Exploring the genetic heterogeneity of Alzheimer's disease: Evidence for genetic subtypes

Jeremy A. Elman, Ph.D.[1,2,*], Nicholas J. Schork[1,3], Aaditya V. Rangan[4] and for the Alzheimer's Disease Neuroimaging Initiative[†]

[1] Department of Psychiatry, University of California San Diego, La Jolla, CA, USA

[2] Center for Behavior Genetics of Aging, University of California San Diego, La Jolla, CA, USA

[3] The Translational Genomics Research Institute, Quantitative Medicine and Systems Biology, Phoenix, AZ, USA

[4] Department of Mathematics, New York University, New York, New York, USA

[†]Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

Running Title: Biclustering of Alzheimer's genetic risk

**Correspondence should be addressed to Jeremy A. Elman, Ph.D., UCSD Department of Psychiatry, 9500 Gilman Drive (MC 0738), La Jolla, CA, USA, 92093. Tel: +1 858-534-6842 Fax: +1 858-822-5856 Email: jaelman@health.ucsd.edu*

ABSTRACT

Background: Alzheimer's disease (AD) exhibits heterogeneity in cognitive impairment, atrophy, and pathological accumulation, suggesting the potential existence of subtypes. AD is under substantial genetic influence, thus identifying systematic variation in genetic risk may provide insights into disease origins.

Objective: We investigated genetic heterogeneity in AD risk through a multi-step analysis.

Methods: We performed principal component analysis (PCA) on AD-associated variants in the UK Biobank (AD cases=2,739, controls=5,478) to assess the presence of structured genetic heterogeneity. Subsequently, a biclustering algorithm searched for distinct disease-specific genetic signatures among subsets of cases. Replication tests were conducted using the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset (AD cases=500, controls=470). We categorized a separate set of ADNI individuals with mild cognitive impairment (MCI; n=399) into genetic subtypes and examined cognitive, amyloid, and tau trajectories.

Results: PCA revealed three distinct clusters ("constellations") within AD-associated variants containing a mixture of cases and controls, reflecting disease-relevant structure. We found two disease-specific biclusters among AD cases. Pathway analysis linked bicluster-associated variants to neuron morphogenesis and outgrowth, including genes related to cellular components and development-modulating factors. Both disease-relevant and disease-specific structure replicated in ADNI. Individuals with genetic signatures resembling bicluster 2 exhibited increased CSF p-tau and cognitive decline over time.

Conclusions: This study unveils a hierarchical structure of AD genetic risk. Disease-relevant constellations may represent differential biological vulnerability that is itself not sufficient to increase risk. Biclusters may represent distinct AD genetic subtypes. This structure replicates in an independent dataset and relates to differential pathological accumulation and cognitive decline over time.

Keywords: Alzheimer’s disease, genetic subtypes, biclustering, genotyping, genetic risk

BACKGROUND

Alzheimer's disease (AD) diagnostic criteria have evolved over the years, but typically diagnosis has been characterized by predominant amnestic impairment that progressively impacts other cognitive domains and everyday functioning. However, clinical presentation is heterogeneous, and non-amnestic predominant variants of AD, often termed "atypical AD", were specifically acknowledged in a 2011 update to AD diagnostic criteria and guidelines (1). Pathological spread and neurodegeneration also tend to proceed in a stereotypical pattern, but mirroring the clinical diversity of AD, AD neuropathological exams and imaging studies have identified marked heterogeneity as well (2-8). Despite this variability, AD is considered a distinct entity due to the overarching clinicopathologic characteristics observed across individuals even though the etiological basis of AD remains unclear. That is, it may have a unitary origin with diverse presentation, a highly heterogenous etiology that converges on a common disease phenotype, or it may be consistent with some intermediate scenario.

Sporadic AD is under considerable genetic influence, with an estimated heritability of 60-80% (9). Thus, examining the genetic architecture of AD risk and the myriad ways in which combinations of variants associate with AD provides a useful foundation to understand its complex etiology. Although *APOE* represents the single largest source of genetic risk for the disease (10) recent GWAS have identified upwards of 75 different risk loci (11-15). Considering variants that do not reach the level of genome-wide-significance but still suggest an association with AD could provide additional insight above and beyond *APOE* (16) and genome-wide significant variants. The highly oligo or polygenic nature of AD risk could reflect underlying etiological heterogeneity across individuals. The evolving definition of AD and subsequent debate over its origins further highlights the complex nature of the disease (17-19). Given this genetic complexity yet commonality among aspects of AD clinical presentations, there is little reason to expect that phenotype-based classifications of sporadic AD (and its subtypes) will

cleanly delineate homogeneous subgroups of genetic risk, providing motivation for a using "genotype-first" approach to identifying subtypes (20-22).

Identifying genetic subtypes or heterogeneity typically involves cluster analysis of some sort. There are many clustering algorithms one can exploit, and these techniques can be applied to various sources of data, including gene expression, GWAS summary statistics, and individual genotype data (23-25). However, clustering approaches to identify genetic subtypes of disease face several difficulties (see Dahl et al. (26) for detailed discussion). First, a fundamental aspect of most clustering algorithms is that they will tend to identify clusters, even if no true clusters are present (i.e., false positive clusters). Second, a common approach is to search for clusters of genes or variants within a group of cases and consider them disease-related when this might not be the case, as it is often unknown whether these same clusters would be found in controls, which could reflect pathways or structures unrelated to disease status. The biclustering method described in Rangan et al. (27) addresses both issues. First, the biclustering method does not assume that a bicluster exists in the data, but rather tests the null hypothesis that one does *not* exist. Second, the technique searches for subsets of SNPs that are present in cases but *not controls*, providing evidence for disease-specificity.

Here we investigated genetic heterogeneity of Alzheimer's disease using a combination of approaches. We first applied a principal component analysis to AD cases and controls from the UK Biobank to identify potential clusters among AD-associated SNPs that may indicate disease-relevant vulnerability across all individuals, regardless of disease status. We then applied a biclustering method to each of these likely vulnerability clusters to search for subsets of cases that harbor distinct genetic signatures that significantly increase AD risk. To validate our findings, we investigate whether the same clusters or patterns of heterogeneity also appear in an independent group of cases and controls from the Alzheimer's Disease Neuroimaging Initiative (ADNI). Taken together, we find a hierarchical structure to an underlying heterogeneity of AD genetic risk, providing further insight into the complex etiology of the disease.

METHODS

Participant characteristics

We used imputed genotyping data from the UK Biobank (UKB) as the discovery dataset (**Table 1**). The UKB is a large-scale biomedical database and research resource containing genetic, lifestyle and health information from half a million UK participants (28, 29). Data from 2,739 Alzheimer's disease cases (30) and 5,478 age- and sex-matched controls with White British ancestry as determined by PCA (29) were included in these analyses.

A replication dataset was obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu; **Table 2**). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and early AD. Our biclustering analyses included genotyping data from 500 individuals with Alzheimer's disease cases and 470 controls from the ADNI-1 (n=520), ADNI-GO/2 (n=225), and ADNI-3 (n=225). Case-control status was based on the ADNI diagnosis given at each participant's last visit available. An additional 399 individuals (ADNI-1=166 , ADNI-GO/2=175, ADNI-3=58) diagnosed with MCI at their latest visit diagnosed with MCI at their latest visit were included for examination of cognitive and biomarker change. Participants were restricted to those with primarily European ancestry (>80%) as determined by SNPweights (31).

Genotyping data and quality control

Genotyping data imputed to the Haplotype Reference Consortium plus UK10K reference haplotype resource were downloaded from the UKB database along with genetic principal components (hereafter referred to as "genome-wide" principal components). Sample QC information provided by the UKB were used to apply exclusionary criteria. Imputed data used in

these analyses included only samples with <10% missingness and biallelic SNPs with >1% minor allele frequency.

Individuals in the ADNI cohorts were genotyped using the following chips: Illumina Human610-Quad BeadChip (ADNI-1), Illumina HumanOmniExpress BeadChip (ADNI-GO/2), and Illumina Infinium Global Screening Array v2 (ADNI-3). Genetic principal components (hereafter referred to as "genome-wide" principal components) were calculated from linkage disequilibrium (LD)-pruned variants in combination with 1000 Genomes data (32) for use as covariates in later analyses. Following standard genotyping QC, imputation was performed on the Michigan Imputation Server (https://imputationserver.sph.umich.edu/) (33) using the 1000 Genomes phase 3 EUR reference panel. Imputed data from all phases were then merged. Imputed data used in this analysis included only samples with <10% missingness and biallelic SNPs with >1% minor allele frequency.

Assessing high-level structure with principal components analysis

We restricted analyses to 486,823 variants in the UKB data that were associated with AD with an uncorrected p-value of <0.05 from the AD GWAS by Kunkle et al. (12) to accommodate potential association signals beyond those exhibiting genome-wide significance. An initial principal components (PC) analysis (PCA) was applied to the allele combinations (coded according to the 3 possible allele combinations for each) of these AD-associated variants across all UKB cases and controls to assess the presence of high-level structure in the data (i.e., a non-Gaussian distribution). Individuals were assigned to clusters, hereafter referred to as "constellations" to disambiguate from the term "bicluster" used in subsequent steps, by applying the ISO-SPLIT algorithm (34) implemented in MATLAB (https://github.com/flatironinstitute/isosplit5) to participant loadings on the first PC.

Although the sample was restricted to White British participants, it is possible that these loadings reflect remaining ancestry-related population substructure. We therefore determined whether the resulting structure was specific to the set of AD-associated variants in two ways.

First, PCA was applied to sets of variants restricted to p-value thresholds ranging from $p<0.05$ to $p<1.0$ in increments of 0.05. Loadings on the first two PCs were plotted at each increment, with individuals labelled according to cluster membership at the $p<0.05$ threshold to visually assess stability of clustering. Second, to determine whether the resulting structure was simply a function of the number of variants analyzed, a PCA was applied to a randomly chosen set of variants with the same size as those included in the analysis restricted to variants with $p<0.05$.

