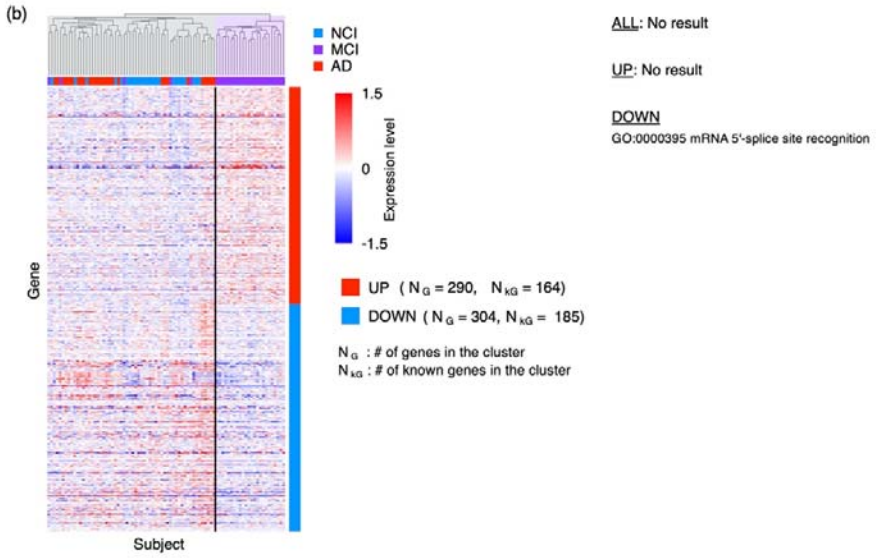
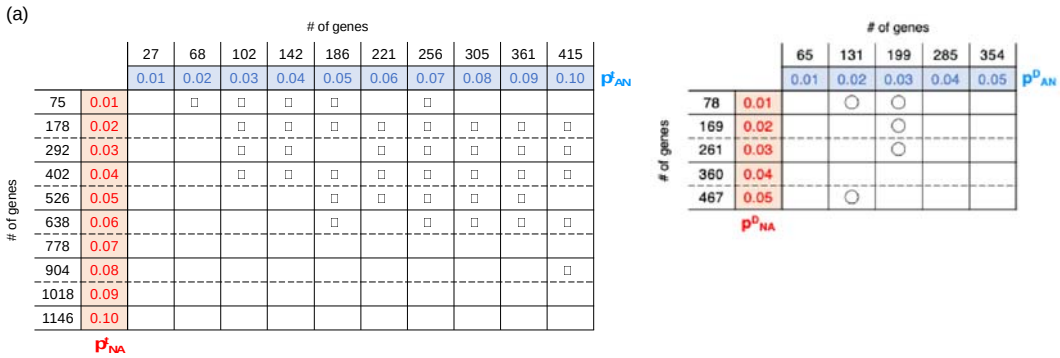
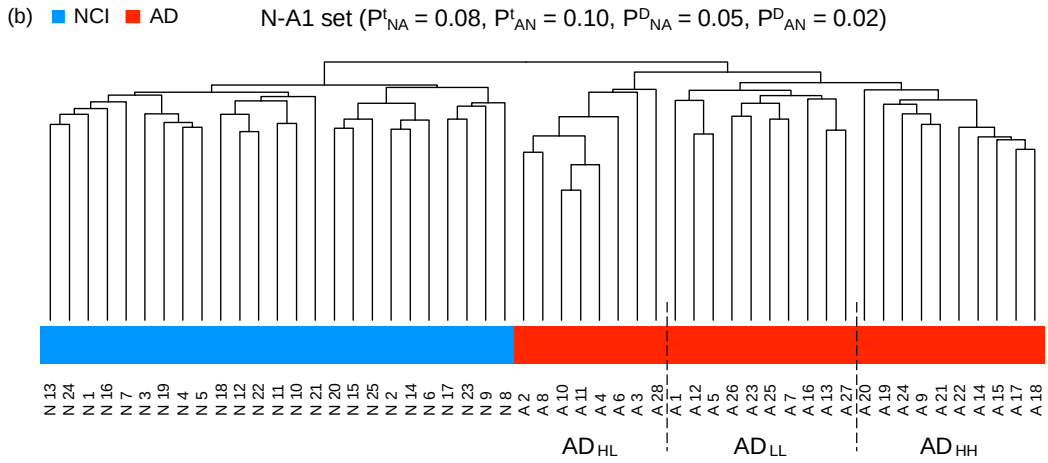
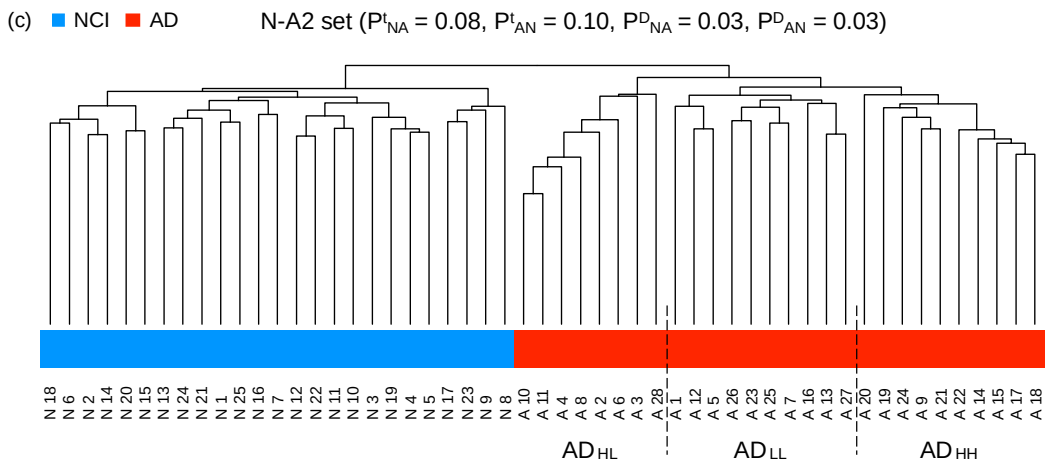


