

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

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

<sup>1</sup> Department of Psychiatry, University of California San Diego, La Jolla, CA, USA

<sup>2</sup> Center for Behavior Genetics of Aging, University of California San Diego, La Jolla, CA, USA

<sup>3</sup> The Translational Genomics Research Institute, Quantitative Medicine and Systems Biology, Phoenix, AZ, USA

<sup>4</sup> Department of Mathematics, New York University, New York, New York, USA

<sup>†</sup>Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database ([adni.loni.usc.edu](http://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](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](mailto:jaelman@health.ucsd.edu)*

## 1 ABSTRACT

2

3 **Background:** Alzheimer's disease (AD) exhibits considerable phenotypic heterogeneity,  
4 suggesting the potential existence of subtypes. AD is under substantial genetic influence, thus  
5 identifying systematic variation in genetic risk may provide insights into disease origins.

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

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

14 **Results:** PCA revealed three distinct clusters ("constellations") driven primarily by different  
15 correlation patterns in a region of strong LD surrounding the *MAPT* locus. Constellations  
16 contained a mixture of cases and controls, reflecting disease-relevant but not disease-specific  
17 structure. We found two disease-specific biclusters among AD cases. Pathway analysis linked  
18 bicluster-associated variants to neuron morphogenesis and outgrowth. Disease-relevant and  
19 disease-specific structure replicated in ADNI, and bicluster 2 exhibited increased CSF p-tau and  
20 cognitive decline over time.

21 **Conclusions:** This study unveils a hierarchical structure of AD genetic risk. Disease-relevant  
22 constellations may represent haplotype structure that does not increase risk directly but may  
23 alter the relative importance of other genetic risk factors. Biclusters may represent distinct AD  
24 genetic subtypes. This structure is replicable and relates to differential pathological  
25 accumulation and cognitive decline over time.

26

27 **Keywords:** Alzheimer's disease, genetic subtypes, biclustering, genotyping, genetic risk

## 28 BACKGROUND

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

41 Sporadic AD is under considerable genetic influence, with an estimated heritability of  
42 60-80% [9]. Thus, examining the genetic architecture of AD risk and the myriad ways in which  
43 combinations of variants associate with AD provides a useful foundation to understand its  
44 complex etiology. Although *APOE* represents the single largest source of genetic risk for the  
45 disease [10] recent GWAS have identified upwards of 75 different risk loci [11-15]. Considering  
46 variants that do not reach the level of genome-wide-significance but still suggest an association  
47 with AD could provide additional insight above and beyond *APOE* [16] and genome-wide  
48 significant variants. The highly oligo or polygenic nature of AD risk could reflect underlying  
49 etiological heterogeneity across individuals. The evolving definition of AD and subsequent  
50 debate over its origins further highlights the complex nature of the disease [17-19]. Given this  
51 genetic complexity yet commonality among aspects of AD clinical presentations, there is little  
52 reason to expect that phenotype-based classifications of sporadic AD (and its subtypes) will

53 cleanly delineate homogeneous subgroups of genetic risk, providing motivation for a using  
54 “genotype-first” approach to identifying subtypes [20-22].

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

70 Here we investigated genetic heterogeneity of Alzheimer’s disease using a combination  
71 of approaches. We first applied a principal component analysis to AD cases and controls from  
72 the UK Biobank to identify potential clusters among AD-associated SNPs that may indicate  
73 disease-relevant vulnerability across all individuals, regardless of disease status. We then  
74 applied a biclustering method to each of these likely vulnerability clusters to search for subsets  
75 of cases that harbor distinct genetic signatures that significantly increase AD risk. To validate  
76 our findings, we investigate whether the same clusters or patterns of heterogeneity also appear  
77 in an independent group of cases and controls from the Alzheimer’s Disease Neuroimaging

78 Initiative (ADNI). Taken together, we find a hierarchical structure to an underlying heterogeneity  
79 of AD genetic risk, providing further insight into the complex etiology of the disease.

## 80 **METHODS**

### 81 ***Participant characteristics***

82 We used imputed genotyping data from the UK Biobank (UKB) as the discovery dataset  
83 (**Table 1**). The UKB is a large-scale biomedical database and research resource containing  
84 genetic, lifestyle and health information from half a million UK participants [28, 29]. Data from  
85 2,739 Alzheimer's disease cases [30] and 5,478 age- and sex-matched controls with White  
86 British ancestry as determined by PCA [29] were included in these analyses.

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

### 100 ***Genotyping data and quality control***

101 Genotyping data imputed to the Haplotype Reference Consortium plus UK10K reference  
102 haplotype resource were downloaded from the UKB database along with genetic principal

103 components (hereafter referred to as “genome-wide” principal components). Sample QC  
104 information provided by the UKB were used to apply exclusionary criteria. Individuals with  
105 excess relatedness were removed. This was defined by the UK Biobank as samples with more  
106 than 10 putative third-degree relatives (KING coefficient between 0.0442 and 0.0884). Imputed  
107 data used in these analyses included only samples with <10% missingness and biallelic SNPs  
108 with >1% minor allele frequency.

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

119 Note that, at this stage we retain all the SNPs for our primary analysis, regardless of the  
120 linkage disequilibrium (LD) relationships that might exist between them. The reason we retain all  
121 the SNPs is that our heterogeneity analysis (described below) will specifically search for  
122 combinations of SNPs that preferentially exhibit correlations across the case-subjects in  
123 contrast to the controls. These correlations drive the heterogeneous structures we are trying to  
124 find and can themselves be thought of as a form of LD that interacts with the disease. Thus, to  
125 carry out our heterogeneity analysis we retain all the relevant SNPs (regardless of LD), and later  
126 correct for ‘population-wide’ LD (e.g., SNP-correlations that are not disease-specific, but  
127 putatively ancestry-related) within our biclustering analysis below.

## 128 ***Assessing high-level structure with principal components analysis***

129 We restricted analyses to 486,823 variants in the UKB data that were associated with  
130 AD with an uncorrected p-value of  $<0.05$  from the AD GWAS by Kunkle et al. [12] to  
131 accommodate potential association signals beyond those exhibiting genome-wide significance.  
132 After applying MAF ( $>1\%$ ) and genotype missingness filters ( $<10\%$ ), a principal components  
133 analysis (PCA) was applied to the allele combinations (dummy-coded according to the 3  
134 possible allele combinations for each) of these AD-associated variants across all UKB cases  
135 and controls to assess the presence of high-level structure in the data (i.e., a non-Gaussian  
136 distribution). For more details on the choice of allele-coding, see Supplemental Methods section  
137 *Rationale for allele-coding*. The PCA was implemented in C using the normalized power  
138 iteration method calculated to a relative error of  $1e-6$ , or single-digit precision. Individuals were  
139 assigned to clusters, hereafter referred to as “constellations” to disambiguate from the term  
140 “bicluster” used in subsequent steps, by applying the ISO-SPLIT algorithm [34] implemented in  
141 MATLAB (<https://github.com/flatironinstitute/isosplit5>) to participant loadings on the first PC.  
142 Classification was based only on the first PC because it most robustly separated constellations  
143 across a range of p-value thresholds.

144 Although the sample was restricted to White British participants, it is possible that these  
145 loadings reflect remaining ancestry-related population substructure. We therefore determined  
146 whether the resulting structure was specific to the set of AD-associated variants in several  
147 ways. First, PCA was applied to sets of variants restricted to p-value thresholds ranging from  
148  $p<0.05$  to  $p<1.0$  in increments of 0.05. Loadings on the first two PCs were plotted at each  
149 increment, with individuals labelled according to cluster membership at the  $p<0.05$  threshold to  
150 visually assess stability of clustering. Second, to determine whether the resulting structure was  
151 simply a function of the number of variants analyzed, a PCA was applied to a randomly chosen  
152 set of variants with the same size as those included in the analysis restricted to variants with  
153  $p<0.05$ . Third, we examined loadings of allele combinations on the primary PC ordered by

154 genomic position as suggested by Privé et al. [35] to assess whether it was driven by regions of  
155 high population-wide LD.

### 156 ***Bicluster analysis of disease-specific structure***

157 Heterogeneity exhibited at the top level of the constellations strongly suggests that there  
158 may be other heterogeneous substructures contained within each. This possibility seems all the  
159 more likely due to the non-Gaussian distribution of subjects within each constellation, and that  
160 the distribution of cases and controls appears to be in different directions across constellations.

161 To search for biclusters we use the half-loop method described in Rangan et al. [36].  
162 This biclustering strategy (along with similar spectral-biclustering strategies) works best when  
163 the combined population of cases and controls is relatively homogenous. Given the clear  
164 structure evident from the PCA and the differential distribution of cases relative to controls  
165 across constellations, we therefore searched for disease-specific heterogeneity within each  
166 constellation separately.

167 Our biclustering strategy involves an iterative process which starts with all the  
168 participants (in this case, all participants in a given constellation) and SNPs, and then  
169 sequentially removes AD cases and allele-combinations from consideration. The Supplementary  
170 Methods contains a detailed description of this procedure but is summarized here. Briefly, we  
171 can measure the fraction of allele combinations that are shared between cases and subtract  
172 from this the fraction that are shared with controls to obtain a disease-related signal strength of  
173 remaining AD cases and controls. This difference can be thought of as a measure of disease-  
174 specific LD within the remaining AD-cases. The subtraction of the control-signal above is a form  
175 of “control-correction”. This control-correction is a critical step because it reduces the influence  
176 of structure present in both cases and controls on clustering. For example, disease non-specific  
177 LD or technical artifacts that are present in the full sample will be controlled for by this step. At  
178 each iteration the cases and allele-combinations with the smallest contributions to this signal are  
179 removed and the process repeats as described above. Recording this value at each step



180 produces a ‘trace’ which indicates the disease-related signal-strength associated with the  
181 remaining AD cases and variants at that iteration.

182 In pursuing this process, we controlled for the first two principal-components extracted  
183 from a GRM of all individuals in our sample (and without thresholding for AD-associated SNPs)  
184 to control for ancestry. When calculating these genome-wide principal-components we did  
185 correct for LD, as our goal was to estimate the principal-components across the whole  
186 population (i.e., assuming homogeneity across subjects).

187 We can use the peak of the trace to delineate the membership of the dominant bicluster  
188 (i.e., which AD cases and allelic-combinations contribute to the disease-specific signal). We  
189 identify the peak by finding the internal maximum, ignoring the initial and final iterations that  
190 include >95% or <5% of AD cases.

191 As a null-hypothesis, we assume that the disease-label (i.e., case vs control) is not  
192 associated with the genetic profile of each subject. Therefore, we randomly permuted the case-  
193 and control-labels across subjects with similar genome-wide principal components (see Rangan  
194 et al. [27] for details) and re-calculated the traces as described above. Here, we use 500  
195 permutations. By comparing the original trace with the distribution of traces drawn from the null-  
196 hypothesis, we can assign a p-value to the observed trace at each iteration. The null distribution  
197 will retain any structure that is uncorrelated with disease label, so provides a second type of  
198 correction against identifying clusters driven by non-disease-related LD. In this case, we only  
199 assess iterations that include >5% of the cases (i.e., the final iterations with very few cases  
200 and/or variants are ignored). To determine whether we have found a statistically significant  
201 bicluster within the original data, termed the ‘dominant’ bicluster, we can examine the p-value of  
202 the highest peak ( $p_{\max}$ ) and the average p-value across iterations ( $p_{\text{avg}}$ ). Depending on the  
203 structure of the bicluster, the various p-values may be quite different, but each may be a useful  
204 metric for identifying significant biclusters. For example, if the original trace has one or more  
205 clear peaks, then there are statistically robust ‘cutpoints’ which can be used to delineate

206 bicluster membership and  $p_{\max}$  is likely to be very small. On the other hand, if the original trace  
207 has a very broad peak or a long plateau, then the bicluster is quite ‘fuzzy’, corresponding to a  
208 continuum of membership. In this case, we might expect  $p_{\text{avg}}$  to be small, but  $p_{\max}$  may be  
209 relatively large. If the dominant bicluster within a data-set is statistically significant, we can  
210 extract it and then search for a secondary bicluster. This is done by scrambling the entries of the  
211 submatrix associated with the bicluster (i.e., entries corresponding to the participants and allele-  
212 combinations that were retained in the bicluster) and running the search algorithm again [27,  
213 37].