Gene set enrichment analysis of constellations

To characterize the biological and molecular functions associated with constellations identified by the PCA, we conducted a GWAS across AD-associated variants comparing individuals in each constellation against those from the other two using PLINK2 (35). Analyses were adjusted for the first two genome-wide principal components (PCs). Variants associated with each constellation (positive log-odds ratio and uncorrected $p<0.05$) were mapped to genes using the *g:Profiler* package in R (36). A gene set enrichment analysis was then applied to annotated genes from each constellation using the *clusterProfiler* package in R (37, 38). The analysis was restricted to gene sets from the Gene Ontology resource (39, 40) containing between 10 and 1,000 genes. A summary of the gene sets significantly enriched in each constellation were visualized with a comparison cloud using the R package *quanteda (41)*. These word clouds show terms that occur more frequently in the descriptions of gene sets associated with one constellation compared to the average frequency across all constellations.

Bicluster analysis of disease-specific structure

Heterogeneity exhibited at the top level of the constellations strongly suggests that there may be other heterogeneous substructures contained within each. This possibility seems all the more likely due to the non-Gaussian distribution of subjects within each constellation, and that the distribution of cases and controls appears to be in different directions across constellations. Therefore, we searched for disease-specific heterogeneity within each constellation separately.

To search for biclusters we use the half-loop method described in Rangan et al. (27). This strategy involves an iterative process which starts with all the participants (in this case, all participants in a given constellation) and SNPs, and then sequentially removes AD cases and allele-combinations from consideration. The Supplementary Methods contains a detailed description of this procedure but is summarized here. Briefly, we can measure the fraction of allele combinations that are shared between cases and subtract from this the fraction that are shared with controls to obtain a disease-related signal strength of remaining AD cases and controls. At each iteration the cases and allele-combinations with the smallest contributions to this signal are removed and the process repeats as described above. Recording this value at each step produces a 'trace' which indicates the disease-related signal-strength associated with the remaining AD cases and variants at that iteration. In pursuing this process, we controlled for the first two principal-components extracted from a GRM of all individuals in our sample to control for ancestry. We can use the peak of the trace to delineate the membership of the dominant bicluster (i.e., which AD cases and allellic-combinations contribute to the disease-specific signal). We identify the peak by finding the internal maximum, ignoring the initial and final iterations that include >95% or <5% of AD cases.

As a null-hypothesis, we assume that the disease-label (i.e., case vs control) is not associated with the genetic profile of each subject. Therefore, we randomly permuted the case- and control-labels across subjects with similar genome-wide principal components (see Rangan et al. (27) for details) and calculating traces as described above. Here, we use 500 permutations. By comparing the original trace with the distribution of traces drawn from the null-hypothesis, we can assign a p-value to the observed trace at each iteration. In this case, we only assess iterations that include >5% of the cases (i.e., the final iterations with very few cases and/or variants are ignored). To determine whether we have found a statistically significant bicluster within the original data (termed the 'dominant' bicluster), we can examine the p-value of the highest peak (p_{max}) and the average p-value across iterations (p_{avg}). Depending on the

structure of the bicluster, the various p-values may be quite different, but each may be a useful metric for identifying significant biclusters. For example, if the original trace has one or more clear peaks, then there are statistically robust 'cutpoints' which can be used to delineate bicluster membership and p_{max} is likely to be very small. On the other hand, if the original trace has a very broad peak or a long plateau, then the bicluster is quite 'fuzzy', corresponding to a continuum of membership. In this case, we might expect p_{avg} to be small, but p_{max} may be relatively large. If the dominant bicluster within a data-set is statistically significant, we can extract it and then search for a secondary bicluster. This is done by scrambling the entries of the submatrix associated with the bicluster (i.e., entries corresponding to the participants and allele-combinations that were retained in the bicluster) and running the search algorithm again (27, 42).

Gene set enrichment analysis of bicluster

We conducted a GWAS across AD-associated SNPs comparing bicluster cases with controls belonging to the same disease-relevant constellation using PLINK2 (35). Analyses adjusted for the first two genome-wide principal components. Variants associated with each constellation (positive log-odds ratio and uncorrected $p<0.05$) were mapped to genes using the *g:Profiler* package in R (36). A gene set enrichment analysis was then applied to annotated genes from each constellation using the *clusterProfiler* package in R (37, 38). The analysis was restricted to gene sets from the Gene Ontology resource (39, 40) containing between 10 and 1,000 genes. The Cytoscape app EnrichmentMap v3.3 (43) was used to visualize gene sets significantly enriched (FDR corrected $p<0.05$) in each bicluster. We constructed a network in which nodes represented gene sets significantly associated with each bicluster (gene sets can be associated with either or both). Edges were defined by the proportion of overlapping genes between gene sets (using a minimum overlap threshold of 0.5). Next, the AutoAnnotate app (http://baderlab.org/Software/AutoAnnotate) was used to cluster and annotate nodes based on degree of overlapping genes using the MCL Cluster algorithm.

Validation of genetic heterogeneity in ADNI

We next investigated whether the disease-relevant and disease-specific heterogeneity found in the UKB would generalize to Alzheimer's disease cases and controls from the ADNI (see Supplementary Methods for full details). First, we determined the subset of AD-associated SNPs common to both the UKB and ADNI datasets. Mirroring the original analysis, we applied a PCA to all cases and controls from the across this set of intersecting SNPs. ADNI participants were projected into the same principal components space using the SNP loadings defined in UKB data, allowing us to assess overlap between datasets. ADNI participants were assigned to the nearest constellation based on participant loadings on the first principal component.

After participants were classified into constellations, we assessed replication of the dominant biclusters separately in the constellations in which they were found (see Supplementary Methods for full details). For a given bicluster, we calculated the first 2 dominant SNP-wise principal components (using only SNPs in common to both datasets) among the UKB cases belonging to the bicluster. The resulting SNP loadings were used to project UKB participants in the constellation (including controls, bicluster cases, and non-bicluster cases). Similarly, we calculated projections for each of the ADNI participants belonging to the constellation using these UKB bicluster-defined loadings. If the bicluster structure is present in the ADNI data, then the distribution of cases and controls in this space should be similar between datasets. That is, for a given participant in the testing set, participants from the training set located nearby should tend to have the same label. We can choose what fraction of the training set participants to compare across, and for a given choice we calculated the proportion of nearest neighbors in the training set that have the same label (i.e., either case or control) as each of the individuals in the test set. We assessed statistical significance using a permutation test in which the average "label match" across participants in the test set is compared to a null distribution constructed from label-shuffled data to calculate an empirical p-value. Here, we ran 500 permutations. We repeated the steps above and calculated an empirical p-value across a

range of nearest neighbors, using up to 50% of the training set participants for calculating label similarity. This allowed us to assess statistical significance across the set of parameter choices. The null distribution built from label-shuffled data also allowed us to calculate means and variances for each fraction of nearest neighbors so that a normalized z-score could be calculated for the original data. The fraction of nearest neighbors producing the maximum z-score indicates the optimal fraction for labelling new data, as described below.

Association of bicluster groups with cognitive and biomarker trajectories

The procedures described above were repeated on the full set of ADNI genotyping data, this time including individuals with MCI in addition to AD cases and controls. This included projecting ADNI data onto PC space defined by the UKB data and assigning to the nearest constellation, and then projecting individuals from constellations 1 and 2 using the bicluster-defined loadings from biclusters 1 and 2, respectively. The number of nearest neighbors with each label (i.e., bicluster case, non-bicluster case, and controls) was then recorded for each ADNI participant. The fraction of training set participants used as nearest neighbors was determined by the fraction with the highest z-score in the validation step described above. Each ADNI participant was assigned a soft label, calculated as the proportion of nearest neighbors in the UKB training set that were bicluster cases. To assign individuals to biclusters 1 and 2, k-means clustering with k=2 was applied to the soft labels (i.e., proportions of nearest neighbors that were bicluster cass) of each ADNI participant belonging to constellations 1 and 2 separately.

Phenotypic analyses were restricted to the ADNI MCI participants who were not included in the bicluster validation analysis. We examined change of cognition, amyloid, and phosphorylated tau (p-tau) over time between individuals labelled as bicluster 1, bicluster 2, or non-bicluster. Cognition was assessed using scores on the Preclinical Alzheimer's Cognitive Composite (PACC) (44, 45). Amyloid was assessed with florbetapir PET data processed according to previously published methods (http://adni.loni.usc.edu/methods) (46, 47).

Specifically, we downloaded mean standardized uptake value ratios (SUVR) from a set of regions including frontal, temporal, parietal and cingulate cortices using whole cerebellum as a reference region. Cerebrospinal fluid p-tau CSF samples were collected on cohort participants and processed as previously described (48). CSF p-tau was measured with the fully automated Elecsys immunoassay (Roche Diagnostics) by the ADNI biomarker core (University of Pennsylvania). Florbetapir PET and CSF p-tau were chosen as the measures of amyloid and p-tau because they provided the highest number of visits with relevant data across our MCI participant group.

Group differences between individuals with MCI labelled as bicluster 1, bicluster 2, or non-bicluster were assessed with linear mixed effects models using the *lme4* (49) and *lmerTest* (50) R packages. For each outcome of interest (cognition, amyloid, or p-tau), all timepoints with available data were included. An interaction between age and group was used to assess differences in the trajectories of cognitive and biomarker measures over time. A random intercept was included for participant.

RESULTS

Identification of disease-relevant constellations

The PCA of AD-associated variants (n=486,823) in the UKB data revealed three distinct clusters, or “constellations”, each containing a mixture of AD cases and controls (**Figure 1**). The plots in **Figure 2** indicate that this structure only emerges when restricting to AD-associated SNPs. When all variants are included in the PCA (e.g., no p-value criterion is used) there is also evidence for substructure, but it is quite different than what is observed when a threshold of $p<0.05$ is used, with individuals from each of the constellations being highly interspersed. The separation of constellations along the dominant component emerges around $p<0.25$, and the constellations begin to further separate along the second PC at $p<0.05$. Results of a PCA applied to a random set of SNPs of the same size as were included at the $p<0.05$ level again

found that individuals from each constellation were interspersed, and the overall structure closely mirrored that seen when all SNPs were included (**Supplementary Figure 1**). These findings suggest that the constellations are *disease-relevant* because they only emerge when considering AD-associated variants. However, they are not *disease-specific* because constellations contain both cases and controls, suggesting cluster membership alone does not discriminate between high and low disease risk. The three constellations were not significantly different on demographic characteristics, including sex, age, education, percent with AD dementia, or number of APOE-e4 alleles (**Supplementary Table 1**).