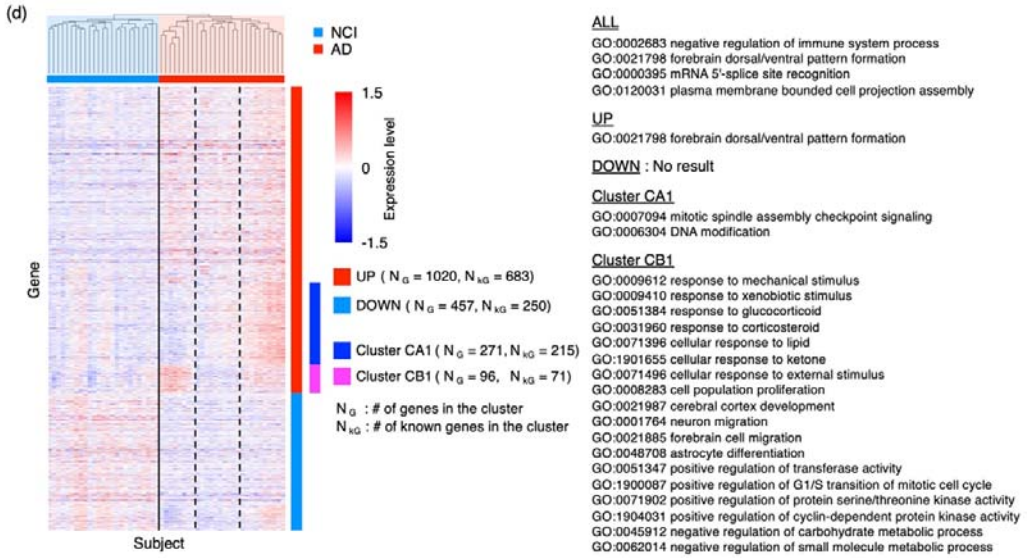
Figure 1: Results of the hierarchical clustering of subjects and enrichment analysis of the M-NA set. Result of the hierarchical clustering of subjects in the case of the M-NA set (P^{A}_{NMA} = 0.04) (a). The expression level of each gene in the M-NA set of each subject was plotted (b, left) (see Table 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 subjects = 0 for each gene. The MCI_{Major} subjects occupy the right region in the dendrogram. GO terms with false discovery rates (FDR) < 0.2 were obtained via enrichment analysis of all genes (ALL) and genes with higher expression levels (UP) and those with lower expression levels (DOWN) in the MCI_{Major} subjects (b, right) (See Table S2b for the detailed results of the enrichment analysis).