#### 214 ***Gene set enrichment analysis of bicluster-associated variants***

215 We conducted a GWAS across AD-associated SNPs comparing bicluster cases with  
216 controls belonging to the same disease-relevant constellation in which the bicluster was found  
217 (e.g., bicluster 1 cases versus controls for constellation 1) using PLINK2 [38]. Analyses were  
218 adjusted for the first two genome-wide principal components. As with the analysis of  
219 constellations, variants positively associated with each bicluster (i.e., positive regression  
220 coefficient and uncorrected  $p < 0.05$ ) were mapped to genes using the *g:Profiler* package in R  
221 [39]. A gene set enrichment analysis was then applied to annotated genes from each bicluster  
222 using the *clusterProfiler* package in R [40, 41]. The analysis was restricted to gene sets from the  
223 Gene Ontology resource [42, 43] containing between 10 and 1,000 genes. The Cytoscape app  
224 EnrichmentMap v3.3 [44] was used to visualize gene sets significantly enriched (FDR corrected  
225  $p < 0.05$ ) in each bicluster. We constructed a network in which nodes represented gene sets  
226 significantly associated with each bicluster (gene sets can be associated with either or both).  
227 Edges were defined by the proportion of overlapping genes between gene sets (using a  
228 minimum overlap threshold of 0.5). Next, the AutoAnnotate app  
229 (<http://baderlab.org/Software/AutoAnnotate>) was used to cluster and annotate nodes based on  
230 degree of overlapping genes using the MCL Cluster algorithm.

231 ***Validation of genetic heterogeneity in ADNI***

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

241 After participants were classified into constellations, we assessed replication of the  
242 dominant biclusters separately in the constellations in which they were found (see  
243 Supplementary Methods section *Replication of disease-specific biclusters in ADNI* for full  
244 details). Thus, we tested replication of bicluster 1 only among individuals belonging to  
245 constellation 1, and tested replication of bicluster 2 only among individuals belonging to  
246 constellation 2. For a given bicluster, we calculated the first 2 dominant SNP-wise principal  
247 components (using only SNPs in common to both datasets) among the UKB cases belonging to  
248 the bicluster. The resulting SNP loadings were used to project UKB participants in the  
249 constellation (including controls, bicluster cases, and non-bicluster cases). Similarly, we  
250 calculated projections for each of the ADNI participants belonging to a given constellation using  
251 these UKB bicluster-defined loadings. If the bicluster structure is present in the ADNI data, then  
252 the distribution of cases and controls in this space should be similar between datasets. That is,  
253 for a given participant in the testing set, participants from the training set located nearby should  
254 tend to have the same label.

255 We can choose what fraction of the training set participants to compare, denoted as  $f$ ,  
256 and for a given choice of  $f$  we calculated the proportion of nearest neighbors in the training set  
257 that have the same label (i.e., either case or control) as each of the individuals in the test set.  
258 We assessed the quality of each choice of  $f$  using a permutation test in which the average  
259 “label match” across participants in the test set is compared to a null distribution constructed  
260 from label-shuffled data. We ran 500 permutations for these analyses. We calculate the average  
261 z-score across the range of  $f$  in  $(f_{lo}, f_{hi})$ , with  $f_{lo}$  corresponding to the lower end of the 95%  
262 confidence-interval for affine-point-matching (i.e.,  $f_{lo} \sim 6\%$ ) and  $f_{hi}$  at 50%. This average is  
263 calculated using a normalizing factor to correct for heteroskedasticity [45]. The variance  
264 determining the normalizing factor is calculated from the analogous z-scores obtained after  
265 alignment of the projections onto principal-components calculated from randomly selected  
266 biclusters (i.e., subsets of cases and allele-combinations) of the same size as the bicluster of  
267 interest. We calculated a global empirical p-value by comparing the average z-score of the  
268 observed data across the range of parameter choices of  $f$  in  $(f_{lo}, f_{hi})$  to the null distribution  
269 across the same range. Given that a particular bicluster is globally significant, values of  $f$   
270 corresponding to high z-scores indicate reasonable values of  $f$  to use when labelling new data,  
271 as described below.

### 272 ***Association of bicluster groups with cognitive and biomarker trajectories***

273 The procedures described above were repeated on the full set of ADNI genotyping data,  
274 this time including individuals with MCI in addition to AD cases and controls. This included  
275 applying the previously described genotyping filters, projecting ADNI data onto PC space  
276 defined by the UKB data and assigning to the nearest constellation, and then projecting  
277 individuals belonging to constellations 1 and 2 using the bicluster-defined loadings from  
278 biclusters 1 and 2, respectively. The number of nearest neighbors with each label (i.e., bicluster  
279 case, non-bicluster case, and controls) was then recorded for each ADNI participant. The

280 fraction of training set participants used as nearest neighbors was determined by the fraction  
281 with the highest z-score in the validation step described above. Each ADNI participant was  
282 assigned a soft label, calculated as the proportion of nearest neighbors in the UKB training set  
283 that were bicluster cases. To assign individuals to biclusters 1 and 2, k-means clustering with  
284  $k=2$  was applied to the soft labels (i.e., proportions of nearest neighbors that were bicluster  
285 cases) of each ADNI participant belonging to constellations 1 and 2 separately.

286 Phenotypic analyses were restricted to the ADNI MCI participants who were not  
287 included in the bicluster validation analysis. We examined change of cognition, amyloid, and  
288 phosphorylated tau (p-tau) over time between individuals labelled as bicluster 1, bicluster 2, or  
289 non-bicluster. Cognition was assessed using scores on the Preclinical Alzheimer's Cognitive  
290 Composite (PACC) [46, 47]. Amyloid was assessed with florbetapir PET data processed  
291 according to previously published methods (<http://adni.loni.usc.edu/methods>) [48, 49].  
292 Specifically, we downloaded mean standardized uptake value ratios (SUVR) from a set of  
293 regions including frontal, temporal, parietal and cingulate cortices using whole cerebellum as a  
294 reference region. Cerebrospinal fluid p-tau CSF samples were collected on cohort participants  
295 and processed as previously described [50]. CSF p-tau was measured with the fully automated  
296 Elecsys immunoassay (Roche Diagnostics) by the ADNI biomarker core (University of  
297 Pennsylvania). Florbetapir PET and CSF p-tau were chosen as the measures of amyloid and p-  
298 tau because they provided the highest number of visits with relevant data across our MCI  
299 participant group.

300 Group differences between individuals with MCI labelled as bicluster 1, bicluster 2, or  
301 non-bicluster were assessed with linear mixed effects models using the *lme4* [51] and *lmerTest*  
302 [52] R packages. For each outcome of interest (cognition, amyloid, or p-tau), all timepoints with  
303 available data were included. An interaction between age and group was used to assess  
304 differences in the trajectories of cognitive and biomarker measures over time. A random  
305 intercept was included for participant.

## 306 RESULTS

### 307 *Identification of disease-relevant constellations*

308 The PCA of AD-associated variants that were retained after applying QC filters  
309 (n=446,700 SNPs; 1,340,100 allele-combinations) in the UKB data revealed three distinct  
310 clusters, or “constellations”, each containing a mixture of AD cases and controls (**Figure 1A**).  
311 The plots in **Figure 1B** indicate that this structure only emerges when restricting to AD-  
312 associated SNPs. When all variants are included in the PCA (e.g., no p-value criterion is used)  
313 there is also evidence for substructure, but it is quite different than what is observed when a  
314 threshold of  $p < 0.05$  is used, with individuals from each of the constellations being highly  
315 interspersed. The separation of constellations along the dominant component emerges around  
316  $p < 0.25$ , and the constellations begin to further separate along the second PC at  $p < 0.05$ . Results  
317 of a PCA applied to a random set of SNPs of the same size as were included at the  $p < 0.05$  level  
318 again found that individuals from each constellation were interspersed, and the overall structure  
319 closely mirrored that seen when all SNPs were included (**Supplementary Figure 1**). Three  
320 similarly distinct clusters emerge if a PCA is applied to additively-coded data as opposed to  
321 allele-coding (**Supplemental Figure 2**).

322 Plotting the SNP-wise loadings of PC1 revealed strong contributions from variants in the  
323 region of chromosome 17q21 (**Figure 2**). This is a known region of extended LD with complex  
324 genomic architecture, including a 900-kb inversion polymorphism surrounding the *MAPT* gene  
325 that defines two haplotypes, H1 and H2, with H1 containing multiple sub-haplotypes [53, 54].  
326 Examining the genotypes of each constellation at the H1/H2 tagging SNP rs8070723, there is a  
327 strong (but not perfect) correspondence between constellations and haplotypes. Constellation 1  
328 is largely homozygous for the H1-associated allele, constellation 3 is largely homozygous for the  
329 H2-associated allele, and constellation 2 is heterozygous (**Supplemental Table 2**). Thus, while  
330 variation from across the genome contributes to this constellation structure, it is driven primarily

331 by several correlation patterns within this LD block, likely corresponding to H1/H2 haplotype  
332 status.

333 Taken together, these findings suggest that the constellations are *disease-relevant*  
334 because they only emerge as dominant modes of variation when restricting consideration to AD-  
335 associated variants, and because the distribution of cases and controls falls along different  
336 directions across constellations. However, they are not *disease-specific* because constellations  
337 contain both cases and controls, suggesting cluster membership alone does not discriminate  
338 between high and low disease risk. The three constellations were not significantly different on  
339 demographic characteristics, including sex, age, education, percent with AD dementia, or  
340 number of APOE-e4 alleles (**Supplementary Table 1**).

341

#### 342 ***Identification of disease-specific biclusters***

343 The stark heterogeneity exhibited at the level of the disease-relevant constellations  
344 strongly suggests that there may be other heterogeneous substructures contained within each.  
345 This possibility seems all the more likely due to the non-Gaussian distribution of subjects within  
346 each constellation. Additionally, the distribution of cases and controls appears to be different  
347 across constellations. Examining **Figure 1A**, it appears that the bias or shift in cases relative to  
348 controls does not fall in the same direction across constellations, thus requiring that the search  
349 for biclusters be carried out separately for each constellation. The results of our bicluster  
350 searches are shown in **Figure 3**. Each subplot corresponds to the search in a different  
351 constellation. Within each subplot the red curve corresponds to the signal-strength of the  
352 dominant bicluster within the data, referred to as a 'trace' in the Methods section. The black  
353 curves indicate the distribution of traces drawn from the null-hypothesis via a permutation-test.  
354 Generally speaking, a red trace that has either a high peak or a high average (relative to the  
355 distribution of black traces) indicates a statistically significant bicluster. For this data set we  
356 detected a statistically significant bicluster in constellation 1 (termed "bicluster 1";  $p_{avg}=0.002$ ,

357  $p_{\max}=0.008$ ) and constellation 2 (termed bicluster 2;  $p_{\text{avg}}=0.028$ ,  $p_{\max}=0.244$ ), but not in  
358 constellation 3 ( $p_{\text{avg}}=0.484$ ,  $p_{\max}=0.102$ ), which is the smallest of the 3 constellations. The shape  
359 of the traces conveys additional information about the structure of the biclusters. For example,  
360 the sharp peaks defining bicluster 1 in **Figure 3A** indicate clear cut points at which membership  
361 can be clearly delineated (i.e., a disease subtype). In contrast, the broader plateau of the trace  
362 in **Figure 3B** reflects a bicluster with “fuzzy” boundaries whereby there is a smoother continuum  
363 of membership in the bicluster. This is also reflected in the different pattern of p-values between  
364 biclusters (i.e., whether the peak or average of the trace tends to be more significant). Thus,  
365 examining the traces informs us whether disease heterogeneity takes the form of distinct  
366 subtypes, a continuous spectrum of risk, or some intermediate structure. After delineating and  
367 removing the dominant bicluster within constellations 1 and 2 (see Methods) we searched once  
368 again for any additional biclusters but did not find a second bicluster that was statistically  
369 significant in either constellation (**Supplementary Figure 3**). However, when searching for  
370 these secondary biclusters, the clearly non-Gaussian distribution of traces drawn from the label-  
371 shuffled null distribution (visualized as clumped strands among the black traces) suggests there  
372 is residual heterogeneity, but we are not powered to identify it in the current sample. The  
373 bicluster participants were not significantly different on demographic characteristics, including  
374 sex, age, education, or number of APOE-e4 alleles (**Supplementary Table 1**). **Supplementary**  
375 **Figure 4** presents a plot of the constellations shown in **Figure 1** with bicluster cases  
376 highlighted.

377

### 378 ***Gene set enrichment results of disease-specific biclusters***

379 Bicluster 1 was enriched for a number of gene sets, including those related to receptor  
380 activity, calcium and sodium ion transport, dendritic structure, GTPase activity, and regulation of  
381 the MAPK cascade. Bicluster 2 was enriched for gene sets related to the MHC protein complex,  
382 regulation of cell size, lipid transport, and tyrosine kinase activity. Both biclusters showed



383 enrichment for cell-projection morphogenesis, and synaptic transmission. **Figure 4** shows  
384 network plots of gene sets enriched in each bicluster. The gene sets are grouped by proportion  
385 of overlapping genes and labelled with dominant terms to illustrate the shared and unique  
386 functions associated with variants in each bicluster.

387       Although many of the enriched gene sets identified in this analysis relate to basic  
388 functions and thus do not appear to be AD-specific, it should be noted that the variants  
389 considered in these analyses were already selected for their association with AD. That is, the  
390 gene sets shown in **Figure 4** are those which are enriched in bicluster variants relative to the  
391 other AD-associated variants, rather than with respect to all other variants, including AD-  
392 nonspecific variants. Pathways that are typically found to be enriched for AD-associated  
393 variants constitute a background signal that is likely shared among many of the cases, including  
394 those that are not part of either bicluster.