Gene set enrichment results of disease-relevant constellations

We sought to identify biological functions associated with each disease-relevant constellation by examining gene set enrichment of variants associated with each. Themes more frequently associated with each constellation were visualized with a comparison cloud, as depicted in **Figure 3**. Constellation 1 was broadly associated with cell signaling. There was also enrichment of gene sets related to transmembrane ion transport, plasma membrane adhesion, regulation of GTPase activity, axon development and morphogenesis, dendritic spine, synaptic vesicle release, kinase activity, and blood vessel development. Constellation 2 featured genes involved in cellular structure, including gene sets related to major cellular components such as axons, synapses, and plasma membrane. Other enriched gene sets included those related to lipid transport. Constellation 3 was generally associated with morphogenesis and regulation of membrane potential through ion homeostasis. This constellation was also associated with gene sets related to axonal development, cell size and growth as well as GTPase-mediated signal transduction. These represent basic processes and cellular components that are not specific to AD but may represent systems that are vulnerable to other genetic and environmental risk factors or insults. It should be noted that these constellations exist in a high-dimensional space, so each shares enrichment for many GO terms with one or the other two of the three constellations (**Supplementary Figure 2**). However, the proportion of overlap seen at the level

of gene set is reduced when looking at the overlap among constituent genes or SNPs (**Supplementary Figure 3**). This indicates that the variants associated with different constellations may have impacts that converge on similar downstream pathways, but the specific perturbations encompassed by each may not be the same.

Identification of disease-specific biclusters

The stark heterogeneity exhibited at the level of the disease-relevant constellations strongly suggests that there may be other heterogeneous substructures contained within each. This possibility seems all the more likely due to the non-Gaussian distribution of subjects within each constellation. Additionally, the distribution of cases and controls appears to be different across constellations. Examining **Figure 1**, it appears that the bias or shift in cases relative to controls does not fall in the same direction across constellations, thus requiring that the search for biclusters be carried out separately for each constellation. The results of our bicluster searches are shown in **Figure 4.** Each subplot corresponds to the search in a different constellation. Within each subplot the red curve corresponds to the signal-strength of the dominant bicluster within the data, referred to as a 'trace' in the Methods section. The black curves indicate the distribution of traces drawn from the null-hypothesis via a permutation-test. Generally speaking, a red trace that has either a high peak or a high average (relative to the distribution of black traces) indicates a statistically significant bicluster. For this data set we detected a statistically significant bicluster in constellation 1 (p_{avg}=0.002, p_{max}=0.008) and constellation 2 (p_{avg}=0.028, p_{max}=0.244), but not in constellation 3 (p_{avg}=0.484, p_{max}=0.102), which is the smallest of the 3 constellations. The shape of the traces conveys additional information about the structure of the biclusters. For example, the sharp peaks defining bicluster 1 in **Figure 4A** indicate clear cut points at which membership can be clearly delineated (i.e., a disease subtype). In contrast, the broader plateau of the trace in **Figure 4B** reflects a bicluster with "fuzzy" boundaries whereby there is a more smooth continuum of membership in the

bicluster. This is also reflected in the different pattern of p-values between biclusters (i.e., whether the peak or average of the trace tends to be more significant). Thus, examining the traces informs us whether disease heterogeneity takes the form of distinct subtypes, a continuous spectrum of risk, or some intermediate structure. After delineating and removing the dominant bicluster within constellations 1 and 2 (see Methods) we searched once again for any additional biclusters but did not find a second bicluster that was statistically significant in either constellation (**Supplementary Figure 4**). However, when searching for these secondary biclusters, the clearly non-Gaussian distribution of traces drawn from the label-shuffled null distribution (visualized as clumped strands among the black traces) suggests there is residual heterogeneity, but we are not powered to identify it in the current sample. The bicluster participants were not significantly different on demographic characteristics, including sex, age, education, or number of APOE-e4 alleles (**Supplementary Table 2**).

Gene set enrichment results of disease-specific biclusters

Bicluster 1 was enriched for a number of gene sets, including those related to receptor activity, calcium and sodium ion transport, dendritic structure, GTPase activity, and regulation of the MAPK cascade. Bicluster 2 was enriched for gene sets related to the MHC protein complex, regulation of cell size, lipid transport, and tyrosine kinase activity. Both biclusters showed enrichment for cell-projection morphogenesis, and synaptic transmission. **Figure 5** shows network plots of gene sets enriched in each bicluster. The gene sets are grouped by proportion of overlapping genes and labelled with dominant terms to illustrate the shared and unique functions associated with variants in each bicluster. Although many of the enriched gene sets identified in this analysis relate to basic functions and thus do not appear to be AD-specific, it should be noted that the variants considered in these analyses were already selected for their association with AD. That is, the gene sets shown in **Figure 5** are those which are enriched in bicluster variants relative to the other AD-associated variants, rather than with respect to all

other variants, including AD-nonspecific variants. Pathways that are typically found to be enriched for AD-associated variants constitute a background signal that is likely shared among many of the cases, including those that are not part of either bicluster.

Disease-relevant constellations are evident in ADNI data.

Using only SNPs common to both datasets, we re-calculated the top principal components of AD-associated variants in the UKB data. These loadings were used to project both the UKB and ADNI data into the same PC space. As seen in **Figure 6**, the grouping of UKB participants into 3 constellations was almost perfectly mirrored in the ADNI data, albeit with a slight shift in positioning. As in the UKB data, each constellation contained a mixture of AD cases and controls, reflecting disease-relevant but not disease-specific clustering.

Disease-specific biclusters replicate in ADNI

Overall, we found evidence that the bicluster structure replicated in the ADNI data in a statistically significant way (i.e., the chance of a random replication is very small). The highest evidence of replication was achieved by calculating label similarity using 31.6% of the sample as nearest neighbors for bicluster 1 ($p<0.002$) and using 27.3% of the sample as nearest neighbors for bicluster 2 ($p=0.028$). **Figure 7** displays the significance of labeling accuracy across a range of nearest neighbors. Note that the shape of these curves can be influenced by the structure of the bicluster. For example, a bicluster with clearly delineated boundaries may be less sensitive to altering the nearest neighbor parameter compared to one with blurrier boundaries. Consistent with this, we see significant replication of bicluster 1 across a large range of nearest neighbor fractions, ranging from about 17% up to our maximum threshold of 50%. In contrast, replication of bicluster 2 is more sensitive to this parameter choice, with

significant label-matching using nearest neighbor fractions from about 26% to 36%. Importantly, these plots provide a range of sensible values one can use to label new data.

Cognitive and biomarker trajectories across bicluster groups in ADNI MCI participants

We found evidence for differential cognitive and biomarker trajectories across bicluster groups in a separate set of ADNI individuals diagnosed with MCI. Individuals assigned to bicluster 2 demonstrated significantly greater decline on the PACC compared to the non-bicluster (β=-0.40, t-value=-2.29, p=0.022) and bicluster 1 groups (β=-0.60, t-value=2.92, p=0.004) (**Figure 8A**). Bicluster 1 demonstrated a somewhat greater increase in florbetapir over time compared to the non-bicluster group, but this difference was not significant groups (β=0.32, t-value=1.52, p=0.129) (**Figure 8B**). Bicluster 2 demonstrated a significantly greater increase of CSF p-tau over time compared to the non-bicluster group (β=0.45, t-value=2.05, p=0.041) (**Figure 8C**).

DISCUSSION

Genetic analyses suggest the etiology of AD is multifactorial (51), but the extent to which the composition of genetic influences on AD risk varies across individuals and whether it is replicable across different data sets has been unclear. We identified evidence for several subsets of individuals that exhibit different patterns of genetically-mediated vulnerability to the risk for AD. Among AD-associated SNPs, there is clear heterogeneity across individuals, and this heterogeneity seemed to follow a hierarchical structure. Importantly, this structure was observed in two independent datasets, indicating that it is not sample-specific, but rather appears to be a generalizable feature of AD genetic risk.

The first level of heterogeneity we identified appears to reflect susceptibility or vulnerability to AD. Essentially, three clusters arose from a PC analysis of the UKB data confined to variants associated with AD based on prior GWAS, which we have termed “constellations”. As noted, the composition of these constellations includes both AD cases and

controls, indicating membership in one or another of these groups alone is not sufficient to distinguish increased likelihood of having an actual AD diagnosis. Similar analyses (i.e., PCA) are often used to identify ancestry-related population structure (52). Despite the fact that the UK Biobank sample is relatively homogeneous when compared to other cohorts of individuals with European ancestry (53), there may still be residual ancestry-related stratification in the UKB data sets (54). However, we found that the constellations only emerged when examining AD-associated SNPs and individuals from each constellation were highly intermixed when analyzed using all SNPs or a random selection of SNPs, suggesting that the constellations do not simply reflect ancestry-related substructure. Thus, since the constellations only arose when restricting analysis to AD-associated variants and include both cases and controls, we believe these constellations are “disease-relevant” but do not reflect variants that only distinguish cases and controls.

While some of the pathways that encompass the genes harboring or associated with the variants used to identify these constellations are intuitive given what is known about AD pathogenesis, it is also evident that many are related to more basic aspects of cellular structure and function. Broadly speaking, constellation 1 was associated with cell signaling, constellation 2 with lipid transport and synaptic structure, and constellation 3 with cellular growth and morphogenesis. In many cases, a given gene set or theme was shared between 2 out of the 3 constellations, indicating vulnerability may exist along a multi-dimensional spectrum and the constellations represent the different axes of this space. It is perhaps not surprising that synaptic structure and function are prominent among the enriched pathways given the central role of synapses in AD. The course of AD involves extensive synaptic degeneration that closely tracks with cognitive decline, and synaptic connections are proposed to be one route of pathological spreading (55-57). It may be that the variant combinations underlying or contributing to the constellations represent ways in which individuals are vulnerable to disease processes, but other factors – whether genetic or environmental – are necessary to induce

these processes and ultimately lead to AD diagnosis. For example, neurons derived from AD patients demonstrated differential levels of Aβ secretion, but when neurons from all patients were treated with comparable levels of extrinsic Aβ they exhibited differential synaptic vulnerability (58) that was independent of their intrinsic Aβ production. In this scenario, the relative vulnerability or resilience of cells does not by itself lead to a disease state, but rather modulates the impact of pathological insult to determine disease progression and severity.