(c)
NCI
AD
N-A2 set (P^t_{NA} = 0.08, P^t_{AN} = 0.10, P^D_{NA} = 0.03, P^D_{AN} = 0.03)
N 18 N 6 N 2 N 14 N 20 N 15 N 13 N 24 N 21 N 1 N 25 N 16 N 7 N 12 N 22 N 11 N 10 N 3 N 19 N 4 N 5 N 17 N 23 N 9 N 8 A 10 A 11 A 4 A 8 A 2 A 6 A 3 A 28 A 1 A 12 A 5 A 26 A 23 A 25 A 7 A 16 A 13 A 27 A 20 A 19 A 24 A 9 A 21 A 22 A 14 A 15 A 17 A 18
AD_{HL}
AD_{LL}
AD_{HH}

716


(d)
NCI
AD
1.5
0
-1.5
Expression level
Gene
Subject
UP (N_G = 1020, N_{kG} = 683)
DOWN (N_G = 457, N_{kG} = 250)
Cluster CA1 (N_G = 271, N_{kG} = 215)
Cluster CB1 (N_G = 96, N_{kG} = 71)
N_G : # of genes in the cluster
N_{kG} : # of known genes in the cluster
ALL
GO:0002683 negative regulation of immune system process
GO:0021798 forebrain dorsal/ventral pattern formation
GO:0000395 mRNA 5'-splice site recognition
GO:0120031 plasma membrane bounded cell projection assembly
UP
GO:0021798 forebrain dorsal/ventral pattern formation
DOWN : No result
Cluster CA1
GO:0007094 mitotic spindle assembly checkpoint signaling
GO:0006304 DNA modification
Cluster CB1
GO:0009612 response to mechanical stimulus
GO:0009410 response to xenobiotic stimulus
GO:0051384 response to glucocorticoid
GO:0031960 response to corticosteroid
GO:0071396 cellular response to lipid
GO:1901655 cellular response to ketone
GO:0071496 cellular response to external stimulus
GO:0008283 cell population proliferation
GO:0021987 cerebral cortex development
GO:0001764 neuron migration
GO:0021885 forebrain cell migration
GO:0048708 astrocyte differentiation
GO:0051347 positive regulation of transferase activity
GO:1900087 positive regulation of G1/S transition of mitotic cell cycle
GO:0071902 positive regulation of protein serine/threonine kinase activity
GO:1904031 positive regulation of cyclin-dependent protein kinase activity
GO:0045912 negative regulation of carbohydrate metabolic process
GO:0062014 negative regulation of small molecule metabolic process