395       There is also a degree of overlap in the gene sets enriched for each bicluster. However,  
396 the proportion of overlap seen at the level of gene sets is reduced when looking at the overlap  
397 among constituent genes or SNPs (**Supplementary Figure 5**). This indicates that the variants  
398 associated with each bicluster may have impacts that converge on similar downstream  
399 pathways, but the specific perturbations encompassed by each may not be the same.

#### 400 ***Disease-relevant constellations are evident in ADNI data.***

401       Using only SNPs common to both datasets, we re-calculated the top principal  
402 components of AD-associated variants in the UKB data. These loadings were used to project  
403 both the UKB and ADNI data into the same PC space. As seen in **Figure 5**, the grouping of  
404 UKB participants into 3 constellations was almost perfectly mirrored in the ADNI data, albeit with  
405 a slight shift in positioning. As in the UKB data, each constellation contained a mixture of AD  
406 cases and controls, reflecting disease-relevant but not disease-specific clustering.

407

408 ***Disease-specific biclusters replicate in ADNI***

409 Overall, we found evidence for significant replication of bicluster 1 in the ADNI data  
410 (global  $p=0.006$ ), and weaker yet significant replication of bicluster 2 (global  $p=0.048$ ). This  
411 pattern is consistent with the initial bicluster search, in which bicluster 1 seemed to have more  
412 distinct boundaries whereas the boundaries of bicluster 2 were more diffuse. The best label  
413 similarity ranking in comparison to the null distribution was achieved using 31.6% of the sample  
414 as nearest neighbors for bicluster 1 and using 27.3% of the sample as nearest neighbors for  
415 bicluster 2. **Figure 6** displays the labeling accuracy across a range of nearest neighbors. Note  
416 that the shape of these curves can be influenced by the structure of the bicluster. For example,  
417 a bicluster with clearly delineated boundaries may be less sensitive to altering the nearest  
418 neighbor parameter compared to one with blurrier boundaries. Consistent with this, we see high  
419 accuracy of label matching in bicluster 1 across a large range of nearest neighbor fractions,  
420 ranging from about 17% up to our maximum threshold of 50%. In contrast, the accuracy of label  
421 matching bicluster 2 is more sensitive to this parameter choice, with high accuracy using  
422 nearest neighbor fractions from about 26% to 36%. Importantly, these plots provide a range of  
423 sensible values one can use to label new data.

424

425 ***Cognitive and biomarker trajectories across bicluster groups in ADNI MCI participants***

426 We found evidence for differential cognitive and biomarker trajectories across bicluster  
427 groups in a separate set of ADNI individuals diagnosed with MCI. Individuals assigned to  
428 bicluster 2 demonstrated significantly greater decline on the PACCC compared to the non-  
429 bicluster ( $\beta=-0.40$ ,  $t$ -value=-2.29,  $p=0.022$ ) and bicluster 1 groups ( $\beta=-0.60$ ,  $t$ -value=2.92,  
430  $p=0.004$ ) (**Figure 7A**). Bicluster 1 demonstrated a somewhat greater increase in florbetapir over  
431 time compared to the non-bicluster group, but this difference was not significant ( $\beta=0.32$ ,  $t$ -  
432 value=1.52,  $p=0.129$ ) (**Figure 7B**). Bicluster 2 demonstrated a significantly greater increase of

433 CSF p-tau over time compared to the non-bicluster group ( $\beta=0.45$ ,  $t\text{-value}=2.05$ ,  $p=0.041$ )  
434 (Figure 7C).

## 435 DISCUSSION

436 Genetic analyses suggest the etiology of AD is multifactorial [55], but the extent to which  
437 the composition of genetic influences on AD risk varies across individuals and whether it is  
438 replicable across different data sets has been unclear. We identified evidence for several  
439 subsets of individuals that exhibit different patterns of genetically-mediated vulnerability to the  
440 risk for AD. Among AD-associated SNPs, there is clear heterogeneity across individuals, and  
441 this heterogeneity seemed to follow a hierarchical structure. Importantly, this structure was  
442 observed in two independent datasets, indicating that it is not sample-specific, but rather  
443 appears to be a generalizable feature of AD genetic risk.

444 The first level of heterogeneity emerged as 3 clusters, which we termed “constellations”,  
445 from a PCA of the UKB data confined to variants associated with AD based on prior GWAS.  
446 Similar analyses (i.e., PCA) are often used to identify ancestry-related population structure [56],  
447 but we found that the constellations only emerged when examining AD-associated SNPs and  
448 individuals from each constellation were highly intermixed when analyzed using all SNPs or a  
449 random selection of SNPs. These results indicate that the constellations do not simply reflect  
450 ancestry-related substructure. We further found a spike in loadings on the first PC used to  
451 define the constellations within the region of chromosome 17q21.31. This is a region of strong  
452 extended LD driven by a 900-kb inversion polymorphism surrounding the *MAPT* gene that  
453 defines two haplotypes, H1 and H2, with the H1 haplotype further dividing into several sub-  
454 haplotypes [53, 54]. The *MAPT* gene codes for tau, the primary component of neurofibrillary  
455 tangles, so it is clearly relevant to AD. However, as a risk gene it is more strongly associated  
456 with primary tauopathies, and investigations of the *MAPT* locus and the H1/H2 haplotypes find  
457 inconsistent evidence for a specific association with AD risk [54, 57]. A stratified GWAS of AD

458 by Strickland, et al. [58] found that there were several variants with haplotype-dependent  
459 associations, which may explain these inconsistencies.

460         It may be that our constellations at least partially reflect the H1/H2 haplotypes (along  
461 with H1 sub-haplotypes) defined by variation in this region. Using the H1/H2 haplotype-tagging  
462 SNP rs8070723, Strickland, et al. [58] found that 39% of their sample were H2 carriers (H1H2 +  
463 H2H2). Based on the genotypes of individuals across our constellations, individuals in  
464 constellations 2 and 3 (who largely have genotypes corresponding to H1H2 and H2H2 carriers,  
465 respectively) accounted for an identical 39% in our sample as well. A block of complete LD in  
466 this region and minimal recombination may further explain the surprising degree of separation  
467 between these constellations [53, 54, 59]. We note an important aspect of these constellations:  
468 they contain both cases and controls and therefore membership is not associated with  
469 increased AD risk. However, the overall distribution of cases and controls is different in each.  
470 Taken together with results from Strickland, et al. [58] showing haplotype-dependent  
471 associations, we refer to the constellation structure are “disease-relevant” in that it may alter the  
472 relative importance of other risk variants (e.g., by associating with protection or vulnerability to  
473 these other risk variants) but is not directly associated with increased AD risk.

474         Cases belonging to the disease-specific biclusters harbored distinct genetic signatures  
475 when compared to other AD cases and may therefore reflect “subtypes” of AD genetic risk. The  
476 variants defining the two disease-specific biclusters were enriched for both similar and unique  
477 pathways when compared to each other and the disease-relevant constellations in which they  
478 were found. It is important to re-emphasize the fact that these analyses were restricted to AD-  
479 associated variants, which are enriched for lipid processing, cholesterol transport, amyloid  
480 precursor (APP) processing and A $\beta$  formation, tau protein binding, and immune response [12].  
481 The bicluster-associated pathways reported here thus exist within this broader context of other  
482 AD-associated pathways. Bicluster 1 was enriched for gene sets involved in regulating the  
483 MAPK cascade, which may contribute to AD progression in several cell-type-dependent ways,

484 such as increasing neuroinflammation, promoting neurofibrillary tangle formation, and  
485 depressing synaptic plasticity [60]. Bicluster 2 demonstrated enrichment in a cluster of gene  
486 sets related to the major histocompatibility complex (MHC), which is consistent with findings that  
487 implicate microglia-mediated immune response as a key player in AD [12, 15, 61, 62]. Both  
488 biclusters – and bicluster 1 in particular – were enriched for gene sets related to synaptic  
489 signaling and signal transduction. Similar enrichment was found among newly prioritized genes  
490 in the most recent largescale GWAS of AD [15]. We also found both biclusters were strongly  
491 enriched for a number of pathways that interact to influence the morphogenesis and outgrowth  
492 of neurons. This includes gene sets related to the cellular components themselves (e.g., cell  
493 bodies, dendritic spines, synapses, axonal projections, cytoskeletal components), but also  
494 factors that modulate their development such as polymerization,  $\beta$ -catenin binding, and GTPase  
495 activity [63-67]. Disruption of these processes may leave structures more vulnerable to insult  
496 resulting in, for example, greater A $\beta$ -mediated synapse loss. APP (the precursor of A $\beta$ ) is  
497 important for regulating axonal and synaptic growth through influencing cytoskeletal remodelling  
498 [68] and tau plays a critical role in stabilizing microtubules [69, 70]. It is therefore possible that  
499 the hallmark pathologies of AD, i.e., A $\beta$  and tau protein build-up, may exert impacts through  
500 toxic effects, but also reflect a loss of function that affects cell structure and growth. Further  
501 work is needed to comprehensively characterize the impacts of these bicluster-associated  
502 variants.

503 We found that both the disease-relevant constellation structure and the disease-specific  
504 biclusters replicated in the ADNI sample. There are several important differences between the  
505 ADNI and UKB cohorts. First, although analyses were restricted to individuals of European  
506 ancestry, the UKB was further restricted to a White British subset, so the ADNI data may  
507 contain relatively more ancestry-related heterogeneity. Second, the studies used different  
508 genotyping chips (and ADNI genotypes were obtained with 3 different chips across phases) and  
509 imputation panels. Third, ADNI participants were recruited with the intent of mirroring a clinical

510 trial population focused on Alzheimer's disease, whereas the UKB attempted to recruit a  
511 broader base of individuals to study multiple outcomes. Replicable structure across these  
512 datasets is thus unlikely to reflect dataset-specific or artifactual confounds that should  
513 differentially affect these datasets.

514 Supporting the relevance of the genetic structure found here, we found that a separate  
515 set of ADNI participants diagnosed with MCI assigned to different bicluster groups  
516 demonstrated differential cognitive and biomarker trajectories. Individuals with genetic  
517 signatures resembling bicluster 2 exhibited greater accumulation of p-tau and a corresponding  
518 steeper decline in cognitive performance over time. The variants associated with this genetic  
519 subtype may contribute to a more aggressive form of AD. We previously found that genetic risk  
520 affecting different biological pathways can preferentially relate to amyloid or tau accumulation  
521 [71]. Thus, it may be that the balance of pathological accumulation is shifted towards p-tau in  
522 these individuals, and p-tau has been shown to be more closely linked to subsequent  
523 neurodegeneration [72] and cognitive decline [73] compared to amyloid. The genetic  
524 heterogeneity observed here may impact not only level but distribution of pathology, which may  
525 be particularly important in the case of tau [3, 7, 74]. Heterogeneity in genetic risk may also  
526 relate to comorbid conditions that exacerbate AD progression. It is therefore clear that more  
527 work is needed to characterize the downstream impacts of genetic heterogeneity on disease  
528 outcomes.

529 Our findings may indicate that there are likely multiple genetically-mediated pathologies  
530 underlying AD that converge on common clinical manifestations. This is not a unique  
531 phenomenon. For example, Charcot-Marie-Tooth disease involves a common set of clinical  
532 symptoms that can arise from separate genetic origins [75]. We did find evidence for overlap in  
533 the pathways enriched across the different constellations and biclusters identified in our  
534 analysis, and this may indicate where convergence begins. However, there are a few caveats.  
535 First, receiving the same clinical diagnosis does not rule out the possibility of meaningful

536 differences across cases. There are atypical forms of clinically-defined AD [76-79] and there is  
537 growing evidence for several biological subtypes of AD characterized by distinct patterns of  
538 pathological spread or neurodegeneration [2, 3, 5]. Further work is needed to determine  
539 whether the genetic heterogeneity identified here is associated with more subtle phenotypic  
540 differences. Second, the degree of overlap among pathways that we observed is driven by  
541 SNPs that were assigned to genes and pathways, but not all SNPs were assigned to genes.  
542 Many AD-associated SNPs are located in non-coding regions, so the overlap of enriched gene  
543 sets ignores substantial numbers of SNPs uniquely associated with each. Differences between  
544 constellations and biclusters may arise from the functional effects of the variants themselves,  
545 including those not assigned to genes. Third, it is possible that the subsets of cases identified by  
546 the biclustering represent misdiagnosed AD (assuming the underlying disease cause is driven  
547 by partially distinct genetic factors). Identifying misdiagnoses or undiagnosed cases in a  
548 heterogeneous population is a potential use-case of the method. However, the number of cases  
549 in each bicluster would require a higher rate of misdiagnosis than one might expect, especially  
550 in ADNI where other dementias were specifically excluded.

