Title: Exploring the genetic heterogeneity of Alzheimer's disease: Evidence for disease-relevant and disease-specific clustering

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

Affiliations:

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

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

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

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

**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*

One Sentence Summary: This study identifies a hierarchical structure of heterogeneity in Alzheimer's disease genetic risk, shedding light on its multifactorial etiology.

Abstract:

Alzheimer's disease (AD) has a complex and multifactorial etiology. It is under significant genetic influence, thus identifying systematic variation in genetic risk represents a potentially useful approach to understanding diverse origins of the disease. Here, we explore genetic heterogeneity underlying AD using a multi-step approach. First, a principal component analysis was applied to AD-associated variants across 2,739 AD cases and 5,478 age and sex-matched controls from the UK Biobank. Three distinct clusters (termed "constellations") were found, each containing a mix of cases and controls. This structure only emerged when restricting analysis to AD-associated variants and therefore appears to be disease-relevant. Next, we applied a recently developed biclustering algorithm that searches for subsets of AD cases and variants that constitute distinct risk groups. We found two significant biclusters, each manifesting disease-specific genetic signatures that increase risk for AD. This pattern of clustering replicated in an independent dataset from the Alzheimer's Disease Neuroimaging Initiative (ADNI). These findings reveal a hierarchy of underlying AD genetic risk. At the first level, disease-relevant constellations may represent differential vulnerability of specific biological systems or pathways that contribute to disease development but are not sufficient to increase disease risk on their own and likely require additional risk factors. At the next level, biclusters may represent disease subtypes, or groups of AD cases with unique combinations of genetic variation that increase their risk of developing AD. More broadly, this study illustrates an approach that can be extended to investigations of genetic heterogeneity underlying other complex diseases.

Main text:

INTRODUCTION

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.

RESULTS

Identification of disease-relevant structure

Genotype data from the UK Biobank (UKB) collected on 2,739 Alzheimer's disease cases and 5,478 age- and sex-matched controls were used as a discovery or training sample. 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*). We considered SNPs with p<0.05 to accommodate potential association signals beyond those exhibiting genome-wide significance. An initial principal components (PC) analysis (PCA) was applied to the allele combinations of all individuals. Examining the first 2 PCs revealed three distinct clusters (hereafter referred to as "constellations" to disambiguate from the term "bicluster" used in subsequent steps), each containing a mixture of AD cases and controls (**Fig. 1**). Although the sample was restricted to White British participants, it is possible that this reflects remaining ancestry-related population substructure. Therefore, we applied PCA restricted to subsets of SNPs across a range of p-value thresholds from the AD GWAS. The plots in **Fig. 2** show that when all variants are included (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 components emerges around $p<0.25$, and the constellations further separate along the second PC at $p<0.05$. To confirm that the structure is not simply a function of the number of variants included, we also applied a PCA to a random set of SNPs of the same size as were included at the $p<0.05$ level. Here again the individuals from each constellation were interspersed, and the overall structure closely mirrored that seen when all SNPs were included (**Supplemental Fig. S1**). 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.

We sought to identify biological functions associated with each disease-relevant constellation by examining gene set enrichment. First, a GWAS tested the relationship of all AD-associated SNPs with membership in each constellation versus the other two. Next, we mapped variants associated with membership in each constellation to genes and performed a gene set enrichment analysis of Gene Ontology terms. Themes more frequently associated with each constellation were visualized with a comparison cloud, as depicted in **Fig. 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 Fig. S2**). 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 Fig. S3**). 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.

Search for 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 **Fig. 1**, it appears that the bias or shift in cases relative to controls does not fall in the same direction across constellations. To further investigate this possibility, we searched for disease-specific heterogeneity within each constellation. To be more specific, we searched within each constellation for subsets of AD-patients that exhibit a genetic signature across a subset of SNPs which is *not* shared by the controls (i.e., is unique among individuals with AD). Each such subset of AD cases, along with the SNPs participating in their disease-specific signature, constitutes what has been referred to as a “bicluster:” a subset of

individuals from a larger sample of individuals that share a subset of features from a larger feature space.

There are many kinds of biclusters, and different methods one can use to search for them. Here, we use an approach introduced by Rangan et al. (*27*) which is specifically designed to search for disease-specific biclusters within GWAS data and can correct for often-used covariates such as (putatively ancestry-related) principal components extracted from a Genetic Relationship Matrix (GRM). As previously described in detail by Rangan et al. (*27*), and summarized in the Methods section, this strategy employs a modified version of spectral biclustering using a permutation-test to assess statistical significance. We use the half-loop method, which involves an iterative process that starts with all the participants and SNPs, and then sequentially removes AD cases and SNP variant-combinations from consideration. At each step of the iteration, the average level of difference between SNP variant profiles of the remaining AD cases and the controls is recorded. 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.

The results of our bicluster searches are shown in **Fig. 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 below. 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

particular 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 **Fig. 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 **Fig. 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 Fig. S4**). 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.

To characterize variants associated with – or defining – each bicluster, we performed gene set enrichment analysis. First, we performed a GWAS comparing bicluster-cases to controls that belong to the same constellation in which a bicluster was found (i.e., constellations 1 and 2). We then mapped variants associated with membership in each bicluster to genes and performed gene set enrichment analysis using GO terms. 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. **Fig. 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 **Fig. 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.

We next investigated whether the disease-relevant and disease-specific heterogeneity found in the UKB would generalize to another dataset. To this end, we examined 476 Alzheimer's disease cases and 576 controls from the Alzheimer's Disease Neuroimaging Initiative (ADNI). 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. Therefore, to the extent that these factors influence the structure identified by our methods, they should have differential effects across datasets.

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 **Fig. 4**, 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

After assigning individuals to their nearest constellation, we assessed replicability of the two biclusters. As when searching for biclusters, participants from each constellation were examined separately. 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 SNP loadings were used to project UKB and ADNI participants onto this bicluster space. 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, but 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

can assess 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. An empirical p-value is calculated across a range of nearest neighbors. This approach allows us to determine what fraction of nearest neighbors from the training set one should average over to approximate the true label of a point in the test-set.

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 10.5% of the sample as nearest neighbors for bicluster 1 ($p<0.002$) and using 30% of the sample as nearest neighbors for bicluster 2 ($p=0.004$). **Fig. 5** 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 8.5% 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 20% to 40%. Importantly, these plots provide a range of sensible values one can use to label new data.

DISCUSSION

Genetic analyses suggest the etiology of AD is multifactorial (*28*), 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 (*29*). Despite the fact that the UK Biobank sample is relatively homogeneous when compared to other cohorts of individuals with European ancestry (*30*), there may still be residual ancestry-related stratification in the UKB data sets (*31*). 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 (*32-34*). 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 (*35*) 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 (*36*). 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, 37, 38*). 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 (*39-43*). 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 (*44*). 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 (*45, 46*). 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.

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 (*47, 48*). 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 (*49*). 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 (*50*). 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 (*51-54*) 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.

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 and thus require different treatment strategies to optimize effectiveness. We also believe that, more broadly, the approach presented here may be extended to elucidate the etiological heterogeneity of other complex diseases.

MATERIALS AND METHODS

Study design

The goal of this study was to explore etiological heterogeneity of Alzheimer's disease (AD) by investigating systematic variation in genetic risk for AD. We hypothesized that there would be subsets of individuals exhibiting unique combinations of genetic variants related to increased likelihood of having an AD clinical diagnosis. To investigate this, we used imputed genotyping data from the UK Biobank (UKB) as the discovery dataset. The UKB is a large-scale biomedical database and research resource containing genetic, lifestyle and health information from half a million UK participants (*55, 56*). Data from 2,739 Alzheimer's disease cases (*57*) and 5,478 age- and sex-matched controls with White British ancestry as determined by PCA (*56*) were included in these analyses. 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.