717

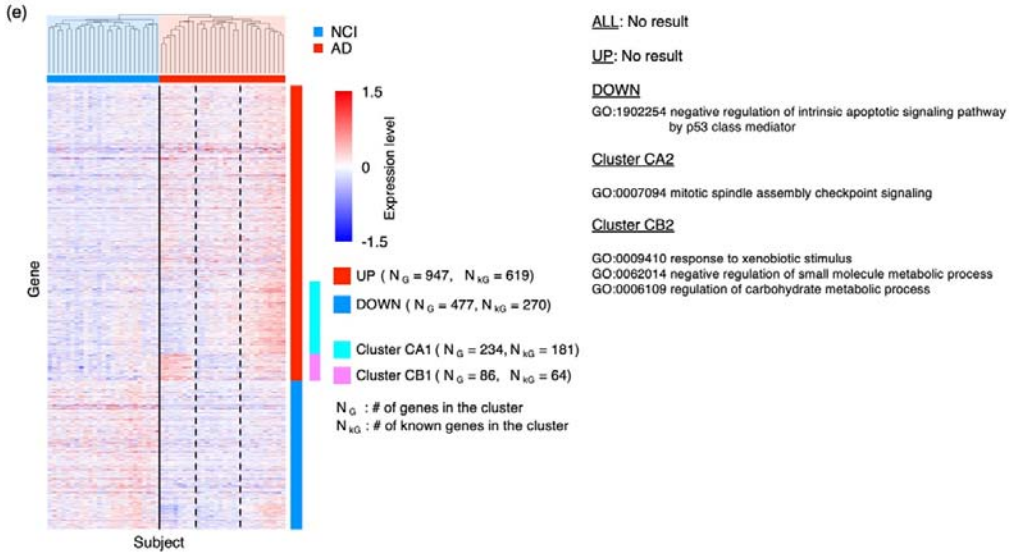


Figure 2: Phase diagrams of the hierarchical clustering of NCI and AD subjects, and the results of the hierarchical clustering of subjects and enrichment analysis of the N-A1 and N-A2 sets. Phase diagram of the results of hierarchical clustering using genes where $p \leq P^{t}_{NA}$ and $p \leq P^{t}_{AN}$ (a, left), and that using genes where $p \leq P^{D}_{NA}$ and $p \leq P^{D}_{AN}$ (a, right). Circles indicate where NCI and AD subjects were completely divided into the different clusters at the first branch of the dendrogram. Shown are the dendrograms obtained using the N-A1 (b) and N-A2 sets (c). The expression level of each gene in the N-A1 set of each subject (d, left) (see Table S3d for more details) and that in the N-A2 set of each subject was plotted (e, left) (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 subjects = 0 for each gene. Broken lines indicate the boundary between the AD_{HL}, AD_{LL}, and AD_{HH}. GO terms with false discovery rates (FDR) < 0.2 were obtained via enrichment analysis (d and e, right). See Table S3f–S3o for the detailed results of the enrichment analysis.