## 551 **CONCLUSIONS**

552 In sum, we found evidence of a hierarchical structure underlying heterogeneity in the  
553 genetic risk of AD. The disease-relevant constellation structure is driven to a large degree by an  
554 extended region of LD on chromosome 17q.21, with constellations potentially reflecting *MAPT*  
555 haplotypes. Membership in a given constellation did not directly increase risk for AD, but may  
556 alter the relative importance of other genetic risk variants for AD. On the other hand, the  
557 biclusters may be considered disease subtypes with distinct genetic signatures compared to the  
558 broader population of AD cases, which may have important implications for treatment efforts.  
559 Despite all cases presenting with a common clinical syndrome, the etiology and path taken to  
560 clinical manifestation may vary across patients. The differential cognitive and biomarker

561 trajectories between genetic subtypes provides some evidence that this heterogeneity has  
562 consequences for downstream disease processes. Identifying subtypes of AD could facilitate  
563 precision medicine approaches that tailor treatment strategies to the individual for increased  
564 effectiveness.



## ***Acknowledgments***

We thank the anonymous reviewers whose insightful comments and suggestions helped improve the manuscript. 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](http://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](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***

Jeremy Elman is an Editorial Board Member of this journal but was not involved in the peer-review process of this article nor had access to any information regarding its peer-review. All other authors have no conflict of interest to report.

## REFERENCES

- [1] McKhann GM, Knopman DS, Chertkow H, Hyman BT, Jack CR, Jr., Kawas CH, Klunk WE, Koroshetz WJ, Manly JJ, Mayeux R, Mohs RC, Morris JC, Rossor MN, Scheltens P, Carrillo MC, Thies B, Weintraub S, Phelps CH (2011) 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.
- [2] Murray ME, Graff-Radford NR, Ross OA, Petersen RC, Duara R, Dickson DW (2011) Neuropathologically defined subtypes of Alzheimer's disease with distinct clinical characteristics: a retrospective study. *Lancet Neurol* **10**, 785-796.
- [3] Vogel JW, Young AL, Oxtoby NP, Smith R, Ossenkuppele R, Strandberg OT, La Joie R, Aksman LM, Grothe MJ, Iturria-Medina Y, Alzheimer's Disease Neuroimaging I, Pontecorvo MJ, Devous MD, Rabinovici GD, Alexander DC, Lyoo CH, Evans AC, Hansson O (2021) Four distinct trajectories of tau deposition identified in Alzheimer's disease. *Nat Med* **27**, 871-881.
- [4] Young AL, Marinescu RV, Oxtoby NP, Bocchetta M, Yong K, Firth NC, Cash DM, Thomas DL, Dick KM, Cardoso J, van Swieten J, Borroni B, Galimberti D, Masellis M, Tartaglia MC, Rowe JB, Graff C, Tagliavini F, Frisoni GB, Laforce R, Jr., Finger E, de Mendonca A, Sorbi S, Warren JD, Crutch S, Fox NC, Ourselin S, Schott JM, Rohrer JD, Alexander DC, Genetic FTDI, Alzheimer's Disease Neuroimaging I (2018) Uncovering the heterogeneity and temporal complexity of neurodegenerative diseases with Subtype and Stage Inference. *Nat Commun* **9**, 4273.
- [5] Ferreira D, Nordberg A, Westman E (2020) Biological subtypes of Alzheimer disease: A systematic review and meta-analysis. *Neurology* **94**, 436-448.
- [6] Ferreira D, Verhagen C, Hernandez-Cabrera JA, Cavallin L, Guo CJ, Ekman U, Muehlboeck JS, Simmons A, Barroso J, Wahlund LO, Westman E (2017) Distinct subtypes of Alzheimer's disease based on patterns of brain atrophy: longitudinal trajectories and clinical applications. *Sci Rep* **7**, 46263.
- [7] Ossenkuppele R, Schonhaut DR, Scholl M, Lockhart SN, Ayakta N, Baker SL, O'Neil JP, Janabi M, Lazaris A, Cantwell A, Vogel J, Santos M, Miller ZA, Bettcher BM, Vessel KA, Kramer JH, Gorno-Tempini ML, Miller BL, Jagust WJ, Rabinovici GD (2016) Tau PET patterns mirror clinical and neuroanatomical variability in Alzheimer's disease. *Brain* **139**, 1551-1567.
- [8] Tijms BM, Gobom J, Reus L, Jansen I, Hong S, Dobricic V, Kilpert F, ten Kate M, Barkhof F, Tsolaki M, Verhey FRJ, Popp J, Martinez-Lage P, Vandenberghe R, Lleó A, Molinuevo JL, Engelborghs S, Bertram L, Lovestone S, Streffer J, Vos S, Bos I, The Alzheimer's Disease Neuroimaging I, Blennow K, Scheltens P, Teunissen CE, Zetterberg H, Visser PJ (2020) Pathophysiological subtypes of Alzheimer's disease based on cerebrospinal fluid proteomics. *Brain*.
- [9] Gatz M, Reynolds CA, Fratiglioni L, Johansson B, Mortimer JA, Berg S, Fiske A, Pedersen NL (2006) Role of genes and environments for explaining Alzheimer disease. *Arch Gen Psychiatry* **63**, 168-174.
- [10] Corder EH, Saunders AM, Strittmatter WJ, Schmechel DE, Gaskell PC, Small GW, Roses AD, Haines JL, Pericak-Vance MA (1993) Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science* **261**, 921-923.
- [11] Harold D, Abraham R, Hollingworth P, Sims R, Gerrish A, Hamshere ML, Pahwa JS, Moskvina V, Dowzell K, Williams A, Jones N, Thomas C, Stretton A, Morgan AR, Lovestone S, Powell J, Proitsi P, Lupton MK, Brayne C, Rubinsztein DC, Gill M, Lawlor B, Lynch A, Morgan K, Brown KS, Passmore PA, Craig D, McGuinness B, Todd S,

- Holmes C, Mann D, Smith AD, Love S, Kehoe PG, Hardy J, Mead S, Fox N, Rossor M, Collinge J, Maier W, Jessen F, Schurmann B, Heun R, van den Bussche H, Heuser I, Kornhuber J, Wiltfang J, Dichgans M, Frolich L, Hampel H, Hull M, Rujescu D, Goate AM, Kauwe JS, Cruchaga C, Nowotny P, Morris JC, Mayo K, Sleegers K, Bettens K, Engelborghs S, De Deyn PP, Van Broeckhoven C, Livingston G, Bass NJ, Gurling H, McQuillin A, Gwilliam R, Deloukas P, Al-Chalabi A, Shaw CE, Tsolaki M, Singleton AB, Guerreiro R, Muhleisen TW, Nothen MM, Moebus S, Jockel KH, Klopp N, Wichmann HE, Carrasquillo MM, Pankratz VS, Younkin SG, Holmans PA, O'Donovan M, Owen MJ, Williams J (2009) Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease. *Nat Genet* **41**, 1088-1093.
- [12] Kunkle BW, Grenier-Boley B, Sims R, Bis JC, Damotte V, Naj AC, Boland A, Vronskaya M, van der Lee SJ, Amlie-Wolf A, Bellenguez C, Frizatti A, Chouraki V, Martin ER, Sleegers K, Badarinarayan N, Jakobsdottir J, Hamilton-Nelson KL, Moreno-Grau S, Ojano T, Raybould R, Chen Y, Kuzma AB, Hiltunen M, Morgan T, Ahmad S, Vardarajan BN, Epelbaum J, Hoffmann P, Boada M, Beecham GW, Garnier JG, Harold D, Fitzpatrick AL, Valladares O, Moutet ML, Gerrish A, Smith AV, Qu L, Bacq D, Denning N, Jian X, Zhao Y, Del Zompo M, Fox NC, Choi SH, Mateo I, Hughes JT, Adams HH, Malamon J, Sanchez-Garcia F, Patel Y, Brody JA, Dombroski BA, Naranjo MCD, Daniilidou M, Eiriksdottir G, Mukherjee S, Wallon D, Uphill J, Aspelund T, Cantwell LB, Garzia F, Galimberti D, Hofer E, Butkiewicz M, Fin B, Scarpini E, Sarnowski C, Bush WS, Meslage S, Kornhuber J, White CC, Song Y, Barber RC, Engelborghs S, Sordon S, Voijnovic D, Adams PM, Vandenbergh R, Mayhaus M, Cupples LA, Albert MS, De Deyn PP, Gu W, Himali JJ, Beekly D, Squassina A, Hartmann AM, Orellana A, Blacker D, Rodriguez-Rodriguez E, Lovestone S, Garcia ME, Doody RS, Munoz-Fernandez C, Sussams R, Lin H, Fairchild TJ, Benito YA, Holmes C, Karamujic-Comic H, Frosch MP, Thonberg H, Maier W, Roshchupkin G, Ghetti B, Giedraitis V, Kawalia A, Li S, Huebinger RM, Kilander L, Moebus S, Hernandez I, Kamboh MI, Brundin R, Turton J, Yang Q, Katz MJ, Concaro L, Lord J, Beiser AS, Keene CD, Helisalmi S, Kloszewska I, Kukull WA, Koivisto AM, Lynch A, Tarraga L, Larson EB, Haapasalo A, Lawlor B, Mosley TH, Lipton RB, Solfrizzi V, Gill M, Longstreth WT, Jr., Montine TJ, Frisardi V, Diez-Fairen M, Rivadeneira F, Petersen RC, Deramecourt V, Alvarez I, Salani F, Ciarabella A, Boerwinkle E, Reiman EM, Fievet N, Rotter JI, Reisch JS, Hanon O, Cupidi C, Andre Uitterlinden AG, Royall DR, Dufouil C, Maletta RG, de Rojas I, Sano M, Brice A, Cecchetti R, George-Hyslop PS, Ritchie K, Tsolaki M, Tsuang DW, Dubois B, Craig D, Wu CK, Soininen H, Avramidou D, Albin RL, Fratiglioni L, Germanou A, Apostolova LG, Keller L, Koutroumani M, Arnold SE, Panza F, Gkatzima O, Asthana S, Hannequin D, Whitehead P, Atwood CS, Caffarra P, Hampel H, Quintela I, Carracedo A, Lannfelt L, Rubinsztein DC, Barnes LL, Pasquier F, Frolich L, Barral S, McGuinness B, Beach TG, Johnston JA, Becker JT, Passmore P, Bigio EH, Schott JM, Bird TD, Warren JD, Boeve BF, Lupton MK, Bowen JD, Proitsi P, Boxer A, Powell JF, Burke JR, Kauwe JSK, Burns JM, Mancuso M, Buxbaum JD, Bonuccelli U, Cairns NJ, McQuillin A, Cao C, Livingston G, Carlson CS, Bass NJ, Carlsson CM, Hardy J, Carney RM, Bras J, Carrasquillo MM, Guerreiro R, Allen M, Chui HC, Fisher E, Masullo C, Crocco EA, DeCarli C, Bisceglia G, Dick M, Ma L, Duara R, Graff-Radford NR, Evans DA, Hodges A, Faber KM, Scherer M, Fallon KB, Riemenschneider M, Fardo DW, Heun R, Farlow MR, Kolsch H, Ferris S, Leber M, Foroud TM, Heuser I, Galasko DR, Giegling I, Gearing M, Hull M, Geschwind DH, Gilbert JR, Morris J, Green RC, Mayo K, Growdon JH, Feulner T, Hamilton RL, Harrell LE, Driche D, Honig LS, Cushion TD, Huentelman MJ, Hollingworth P, Hulette CM, Hyman BT, Marshall R, Jarvik GP, Meggy A, Abner E, Menzies GE, Jin LW, Leonenko G, Real LM, Jun GR, Baldwin CT, Grozeva D, Karydas A, Russo G, Kaye JA, Kim R, Jessen F, Kowall NW, Vellas B, Kramer JH, Vardy E, LaFerla FM, Jockel KH,