A replication dataset was obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). 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 analyses included genotyping data from 476 individuals with Alzheimer's disease cases and 576 controls from the

ADNI-1 (n=561), ADNI-GO/2 (n=253), and ADNI-3 (n=238). Diagnoses were assigned at the baseline visit. 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). Participants were restricted to those with primarily European ancestry (>80%) as determined by SNPweights (*58*). 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 (*59*) for use as covariates in later analyses. Following standard genotyping QC, imputation was performed on the Michigan Imputation Server (https://imputationserver.sph.umich.edu/) (*60*) 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.

Defining disease-relevant constellations

Principal component analysis (PCA) was applied to different sets of variants (coded according to the 3 possible allele combinations for each) possessed by UKB AD cases and controls. Variants were restricted to those with a p-value<0.05 from the Kunkle et al. GWAS of AD (i.e., "AD-associated variants"). Individuals were assigned to clusters by applying the ISO-SPLIT algorithm (*61*) implemented in MATLAB to participant loadings on the first principal component. We determined whether the resulting structure was specific to the included 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 $p<0.05$. Results were visualized as described above.

Gene set enrichment analysis of constellations

We conducted a GWAS across AD-associated variants comparing individuals in each constellation against those from the other two using PLINK2 (*62*). 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 used to characterize constellations. These variants were mapped to genes using the *g:Profiler* package in R (*63*). A gene set enrichment analysis was then applied to annotated genes from each constellation using the *clusterProfiler* package in R (*64, 65*). The analysis was restricted to gene sets from the Gene Ontology resource (*66, 67*) 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 (68)*. 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.

Searching for disease-specific biclusters

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. Briefly, we can take an $M_D \times N$ matrix of cases D and an $M_X \times N$ matrix of controls X. The entry $D_{j,k}$ records the presence or absence of the k-th allele-combination for the jth case, and the entry $X_{j,k}$ records the presence or absence of the kth allele-combination for the jth control. For each case j we can measure the fraction of cases (i.e., restricted to matrix D) that share the allele-combination k and call this $\left[D \leftrightarrow D\right]_{j,k}$. Similarly, we can calculate the fraction of controls that share the allele-combination k and call this $\left[D \leftrightarrow D\right]_{j,k}$ (i.e., extending into matrix X). We can then calculate $Z_{j,k} = \left[D \leftrightarrow D\right]_{j,k} - \left[D \leftrightarrow X\right]_{j,k}$, which represents the difference in fraction of shared allele-combinations between the jth case and other remaining cases compared to the fraction shared with controls for the kth allele-combination. Averaging over the remaining allele-combinations will give us Z_j, referred to as row-scores in our algorithm. Finally, averaging over the remaining cases will leave us with Z, or the row trace, which indicates the disease-related signal strength for the remaining subset of AD cases and allele-combinations. At each iteration the cases and allele-combinations with the smallest contributions to Z are removed and the row trace is recalculated as described above.

This process is conducted while controlling for the first two genetic principal-components. Covariate correction is also described in detail in Rangan et al. (*27*), but essentially weights the contributions of cases and allele-combinations such that the row score and row trace (as well as analogous column-scores and column trace) will be shrunk if drawn from an imbalanced distribution of the covariate space. That is, the biclustering algorithm will tend to ignore these structures in favor of structures that are more evenly distributed in covariate space.

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. We can draw samples from this null-hypothesis by randomly permuting the case- and control-labels across subjects with similar genome-wide principal components (see Rangan et al. (*27*) for details). 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 Z 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 (as shown in **Fig. 3A**), 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 (as shown in **Fig. 3B**), 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. Moreover, we can use the peak of the original 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.

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, 69*).

Gene set enrichment analysis of biclusters

We conducted a GWAS across AD-associated SNPs comparing bicluster cases with controls belonging to the same disease-relevant constellation using PLINK2 (*62*). 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 (*63*). A gene set enrichment analysis was then applied to annotated genes from each constellation using the *clusterProfiler* package in R (*64, 65*). The analysis was restricted to gene sets from the Gene Ontology resource (*66, 67*) containing between 10 and 1,000 genes. The Cytoscape app EnrichmentMap v3.3 (*70*) 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.

Replication of disease-relevant constellations in ADNI

We assessed the extent to which the two levels of structure found in the UKB data were also present in the ADNI data. The first level of the hierarchical structure observed in the UKB data involved the constellations shown in **Fig. 1**. To assess the replication of these structures

within the ADNI data, we determined the subset of SNPs which lie in the intersection of the UKB and ADNI datasets, and then further restricted this subset of SNPs to include only those SNPs which had Kunkle p-values of <0.05, denoting this restricted subset by $S_{\cap}$. Mirroring the original analysis, we then re-calculated the dominant 2 principal-components for the UKB data across $S_{\cap}$ to obtain loadings u_1^{UKB} and u_2^{UKB} or each participant, and loadings $v1$ and $v2$ for each SNP. Next, we use $v1$ and $v2$ to calculate projections u_1^{ADNI} and u_2^{ADNI} for each ADNI participant. We can then compare the distribution of u_1^{UKB} and u_2^{UKB} with the distribution of u_1^{ADNI} and u_2^{ADNI} to assess overlap between datasets as shown in **Fig. 4.** ADNI participants were assigned to the nearest constellation based on participant loadings on the first principal component.

Replication of disease-specific biclusters in ADNI

After participants were classified into constellations, we assessed replication of the dominant bicluster separately in the two constellations in which they were found. First, we restrict the subset of SNPs delineating the bicluster to those that overlap with the ADNI data, referring to that (restricted) subset as $\mathcal{K}$, which is itself a subset of $S_{\cap}$. We can organize the data into a $|\mathcal{J}| \times |\mathcal{K}|$ array A^{UKB}, where $\mathcal{J}$ represents only UKB participants from a given constellation. Likewise, the ADNI data can be organized into a $|\mathcal{J}'| \times |\mathcal{K}|$ array where $\mathcal{J}'$ represents the ADNI participants that have been assigned to the constellation of interest. The bicluster itself, denoted by B, is a $|\mathcal{J}_B| \times |\mathcal{K}|$ submatrix of A^{UKB}. We then calculate the dominant SNP-wise principal components v_1 and v_2 of B across $\mathcal{K}$. We use SNP loadings v_1 and v_2 to calculate projections for each of the $\mathcal{J}$ UKB participants in the constellation and defined u_1^{UKB} and u_2^{UKB} as the first two columns (i.e., participant-wise principal components).

Similarly, we can calculate u_1^{ADNI} and u_2^{ADNI} for each of the $\mathcal{J}'$ ADNI participants belonging to the constellation.

After projecting UKB and ADNI participants onto the dominant principal components of $\mathcal{K}$ we must first align the distributions of u_1^{ADNI} and u_2^{ADNI} with u_1^{UKB} and u_2^{UKB}. To do so, we use a point-set registration algorithm described in the section "*Affine point-set registration*" to calculate an affine transform that can be applied to perform this alignment.