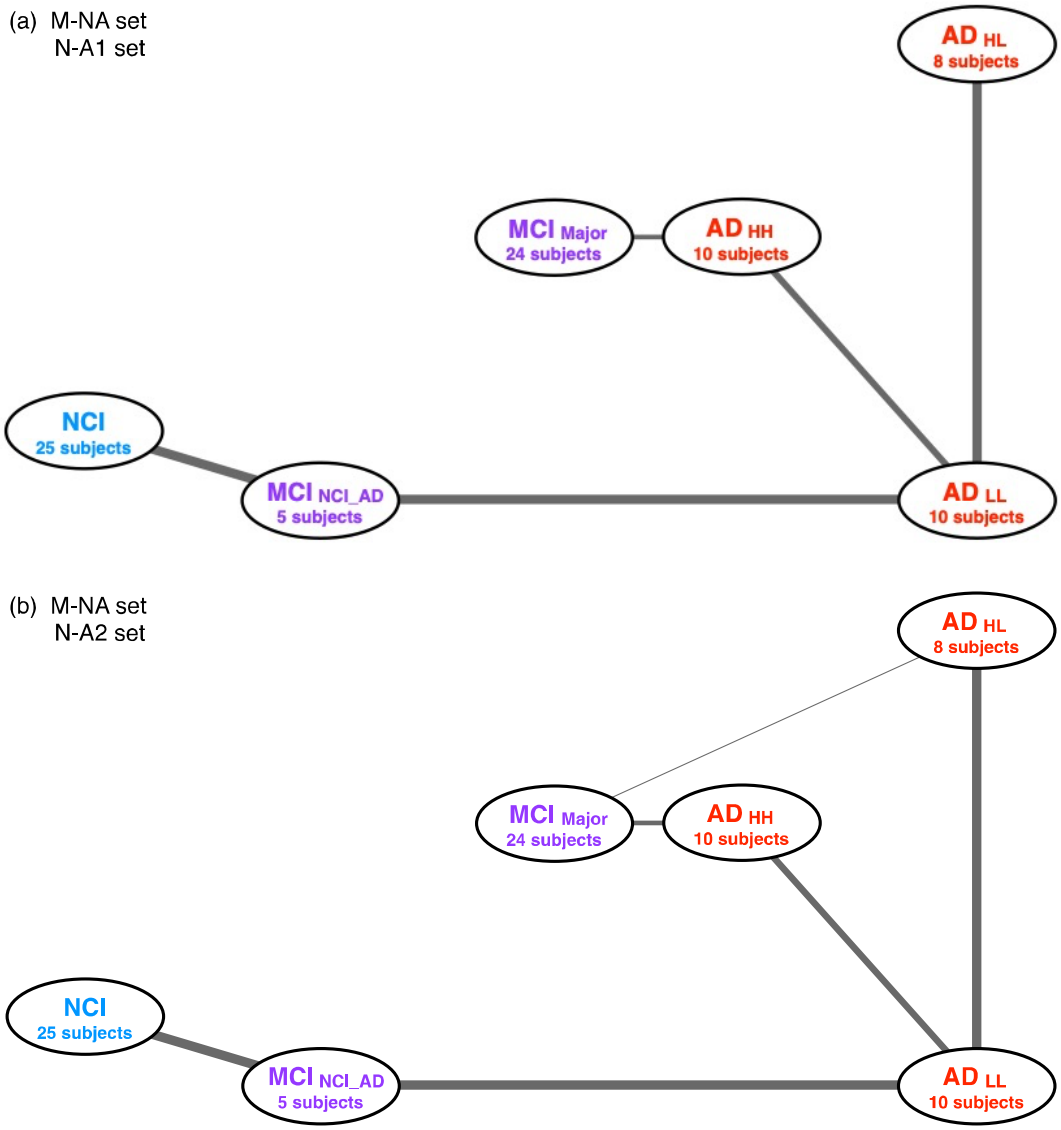


Figure 3: Adjacency network among substages. Shown are the networks constructed using the M-NA and N-A1 sets (a), and the M-NA and N-A2 sets (b). The width of each edge indicates the confidence in the presence of a connection between each pair of connected nodes.

(a)

# 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

# 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

(b)