- Lah JJ, Dichgans M, Leverenz JB, Mann D, Levey AI, Pickering-Brown S, Lieberman AP, Klopp N, Lunetta KL, Wichmann HE, Lyketsos CG, Morgan K, Marson DC, Brown K, Martiniuk F, Medway C, Mash DC, Nothen MM, Masliah E, Hooper NM, McCormick WC, Daniele A, McCurry SM, Bayer A, McDavid AN, Gallacher J, McKee AC, van den Bussche H, Mesulam M, Brayne C, Miller BL, Riedel-Heller S, Miller CA, Miller JW, Al-Chalabi A, Morris JC, Shaw CE, Myers AJ, Wiltfang J, O'Bryant S, Olichney JM, Alvarez V, Parisi JE, Singleton AB, Paulson HL, Collinge J, Perry WR, Mead S, Peskind E, Cribbs DH, Rossor M, Pierce A, Ryan NS, Poon WW, Nacmias B, Potter H, Sorbi S, Quinn JF, Sacchinelli E, Raj A, Spalletta G, Raskind M, Caltagirone C, Bossu P, Orfei MD, Reisberg B, Clarke R, Reitz C, Smith AD, Ringman JM, Warden D, Roberson ED, Wilcock G, Rogaeva E, Bruni AC, Rosen HJ, Gallo M, Rosenberg RN, Ben-Shlomo Y, Sager MA, Mecocci P, Saykin AJ, Pastor P, Cuccaro ML, Vance JM, Schneider JA, Schneider LS, Slifer S, Seeley WW, Smith AG, Sonnen JA, Spina S, Stern RA, Swerdlow RH, Tang M, Tanzi RE, Trojanowski JQ, Troncoso JC, Van Deerlin VM, Van Eldik LJ, Vinters HV, Vonsattel JP, Weintraub S, Welsh-Bohmer KA, Wilhelmsen KC, Williamson J, Wingo TS, Woltjer RL, Wright CB, Yu CE, Yu L, Saba Y, Pilotto A, Bullido MJ, Peters O, Crane PK, Bennett D, Bosco P, Coto E, Boccardi V, De Jager PL, Lleo A, Warner N, Lopez OL, Ingelsson M, Deloukas P, Cruchaga C, Graff C, Gwilliam R, Fornage M, Goate AM, Sanchez-Juan P, Kehoe PG, Amin N, Ertekin-Taner N, Berr C, Debette S, Love S, Launer LJ, Younkin SG, Dartigues JF, Corcoran C, Ikram MA, Dickson DW, Nicolas G, Champion D, Tschanz J, Schmidt H, Hakonarson H, Clarimon J, Munger R, Schmidt R, Farrer LA, Van Broeckhoven C, M COD, DeStefano AL, Jones L, Haines JL, Deleuze JF, Owen MJ, Gudnason V, Mayeux R, Escott-Price V, Psaty BM, Ramirez A, Wang LS, Ruiz A, van Duijn CM, Holmans PA, Seshadri S, Williams J, Amouyel P, Schellenberg GD, Lambert JC, Pericak-Vance MA, Alzheimer Disease Genetics C, European Alzheimer's Disease I, Cohorts for H, Aging Research in Genomic Epidemiology C, Genetic, Environmental Risk in Ad/Defining Genetic P, Environmental Risk for Alzheimer's Disease C (2019) Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Abeta, tau, immunity and lipid processing. *Nat Genet* **51**, 414-430.
- [13] Lambert JC, Ibrahim-Verbaas CA, Harold D, Naj AC, Sims R, Bellenguez C, DeStefano AL, Bis JC, Beecham GW, Grenier-Boley B, Russo G, Thornton-Wells TA, Jones N, Smith AV, Chouraki V, Thomas C, Ikram MA, Zelenika D, Vardarajan BN, Kamatani Y, Lin CF, Gerrish A, Schmidt H, Kunkle B, Dunstan ML, Ruiz A, Bihoreau MT, Choi SH, Reitz C, Pasquier F, Cruchaga C, Craig D, Amin N, Berr C, Lopez OL, De Jager PL, Deramecourt V, Johnston JA, Evans D, Lovestone S, Letenneur L, Moron FJ, Rubinsztein DC, Eiriksdottir G, Sleegers K, Goate AM, Fievet N, Huentelman MW, Gill M, Brown K, Kamboh MI, Keller L, Barberger-Gateau P, McGuinness B, Larson EB, Green R, Myers AJ, Dufouil C, Todd S, Wallon D, Love S, Rogaeva E, Gallacher J, St George-Hyslop P, Clarimon J, Lleo A, Bayer A, Tsuang DW, Yu L, Tsolaki M, Bossu P, Spalletta G, Proitsi P, Collinge J, Sorbi S, Sanchez-Garcia F, Fox NC, Hardy J, Deniz Naranjo MC, Bosco P, Clarke R, Brayne C, Galimberti D, Mancuso M, Matthews F, European Alzheimer's Disease I, Genetic, Environmental Risk in Alzheimer's D, Alzheimer's Disease Genetic C, Cohorts for H, Aging Research in Genomic E, Moebus S, Mecocci P, Del Zompo M, Maier W, Hampel H, Pilotto A, Bullido M, Panza F, Caffarra P, Nacmias B, Gilbert JR, Mayhaus M, Lannfelt L, Hakonarson H, Pichler S, Carrasquillo MM, Ingelsson M, Beekly D, Alvarez V, Zou F, Valladares O, Younkin SG, Coto E, Hamilton-Nelson KL, Gu W, Razquin C, Pastor P, Mateo I, Owen MJ, Faber KM, Jonsson PV, Combarros O, O'Donovan MC, Cantwell LB, Soininen H, Blacker D, Mead S, Mosley TH, Jr., Bennett DA, Harris TB, Fratiglioni L, Holmes C, de Bruijn RF, Passmore P, Montine TJ, Bettens K, Rotter JI, Brice A, Morgan K, Foroud TM, Kukull

- WA, Hannequin D, Powell JF, Nalls MA, Ritchie K, Lunetta KL, Kauwe JS, Boerwinkle E, Riemenschneider M, Boada M, Hiltunen M, Martin ER, Schmidt R, Rujescu D, Wang LS, Dartigues JF, Mayeux R, Tzourio C, Hofman A, Nothen MM, Graff C, Psaty BM, Jones L, Haines JL, Holmans PA, Lathrop M, Pericak-Vance MA, Launer LJ, Farrer LA, van Duijn CM, Van Broeckhoven C, Moskvina V, Seshadri S, Williams J, Schellenberg GD, Amouyel P (2013) Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat Genet* **45**, 1452-1458.
- [14] Jansen IE, Savage JE, Watanabe K, Bryois J, Williams DM, Steinberg S, Sealock J, Karlsson IK, Hagg S, Athanasiu L, Voyle N, Proitsi P, Witoelar A, Stringer S, Aarsland D, Almdahl IS, Andersen F, Bergh S, Bettella F, Bjornsson S, Braekhus A, Brathen G, de Leeuw C, Desikan RS, Djurovic S, Dumitrescu L, Fladby T, Hohman TJ, Jonsson PV, Kiddle SJ, Rongve A, Saltvedt I, Sando SB, Selbaek G, Shoai M, Skene NG, Snaedal J, Stordal E, Ulstein ID, Wang Y, White LR, Hardy J, Hjerling-Leffler J, Sullivan PF, van der Flier WM, Dobson R, Davis LK, Stefansson H, Stefansson K, Pedersen NL, Ripke S, Andreassen OA, Posthuma D (2019) Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. *Nat Genet* **51**, 404-413.
- [15] Bellenguez C, Kucukali F, Jansen IE, Kleindam L, Moreno-Grau S, Amin N, Naj AC, Campos-Martin R, Grenier-Boley B, Andrade V, Holmans PA, Boland A, Damotte V, van der Lee SJ, Costa MR, Kuulasmaa T, Yang Q, de Rojas I, Bis JC, Yaqub A, Prokic I, Chapuis J, Ahmad S, Giedraitis V, Aarsland D, Garcia-Gonzalez P, Abdelnour C, Alarcon-Martin E, Alcolea D, Alegret M, Alvarez I, Alvarez V, Armstrong NJ, Tsolaki A, Antunez C, Appollonio I, Arcaro M, Archetti S, Pastor AA, Arosio B, Athanasiu L, Bailly H, Banaj N, Baquero M, Barral S, Beiser A, Pastor AB, Below JE, Benček P, Benussi L, Berr C, Besse C, Bessi V, Binetti G, Bizarro A, Blesa R, Boada M, Boerwinkle E, Borroni B, Boschi S, Bossu P, Brathen G, Bressler J, Bresner C, Brodaty H, Brookes KJ, Brusco LI, Buiza-Rueda D, Burger K, Burholt V, Bush WS, Calero M, Cantwell LB, Chene G, Chung J, Cuccaro ML, Carracedo A, Cecchetti R, Cervera-Carles L, Charbonnier C, Chen HH, Chillotti C, Ciccone S, Claassen J, Clark C, Conti E, Corma-Gomez A, Costantini E, Custodero C, Daian D, Dalmaso MC, Daniele A, Dardiotis E, Dartigues JF, de Deyn PP, de Paiva Lopes K, de Witte LD, Debette S, Deckert J, Del Ser T, Denning N, DeStefano A, Dichgans M, Diehl-Schmid J, Diez-Fairen M, Rossi PD, Djurovic S, Duron E, Duzel E, Dufouil C, Eiriksdottir G, Engelborghs S, Escott-Price V, Espinosa A, Ewers M, Faber KM, Fabrizio T, Nielsen SF, Fardo DW, Farotti L, Fenoglio C, Fernandez-Fuertes M, Ferrari R, Ferreira CB, Ferri E, Fin B, Fischer P, Fladby T, Fliessbach K, Fongang B, Fornage M, Fortea J, Foroud TM, Fostinelli S, Fox NC, Franco-Macias E, Bullido MJ, Frank-Garcia A, Froelich L, Fulton-Howard B, Galimberti D, Garcia-Alberca JM, Garcia-Gonzalez P, Garcia-Madrona S, Garcia-Ribas G, Ghidoni R, Giegling I, Giorgio G, Goate AM, Goldhardt O, Gomez-Fonseca D, Gonzalez-Perez A, Graff C, Grande G, Green E, Grimmer T, Grunblatt E, Grunin M, Gudnason V, Guetta-Baranes T, Haapasalo A, Hadjigeorgiou G, Haines JL, Hamilton-Nelson KL, Hampel H, Hanon O, Hardy J, Hartmann AM, Hausner L, Harwood J, Heilmann-Heimbach S, Helisalmi S, Heneka MT, Hernandez I, Herrmann MJ, Hoffmann P, Holmes C, Holstege H, Vilas RH, Hulsman M, Humphrey J, Biessels GJ, Jian X, Johansson C, Jun GR, Kastumata Y, Kauwe J, Kehoe PG, Kilander L, Stahlbom AK, Kivipelto M, Koivisto A, Kornhuber J, Kosmidis MH, Kukull WA, Kuksa PP, Kunkle BW, Kuzma AB, Lage C, Laukka EJ, Launer L, Lauria A, Lee CY, Lehtisalo J, Lerch O, Lleo A, Longstreth W, Jr., Lopez O, de Munain AL, Love S, Lowemark M, Luckcuck L, Lunetta KL, Ma Y, Macias J, MacLeod CA, Maier W, Mangialasche F, Spallazzi M, Marquie M, Marshall R, Martin ER, Montes AM, Rodriguez CM, Masullo C, Mayeux R, Mead S, Mecocci P, Medina M, Meggy A, Mehrabian S, Mendoza S, Menendez-Gonzalez M, Mir P, Moebus S, Mol M, Molina-Porcel L, Montreal L, Morelli L, Moreno F, Morgan K, Mosley T,