Following alignment of the two distributions, we use the methods described in the section "*Calculating label similarity*" to calculate case-control label similarity of participants in the ADNI dataset to nearby participants in the UKB dataset, $\tilde{\sigma}(f;\{y_j\},\{l_j\},\{y'_{j'}\},\{l'_{j'}\})$. In this expression, y_j and l_j refer to the projected data (u_1^{UKB}, u_2^{UKB}) and case-control label for UKB participant j, respectively. The terms $y'_{j'}$ and $l'_{j'}$ refer to the projected and aligned data (u_1^{ADNI}, u_2^{ADNI}) and case-control label for ADNI participant j', respectively. The value f represents the fraction of individuals from the training set (i.e., UKB data) one should use as nearest neighbors when averaging to approximate the true label of a point in the testing set (e.g., ADNI data).

We can assess the statistical significance of the similarity measure $\tilde{\sigma}\left(f;\{y_j\},\{l_j\},\{y'_{j'}\},\{l'_{j'}\}\right)$ with a permutation test. The null-hypothesis is that the labels are independent from the position of each point in projected space. Thus, we can draw a sample from the null-hypothesis by (i) randomly permuting the labels l_j within the UKB data, (ii) randomly permuting the labels $l'_{j'}$ within the ADNI data, and then (iii) recalculating the above steps for this label-shuffled data, including the affine point-set registration. Here, we ran 500 permutations. After accumulating a distribution of values for $\tilde{\sigma}(f)$ under the null-hypothesis, we can compare the original $\tilde{\sigma}(f)$ to this distribution. This allows us to examine the statistical significance across

a range of choices for f. This approach can also be used to label participants in a new dataset. The range over which f remains highly significant provides a set of reasonable values to use when constructing these labels.

Affine point-set registration

Prior to calculating label similarity between datasets, we perform affine point set registration. In the scenario where the dimension of the data (i.e., number of SNPs) is much larger than the number of samples, the principal components may be overfit to the training dataset. That is, the principal components are calculated to maximize the variance in the training set but not the test set. When projecting the test set data onto the principal components of the training set, the participant data from the test set will almost certainly have a smaller amplitude than the participant data in the training set. Technical artifacts, such as might be introduced by different genotyping chips across studies, may also lead to shifts or distortions in the data distributions. Given two collections of points, say in $\mathbb{R}^2$, we would therefore like to construct a transformation $T : \mathbb{R}^2 \to \mathbb{R}^2$ which attempts to align one set of points to the other. The strategy we employ (described below) is designed to be part of our bicluster replication pipeline. Thus, we restrict ourselves to affine linear transformations, with the goal of registering the projected values of one dataset to the projected values of another, while controlling for possible differences in the number of cases and controls between these two datasets. While our algorithm may not be 'optimal', it has certain advantages. Specifically, our algorithm is 'automatic', and there are no free parameters which need to be tuned.

To describe our algorithm, we first introduce some notation. Consider a collection of J vectors $\{u_1, \dots, u_J\}$, each in $\mathbb{R}^2$. These u_j are each endowed with a label $l_j \in \{1, \dots, L\}$,

indicating which category they are in – cases or controls. Also consider a second collection of J' vectors $\{u'_1, \dots, u'_{J'}\}$, also in $\mathbb{R}^2$. The $u'_{j'}$ are also endowed with labels $l'_j \in \{1, \dots, L\}$. Given this notation, our goal will be to define an affine transformation $T : \mathbb{R}^2 \to \mathbb{R}^2$ which aligns the $\{u'_{j'}\}$ to the $\{u_J\}$ while correcting for the labels $\{l'_{j'}\}$ and $\{l_j\}$. This reduces the influence that the proportion of cases and controls can have on the transformation.

To begin, we'll first define the k-element vector $\mathrm{d}(\mathrm{k}; \{\mathrm{y_j}\}, \mathrm{y}') \in \mathbb{R}^{\mathrm{k}}$, defined for a generic set $\{\mathrm{y_j}\}$ and vector y' as follows. We first find the k nearest neighbors of y' in the set $\{\mathrm{y_j}\}$. Then we measure the distance between y' and each of these k nearest neighbors. Finally, we define $\mathrm{d}(\mathrm{k}; \{\mathrm{y_j}\}, \mathrm{y}')$ to be the k-element vector of these distances.

Using d we can define the following symmetric error E as follows:

$$\mathrm{E}\left(\mathrm{k}; \{\mathrm{y_j}\}, \{\mathrm{y'_j}\}\right) = \frac{1}{\mathrm{k}|\{\mathrm{y_{j'}}\}|}\sum_{\mathrm{j'}} \left\|\mathrm{d}\left(\mathrm{k}; \{\mathrm{y_j}\}, \mathrm{y'_{j'}}\right)\right\|^2 + \frac{1}{\mathrm{k}|\{\mathrm{y_j}\}|}\sum_{\mathrm{j}} \left\|\mathrm{d}\left(\mathrm{k}; \{\mathrm{y'_{j'}}\}, \mathrm{y_j}\right)\right\|^2,$$

where $|\{\cdot\}|$ refers to the size of set $\{\cdot\}$, and each $\|\mathrm{d}(\mathrm{k})\|^2$ is the sum of squares of the k entries in that $\mathrm{d}(\mathrm{k})$ (i.e., the squared frobenius-norm of $\mathrm{d}(\mathrm{k})$).

Using E and the labels $\{\mathrm{l_j}\}$ and $\{\mathrm{l'_{j'}}\}$, we can define a T-dependent error F as follows:

$$\mathrm{F}\left(\mathrm{T}, \mathrm{k}; \{\mathrm{x_j}\}, \{\mathrm{l_j}\}, \{\mathrm{x'_{j'}}\}, \{\mathrm{l_{j'}}\}\right) = \frac{1}{\mathrm{L}}\sum_{\mathrm{l}} \mathrm{E}\left(\mathrm{k}; \{\mathrm{x_j} \,|\, \mathrm{l_j} = \mathrm{l}\}, \{\mathrm{T} \circ x'_{\mathrm{j'}} \,|\, \mathrm{l_{j'}} = \mathrm{l}\}\right)$$

Because the sets $\{\mathrm{x_j}\}$, $\{\mathrm{l_j}\}$, $\{\mathrm{x'_{j'}}\}$ and $\{\mathrm{l_{j'}}\}$ are fixed, we'll refer to F below simply as $\mathrm{F}(\mathrm{T}, \mathrm{k})$.

Using this notion of error, we can define T as follows:

Initialize: Set T to be the identity transformation and set k to be equal to $\max(\mathrm{J}, \mathrm{J}')$.

Minimize: Starting with the current estimate for T, use Nelder-Meade optimization to update T by minimizing $\mathrm{F}(\mathrm{T}, \mathrm{k})$.

Iterate: Now reduce k by one and return to the previous step.

Terminate: Stop when $k = 1$.

In practice we often observe that the final transformation T is unchanged if the initial k is sufficiently large. Here, we set $k = 32$. Similarly, the final transformation T is often unchanged if k is reduced by more than 1 each iteration.