Cases belonging to the disease-specific biclusters harbored distinct genetic signatures when compared to other AD cases and may therefore reflect “subtypes” of AD genetic risk. The variants defining the two disease-specific biclusters were enriched for both similar and unique pathways when compared to each other and the disease-relevant constellations in which they were found. It is important to re-emphasize the fact that these analyses were restricted to AD-associated variants, which are enriched for lipid processing, cholesterol transport, amyloid precursor (APP) processing and Aβ formation, tau protein binding, and immune response (12). The bicluster-associated pathways reported here thus exist within this broader context of other AD-associated pathways. Bicluster 1 was enriched for gene sets involved in regulating the MAPK cascade, which may contribute to AD progression in several cell-type-dependent ways, such as increasing neuroinflammation, promoting neurofibrillary tangle formation, and depressing synaptic plasticity (59). Bicluster 2 demonstrated enrichment in a cluster of gene sets related to the major histocompatibility complex (MHC), which is consistent with findings that implicate microglia-mediated immune response as a key player in AD (12, 15, 60, 61). Both biclusters – and bicluster 1 in particular – were enriched for gene sets related to synaptic signaling and signal transduction. Similar enrichment was found among newly prioritized genes in the most recent largescale GWAS of AD (15). We also found both biclusters were strongly enriched for a number of pathways that interact to influence the morphogenesis and outgrowth of neurons. This includes gene sets related to the cellular components themselves (e.g., cell bodies, dendritic spines, synapses, axonal projections, cystoskeletal components), but also

factors that modulate their development such as polymerization, β-catenin binding, and GTPase activity (62-66). Disruption of these processes may leave structures more vulnerable to insult resulting in, for example, greater Aβ-mediated synapse loss. The hallmark pathologies of AD, i.e., Aβ and tau protein build-up, may impact development and structural integrity through their toxic effects, but also through a loss of function. For example, APP is important for regulating axonal and synaptic growth through influencing cytoskeletal remodelling (67). It may be that a shift towards the amyloidogenic pathway of APP metabolism results in increased Aβ production, but also reduced function of APP in regulating axonal and synaptic growth. Similarly, tau plays a critical role in stabilizing microtubules and, when hyperphosphorylated, becomes unbound. This may allow increased aggregation into neurofibrillary tangles but also results in loss of cytoskeletal integrity (68, 69). Further work is needed to comprehensively characterize the impacts of these bicluster-associated variants, but their association with fundamental aspects of cellular development suggests AD risks may manifest relatively early in life.

We found the both the disease-relevant constellation structure and the disease-specific biclusters replicated in the ADNI sample. There are several important differences between the ADNI and UKB cohorts. First, although analyses were restricted to individuals of European ancestry, the UKB was further restricted to a White British subset, so the ADNI data may contain relatively more ancestry-related heterogeneity. Second, the studies used different genotyping chips (and ADNI genotypes were obtained with 3 different chips across phases) and imputation panels. Third, ADNI participants were recruited with the intent of mirroring a clinical trial population focused on Alzheimer's disease, whereas the UKB attempted to recruit a broader base of individuals to study multiple outcomes. Replicable structure across these datasets is thus unlikely to reflect dataset-specific or artifactual confounds that should differentially affect these datasets.

Supporting the relevance of the genetic structure found here, we found that a separate set of ADNI participants diagnosed with MCI assigned to different bicluster groups

demonstrated differential cognitive and biomarker trajectories. Individuals with genetic signatures resembling bicluster 2 exhibited greater accumulation of p-tau and a corresponding steeper decline in cognitive performance over time. The variants associated with this genetic subtype may contribute to a more aggressive form of AD. We previously found that genetic risk affecting different biological pathways can preferentially relate to amyloid or tau accumulation (70). Thus, it may be that the balance of pathological accumulation is shifted towards p-tau in these individuals, and p-tau has been shown to be more closely linked to subsequent neurodegeneration (71) and cognitive decline (72) compared to amyloid. The genetic heterogeneity observed here may impact not only level but distribution of pathology, which may be particularly important in the case of tau (3, 7, 73). Heterogeneity in genetic risk may also relate to comorbid conditions that exacerbate AD progression. It is therefore clear that more work is needed to characterize the downstream impacts of genetic heterogeneity on disease outcomes.

The 2018 National Institute on Aging and Alzheimer's Association research framework represented a shift towards defining AD as a biological construct (17). One goal of this effort is to focus research on disease mechanisms that may guide development of more pathology-specific interventions, but many questions regarding the etiology of AD remain. For example, there is evidence that Aβ and tau, the defining features of this framework, impact downstream events such as inflammatory response and subsequent neuronal loss (74, 75). However, there may be additional drivers that represent potential therapeutic targets, including factors upstream of Aβ and tau accumulation. Identifying subtypes of AD could facilitate precision medicine approaches that tailor treatment to the individual. Unfortunately, determining which processes may be causes of pathology and which may result from an underlying pathology is not trivial; for example, gene expression differences between individuals with and without AD may reflect both causal factors and outcomes of disease processes (76). A benefit of analyzing genetic data is

that the direction of causation can more easily be inferred since one's genome – or more specifically germline genome – is fixed at fertilization.

Our findings may indicate that there are likely multiple genetically-mediated pathologies underlying AD that converge on common clinical manifestations. This is not a unique phenomenon. For example, Charcot-Marie-Tooth disease involves a common set of clinical symptoms that can arise from separate genetic origins (77). We did find evidence for overlap in the pathways enriched across the different constellations and biclusters identified in our analysis, and this may indicate where convergence begins. However, there are a few caveats. First, receiving the same clinical diagnosis does not rule out the possibility of meaningful differences across cases. There are atypical forms of clinically-defined AD (78-81) and there is growing evidence for several biological subtypes of AD characterized by distinct patterns of pathological spread or neurodegeneration (2, 3, 5). Further work is needed to determine whether the genetic heterogeneity identified here is associated with more subtle phenotypic differences. Second, the degree of overlap among pathways that we observed is driven by SNPs that were assigned to genes and pathways, but not all SNPs were assigned to genes. Many AD-associated SNPs are located in non-coding regions, so the overlap of enriched gene sets ignores substantial numbers of SNPs uniquely associated with each. Differences between constellations and biclusters may arise from the functional effects of the variants themselves, including those not assigned to genes.

CONCLUSIONS

In sum, we found evidence of a hierarchical structure underlying heterogeneity in the genetic risk of AD. The disease-relevant constellations we identified may represent vulnerability of particular systems and processes throughout the lifespan that confers susceptibility to diseases beyond just Alzheimer's. On the other hand, the biclusters may be considered disease subtypes with distinct genetic signatures compared to the broader population of AD cases,

which may have important implications for treatment efforts. Despite all cases presenting with a common clinical syndrome, the etiology and path taken to clinical manifestation may vary across patients. The differential cognitive and biomarker trajectories between genetic subtypes provides some evidence that this heterogeneity has consequences for downstream disease processes. In sum, identifying subtypes of AD could facilitate precision medicine approaches that tailor treatment strategies to the individual for increased effectiveness.

Acknowledgments

This research has been conducted using data from UK Biobank, a major biomedical database (http://www.ukbiobank.ac.uk/). Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

Funding

This work was supported by National Institute on Aging grants K01 AG063805 (JAE), UH3-AG064706 (NJS), U19 AG023122 (NJS), U19 AG065169-01A1 (NJS) and U24AG078753 (NJS).

Data availability

Data used in preparation of this article were obtained from the UK Biobank (application ID 63648) database at http://www.ukbiobank.ac.uk/ and the Alzheimer's Disease Neuroimaging Initiative (ADNI) database at https://adni.loni.usc.edu/. These data may be obtained upon application to the respective studies and completion of data use and/or material transfer

agreements. Code implementing the bicluster search is available at https://github.com/adirangan/lakcluster_matlab. Additional code used in the analysis is available upon request to the corresponding author (J.A.E.).

Conflict of interest

The authors declare that they have no competing interests.

REFERENCES

1. McKhann GM, Knopman DS, Chertkow H, Hyman BT, Jack CR, Jr., Kawas CH, et al. The diagnosis of dementia due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimers Dement. 2011;7(3):263-9.
2. Murray ME, Graff-Radford NR, Ross OA, Petersen RC, Duara R, Dickson DW. Neuropathologically defined subtypes of Alzheimer's disease with distinct clinical characteristics: a retrospective study. Lancet Neurol. 2011;10(9):785-96.
3. Vogel JW, Young AL, Oxtoby NP, Smith R, Ossenkoppele R, Strandberg OT, et al. Four distinct trajectories of tau deposition identified in Alzheimer’s disease. Nature Medicine. 2021.
4. Young AL, Marinescu RV, Oxtoby NP, Bocchetta M, Yong K, Firth NC, et al. Uncovering the heterogeneity and temporal complexity of neurodegenerative diseases with Subtype and Stage Inference. Nat Commun. 2018;9(1):4273.
5. Ferreira D, Nordberg A, Westman E. Biological subtypes of Alzheimer disease: A systematic review and meta-analysis. Neurology. 2020;94(10):436-48.
6. Ferreira D, Verhagen C, Hernandez-Cabrera JA, Cavallin L, Guo CJ, Ekman U, et al. Distinct subtypes of Alzheimer's disease based on patterns of brain atrophy: longitudinal trajectories and clinical applications. Scientific reports. 2017;7(1):46263.
7. Ossenkoppele R, Schonhaut DR, Scholl M, Lockhart SN, Ayakta N, Baker SL, et al. Tau PET patterns mirror clinical and neuroanatomical variability in Alzheimer's disease. Brain. 2016;139(Pt 5):1551-67.
8. Tijms BM, Gobom J, Reus L, Jansen I, Hong S, Dobricic V, et al. Pathophysiological subtypes of Alzheimer’s disease based on cerebrospinal fluid proteomics. Brain. 2020.
9. Gatz M, Reynolds CA, Fratiglioni L, Johansson B, Mortimer JA, Berg S, et al. Role of genes and environments for explaining Alzheimer disease. Arch Gen Psychiatry. 2006;63(2):168-74.
10. Corder EH, Saunders AM, Strittmatter WJ, Schmechel DE, Gaskell PC, Small GW, et al. Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. Science. 1993;261(5123):921-3.
11. Harold D, Abraham R, Hollingworth P, Sims R, Gerrish A, Hamshere ML, et al. Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease. Nat Genet. 2009;41(10):1088-93.
12. Kunkle BW, Grenier-Boley B, Sims R, Bis JC, Damotte V, Naj AC, et al. Genetic meta-analysis of diagnosed Alzheimer’s disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing. Nature Genetics. 2019;51(3):414-30.
13. Lambert JC, Ibrahim-Verbaas CA, Harold D, Naj AC, Sims R, Bellenguez C, et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. Nat Genet. 2013;45(12):1452-8.
14. Jansen IE, Savage JE, Watanabe K, Bryois J, Williams DM, Steinberg S, et al. Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. Nat Genet. 2019.
15. Bellenguez C, Küçükali F, Jansen IE, Kleineidam L, Moreno-Grau S, Amin N, et al. New insights into the genetic etiology of Alzheimer’s disease and related dementias. Nature Genetics. 2022.
16. Escott-Price V, Sims R, Bannister C, Harold D, Vronskaya M, Majounie E, et al. Common polygenic variation enhances risk prediction for Alzheimer's disease. Brain. 2015;138(Pt 12):3673-84.