# 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

# 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

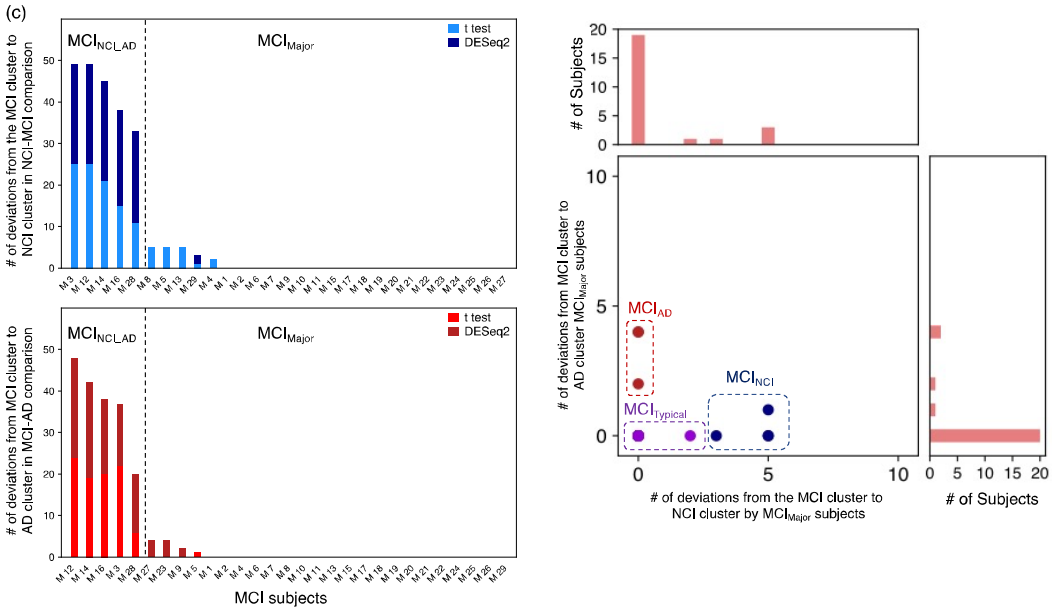


Figure 4: Detailed classification of the MCI subjects. Number of NCI subjects who belong to the MCI cluster (left value) and that of MCI subjects who belong to the NCI cluster (right value) as a function of P^t_{NM} and P^t_{MN} (a, left), and as a function of P^D_{NM} and P^D_{MN} (a, right). Number of MCI subjects who belong to the AD cluster (left value) and that of AD subjects who belong to the MCI cluster (right value) as a function of P^t_{MA} and P^t_{AM} (b, left), and as a function of P^D_{MA} and P^D_{AM} (b, right). Number of deviations of each MCI subject from the MCI cluster to the NCI cluster (c, upper left) and to the AD cluster (c, lower left). Broken lines indicate boundaries between the MCI_{NCI_AD} and MCI_{Major} subjects. Scatter plot of the MCI_{Major} subjects as a function of the number of times that

each subject belonged to the NCI cluster and the AD cluster (c, right). Using k-means clustering (Fig. S4c), the MCI_{Major} subjects were divided into $MCI_{Typical}$, MCI_{NCI}, and MCI_{AD}.

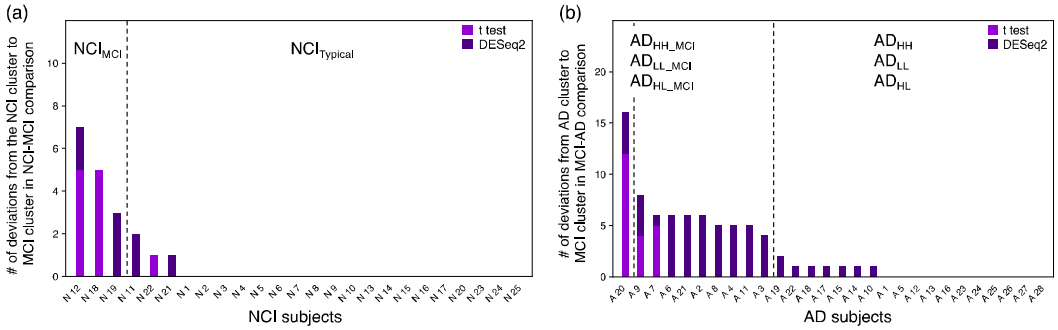
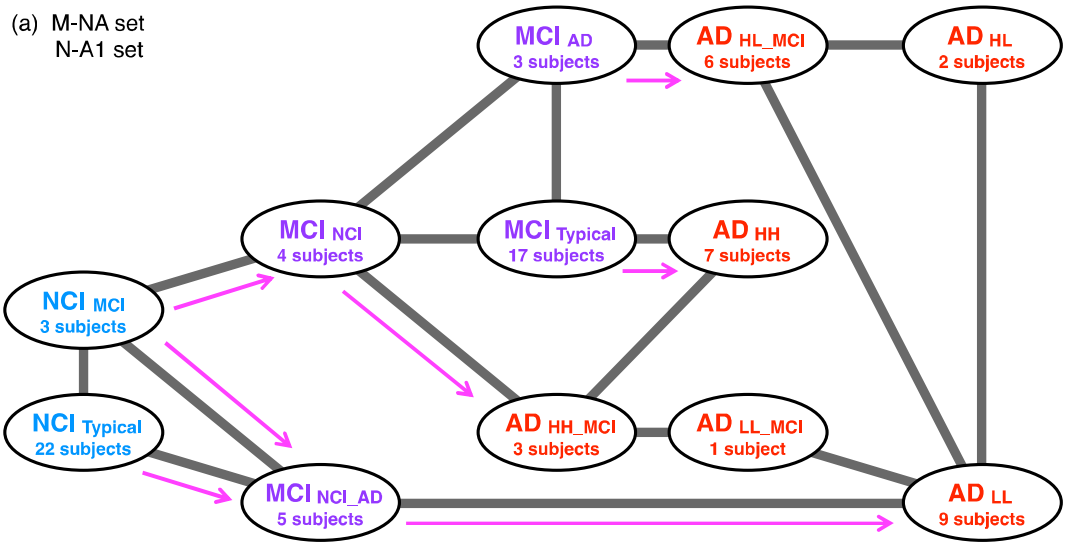