Calculating label similarity

To determine label similarity, we can take a particular point $y'_{j'}$ (i.e., an ADNI participant) with case-control label $l'_{j'}$, and compare it to the k nearest neighbors in the set $\{y_j\}$ (i.e., UKB participants' data). The number of nearest neighbors with a matching label is denoted as $n(j,k)$. Then, we define $p(j,k) = n(j,k)/k$ to represent the fraction of nearest neighbors with a matching label. Then, we can define $\sigma(k;\{y_j\},\{l_j\},\{y'_{j'}\},\{l'_{j'}\})$ as the average of $p(j,k)$ over j. That is, the average fraction of the k-nearest neighbors with the same label. We use linear interpolation to extend this to a continuous function in the interval $[0,|\{y_j\}|]$. Analogously, we can determine the similarity of UKB labels to ADNI labels, $\sigma(k';\ \{y'_{j'}\},\{l'_{j'}\},\{y_j\},\{l_j\})$. Note that $\sigma(k';\ \{y'_{j'}\},\{l'_{j'}\},\{y_j\},\{l_j\})$ is defined for all k' in the interval $[0,|\{y'_{j'}\}|]$. Finally, for all $f \in [0,1]$, define we define the symmetric similarity as follows:

$$\tilde{\sigma}(f;\{y_j\},\{l_j\},\{y'_{j'}\},\{l'_{j'}\})$$
$$= \frac{1}{2}\sigma(f|\{y_j\}|;\{y_j\},\{l_j\},\{y'_{j'}\},\{l'_{j'}\}) + \frac{1}{2}\sigma(f|\{y'_j\}|;\{y'_{j'}\},\{l'_{j'}\},\{y_j\},\{l_j\})$$

Statistical analysis

All statistical analyses were conducted using MATLAB 2021a and R v4.2.3. Custom scripts written in MATLAB and C based on previously described methods (*27*) were used to search for biclusters. Classification of individuals into disease-relevant constellations was conducted using the ISO-SPLIT algorithm (*61*) implemented in MATLAB and available at

https://github.com/flatironinstitute/isosplit5. Analyses to characterize constellations and biclusters utilized PLINK2 (*62*), and R packages *g:Profiler* (*63*) and *clusterProfiler* (*64, 65*). Visualization of gene set enrichment results used the R package *quanteda (68)* and Cytoscape app EnrichmentMap v3.3 (*70*).

List of Supplementary Materials

Fig. S1. Principal component analysis of a random subset of variants.

Fig. S2. Network plot of gene sets significantly enriched among disease-relevant constellations.

Fig. S3. Overlap in gene sets, genes, and SNPs between constellations and biclusters.

Fig. S4. Traces of search for primary and secondary biclusters.

References and Notes

1. G. M. McKhann *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* **7**, 263-269 (2011).
2. M. E. Murray *et al.*, Neuropathologically defined subtypes of Alzheimer's disease with distinct clinical characteristics: a retrospective study. *Lancet Neurol* **10**, 785-796 (2011).
3. J. W. Vogel *et al.*, Four distinct trajectories of tau deposition identified in Alzheimer's disease. *Nature Medicine*, (2021).
4. A. L. Young *et al.*, Uncovering the heterogeneity and temporal complexity of neurodegenerative diseases with Subtype and Stage Inference. *Nat Commun* **9**, 4273 (2018).
5. D. Ferreira, A. Nordberg, E. Westman, Biological subtypes of Alzheimer disease: A systematic review and meta-analysis. *Neurology* **94**, 436-448 (2020).
6. D. Ferreira *et al.*, Distinct subtypes of Alzheimer's disease based on patterns of brain atrophy: longitudinal trajectories and clinical applications. *Scientific reports* **7**, 46263 (2017).
7. R. Ossenkoppele *et al.*, Tau PET patterns mirror clinical and neuroanatomical variability in Alzheimer's disease. *Brain* **139**, 1551-1567 (2016).
8. B. M. Tijms *et al.*, Pathophysiological subtypes of Alzheimer's disease based on cerebrospinal fluid proteomics. *Brain*, (2020).
9. M. Gatz *et al.*, Role of genes and environments for explaining Alzheimer disease. *Arch Gen Psychiatry* **63**, 168-174 (2006).
10. E. H. Corder *et al.*, Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science* **261**, 921-923 (1993).
11. D. Harold *et al.*, Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease. *Nat Genet* **41**, 1088-1093 (2009).
12. B. W. Kunkle *et al.*, Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing. *Nature Genetics* **51**, 414-430 (2019).
13. J. C. Lambert *et al.*, Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat Genet* **45**, 1452-1458 (2013).
14. I. E. Jansen *et al.*, Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. *Nat Genet*, (2019).
15. C. Bellenguez *et al.*, New insights into the genetic etiology of Alzheimer's disease and related dementias. *Nature Genetics*, (2022).
16. V. Escott-Price *et al.*, Common polygenic variation enhances risk prediction for Alzheimer's disease. *Brain* **138**, 3673-3684 (2015).
17. C. R. Jack, Jr. *et al.*, NIA-AA Research Framework: Toward a biological definition of Alzheimer's disease. *Alzheimers Dement* **14**, 535-562 (2018).
18. B. Dubois *et al.*, Clinical diagnosis of Alzheimer's disease: recommendations of the International Working Group. *The Lancet Neurology*, (2021).
19. B. Dubois *et al.*, Revising the definition of Alzheimer's disease: a new lexicon. *Lancet Neurol* **9**, 1118-1127 (2010).
20. A. M. Fiksinski, G. D. Hoftman, J. A. S. Vorstman, C. E. Bearden, A genetics-first approach to understanding autism and schizophrenia spectrum disorders: the 22q11.2 deletion syndrome. *Mol Psychiatry* **28**, 341-353 (2023).
21. H. A. Stessman, R. Bernier, E. E. Eichler, A genotype-first approach to defining the subtypes of a complex disease. *Cell* **156**, 872-877 (2014).

22. C. M. Wilczewski *et al.*, Genotype first: Clinical genomics research through a reverse phenotyping approach. *Am J Hum Genet* **110**, 3-12 (2023).
23. R. A. Neff *et al.*, Molecular subtyping of Alzheimer's disease using RNA sequencing data reveals novel mechanisms and targets. *Sci Adv* **7**, eabb5398 (2021).
24. M. S. Udler *et al.*, Type 2 diabetes genetic loci informed by multi-trait associations point to disease mechanisms and subtypes: A soft clustering analysis. *PLoS Med* **15**, e1002654 (2018).
25. J. Arnedo *et al.*, Uncovering the hidden risk architecture of the schizophrenias: confirmation in three independent genome-wide association studies. *Am J Psychiatry* **172**, 139-153 (2015).
26. A. Dahl, N. Zaitlen, Genetic Influences on Disease Subtypes. *Annu Rev Genomics Hum Genet* **21**, 413-435 (2020).
27. A. V. Rangan *et al.*, A loop-counting method for covariate-corrected low-rank biclustering of gene-expression and genome-wide association study data. *PLoS Comput Biol* **14**, e1006105 (2018).
28. R. Sims, M. Hill, J. Williams, The multiplex model of the genetics of Alzheimer's disease. *Nat Neurosci* **23**, 311-322 (2020).
29. A. L. Price *et al.*, Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**, 904-909 (2006).
30. J. J. Berg *et al.*, Reduced signal for polygenic adaptation of height in UK Biobank. *Elife* **8**, (2019).
31. S. Haworth *et al.*, Apparent latent structure within the UK Biobank sample has implications for epidemiological analysis. *Nat Commun* **10**, 333 (2019).
32. M. Tzioras, R. I. McGeachan, C. S. Durrant, T. L. Spires-Jones, Synaptic degeneration in Alzheimer disease. *Nature reviews. Neurology* **19**, 19-38 (2023).
33. Tara L. Spires-Jones, Bradley T. Hyman, The Intersection of Amyloid Beta and Tau at Synapses in Alzheimer’s Disease. *Neuron* **82**, 756-771 (2014).
34. A. de Calignon *et al.*, Propagation of tau pathology in a model of early Alzheimer's disease. *Neuron* **73**, 685-697 (2012).
35. B. Ng *et al.*, Neurons derived from individual early Alzheimer's disease patients reflect their clinical vulnerability. *Brain Commun* **4**, fcac267 (2022).
36. L. Munoz, A. J. Ammit, Targeting p38 MAPK pathway for the treatment of Alzheimer's disease. *Neuropharmacology* **58**, 561-568 (2010).
37. C. E. G. Leyns, D. M. Holtzman, Glial contributions to neurodegeneration in tauopathies. *Mol Neurodegener* **12**, 50 (2017).
38. I. Tooyama, H. Kimura, H. Akiyama, P. L. McGeer, Reactive microglia express class I and class II major histocompatibility complex antigens in Alzheimer's disease. *Brain Res* **523**, 273-280 (1990).
39. C. D. Nobes, A. Hall, Rho GTPases control polarity, protrusion, and adhesion during cell movement. *J Cell Biol* **144**, 1235-1244 (1999).
40. J. Settleman, in *Cytoskeleton and Small G Proteins,* P. Jeanteur, Ed. (Springer Berlin Heidelberg, Berlin, Heidelberg, 1999), pp. 201-229.
41. P. Hotulainen, C. C. Hoogenraad, Actin in dendritic spines: connecting dynamics to function. *J Cell Biol* **189**, 619-629 (2010).
42. Y. Cabrales Fontela *et al.*, Multivalent cross-linking of actin filaments and microtubules through the microtubule-associated protein Tau. *Nat Commun* **8**, 1981 (2017).
43. T. Valenta, G. Hausmann, K. Basler, The many faces and functions of beta-catenin. *EMBO J* **31**, 2714-2736 (2012).
44. A. Soldano, B. A. Hassan, Beyond pathology: APP, brain development and Alzheimer's disease. *Curr Opin Neurobiol* **27**, 61-67 (2014).