17. Jack CR, Jr., Bennett DA, Blennow K, Carrillo MC, Dunn B, Haeberlein SB, et al. NIA-AA Research Framework: Toward a biological definition of Alzheimer's disease. Alzheimers Dement. 2018;14(4):535-62.
18. Dubois B, Villain N, Frisoni GB, Rabinovici GD, Sabbagh M, Cappa S, et al. Clinical diagnosis of Alzheimer's disease: recommendations of the International Working Group. The Lancet Neurology. 2021.
19. Dubois B, Feldman HH, Jacova C, Cummings JL, Dekosky ST, Barberger-Gateau P, et al. Revising the definition of Alzheimer's disease: a new lexicon. Lancet Neurol. 2010;9(11):1118-27.
20. Fiksinski AM, Hoftman GD, Vorstman JAS, Bearden CE. A genetics-first approach to understanding autism and schizophrenia spectrum disorders: the 22q11.2 deletion syndrome. Mol Psychiatry. 2023;28(1):341-53.
21. Stessman HA, Bernier R, Eichler EE. A genotype-first approach to defining the subtypes of a complex disease. Cell. 2014;156(5):872-7.
22. Wilczewski CM, Obasohan J, Paschall JE, Zhang S, Singh S, Maxwell GL, et al. Genotype first: Clinical genomics research through a reverse phenotyping approach. Am J Hum Genet. 2023;110(1):3-12.
23. Neff RA, Wang M, Vatansever S, Guo L, Ming C, Wang Q, et al. Molecular subtyping of Alzheimer's disease using RNA sequencing data reveals novel mechanisms and targets. Sci Adv. 2021;7(2):eabb5398.
24. Udler MS, Kim J, von Grotthuss M, Bonas-Guarch S, Cole JB, Chiou J, et al. Type 2 diabetes genetic loci informed by multi-trait associations point to disease mechanisms and subtypes: A soft clustering analysis. PLoS Med. 2018;15(9):e1002654.
25. Arnedo J, Svrakic DM, Del Val C, Romero-Zaliz R, Hernandez-Cuervo H, Molecular Genetics of Schizophrenia C, et al. Uncovering the hidden risk architecture of the schizophrenias: confirmation in three independent genome-wide association studies. Am J Psychiatry. 2015;172(2):139-53.
26. Dahl A, Zaitlen N. Genetic Influences on Disease Subtypes. Annu Rev Genomics Hum Genet. 2020;21(1):413-35.
27. Rangan AV, McGrouther CC, Kelsoe J, Schork N, Stahl E, Zhu Q, et al. A loop-counting method for covariate-corrected low-rank biclustering of gene-expression and genome-wide association study data. PLoS Comput Biol. 2018;14(5):e1006105.
28. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med. 2015;12(3):e1001779.
29. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK Biobank resource with deep phenotyping and genomic data. Nature. 2018;562(7726):203-9.
30. Wilkinson T, Schnier C, Bush K, Rannikmae K, Henshall DE, Lerpiniere C, et al. Identifying dementia outcomes in UK Biobank: a validation study of primary care, hospital admissions and mortality data. Eur J Epidemiol. 2019;34(6):557-65.
31. Chen CY, Pollack S, Hunter DJ, Hirschhorn JN, Kraft P, Price AL. Improved ancestry inference using weights from external reference panels. Bioinformatics. 2013;29(11):1399-406.
32. The 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature. 2015;526(7571):68-74.
33. Das S, Forer L, Schonherr S, Sidore C, Locke AE, Kwong A, et al. Next-generation genotype imputation service and methods. Nat Genet. 2016;48(10):1284-7.
34. Chung JE, Magland JF, Barnett AH, Tolosa VM, Tooker AC, Lee KY, et al. A Fully Automated Approach to Spike Sorting. Neuron. 2017;95(6):1381-94 e6.
35. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015;4(1):7.

36. Raudvere U, Kolberg L, Kuzmin I, Arak T, Adler P, Peterson H, et al. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). Nucleic Acids Res. 2019;47(W1):W191-W8.
37. Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, et al. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. Innovation (Camb). 2021;2(3):100141.
38. Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS. 2012;16(5):284-7.
39. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000;25(1):25-9.
40. Gene Ontology Consortium. The Gene Ontology resource: enriching a GOld mine. Nucleic Acids Res. 2021;49(D1):D325-D34.
41. Benoit K, Watanabe K, Wang H, Nulty P, Obeng A, Müller S, et al. quanteda: An R package for the quantitative analysis of textual data. Journal of Open Source Software. 2018;3(30):774.
42. Zhou H, Lin W, Labra SR, Lipton SA, Schork NJ, Rangan AV. Detecting boolean asymmetric relationships with a loop counting technique and its implications for analyzing heterogeneity within gene expression datasets. bioRxiv. 2022:2022.08.04.502792.
43. Merico D, Isserlin R, Stueker O, Emili A, Bader GD. Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. PLoS One. 2010;5(11):e13984.
44. Donohue MC, Sperling RA, Petersen R, Sun CK, Weiner MW, Aisen PS, et al. Association Between Elevated Brain Amyloid and Subsequent Cognitive Decline Among Cognitively Normal Persons. JAMA. 2017;317(22):2305-16.
45. Donohue MC, Sperling RA, Salmon DP, Rentz DM, Raman R, Thomas RG, et al. The preclinical Alzheimer cognitive composite: measuring amyloid-related decline. JAMA Neurol. 2014;71(8):961-70.
46. Landau SM, Breault C, Joshi AD, Pontecorvo M, Mathis CA, Jagust WJ, et al. Amyloid-beta imaging with Pittsburgh compound B and florbetapir: comparing radiotracers and quantification methods. Journal of nuclear medicine : official publication, Society of Nuclear Medicine. 2013;54(1):70-7.
47. Landau SM, Marks SM, Mormino EC, Rabinovici GD, Oh H, O'Neil JP, et al. Association of lifetime cognitive engagement and low β-amyloid deposition. Archives of Neurology. 2012;69(5):623-9.
48. Shaw LM, Vanderstichele H, Knapik-Czajka M, Clark CM, Aisen PS, Petersen RC, et al. Cerebrospinal fluid biomarker signature in Alzheimer's disease neuroimaging initiative subjects. Ann Neurol. 2009;65(4):403-13.
49. Bates D, Machler M, Bolker BM, Walker SC. Fitting Linear Mixed-Effects Models Using lme4. J Stat Softw. 2015;67(1):1-48.
50. Kuznetsova A, Brockhoff PB, Christensen RHB. lmerTest Package: Tests in Linear Mixed Effects Models. J Stat Softw. 2017;82(13):1-26.
51. Sims R, Hill M, Williams J. The multiplex model of the genetics of Alzheimer's disease. Nat Neurosci. 2020;23(3):311-22.
52. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006;38(8):904-9.
53. Berg JJ, Harpak A, Sinnott-Armstrong N, Joergensen AM, Mostafavi H, Field Y, et al. Reduced signal for polygenic adaptation of height in UK Biobank. Elife. 2019;8.
54. Haworth S, Mitchell R, Corbin L, Wade KH, Dudding T, Budu-Aggrey A, et al. Apparent latent structure within the UK Biobank sample has implications for epidemiological analysis. Nat Commun. 2019;10(1):333.
55. Tzioras M, McGeachan RI, Durrant CS, Spires-Jones TL. Synaptic degeneration in Alzheimer disease. Nature reviews Neurology. 2023;19(1):19-38.

