INTRODUCTION

Late-Onset Alzheimer’s Disease (LOAD) is the most common and detrimental cause of cognitive decline with age, inevitably leading to loss of independence and premature death [1]. As with other complex diseases, the susceptibility and progression of LOAD are affected by a myriad of genetic and environmental factors and interactions, resulting in a heterogenous presentation and outcome [2]–[4]. Therefore, although LOAD’s heritability can be as high as 80% [3], it has been challenging to identify genetic risk variants and therapeutic targets using genome-wide associations with a simplified AD-Control phenotype. In this study we attempt to characterize LOAD heterogeneity based on transcriptomic data from post-mortem brain tissues of elderly individuals. Instead of using the commonly-used cluster methods that focus on finding averaged subgroups, we use archetype analysis [REFS##], which aims to identify end-members that highlight the distinct attributes of subgroups.

Several studies have recently utilized molecular data to identify LOAD stages and subtypes in order to better characterize LOAD phenotypes and thus refine genome-wide associations [5]–[11]. Both transcriptomic and proteomic data consistently reveal co-expression modules that are enriched for different brain cell types and known LOAD GWAS candidates, and are associated with different disease stages and subtypes. For example, a proteomic study has found that oligodendrocytes and microglia markers linked to inflammation have been implicated in the symptomatic stages of the disease, while neuronal and synaptic loss may precede the onset of symptoms [6]. Similarly, transcriptomic studies have shown differential expression of genes associated with loss of basic cell functions at earlier stages including neuronal and synaptic function, and increased immune response at later stages [7], [8], [10]. Clustering methods of transcriptomic data have revealed 2-5 subtypes underlying population heterogeneity in different brain regions [5]. Using these modules and subtypes as quantitative phenotypes has revealed both known and new candidate genes specifically relevant to LOAD, illustrating the importance of accounting for population heterogeneity. Since archetype analysis focuses on end-members rather than averaged clusters, it could potentially reveal even clearer molecular signals and more distinct subgroups in the data.

Archetype analysis has been used in economics and engineering for a while. Recently, it has been adapted for the study of single-cell RNA-seq data.

METHODS

**Data**

We obtained clinical variables, whole-genome sequencing, and RNA sequencing (RNA-Seq) data from Synapse (<https://www.synapse.org/>; syn3191087, syn11707420, and syn8456719, respectively). The Rush University's Religious Orders Study and Memory and Aging Project (ROSMAP) cohort consists of 623 dorsolateral prefrontal cortex (DLPFC) samples of individuals from 40 groups of religious orders from across the United States (ROS) and older adults in retirement communities in the Chicago area (MAP), characterized as LOAD, elderly control, and other dementia [12]. See Table S1 for a summary of samples by sex and diagnosis, and Table S2 for Metadata. Sex, age of death, and batch were used as covariates for normalization of RNA-seq data, starting with log-transformed CPM (see [syn8456629](https://www.synapse.org/#!Synapse:syn8456629) for details). For comparison, we obtained subtype specificity for each sample and co-expression modules and submodules generated by [5] for the same cohort, based on the same RNA-seq data. Subtyping in that study was generated using k-means clustering and co-expression modules where generated using iterative WGCNA based on modules from [11].

**Archetype fitting**

We estimated the optimal number of archetypes and their position in the gene expression space using Principal Convex Hull algorithm specifically designed for archetypal analysis by [13], as implemented in the R package paretoTI [14]. Statistical significance of the goodness of fit of different polytopes is determined in this package by permutations of the original data that disrupt relationships between variables but maintains their specific distributions. Bootstrapping is implemented to determine variation around the inferred position of each vertex (i.e., archetype), providing another measure for how well different polytopes fit the data. We tested polytopes ranging from 2 to 6 vertices.

We used only samples identified as AD patients for the fitting process. However, all downstream analyses included control and “other” samples as well. Principal component analysis indicated that the first 20 PC’s were significantly higher than expected by chance using a broken stick model, explaining altogether 17% of the variance in the original data (Fig. S1). Fitting archetypes to 400 PC’s (80% of the variance) yielded essentially the same results (Fig. S2). The final fitting process and all further analyses were therefore based on the first 20 PC’s.

**Gene Expression Annotation of Archetypes**

We reconstructed gene expression profiles for each archetype by finding the genes whose expression decreases significantly with distance from that archetype, using Wilcox test, also implemented in the R package ParetoTI. We analyzed the resulting gene sets for their GO terms enrichment using the R package clusterProfiler [15]. In addition, we tested the genes sets for their enrichment in cell-specific markers as provided by [16]. The cell-specificity for a given gene was calculated as the minimum log fold change of that gene’s expression level in a given cell type compared with its expression level in all other cell types.