45. N. W. Kowall, K. S. Kosik, Axonal disruption and aberrant localization of tau protein characterize the neuropil pathology of Alzheimer's disease. *Ann Neurol* **22**, 639-643 (1987).
46. K. Iqbal, F. Liu, C. X. Gong, I. Grundke-Iqbal, Tau in Alzheimer disease and related tauopathies. *Curr Alzheimer Res* **7**, 656-664 (2010).
47. F. Leng, P. Edison, Neuroinflammation and microglial activation in Alzheimer disease: where do we go from here? *Nature reviews. Neurology* **17**, 157-172 (2021).
48. M. T. Heneka, M. K. O'Banion, Inflammatory processes in Alzheimer's disease. *J Neuroimmunol* **184**, 69-91 (2007).
49. E. Porcu *et al.*, Differentially expressed genes reflect disease-induced rather than disease-causing changes in the transcriptome. *Nat Commun* **12**, 5647 (2021).
50. A. S. Saporta *et al.*, Charcot-Marie-Tooth disease subtypes and genetic testing strategies. *Ann Neurol* **69**, 22-33 (2011).
51. S. J. Crutch *et al.*, Consensus classification of posterior cortical atrophy. *Alzheimers Dement* **13**, 870-884 (2017).
52. M. L. Gorno-Tempini *et al.*, Classification of primary progressive aphasia and its variants. *Neurology* **76**, 1006-1014 (2011).
53. R. A. Townley *et al.*, Progressive dysexecutive syndrome due to Alzheimer's disease: a description of 55 cases and comparison to other phenotypes. *Brain Commun* **2**, fcaa068 (2020).
54. R. Ossenkoppele *et al.*, Research Criteria for the Behavioral Variant of Alzheimer Disease: A Systematic Review and Meta-analysis. *JAMA Neurol* **79**, 48-60 (2022).
55. C. Sudlow *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* **12**, e1001779 (2015).
56. C. Bycroft *et al.*, The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203-209 (2018).
57. T. Wilkinson *et al.*, Identifying dementia outcomes in UK Biobank: a validation study of primary care, hospital admissions and mortality data. *Eur J Epidemiol* **34**, 557-565 (2019).
58. C. Y. Chen *et al.*, Improved ancestry inference using weights from external reference panels. *Bioinformatics* **29**, 1399-1406 (2013).
59. The 1000 Genomes Project Consortium, A global reference for human genetic variation. *Nature* **526**, 68-74 (2015).
60. S. Das *et al.*, Next-generation genotype imputation service and methods. *Nat Genet* **48**, 1284-1287 (2016).
61. J. E. Chung *et al.*, A Fully Automated Approach to Spike Sorting. *Neuron* **95**, 1381-1394 e1386 (2017).
62. C. C. Chang *et al.*, Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
63. U. Raudvere *et al.*, g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res* **47**, W191-W198 (2019).
64. T. Wu *et al.*, clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation (Camb)* **2**, 100141 (2021).
65. G. Yu, L. G. Wang, Y. Han, Q. Y. He, clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* **16**, 284-287 (2012).
66. M. Ashburner *et al.*, Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25-29 (2000).
67. Gene Ontology Consortium, The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Res* **49**, D325-D334 (2021).
68. K. Benoit *et al.*, quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software* **3**, 774 (2018).

69. H. Zhou *et al.*, Detecting boolean asymmetric relationships with a loop counting technique and its implications for analyzing heterogeneity within gene expression datasets. *bioRxiv*, 2022.2008.2004.502792 (2022).
70. D. Merico, R. Isserlin, O. Stueker, A. Emili, G. D. Bader, Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. *PLoS One* **5**, e13984 (2010).

Acknowledgments:

This research has been conducted using data from UK Biobank, a major biomedical database (http://www.ukbiobank.ac.uk/). 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.

Figures and tables

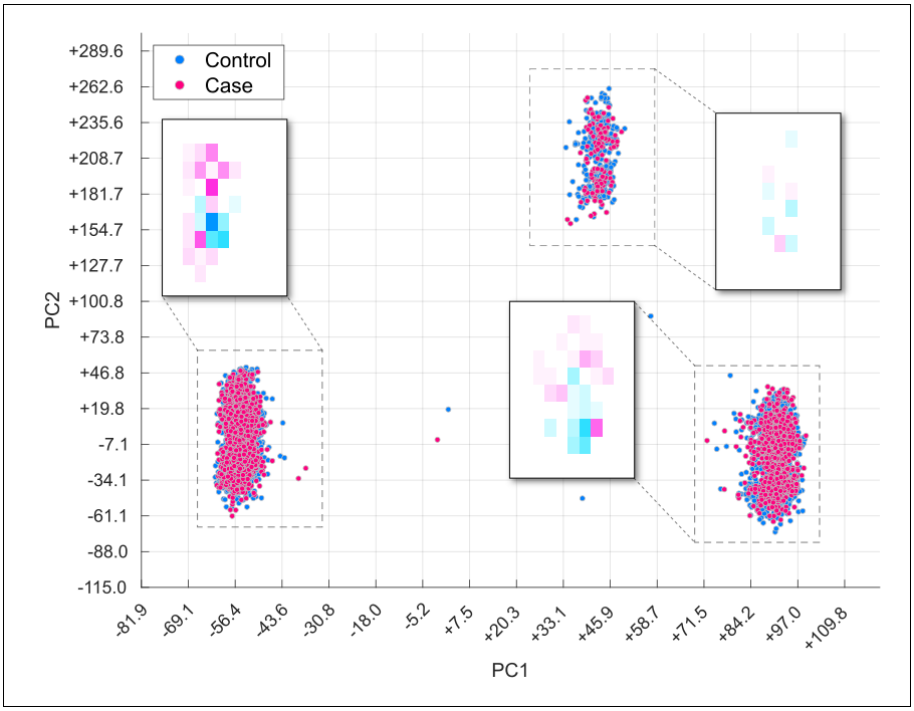

Fig. 1. Principal component analysis of UKB data restricted to Alzheimer's disease-associated variants reveals three constellations. Principal component analysis was applied to allele combinations of UK Biobank cases and controls. Variants were restricted to those with a p-value<0.05 in the Kunkle et 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.