56. Spires-Jones Tara L, Hyman Bradley T. The Intersection of Amyloid Beta and Tau at Synapses in Alzheimer’s Disease. Neuron. 2014;82(4):756-71.
57. de Calignon A, Polydoro M, Suarez-Calvet M, William C, Adamowicz DH, Kopeikina KJ, et al. Propagation of tau pathology in a model of early Alzheimer's disease. Neuron. 2012;73(4):685-97.
58. Ng B, Rowland HA, Wei T, Arunasalam K, Hayes EM, Koychev I, et al. Neurons derived from individual early Alzheimer's disease patients reflect their clinical vulnerability. Brain Commun. 2022;4(6):fcac267.
59. Munoz L, Ammit AJ. Targeting p38 MAPK pathway for the treatment of Alzheimer's disease. Neuropharmacology. 2010;58(3):561-8.
60. Leyns CEG, Holtzman DM. Glial contributions to neurodegeneration in tauopathies. Mol Neurodegener. 2017;12(1):50.
61. Tooyama I, Kimura H, Akiyama H, McGeer PL. Reactive microglia express class I and class II major histocompatibility complex antigens in Alzheimer's disease. Brain Res. 1990;523(2):273-80.
62. Nobes CD, Hall A. Rho GTPases control polarity, protrusion, and adhesion during cell movement. J Cell Biol. 1999;144(6):1235-44.
63. Settleman J. Rho GTPases in Development. In: Jeanteur P, editor. Cytoskeleton and Small G Proteins. Berlin, Heidelberg: Springer Berlin Heidelberg; 1999. p. 201-29.
64. Hotulainen P, Hoogenraad CC. Actin in dendritic spines: connecting dynamics to function. J Cell Biol. 2010;189(4):619-29.
65. Cabrales Fontela Y, Kadavath H, Biernat J, Riedel D, Mandelkow E, Zweckstetter M. Multivalent cross-linking of actin filaments and microtubules through the microtubule-associated protein Tau. Nat Commun. 2017;8(1):1981.
66. Valenta T, Hausmann G, Basler K. The many faces and functions of beta-catenin. EMBO J. 2012;31(12):2714-36.
67. Soldano A, Hassan BA. Beyond pathology: APP, brain development and Alzheimer's disease. Curr Opin Neurobiol. 2014;27:61-7.
68. Kowall NW, Kosik KS. Axonal disruption and aberrant localization of tau protein characterize the neuropil pathology of Alzheimer's disease. Ann Neurol. 1987;22(5):639-43.
69. Iqbal K, Liu F, Gong CX, Grundke-Iqbal I. Tau in Alzheimer disease and related tauopathies. Curr Alzheimer Res. 2010;7(8):656-64.
70. Schork NJ, Elman JA. Pathway-specific polygenic risk scores correlate with clinical status and Alzheimer's-related biomarkers. Res Sq. 2023.
71. La Joie R, Visani AV, Baker SL, Brown JA, Bourakova V, Cha J, et al. Prospective longitudinal atrophy in Alzheimer's disease correlates with the intensity and topography of baseline tau-PET. Sci Transl Med. 2020;12(524):eaau5732.
72. Hanseeuw BJ, Betensky RA, Jacobs HIL, Schultz AP, Sepulcre J, Becker JA, et al. Association of Amyloid and Tau With Cognition in Preclinical Alzheimer Disease: A Longitudinal Study. JAMA Neurol. 2019;76(8):915-24.
73. Ossenkoppele R, Lyoo CH, Sudre CH, van Westen D, Cho H, Ryu YH, et al. Distinct tau PET patterns in atrophy-defined subtypes of Alzheimer's disease. Alzheimer's & Dementia. 2019.
74. Leng F, Edison P. Neuroinflammation and microglial activation in Alzheimer disease: where do we go from here? Nature reviews Neurology. 2021;17(3):157-72.
75. Heneka MT, O'Banion MK. Inflammatory processes in Alzheimer's disease. J Neuroimmunol. 2007;184(1-2):69-91.
76. Porcu E, Sadler MC, Lepik K, Auwerx C, Wood AR, Weihs A, et al. Differentially expressed genes reflect disease-induced rather than disease-causing changes in the transcriptome. Nat Commun. 2021;12(1):5647.

77. Saporta AS, Sottile SL, Miller LJ, Feely SM, Siskind CE, Shy ME. Charcot-Marie-Tooth disease subtypes and genetic testing strategies. Ann Neurol. 2011;69(1):22-33.
78. Crutch SJ, Schott JM, Rabinovici GD, Murray M, Snowden JS, van der Flier WM, et al. Consensus classification of posterior cortical atrophy. Alzheimers Dement. 2017;13(8):870-84.
79. Gorno-Tempini ML, Hillis AE, Weintraub S, Kertesz A, Mendez M, Cappa SF, et al. Classification of primary progressive aphasia and its variants. Neurology. 2011;76(11):1006-14.
80. Townley RA, Graff-Radford J, Mantyh WG, Botha H, Polsinelli AJ, Przybelski SA, et al. Progressive dysexecutive syndrome due to Alzheimer's disease: a description of 55 cases and comparison to other phenotypes. Brain Commun. 2020;2(1):fcaa068.
81. Ossenkoppele R, Singleton EH, Groot C, Dijkstra AA, Eikelboom WS, Seeley WW, et al. Research Criteria for the Behavioral Variant of Alzheimer Disease: A Systematic Review and Meta-analysis. JAMA Neurol. 2022;79(1):48-60.

TABLES

Table 1. Sample characteristics of UK Biobank dataset.

	Overall	CU	AD
n	8217	5478	2739
Sex, n male (%)	3993 (48.6)	2662 (48.6)	1331 (48.6)
Age, years (SD)	64.74 (4.21)	64.74 (4.21)	64.74 (4.21)
Education, years (SD)	11.95 (4.96)	12.21 (5.01)	11.41 (4.80)
APOE-e4 alleles, n (%)			
0	5082 (61.8)	4063 (74.2)	1019 (37.2)
1	2647 (32.2)	1319 (24.1)	1328 (48.5)
2	488 (5.9)	96 (1.8)	392 (14.3)
APOE-e4 carriers, n (%)	3135 (38.2)	1415 (25.8)	1720 (62.8)

CU = Cognitively unimpaired, AD = Alzheimer's disease dementia

Table 2. Sample characteristics of ADNI dataset. The Alzheimer's disease (AD) and cognitively unimpaired (CU) groups were included in validation analyses of genetic heterogeneity found in the UK Biobank. The mild cognitive impairment (MCI) group was included in analyses of cognitive and biomarker trajectories.

	Overall	CN	Dementia	MCI
n	1369	470	500	399
Sex, n male (%)	765 (55.9)	225 (47.9)	297 (59.4)	243 (60.9)
Age, years (SD)	73.70 (7.11)	72.47 (6.31)	74.61 (7.43)	74.00 (7.38)
Education, years (SD)	16.01 (2.76)	16.67 (2.44)	15.51 (2.89)	15.87 (2.79)
APOE-e4 alleles, n (%)				
0	750 (54.8)	334 (71.1)	174 (34.8)	242 (60.7)
1	493 (36.0)	122 (26.0)	245 (49.0)	126 (31.6)
2	126 (9.2)	14 (3.0)	81 (16.2)	31 (7.8)
APOE-e4 carriers, n (%)	619 (45.2)	136 (28.9)	326 (65.2)	157 (39.3)

FIGURES

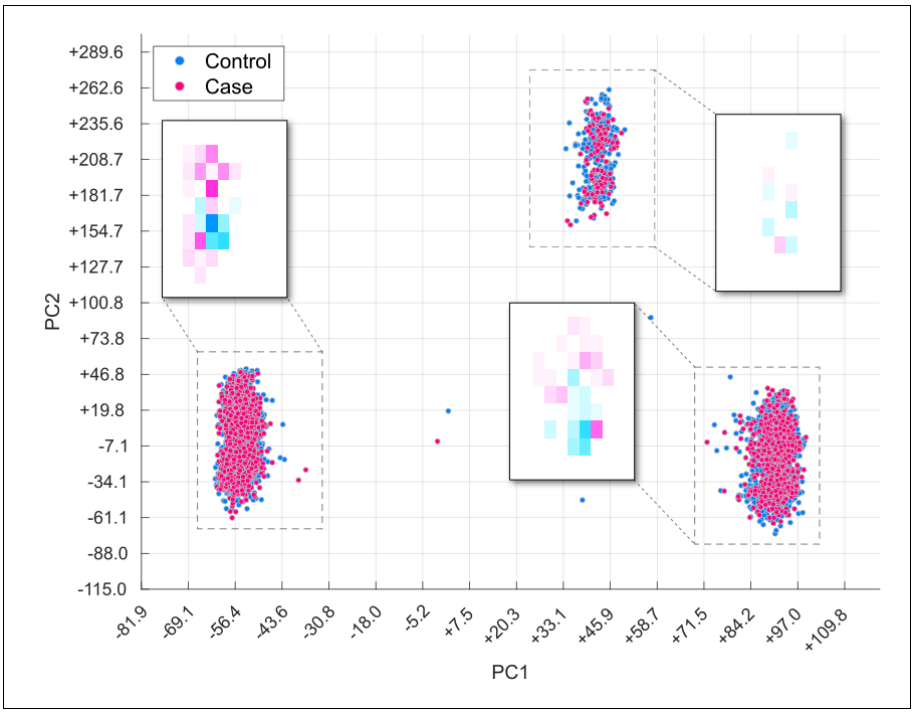

Figure 1. Principal component analysis of UKB data restricted to Alzheimer's disease-associated variants reveals three constellations. Principal component analysis was applied to allele combinations of UK Biobank cases and controls. Variants were restricted to those with a p-value<0.05 in the Kunkle et al. Alzheimer's GWAS (12). The scatter plot displays participant loading on the first two principal components (PC1 and PC2) and colored by Alzheimer's disease case-control status. Three distinct clusters, or constellations, are clearly present and each contains a mix of cases and controls. Heatmaps displaying the density of cases and controls in each constellation are also shown to demonstrate that, despite substantial overlap between the groups, there is some offset in the distributions. However, the directional bias is not consistent across constellations.