- Nothen MM, Muchnik C, Mukherjee S, Nacmias B, Ngandu T, Nicolas G, Nordestgaard BG, Olaso R, Orellana A, Orsini M, Ortega G, Padovani A, Paolo C, Papenberg G, Parnetti L, Pasquier F, Pastor P, Peloso G, Perez-Cordon A, Perez-Tur J, Pericard P, Peters O, Pijnenburg YAL, Pineda JA, Pinol-Ripoll G, Pisanu C, Polak T, Popp J, Posthuma D, Priller J, Puerta R, Quenez O, Quintela I, Thomassen JQ, Rabano A, Rainero I, Rajabli F, Ramakers I, Real LM, Reinders MJT, Reitz C, Reyes-Dumeyer D, Ridge P, Riedel-Heller S, Riederer P, Roberto N, Rodriguez-Rodriguez E, Rongve A, Allende IR, Rosende-Roca M, Royo JL, Rubino E, Rujescu D, Saez ME, Sakka P, Saltvedt I, Sanabria A, Sanchez-Arjona MB, Sanchez-Garcia F, Juan PS, Sanchez-Valle R, Sando SB, Sarnowski C, Satizabal CL, Scamosci M, Scarmeas N, Scarpini E, Scheltens P, Scherbaum N, Scherer M, Schmid M, Schneider A, Schott JM, Selbaek G, Seripa D, Serrano M, Sha J, Shadrin AA, Skrobot O, Slifer S, Sniijders GJL, Soininen H, Solfrizzi V, Solomon A, Song Y, Sorbi S, Sotolongo-Grau O, Spalletta G, Spottke A, Squassina A, Stordal E, Tartan JP, Tarraga L, Tesi N, Thalamuthu A, Thomas T, Tosto G, Traykov L, Tremolizzo L, Tybjaerg-Hansen A, Uitterlinden A, Ullgren A, Ulstein I, Valero S, Valladares O, Broeckhoven CV, Vance J, Vardarajan BN, van der Lugt A, Dongen JV, van Rooij J, van Swieten J, Vandenberghe R, Verhey F, Vidal JS, Vogelgsang J, Vyhnaek M, Wagner M, Wallon D, Wang LS, Wang R, Weinhold L, Wiltfang J, Windle G, Woods B, Yannakoulia M, Zare H, Zhao Y, Zhang X, Zhu C, Zulaica M, Eadb, Gr@Ace, Degesco, Eadi, Gerad, Demgene, FinnGen, Adgc, Charge, Farrer LA, Psaty BM, Ghanbari M, Raj T, Sachdev P, Mather K, Jessen F, Ikram MA, de Mendonca A, Hort J, Tsolaki M, Pericak-Vance MA, Amouyel P, Williams J, Frikke-Schmidt R, Clarimon J, Deleuze JF, Rossi G, Seshadri S, Andreassen OA, Ingelsson M, Hiltunen M, Sleegers K, Schellenberg GD, van Duijn CM, Sims R, van der Flier WM, Ruiz A, Ramirez A, Lambert JC (2022) New insights into the genetic etiology of Alzheimer's disease and related dementias. *Nat Genet* **54**, 412-436.
- [16] Escott-Price V, Sims R, Bannister C, Harold D, Vronskaya M, Majounie E, Badarinarayan N, Gerad/Perades, consortia I, Morgan K, Passmore P, Holmes C, Powell J, Brayne C, Gill M, Mead S, Goate A, Cruchaga C, Lambert JC, van Duijn C, Maier W, Ramirez A, Holmans P, Jones L, Hardy J, Seshadri S, Schellenberg GD, Amouyel P, Williams J (2015) Common polygenic variation enhances risk prediction for Alzheimer's disease. *Brain* **138**, 3673-3684.
- [17] Jack CR, Jr., Bennett DA, Blennow K, Carrillo MC, Dunn B, Haeberlein SB, Holtzman DM, Jagust W, Jessen F, Karlawish J, Liu E, Molinuevo JL, Montine T, Phelps C, Rankin KP, Rowe CC, Scheltens P, Siemers E, Snyder HM, Sperling R, Contributors (2018) NIA-AA Research Framework: Toward a biological definition of Alzheimer's disease. *Alzheimers Dement* **14**, 535-562.
- [18] Dubois B, Villain N, Frisoni GB, Rabinovici GD, Sabbagh M, Cappa S, Bejanin A, Bombois S, Epelbaum S, Teichmann M, Habert M-O, Nordberg A, Blennow K, Galasko D, Stern Y, Rowe CC, Salloway S, Schneider LS, Cummings JL, Feldman HH (2021) Clinical diagnosis of Alzheimer's disease: recommendations of the International Working Group. *The Lancet Neurology*.
- [19] Dubois B, Feldman HH, Jacova C, Cummings JL, Dekosky ST, Barberger-Gateau P, Delacourte A, Frisoni G, Fox NC, Galasko D, Gauthier S, Hampel H, Jicha GA, Meguro K, O'Brien J, Pasquier F, Robert P, Rossor M, Salloway S, Sarazin M, de Souza LC, Stern Y, Visser PJ, Scheltens P (2010) Revising the definition of Alzheimer's disease: a new lexicon. *Lancet Neurol* **9**, 1118-1127.
- [20] Fiksinski AM, Hoftman GD, Vorstman JAS, Bearden CE (2023) A genetics-first approach to understanding autism and schizophrenia spectrum disorders: the 22q11.2 deletion syndrome. *Mol Psychiatry* **28**, 341-353.

- [21] Stessman HA, Bernier R, Eichler EE (2014) A genotype-first approach to defining the subtypes of a complex disease. *Cell* **156**, 872-877.
- [22] Wilczewski CM, Obasohan J, Paschall JE, Zhang S, Singh S, Maxwell GL, Similuk M, Wolfsberg TG, Turner C, Biesecker LG, Katz AE (2023) Genotype first: Clinical genomics research through a reverse phenotyping approach. *Am J Hum Genet* **110**, 3-12.
- [23] Neff RA, Wang M, Vatansever S, Guo L, Ming C, Wang Q, Wang E, Horgusluoglu-Moloch E, Song WM, Li A, Castranio EL, Tcw J, Ho L, Goate A, Fossati V, Noggle S, Gandy S, Ehrlich ME, Katsel P, Schadt E, Cai D, Brennand KJ, Haroutunian V, Zhang B (2021) Molecular subtyping of Alzheimer's disease using RNA sequencing data reveals novel mechanisms and targets. *Sci Adv* **7**, eabb5398.
- [24] Udler MS, Kim J, von Grotthuss M, Bonas-Guarch S, Cole JB, Chiou J, Christopher DAoboM, the I, Boehnke M, Laakso M, Atzmon G, Glaser B, Mercader JM, Gaulton K, Flannick J, Getz G, Florez JC (2018) Type 2 diabetes genetic loci informed by multi-trait associations point to disease mechanisms and subtypes: A soft clustering analysis. *PLoS Med* **15**, e1002654.
- [25] Arnedo J, Svrakic DM, Del Val C, Romero-Zaliz R, Hernandez-Cuervo H, Molecular Genetics of Schizophrenia C, Fanous AH, Pato MT, Pato CN, de Erausquin GA, Cloninger CR, Zwir I (2015) Uncovering the hidden risk architecture of the schizophrenias: confirmation in three independent genome-wide association studies. *Am J Psychiatry* **172**, 139-153.
- [26] Dahl A, Zaitlen N (2020) Genetic Influences on Disease Subtypes. *Annu Rev Genomics Hum Genet* **21**, 413-435.
- [27] Rangan AV, McGrouther CC, Kelsoe J, Schork N, Stahl E, Zhu Q, Krishnan A, Yao V, Troyanskaya O, Bilaloglu S, Raghavan P, Bergen S, Jureus A, Landen M, Bipolar Disorders Working Group of the Psychiatric Genomics C (2018) A loop-counting method for covariate-corrected low-rank biclustering of gene-expression and genome-wide association study data. *PLoS Comput Biol* **14**, e1006105.
- [28] Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, Downey P, Elliott P, Green J, Landray M, Liu B, Matthews P, Ong G, Pell J, Silman A, Young A, Sprosen T, Peakman T, Collins R (2015) 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.
- [29] Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, Motyer A, Vukcevic D, Delaneau O, O'Connell J, Cortes A, Welsh S, Young A, Effingham M, McVean G, Leslie S, Allen N, Donnelly P, Marchini J (2018) The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203-209.
- [30] Wilkinson T, Schnier C, Bush K, Rannikmae K, Henshall DE, Lerpiniere C, Allen NE, Flaig R, Russ TC, Bathgate D, Pal S, O'Brien JT, Sudlow CLM, Dementias Platform UK, Biobank UK (2019) Identifying dementia outcomes in UK Biobank: a validation study of primary care, hospital admissions and mortality data. *Eur J Epidemiol* **34**, 557-565.
- [31] Chen CY, Pollack S, Hunter DJ, Hirschhorn JN, Kraft P, Price AL (2013) Improved ancestry inference using weights from external reference panels. *Bioinformatics* **29**, 1399-1406.
- [32] The 1000 Genomes Project Consortium (2015) A global reference for human genetic variation. *Nature* **526**, 68-74.
- [33] Das S, Forer L, Schonherr S, Sidore C, Locke AE, Kwong A, Vrieze SI, Chew EY, Levy S, McGue M, Schlessinger D, Stambolian D, Loh PR, Iacono WG, Swaroop A, Scott LJ, Cucca F, Kronenberg F, Boehnke M, Abecasis GR, Fuchsberger C (2016) Next-generation genotype imputation service and methods. *Nat Genet* **48**, 1284-1287.



- [34] Chung JE, Magland JF, Barnett AH, Tolosa VM, Tooker AC, Lee KY, Shah KG, Felix SH, Frank LM, Greengard LF (2017) A Fully Automated Approach to Spike Sorting. *Neuron* **95**, 1381-1394 e1386.
- [35] Prive F, Luu K, Blum MGB, McGrath JJ, Vilhjalmsson BJ (2020) Efficient toolkit implementing best practices for principal component analysis of population genetic data. *Bioinformatics* **36**, 4449-4457.
- [36] (!!! INVALID CITATION !!! (27)).
- [37] Zhou H, Lin W, Labra SR, Lipton SA, Schork NJ, Rangan AV (2022) Detecting boolean asymmetric relationships with a loop counting technique and its implications for analyzing heterogeneity within gene expression datasets. *bioRxiv*, 2022.2008.2004.502792.
- [38] Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7.
- [39] Raudvere U, Kolberg L, Kuzmin I, Arak T, Adler P, Peterson H, Vilo J (2019) g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res* **47**, W191-W198.
- [40] Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, Feng T, Zhou L, Tang W, Zhan L, Fu X, Liu S, Bo X, Yu G (2021) clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation (Camb)* **2**, 100141.
- [41] Yu G, Wang LG, Han Y, He QY (2012) clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* **16**, 284-287.
- [42] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25-29.
- [43] Gene Ontology C (2021) The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Res* **49**, D325-D334.
- [44] Merico D, Isserlin R, Stueker O, Emili A, Bader GD (2010) Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. *PLoS One* **5**, e13984.
- [45] Febrero M, Galeano P, González - Manteiga W (2007) Outlier detection in functional data by depth measures, with application to identify abnormal NOx levels. *Environmetrics* **19**, 331-345.
- [46] Donohue MC, Sperling RA, Petersen R, Sun CK, Weiner MW, Aisen PS, Alzheimer's Disease Neuroimaging I (2017) Association Between Elevated Brain Amyloid and Subsequent Cognitive Decline Among Cognitively Normal Persons. *JAMA* **317**, 2305-2316.
- [47] Donohue MC, Sperling RA, Salmon DP, Rentz DM, Raman R, Thomas RG, Weiner M, Aisen PS, Australian Imaging B, Lifestyle Flagship Study of A, Alzheimer's Disease Neuroimaging I, Alzheimer's Disease Cooperative S (2014) The preclinical Alzheimer cognitive composite: measuring amyloid-related decline. *JAMA Neurol* **71**, 961-970.
- [48] Landau SM, Breault C, Joshi AD, Pontecorvo M, Mathis CA, Jagust WJ, Mintun MA, Alzheimer's Disease Neuroimaging I (2013) Amyloid-beta imaging with Pittsburgh compound B and florbetapir: comparing radiotracers and quantification methods. *J Nucl Med* **54**, 70-77.
- [49] Landau SM, Marks SM, Mormino EC, Rabinovici GD, Oh H, O'Neil JP, Wilson RS, Jagust WJ (2012) Association of lifetime cognitive engagement and low  $\beta$ -amyloid deposition. *Archives of Neurology* **69**, 623-629.
- [50] Shaw LM, Vanderstichele H, Knapik-Czajka M, Clark CM, Aisen PS, Petersen RC, Blennow K, Soares H, Simon A, Lewczuk P, Dean R, Siemers E, Potter W, Lee VM, Trojanowski JQ, Alzheimer's Disease Neuroimaging I (2009) Cerebrospinal fluid

- biomarker signature in Alzheimer's disease neuroimaging initiative subjects. *Ann Neurol* **65**, 403-413.
- [51] Bates D, Machler M, Bolker BM, Walker SC (2015) Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* **67**, 1-48.
- [52] Kuznetsova A, Brockhoff PB, Christensen RHB (2017) lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software* **82**, 1-26.
- [53] Stefansson H, Helgason A, Thorleifsson G, Steinthorsdottir V, Masson G, Barnard J, Baker A, Jonasdottir A, Ingason A, Gudnadottir VG, Desnica N, Hicks A, Gylfason A, Gudbjartsson DF, Jonsdottir GM, Sainz J, Agnarsson K, Birgisdottir B, Ghosh S, Olafsdottir A, Cazier JB, Kristjansson K, Frigge ML, Thorgeirsson TE, Gulcher JR, Kong A, Stefansson K (2005) A common inversion under selection in Europeans. *Nat Genet* **37**, 129-137.
- [54] Pittman AM, Fung HC, de Silva R (2006) Untangling the tau gene association with neurodegenerative disorders. *Hum Mol Genet* **15 Spec No 2**, R188-195.
- [55] Sims R, Hill M, Williams J (2020) The multiplex model of the genetics of Alzheimer's disease. *Nat Neurosci* **23**, 311-322.
- [56] Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**, 904-909.
- [57] Allen M, Kachadoorian M, Quicksall Z, Zou F, Chai HS, Younkin C, Crook JE, Pankratz VS, Carrasquillo MM, Krishnan S, Nguyen T, Ma L, Malphrus K, Lincoln S, Biscoglio G, Kolbert CP, Jen J, Mukherjee S, Kauwe JK, Crane PK, Haines JL, Mayeux R, Pericak-Vance MA, Farrer LA, Schellenberg GD, Parisi JE, Petersen RC, Graff-Radford NR, Dickson DW, Younkin SG, Ertekin-Taner N (2014) Association of MAPT haplotypes with Alzheimer's disease risk and MAPT brain gene expression levels. *Alzheimers Res Ther* **6**, 39.
- [58] Strickland SL, Reddy JS, Allen M, N'Songo A, Burgess JD, Corda MM, Ballard T, Wang X, Carrasquillo MM, Biernacka JM, Jenkins GD, Mukherjee S, Boehme K, Crane P, Kauwe JS, Ertekin-Taner N, Alzheimer's Disease Genetics C (2020) MAPT haplotype-stratified GWAS reveals differential association for AD risk variants. *Alzheimers Dement* **16**, 983-1002.
- [59] Baker M, Litvan I, Houlden H, Adamson J, Dickson D, Perez-Tur J, Hardy J, Lynch T, Bigio E, Hutton M (1999) Association of an extended haplotype in the tau gene with progressive supranuclear palsy. *Hum Mol Genet* **8**, 711-715.
- [60] Munoz L, Ammit AJ (2010) Targeting p38 MAPK pathway for the treatment of Alzheimer's disease. *Neuropharmacology* **58**, 561-568.
- [61] Leyns CEG, Holtzman DM (2017) Glial contributions to neurodegeneration in tauopathies. *Mol Neurodegener* **12**, 50.
- [62] Tooyama I, Kimura H, Akiyama H, McGeer PL (1990) Reactive microglia express class I and class II major histocompatibility complex antigens in Alzheimer's disease. *Brain Res* **523**, 273-280.
- [63] Nobes CD, Hall A (1999) Rho GTPases control polarity, protrusion, and adhesion during cell movement. *J Cell Biol* **144**, 1235-1244.
- [64] Settleman J (1999) Rho GTPases in Development In *Cytoskeleton and Small G Proteins*, Jeanteur P, ed. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 201-229.
- [65] Hotulainen P, Hoogenraad CC (2010) Actin in dendritic spines: connecting dynamics to function. *J Cell Biol* **189**, 619-629.
- [66] Cabrales Fontela Y, Kadavath H, Biernat J, Riedel D, Mandelkow E, Zweckstetter M (2017) Multivalent cross-linking of actin filaments and microtubules through the microtubule-associated protein Tau. *Nat Commun* **8**, 1981.