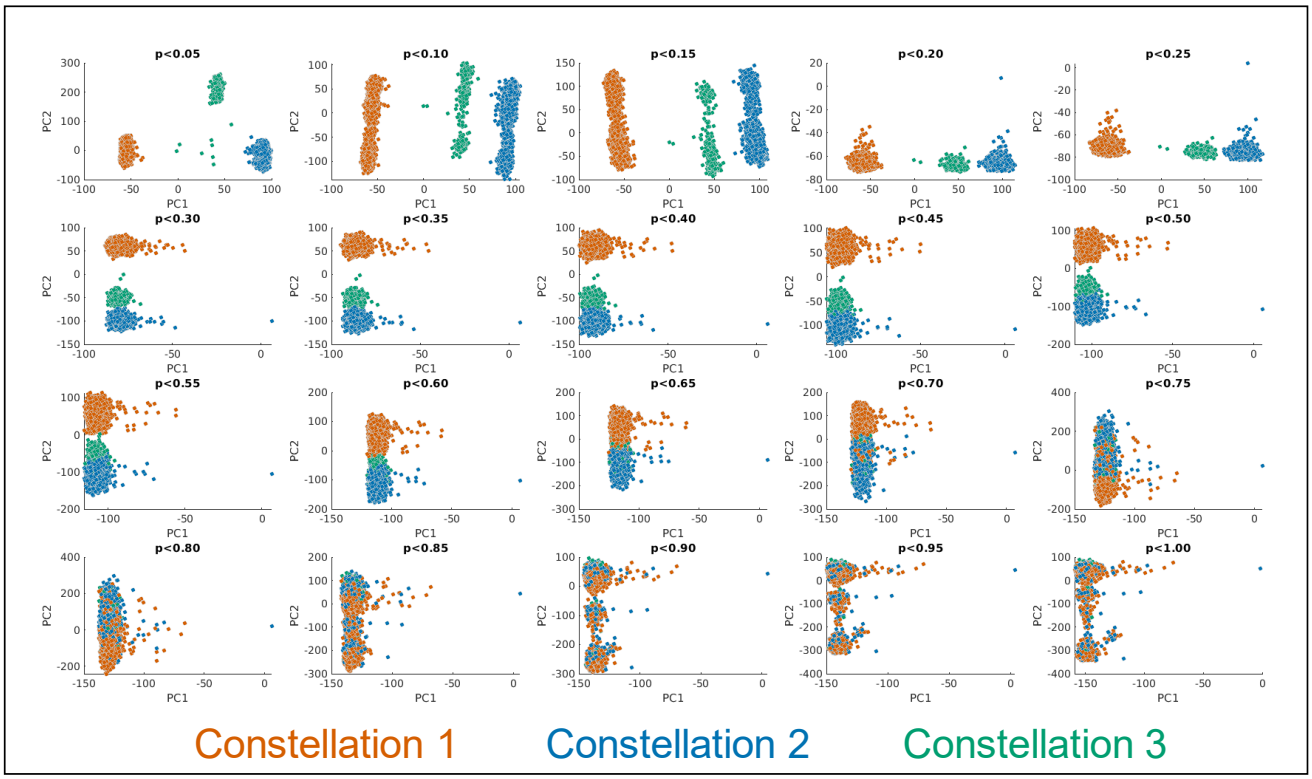

Fig. 2. Principal component analysis across varying thresholds of Alzheimer's associated variants. Principal component analysis was applied to allele combinations of UK Biobank cases and controls. Variants were restricted across a range of p-value thresholds from the Kunkle et al. Alzheimer's GWAS (*12*). 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.

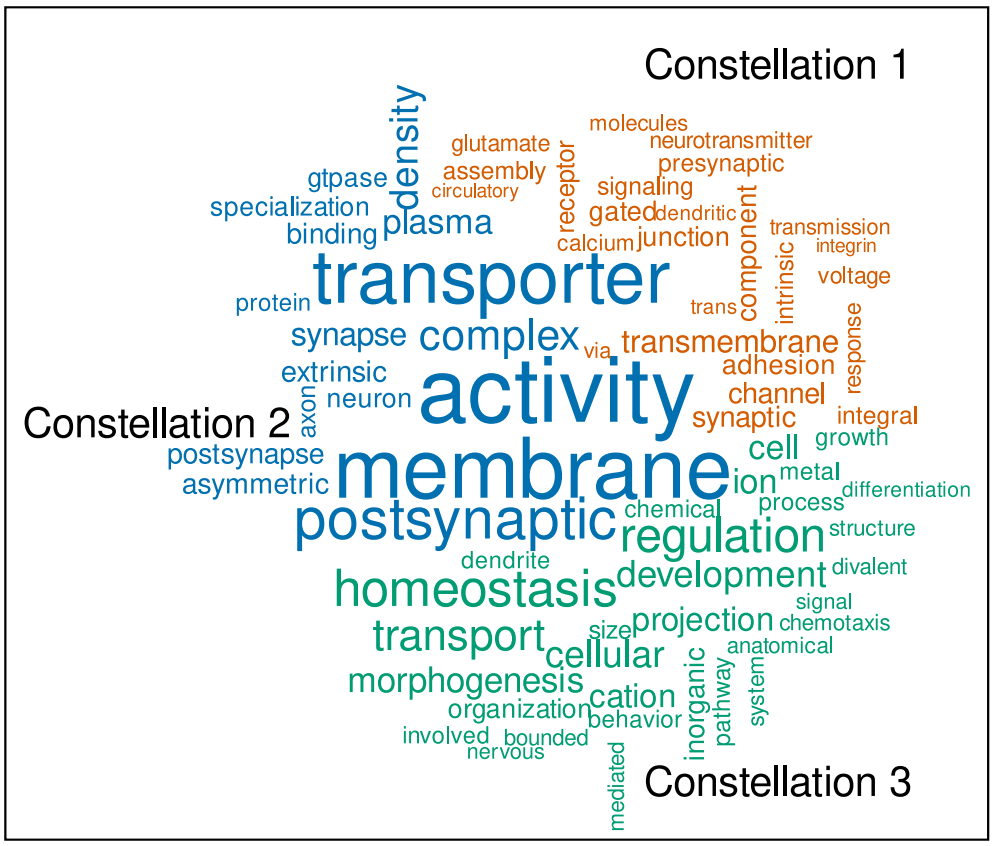

Fig. 3. Comparison cloud of gene sets associated with disease-relevant constellations. Visualization of term frequency from descriptions of gene sets associated with each constellation. Variants associated with each constellation were mapped to genes and gene set enrichment was applied to each constellation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