**Single-Variant Mapping of Archetype Specificity**

We generated 4 new phenotypes for each patient based on the archetype analysis: three of the phenotypes were calculated as the Euclidean distance of each patient from each of the archetype in the reduced PC space of the gene expression data; the forth phenotype was the categorical assignment of each sample to the nearest archetype.

We mapped each of these four phenotypes to the whole-genome sequencing data using EMMAX, a variance-component linear mixed model [17]. QQ plots indicate minimal effects of genomic inflation and population substructure on our analyses (Figure S7). While we did not find significant snps at p-value < 5x10-8 level, many were suggestive at p-value < 1x10-5 level. We mapped these variants to known genes using FUMA (<https://fuma.ctglab.nl/>), and annotated them by querying the GWAS Catalog <https://www.ebi.ac.uk/gwas/> using the R package gwasrapidd [18], as well as manually querying the Human Gene Database GeneCards (<https://www.genecards.org/>).

In order to visualize the relationships between our archetype specificity phenotypes and their associated loci, we built a directed network using Cytoscape version 3.7 (https://cytoscape.org/) [19]. For comparison, we included in this network results of single-variant association of other relevant traits, including subtype specificity metrics generated by [5], as well as Braak stage, CERAD scores, cognitive diagnosis, and case-control diagnosis (using EMMAX [17] as for the other traits).

RESULTS

**Archetypes Fitting and Annotation**

In order to find the number of archetypes that best fit the data, we fitted polygons with an increasing number of vertices (i.e., archetypes), using the first 20 PC’s of the gene expression data. As expected, the proportion of variance explained by the polygon increased with the number of vertices (Fig. 1A). However, the added benefit of adding more vertices decreased substantially past three (Fig. 1B). In addition, the variance in the inferred position of each archetype, estimated based on 200 bootstrapped samples of the original space, increased substantially for polygons with more than 3 vertices. Therefore, we concluded that the optimal number of archetypes for this dataset is 3, represented by a triangle (see also Figure 2). Table S2 provides the archetype specificity metrics for each patient, calculated as the Euclidean distance from each patient to each archetype, as well as the categorical assignment of each patient to the nearest archetype.

Previously published classification of the same cohort by [5], using k-means clustering, yielded two subtypes that largely corresponded to our archetypes, with subtype A largely overlapping with archetypes 1 and 2, and subtype B largely overlapping with archetype 3 (Figure 2). Archetype 2 is closer to subtype B than archetype 1 is. Chi2 test confirmed that these observations were highly significant (chi2=65.66, df=6, p = 3.1613-12) and that archetype 2 was more of a mix between subtype A and B, compared with archetype 1 and 3, including more control cases as well (Fig S3).

We tested the association between archetypes, sex, and various clinical variables (clinical diagnosis, Braak score, CERAD, and Apoe4 status). Sex in itself was not associated with the archetypes (chi2=0.255, df=2, p-value=0.88), and neither CERAD and Apoe4 status (Figs. S4-S5). Cognitive score was somewhat significant without conditioning on sex, but the deviations were rather small (Fig 3A). Only Braak score was significant with and without conditioning on sex (Fig. 3B; Cochran-Mantel-Haenszel M2=23.834, df=12, p=0.021; chi2=23.6, df=12, p=0.023, respectively). Both males and female AD patients were somewhat over-represented in archetype 1, based on both cognitive and Braak scores (Fig. 3). Archetype 3 included more control and MCI’s, with females particularly over-represented in cognitive score and males in Braak score. Archetype 2 is somewhat in between, being again more of an equal mix. Although the deviations were small and non-significant, associations of CERAD and Apoe4 with sex and archetypes were in accord with the observation that archetype 1 included more cases of advanced and high-risk AD patients, particularly females, archetype 3 included more cases of early and low-risk patients, while archetype 2 was a mix (Figs. S4-S5).

To further explore the association of the archetypes with levels of disease, we compared our results with previously-published pseudotime estimates [8]. Figure 4 shows the association between pseudotime estimates and the categorical assignments of samples to the nearest archetype. Figure S6 shows the same association using distances of samples from each archetype. Using either discreet assignment or distances, we found that archetype 1 was associated with higher pseudotime estimates, presumably reflecting more advances stages of AD, archetype 3 was associated with lower pseudotime estimates, while archetype 2 was in between.

**Annotation of Genes Associated with the Archetypes**

In order to study the differences between archetypes in their gene expression, we looked for genes whose expression level was decreasing significantly as a function of distance from each archetype. Out of about 15000 genes included in the original dataset, we found 626 genes associated with archetype 1, 551 with archetype 2, and 1538 with archetype 3, with minimal overlap between the gene sets. Only 8 genes overlapped between archetype 1 and 2, one gene overlapped between archetype 2 and 3, and none between archetype 1 and 3.