- [67] Valenta T, Hausmann G, Basler K (2012) The many faces and functions of beta-catenin. *EMBO J* **31**, 2714-2736.
- [68] Soldano A, Hassan BA (2014) Beyond pathology: APP, brain development and Alzheimer's disease. *Curr Opin Neurobiol* **27**, 61-67.
- [69] Kowall NW, Kosik KS (1987) Axonal disruption and aberrant localization of tau protein characterize the neuropil pathology of Alzheimer's disease. *Ann Neurol* **22**, 639-643.
- [70] Iqbal K, Liu F, Gong CX, Grundke-Iqbal I (2010) Tau in Alzheimer disease and related tauopathies. *Curr Alzheimer Res* **7**, 656-664.
- [71] Schork NJ, Elman JA (2023) Pathway-specific polygenic risk scores correlate with clinical status and Alzheimer's-related biomarkers. *Res Sq*.
- [72] La Joie R, Visani AV, Baker SL, Brown JA, Bourakova V, Cha J, Chaudhary K, Edwards L, Iaccarino L, Janabi M, Lesman-Segev OH, Miller ZA, Perry DC, O'Neil JP, Pham J, Rojas JC, Rosen HJ, Seeley WW, Tsai RM, Miller BL, Jagust WJ, Rabinovici GD (2020) Prospective longitudinal atrophy in Alzheimer's disease correlates with the intensity and topography of baseline tau-PET. *Sci Transl Med* **12**, eaau5732.
- [73] Hanseeuw BJ, Betensky RA, Jacobs HIL, Schultz AP, Sepulcre J, Becker JA, Cosio DMO, Farrell M, Quiroz YT, Mormino EC, Buckley RF, Papp KV, Amariglio RA, Dewachter I, Ivanoiu A, Huijbers W, Hedden T, Marshall GA, Chhatwal JP, Rentz DM, Sperling RA, Johnson K (2019) Association of Amyloid and Tau With Cognition in Preclinical Alzheimer Disease: A Longitudinal Study. *JAMA Neurol* **76**, 915-924.
- [74] Ossenkoppele R, Lyoo CH, Sudre CH, van Westen D, Cho H, Ryu YH, Choi JY, Smith R, Strandberg O, Palmqvist S, Westman E, Tsai R, Kramer J, Boxer AL, Gorno-Tempini ML, La Joie R, Miller BL, Rabinovici GD, Hansson O (2020) Distinct tau PET patterns in atrophy-defined subtypes of Alzheimer's disease. *Alzheimers Dement* **16**, 335-344.
- [75] Saporta AS, Sottile SL, Miller LJ, Feely SM, Siskind CE, Shy ME (2011) Charcot-Marie-Tooth disease subtypes and genetic testing strategies. *Ann Neurol* **69**, 22-33.
- [76] Crutch SJ, Schott JM, Rabinovici GD, Murray M, Snowden JS, van der Flier WM, Dickerson BC, Vandenberghe R, Ahmed S, Bak TH, Boeve BF, Butler C, Cappa SF, Ceccaldi M, de Souza LC, Dubois B, Felician O, Galasko D, Graff-Radford J, Graff-Radford NR, Hof PR, Krolak-Salmon P, Lehmann M, Magnin E, Mendez MF, Nestor PJ, Onyike CU, Pelak VS, Pijnenburg Y, Primativo S, Rossor MN, Ryan NS, Scheltens P, Shakespeare TJ, Suarez Gonzalez A, Tang-Wai DF, Yong KXX, Carrillo M, Fox NC, Alzheimer's Association IAAsD, Associated Syndromes Professional Interest A (2017) Consensus classification of posterior cortical atrophy. *Alzheimers Dement* **13**, 870-884.
- [77] Gorno-Tempini ML, Hillis AE, Weintraub S, Kertesz A, Mendez M, Cappa SF, Ogar JM, Rohrer JD, Black S, Boeve BF, Manes F, Dronkers NF, Vandenberghe R, Rascovsky K, Patterson K, Miller BL, Knopman DS, Hodges JR, Mesulam MM, Grossman M (2011) Classification of primary progressive aphasia and its variants. *Neurology* **76**, 1006-1014.
- [78] Townley RA, Graff-Radford J, Mantyh WG, Botha H, Polsinelli AJ, Przybelski SA, Machulda MM, Makhlof AT, Senjem ML, Murray ME, Reichard RR, Savica R, Boeve BF, Drubach DA, Josephs KA, Knopman DS, Lowe VJ, Jack CR, Jr., Petersen RC, Jones DT (2020) Progressive dysexecutive syndrome due to Alzheimer's disease: a description of 55 cases and comparison to other phenotypes. *Brain Commun* **2**, fcaa068.
- [79] Ossenkoppele R, Singleton EH, Groot C, Dijkstra AA, Eikelboom WS, Seeley WW, Miller B, Laforce RJ, Scheltens P, Papma JM, Rabinovici GD, Pijnenburg YAL (2022) Research Criteria for the Behavioral Variant of Alzheimer Disease: A Systematic Review and Meta-analysis. *JAMA Neurol* **79**, 48-60.
- [80] (!!! INVALID CITATION !!! (12)).



## TABLES

**Table 1. Sample characteristics of UK Biobank dataset.**

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

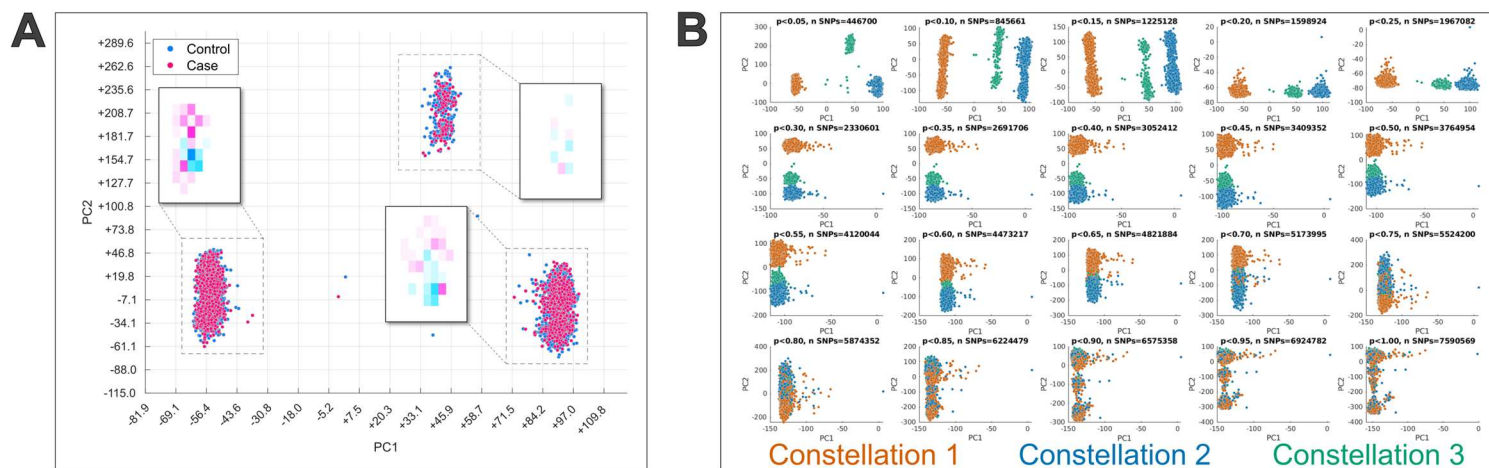
analyses of genetic heterogeneity found in the UK Biobank. The mild cognitive

impairment (MCI) group was included in analyses of cognitive and biomarker

trajectories.

	<b>Overall</b>	<b>CN</b>	<b>Dementia</b>	<b>MCI</b>	
<b>n</b>	1369	470	500		399
<b>Sex, n male (%)</b>	765 (55.9)	225 (47.9)	297 (59.4)	243 (60.9)	
<b>Age, years (SD)</b>	73.70 (7.11)	72.47 (6.31)	74.61 (7.43)	74.00 (7.38)	
<b>Education, years (SD)</b>	16.01 (2.76)	16.67 (2.44)	15.51 (2.89)	15.87 (2.79)	
<b>APOE-e4 alleles, n (%)</b>					
<b>0</b>	750 (54.8)	334 (71.1)	174 (34.8)	242 (60.7)	
<b>1</b>	493 (36.0)	122 (26.0)	245 (49.0)	126 (31.6)	
<b>2</b>	126 (9.2)	14 (3.0)	81 (16.2)	31 (7.8)	
<b>APOE-e4 carriers, n (%)</b>	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. A)** Principal component analysis was

applied to allele combinations of UK Biobank cases and controls restricted to variants with a p-value < 0.05 in the Kunkle et al. Alzheimer's GWAS [80]. 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. **B)** Principal component analysis results when variants were

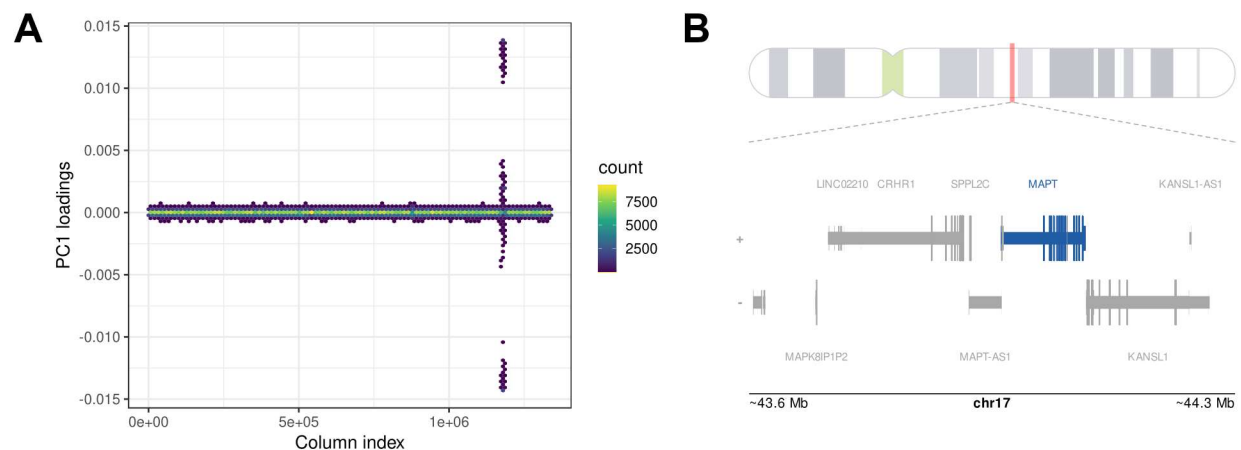
restricted across a range of p-value thresholds from the Kunkle et al. Alzheimer's GWAS [80].

The scatter plots are colored by constellation labels defined at the p < 0.05 threshold.