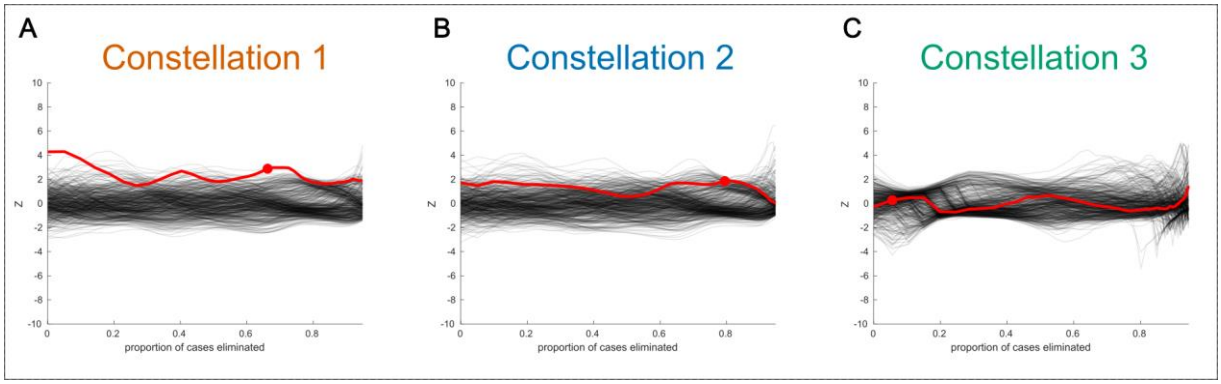

Fig. 4. Bicluster traces of observed data versus label-shuffled data. The disease-related signal-strength associated with the remaining UKB Alzheimer's cases relative to controls is plotted on the y-axis. At each iteration, allele combinations and cases that contribute least to this difference are removed. 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.

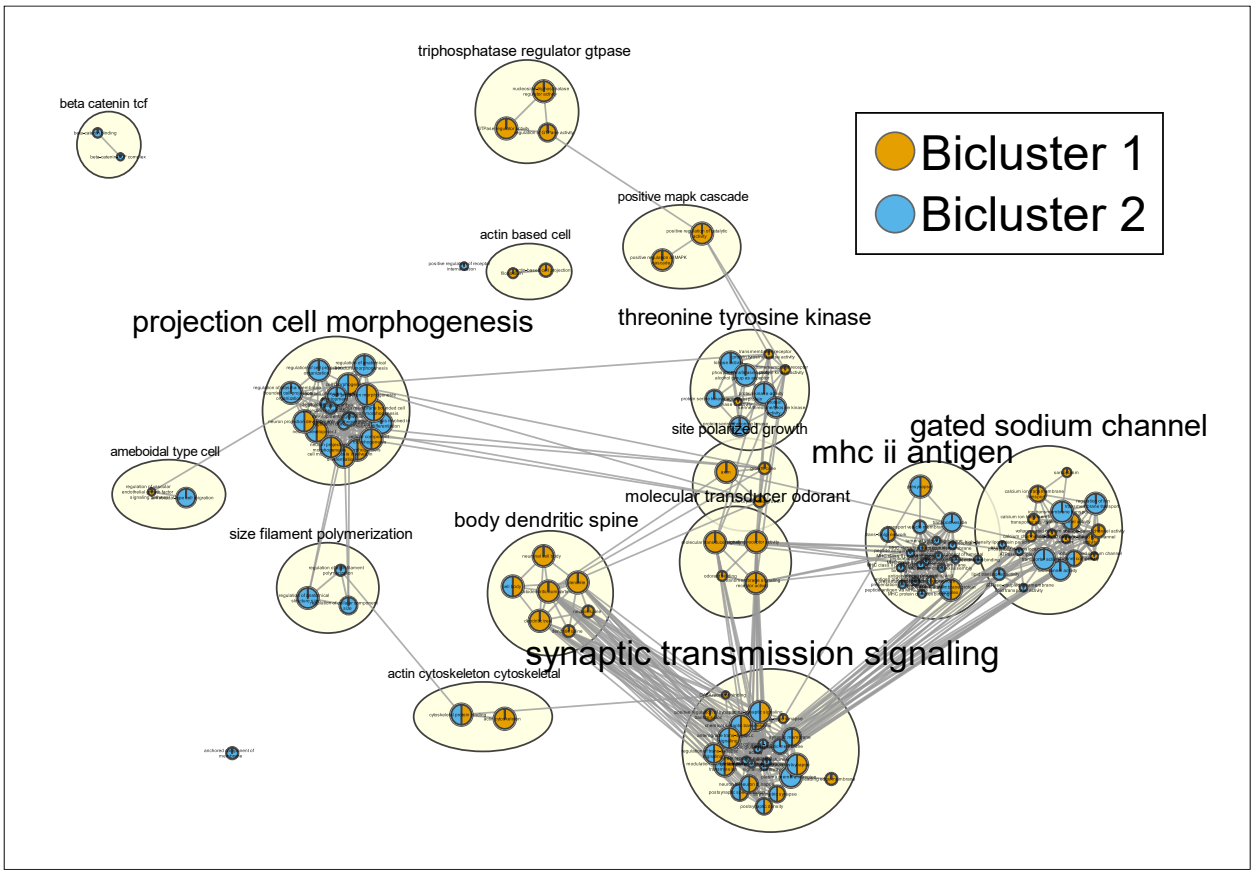

Fig. 5. Network plot of gene sets enriched among bicluster genes. Network visualization of gene set enrichment results for biclusters 1 and 2. Nodes represent significantly enriched gene sets with color indicating the bicluster they are associated with. Edges represent the overlap in genes belonging 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.

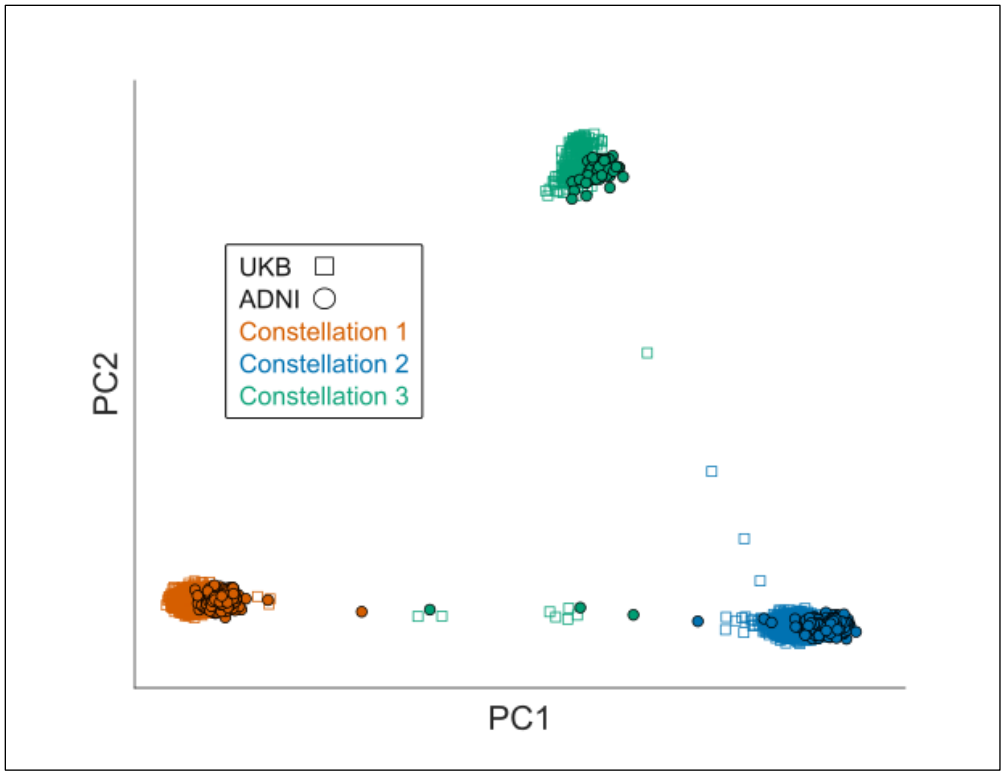

Fig. 6. ADNI data projected along principal components defined in UKB data replicate distribution of disease-relevant constellations. Principal components were recalculated in UKB data using only AD-associated variants common to both datasets. Participants from both datasets were then plotted by participants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