Figure 5: Detailed classification of the NCI subjects. Number of deviations of each NCI subject from the NCI cluster to the MCI cluster (a), and that of each AD subject from the AD cluster to the MCI cluster (b). Broken lines indicate boundaries among clusters of subjects obtained via k-means clustering (Fig. S5a and S5b).



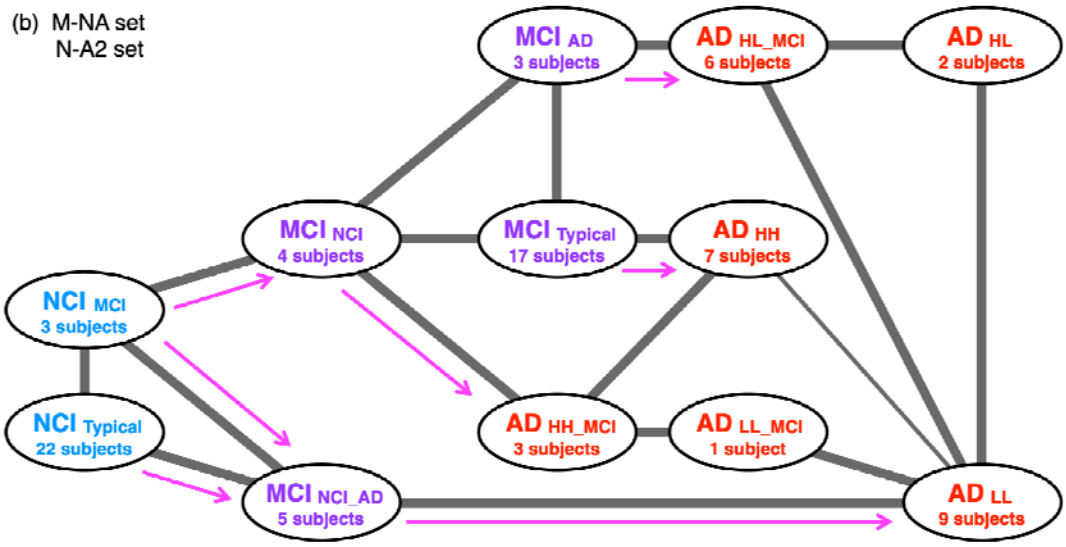
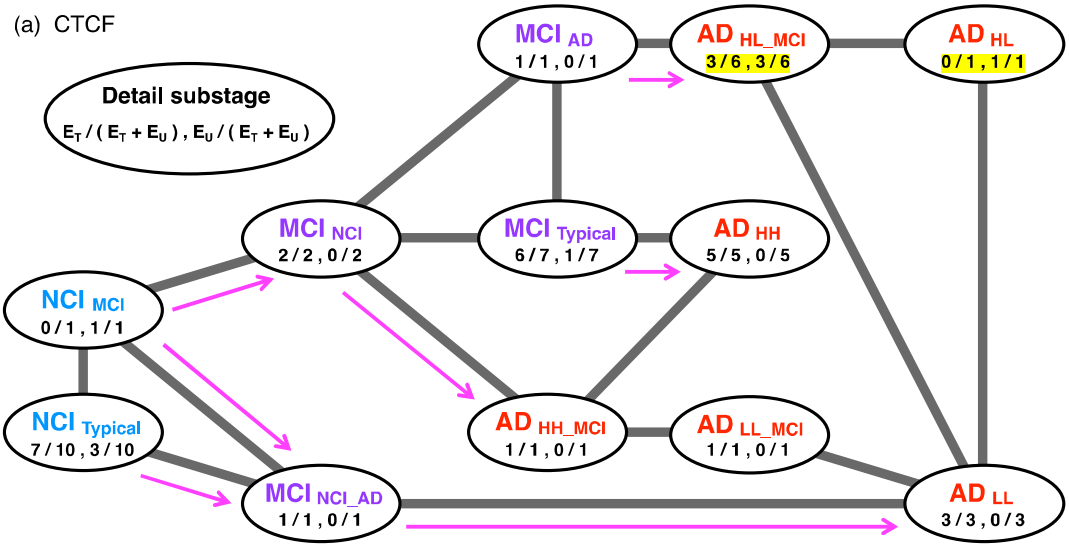