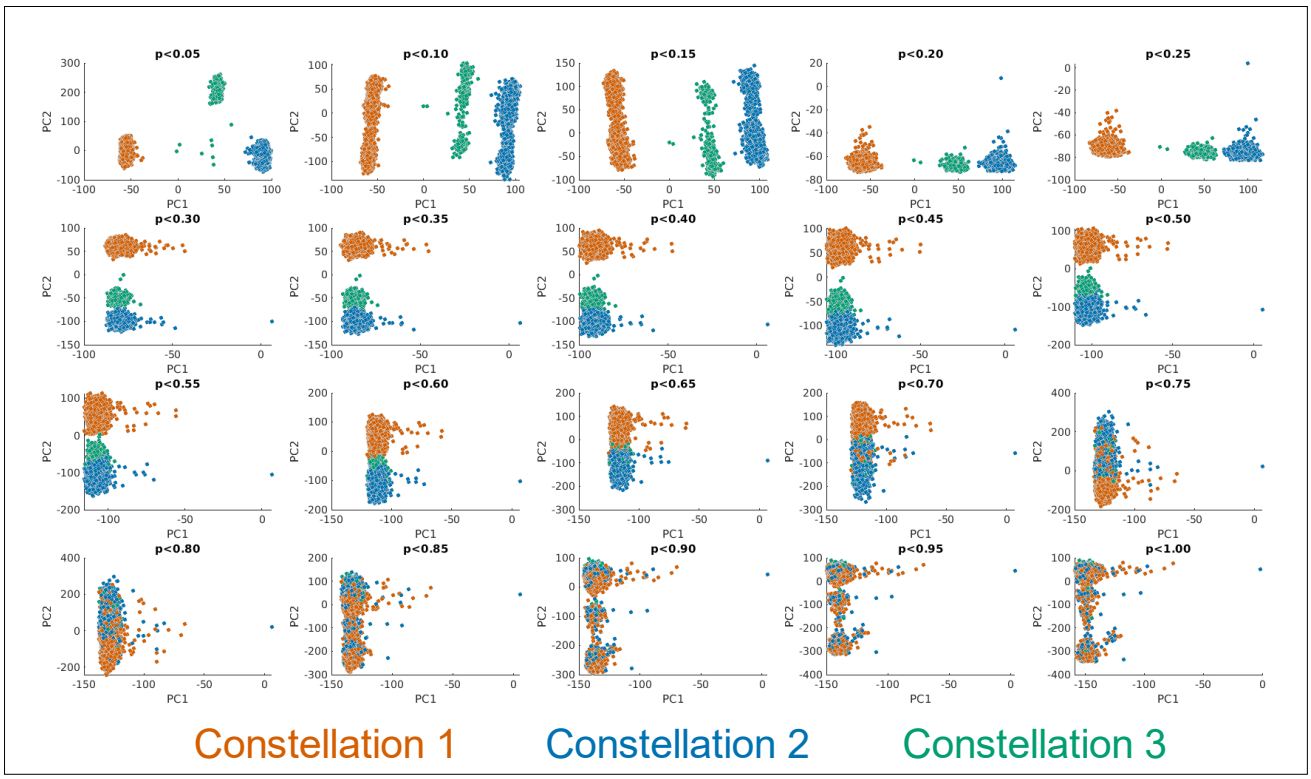

Figure 2. Principal component analysis across varying thresholds of Alzheimer's associated variants. Principal component analysis was applied to allele combinations of UK Biobank cases and controls. Variants were restricted across a range of p-value thresholds from the Kunkle et al. Alzheimer's GWAS (12). The scatter plots display participant loadings on the first two principal components (PC1 and PC2) and colored by constellation labels defined at the p<0.05 threshold. Participants from all three constellations are highly mixed when all variants (p<1.0) are included. The constellation structure begins to emerge along the first principal component at p<0.25, and further separate along the second principal component at p<0.05. The emergence of structure only at smaller p-values suggests these constellations are disease-relevant.

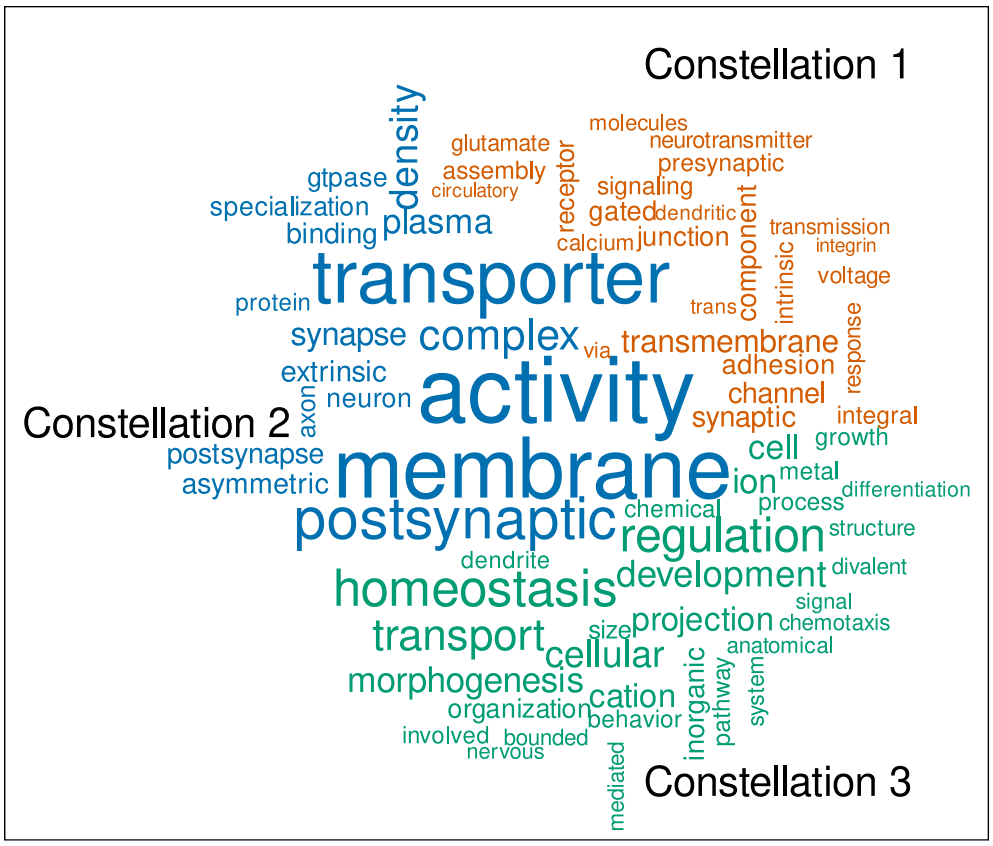

Figure 3. Comparison cloud of gene sets associated with disease-relevant constellations. Visualization of term frequency from descriptions of gene sets associated with each constellation. Variants associated with each constellation were mapped to genes and gene set enrichment was applied to each constellation using GO gene sets. Word size represents the deviation of frequency in each constellation compared to the average frequency across constellations.

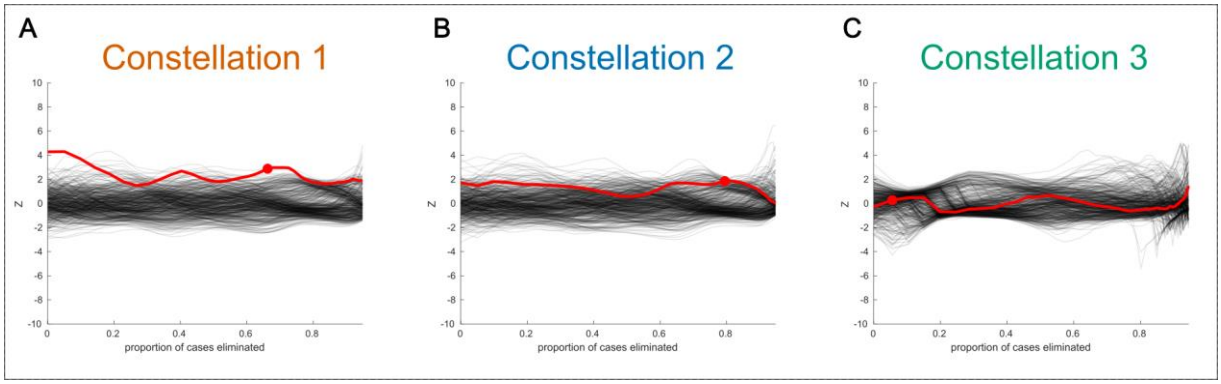

Figure 4. Bicluster traces of observed data versus label-shuffled data. The disease-related signal-strength associated with the remaining UKB Alzheimer's cases relative to controls is plotted on the y-axis. At each iteration, allele combinations and cases that contribute least to this difference are removed. The proportion of remaining cases is shown on the x-axis. The red trace represents the original data and black traces represent label-shuffled data, corresponding to a null distribution. A red dot indicates the iteration with the maximum separation between cases and controls (ignoring signal in the first iterations), and is used to define the bicluster. The sharper peak of constellation 1 indicates that this bicluster has more distinct boundaries, whereas the bicluster in constellation 2 has "fuzzier" boundaries as indicated by the broad yet lower peak.