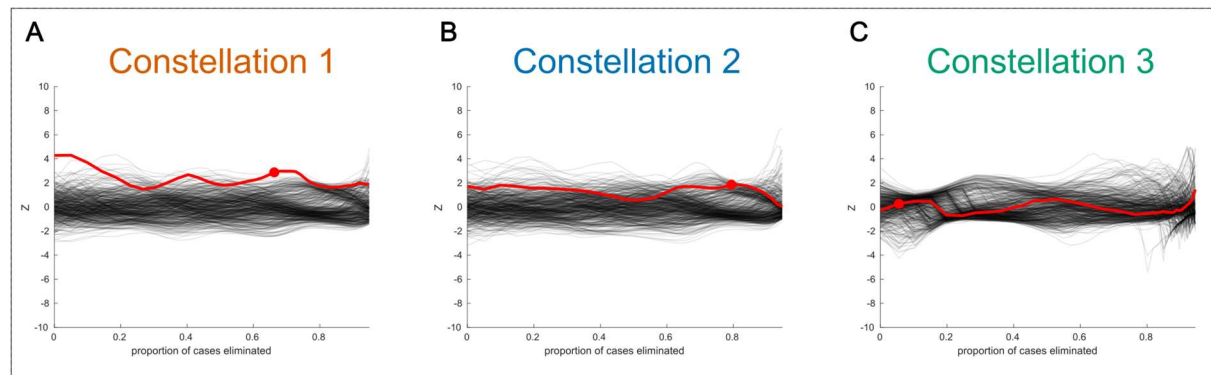
Participants from all three constellations are highly mixed when all variants (p < 1.0) are included.

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.

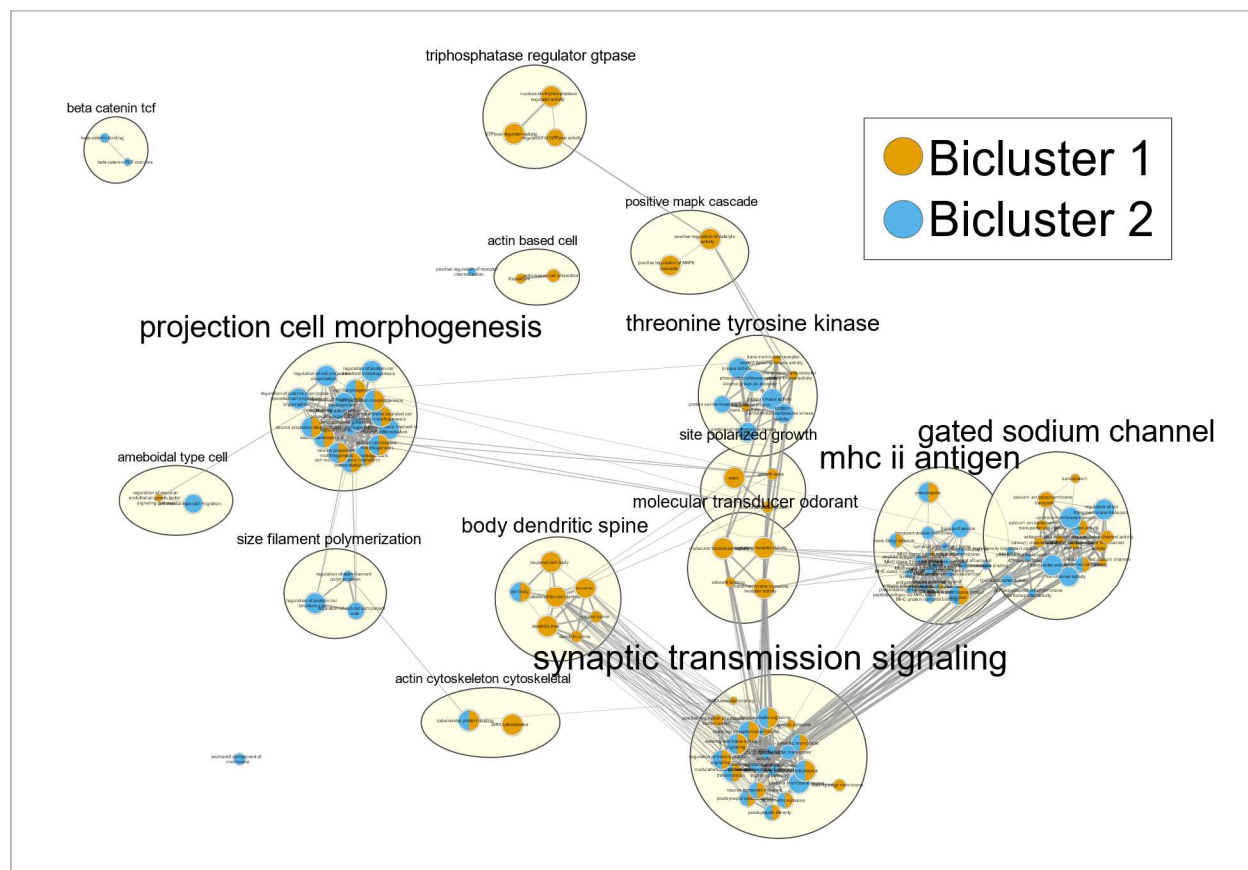


**Figure 2. A)** Hex bin plot of loadings across allele combinations on first principal component from the PCA on UKB cases and controls restricted to variants with a p-value < 0.05 in the Kunkle et al. Alzheimer's GWAS [12]. Allele combinations are ordered by chromosome and base position across the x-axis. Color represents density of data points that fall within a given hex. **B)** The location of peak loadings on PC1 is shown in greater detail along with gene annotations from the UCSC database. The peak loadings occur in the region of 17q21.31, overlapping with a known region of extended LD surrounding the *MAPT* locus.

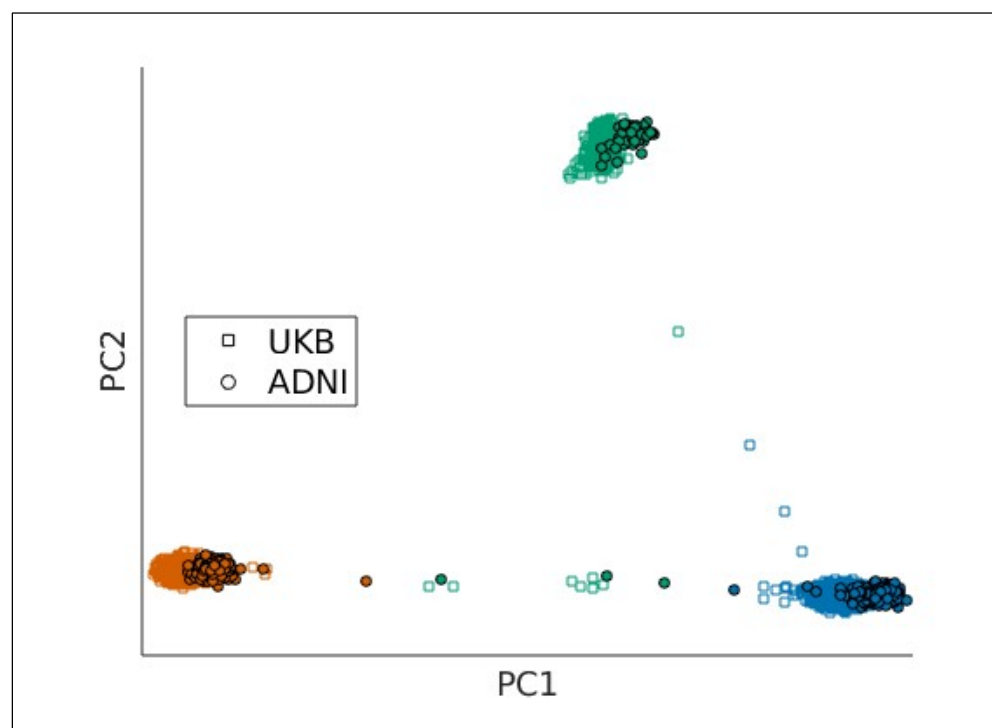


**Figure 3. Bicluster traces of observed data versus label-shuffled data.** The disease-related signal-strength associated with the remaining UKB Alzheimer’s cases relative to controls is plotted on the y-axis. At each iteration, allele combinations and cases that contribute least to this difference 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.

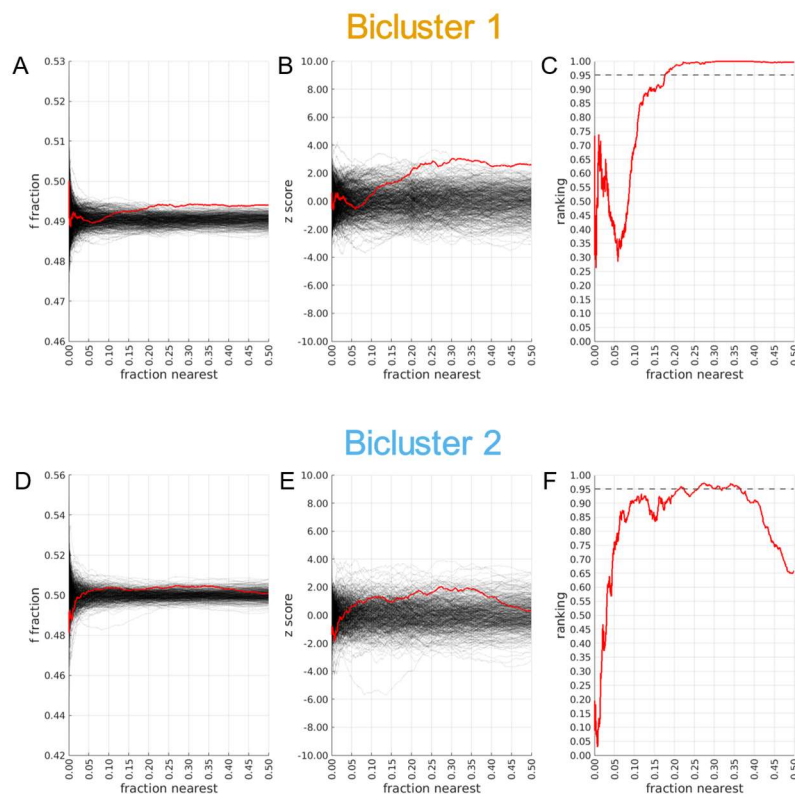




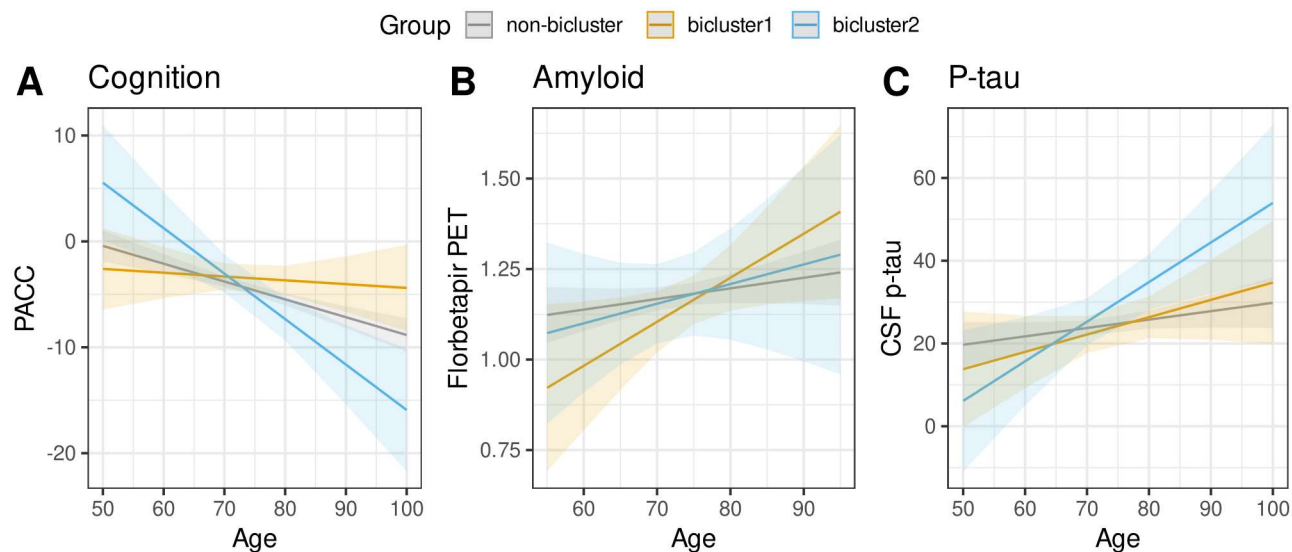
**Figure 4. Network plot of gene sets enriched among bicluster genes.** Network visualization of gene set enrichment results for biclusters 1 and 2. First, separate GWAS compared individuals in each bicluster to all controls from the same constellation in which the bicluster was found (e.g., bicluster 1 cases versus all controls from constellation 1). SNPs that were nominally associated with a given bicluster (i.e., positive regression coefficient and  $p < 0.05$ ) were mapped to genes based on position. Over-representation analysis of gene lists for each bicluster was used to identify gene sets associated with each bicluster. Nodes represent significantly enriched gene sets ( $p_{FDR} < 0.05$ ) 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 5. 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 6. 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 compared to the null distribution is shown in Panels (C) and (F).



**Figure 7. 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 longitudinal cognitive and biomarker trajectories 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 measured using florbetapir PET. C) Phosphorylated tau (p-tau) was measured using CSF p-tau.