Figure 6: Adjacency networks among substages obtained using the data in the M-NA and N-A1 sets (a) and in the M-NA set and N-A2 sets (b). The width of each edge indicates the confidence in the presence of a connection between each pair of connected nodes. Arrows indicated the inferred direction of disease progression.



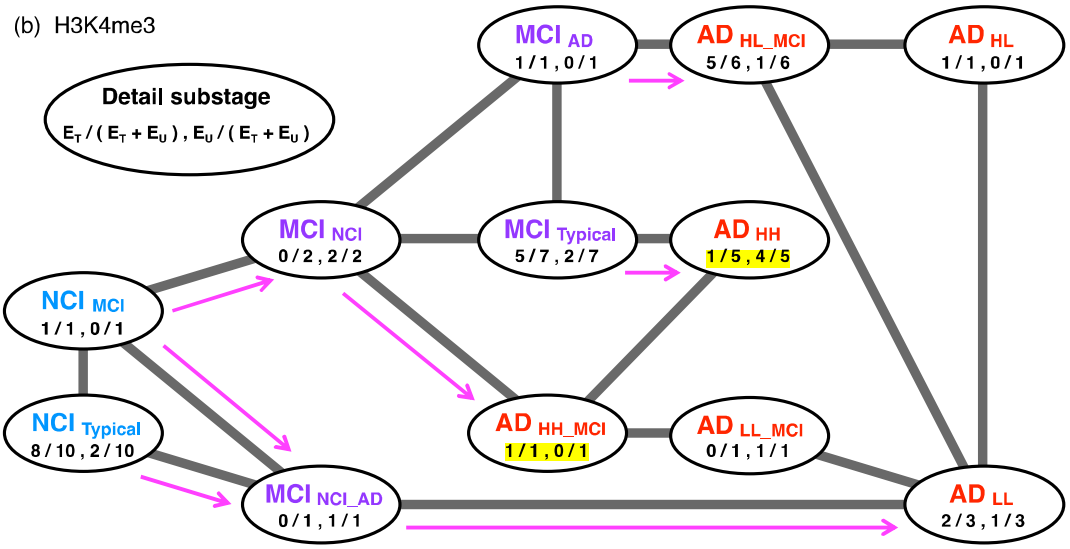


Figure 7: Adjacency network among substages showing the ratio of subjects with typical and untypical genome-wide distributions of CTCF (a) and H3K4me3 (b). A network obtained using data in the M-NA and N-A1 sets was used as an example for illustration. E_T and E_U indicate the number of subjects that exhibited typical and untypical genome-wide epigenetic states, respectively. Here, the state exhibited by a large number of subjects was regarded as the typical genome-wide epigenetic state (Details are shown in Fig. S6 and Table S6.). Data highlighted in yellow indicate that the proportion of E_U is higher in that substage than in others.