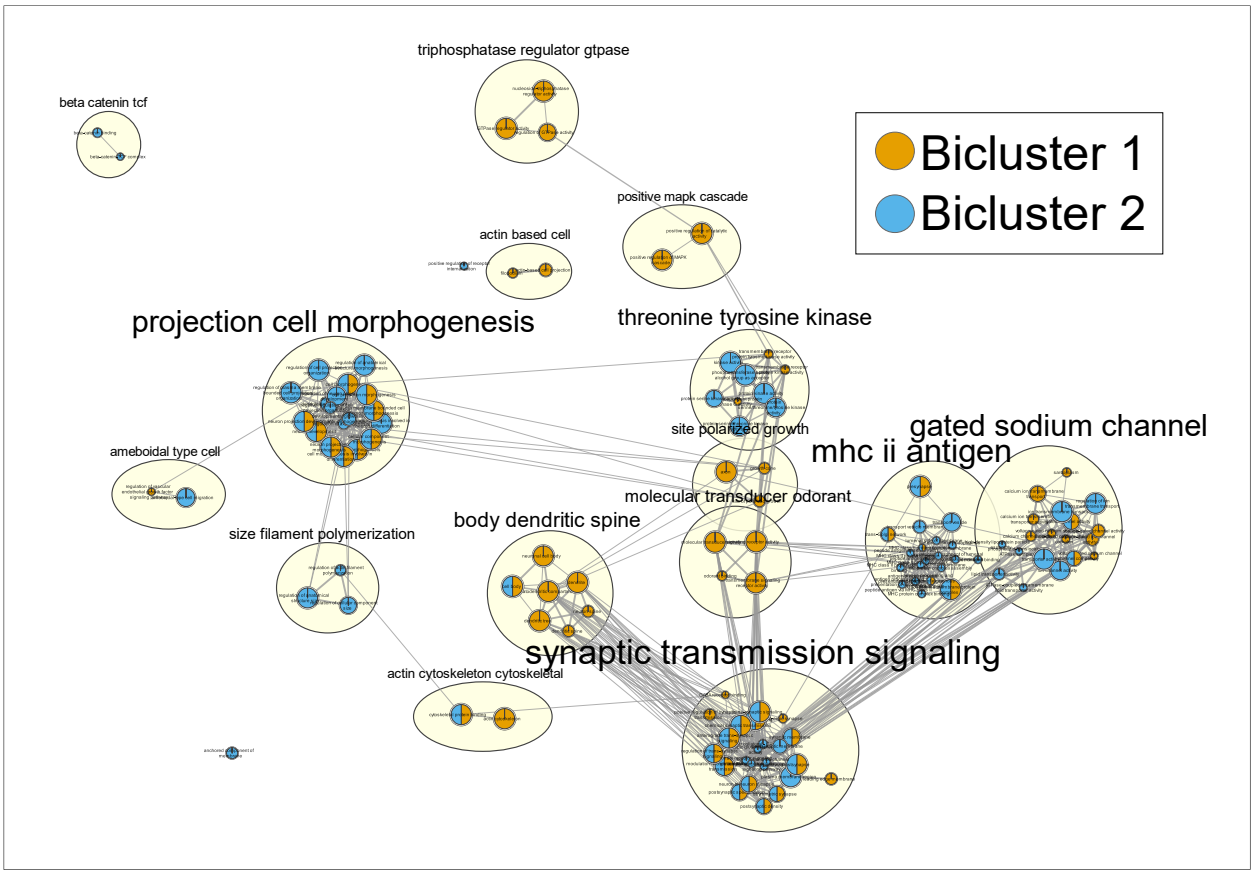

Figure 5. Network plot of gene sets enriched among bicluster genes. Network visualization of gene set enrichment results for biclusters 1 and 2. Nodes represent significantly enriched gene sets with color indicating the bicluster they are associated with. Edges represent the overlap in genes belonging to gene sets using a threshold of 0.5. Gene sets were clustered based on overlaps and automatically annotated based on the descriptions of each gene set cluster.

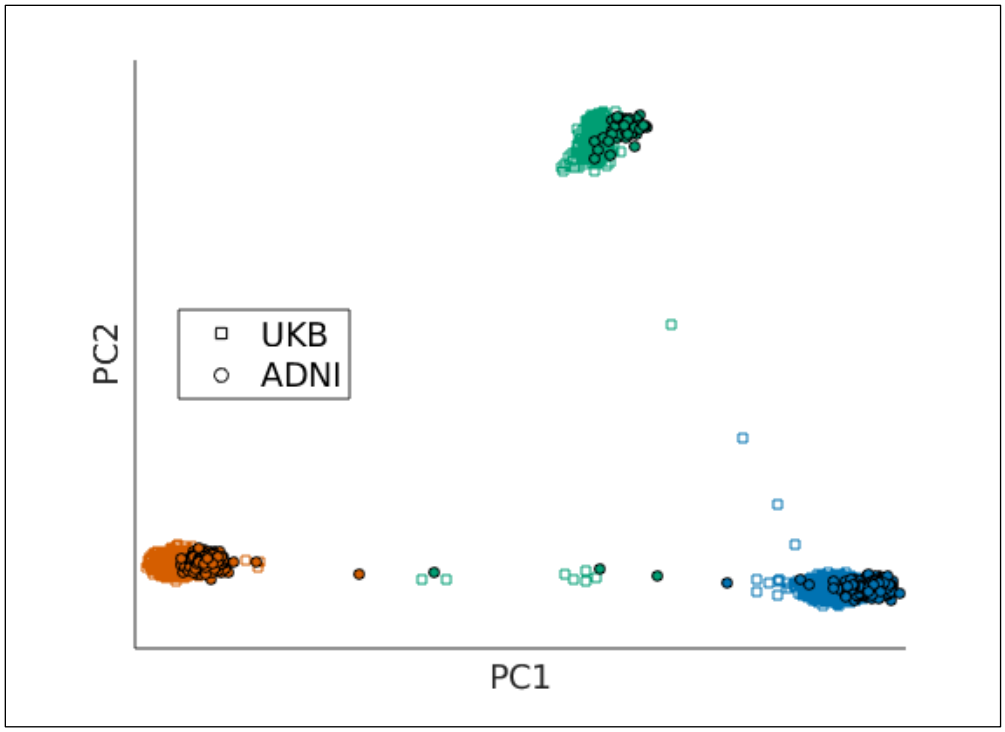

Figure 6. ADNI data projected along principal components defined in UKB data replicate distribution of disease-relevant constellations. Principal components were recalculated in UKB data using only AD-associated variants common to both datasets. Participants from both datasets were then plotted by participants loadings on the first two principal components. Colors represent constellations, UKB participants are plotted with squares, and ADNI participants are plotted with circles.

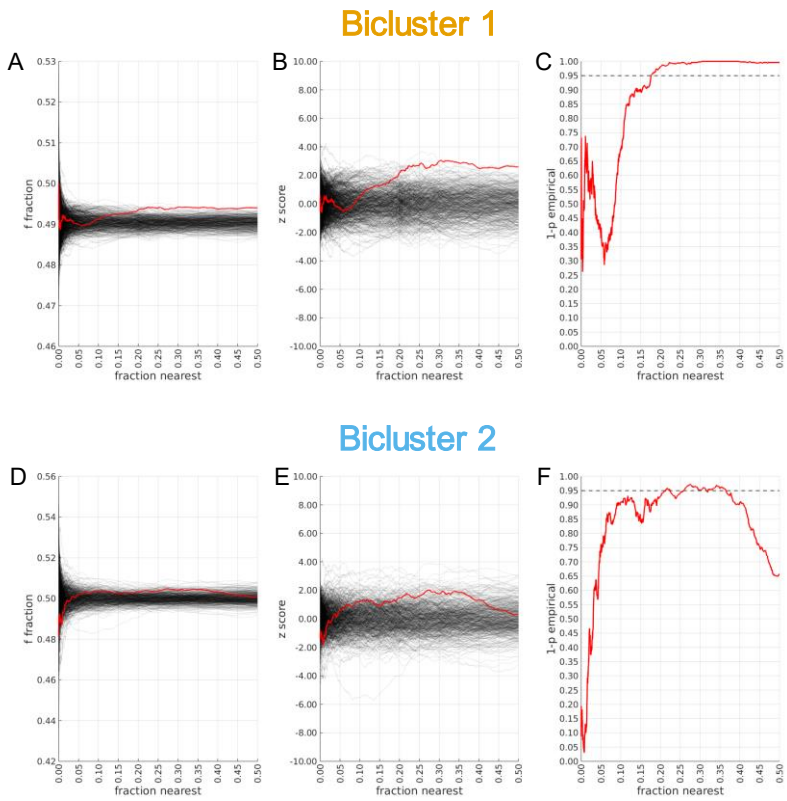

Figure 7. Replication of disease-specific biclusters in ADNI data. The similarity between case-control labels of individuals in the test set (ADNI) and the most frequent label among nearest neighbors in the training set (UKB) was used to assess replication of biclusters. Only individuals belonging to the constellation in which the given bicluster was found were considered. The fraction of individuals from the training set (i.e., UKB) considered as nearest neighbors is plotted along the x-axis. The y-axis in Panels (**A**) and (**D**) shows the average fraction of nearest neighbors with a matching label. The red line in Panels (**A**) and (**D**) shows values from the original data, while the black lines show values from label-shuffled trials drawn from the null-distribution. A trial-wise mean and variance can be defined from the null distribution to normalize values, with the associated z-scores shown in Panels (**B**) and (**E**). A rank-normalization of the scores is shown in Panels (**C**) and (**F**), from which can be read the empirical p-value for the label-matching. Values that lie above the 95th percentile – as indicated by the dashed line – would correspond to values of $p<0.05$ for that fraction of nearest neighbors.

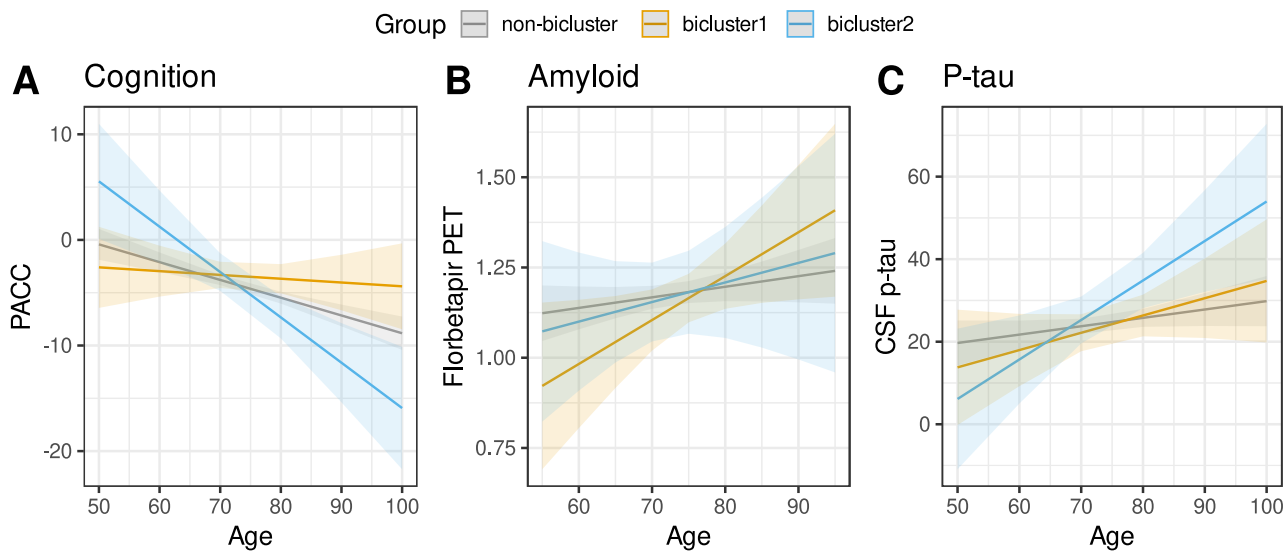

Figure 8. Association of bicluster groups in ADNI MCI sample. ADNI Individuals diagnosed with MCI were assigned to bicluster groups (bicluster 1, bicluster 2, or non-bicluster). Differences in cognitive and biomarker trajectories over time were between groups were tested with age x group interactions in linear mixed effects models. Model predicted values are shown in the figure. A) Cognition was measured with the Preclinical Alzheimer's Cognitive Composite (PACC). B) Amyloid was measures using florbetapir PET. C) Phosphorylated tau (p-tau) was measured using CSF p-tau.