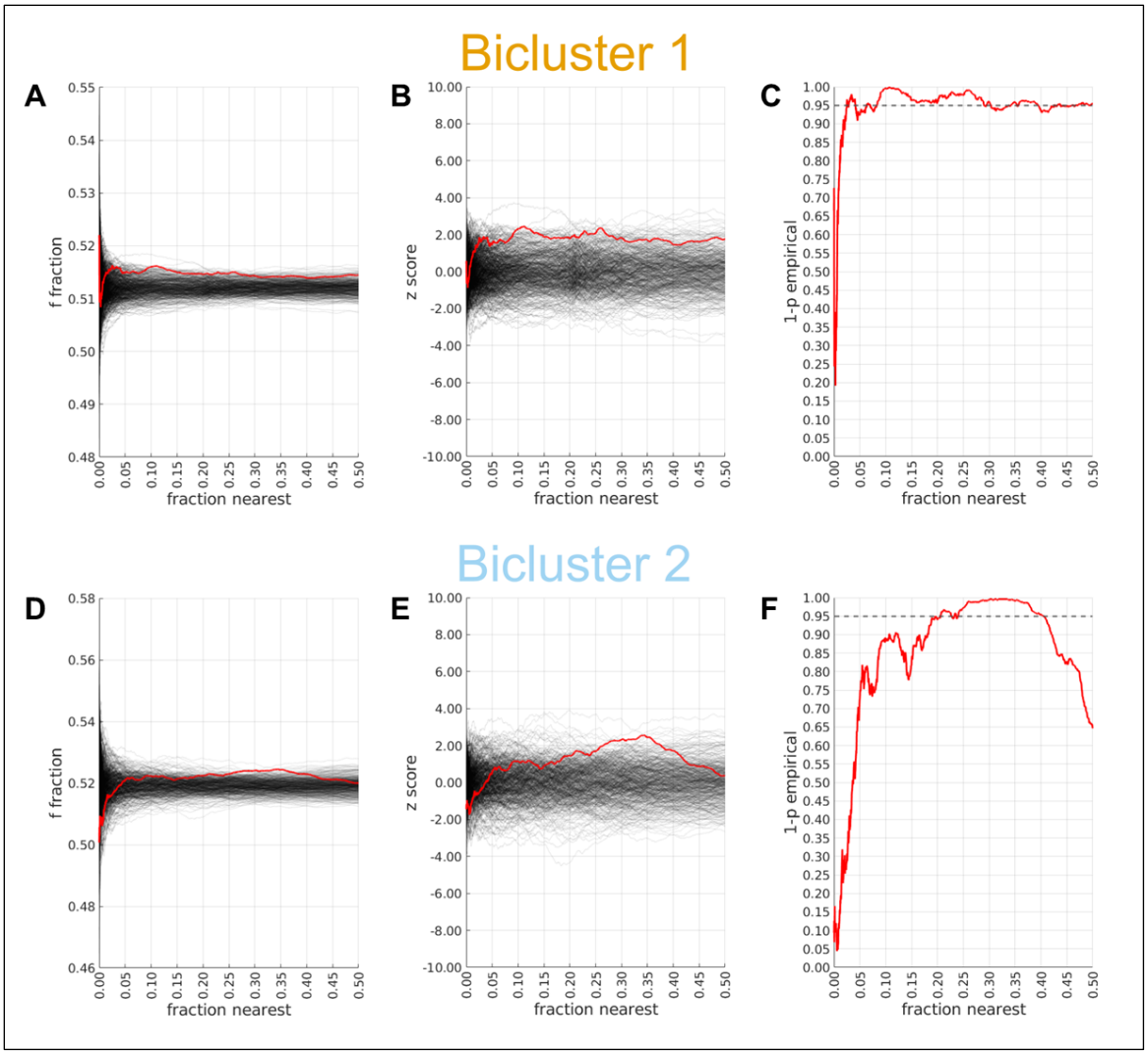

Fig. 7. Replication of disease-specific biclusters in ADNI data. The similarity between case-control labels of individuals in the test set (ADNI) and the most frequent label among nearest neighbors in the training set (UKB) was used to assess replication of biclusters. Only individuals belonging to the constellation in which the given bicluster was found were considered. The proportion of individuals considered as nearest neighbors (i.e., f from the Methods section "*Replication of disease-specific biclusters in ADNI*") 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 (i.e., $\tilde{\sigma}(f)$). The red line in Panels (**A**) and (**D**) shows $\tilde{\sigma}(f)$ from the original data, while the black

lines show the $\tilde{\sigma}(f)$ from label-shuffled trials drawn from the null-distribution. These label-shuffled trials can be used to define a trial-wise mean and variance for $\tilde{\sigma}(f)$ at each f. These trial-wise means and variances can then be used to normalize $\tilde{\sigma}(f)$, with the associated z-scores shown in Panels (**B**) and (**E**). A rank-normalization of the $\tilde{\sigma}(f)$ is shown in Panels (**C**) and (**F**), from which can be read the empirical p-value for the label-matching (i.e., values of $\tilde{\sigma}(f)$ that lie above the 95th percentile – as indicated by the dashed line – would correspond to values of $p<0.05$ for that value of f).

Table 1. Sample characteristics. Descriptive statistics for the UK Biobank and Alzheimer's Disease Neuroimaging Initiave (ADNI) datasets. Characteristics are presented for the overall daset and stratified by clinical status. CU = Cognitively unimpaired, AD = Alzheimer's disease.

	UK Biobank			**ADNI**		
	Overall	**CU**	**AD**	**Overall**	**CU**	**AD**
N	8217	5478	2739	1052	576	476
Sex, n male (%)	3993 (48.6%)	2662 (48.6%)	1331 (48.6%)	571 (54.3%)	283 (49.1%)	288 (60.5%)
Age, years (sd)	64.74 (4.21)	64.74 (4.21)	64.74 (4.21)	73.72 (6.96)	73.05 (6.35)	74.54 (7.55)
Education, years (sd)	11.95 (4.96)	12.21 (5.01)	11.41 (4.80)	16.05 (2.71)	16.52 (2.46)	15.47 (2.89_
APOE-e4 carriers, n (%)	3135 (38.2%)	1415 (25.8%)	1720 (62.8%)	486 (46.2%)	172 (29.9%)	314 (66.0%)

Funding:

National Institute on Aging grant K01 AG063805 to JAE

National Institute on Aging grants UH3-AG064706, U19 AG023122, U19 AG065169-01A1 and U24AG078753 to NJS

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).

Author contributions:

Conceptualization: JAE, AVR, NJS
Methodology: JAE, AVR
Investigation: JAE, AVR
Visualization: JAE, AVR
Funding acquisition: JAE, AVR, NJS
Project administration: JAE
Supervision: JAE
Writing – original draft: JAE, AVR
Writing – review & editing: JAE, AVR, NJS

Competing interests:

The authors declare that they have no competing interests.

Data and materials 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 authors.