GO enrichment analysis revealed a distinct functional enrichment signal for each archetype (Figure 5, table S#). The gene set of archetype 1 was enriched for GO terms related to immune function and inflammation. Archetype 2 had a mix of functions over-represented in its gene set, including basic cell functions and metabolic processes associated specifically with the generation and maintenance of glial cells. Archetype 3 was clearly associated with neuronal function and trans-synaptic signaling.

We further annotated the gene sets of each archetype by comparing them to previously-published cell-specific markers [16]. Again, we found a clear and significant distinction between the archetypes (Figure 6). In accord with the GO enrichment analysis, archetype 1 was highly associated with endothelial cells and microglia, archetype 2 with oligodendrocytes and astrocytes, and archetype 3 with neurons.

The archetype gene set overlapped greatly with gene expression modules and submodules previously estimated for the same cohort by [5]. 80% of the archetype genes were included in one of the modules, and 62% in one of the submodules. In accord with their independent annotation, archetype 1 included mostly inflammatory-related submodules, archetype 2 included generation and maintenance of glial cells, and archetype 3 comprised largely of one module related to neuronal function and synaptic signaling (Figure 7).

**Single-Variant Mapping of Archetype Specificity**

We generated 4 new phenotypes for each patient based on the archetype analysis, specifying their distance from each of the archetype as well as their classification to the nearest archetype, and mapped each of them to their whole-genome sequence data using EMMAX [17]. QQ plots indicate minimal effects of genomic inflation and population substructure on our analyses (Figure S7). While we did not find snps significant at p-value < 5x10-8 level, numerous snps were suggestive at p-value < 1x10-5 level, many of which mapped to genes that are involved in the development and function of the central nervous system (Tables S### FUMA+NHGRI).

The distance from Archetype 1 was associated with genes such as ERBB4 and CNTN4, both involved in CNS development and plasticity; GRIK2, which codes for glutamate receptor, the main excitatory neurotransmitter receptor in the mammalian brain; and PSENEN/PROSER3, which is required for intramembranous processing of the Notch proteins and APP and has been widely associated with Alzheimer’s disease. Variants of ERBB4 and CNTN4 have been associated with numerous behavioral and cognitive conditions, as well as metabolic traits and sleep regulation. Variants of GRIK2 have been associated with behavioral and cognitive conditions as well as blood and heart conditions.

Most genes associated with Archetype 2 were directly related to the nervous system development and function, including most notably RGMB and NRG3. RGMB is a repulsive guidance molecule that contributes to the patterning of the developing nervous system as well as neuronal adhesion. NRG3 activates tyrosine phosphorylation and is involved in neuroblast proliferation, migration, and differentiation, and has been associated with various behavioral and metabolic conditions.

The genes associated with Archetype 3 included TMEM106B, KCNA4, and CER6. TMEM106N is a transmembrane protein involved mainly with dendrite morphogenesis and maintenance. It has been well established in association with neurodegenerative diseases and cognitive aging, including protective variants that interact with APOE (Table S###). KCNA4 is a potassium channels member, involved with a variety of functions including neurotransmitter release and neuronal excitability. CER6 is a ceramid synthase that regulates metabolism and hepatic lipid accumulation, and has been associated with inflammatory response and autonomic neuropathy.

Mapping the categorical assignment to the nearest archetype revealed several notable genes. SGIP1 is an endocytic protein that affects neuronal signaling involved in energy homeostasis. SYNE1 is a nuclear envelope protein particularly critical for neuronal signaling in the cerebellum. NTM promotes neurite outgrowth and adhesion. All three have been associated with various mental and metabolic disorders.

Figure 8 illustrates the genetic architecture underlying these phenotypes in terms of overlap in their associated genetic variants. For comparison, we included also subtypes generated for the same dataset using k-means clustering [5] as well as clinical phenotypes of that cohort, all mapped to snps using the same protocol in FUMA. The three archetypes do not overlap in their associated genetic loci. In accord with their gene expression profiles, archetypes 1 and 2 overlap with subtype A, while archetype 3 overlaps with subtype B. This suggests that the archetypes provide a way to refine subtype A in terms of the underlying genetics as well as function. In addition, the archetypes reveal two more genes that connect the gene expression profiles with clinical variables: archetype 2 is connected to clinical diagnosis through CSMD1, known to be highly expressed in the brain in relation to glucose homeostasis and startle response among other things, and archetype 3 is connected to CERAD score through BNC2, a zinc finger protein related to epithelial cells. Like subtype B, archetype 3 is also connected to BRAAK score through DLC1, a gene related to cytoskeletal reorganization and tumor suppression. In comparison, the same analysis using the submodules as phenotypes revealed prevalent overlap among the submodules and between submodules and subtypes (Fig. S9; see also [5]), suggesting that the archetypes could provide a more informative tool for patient stratification.

REFERENCES

[1] R. Mayeux and Y. Stern, “Epidemiology of Alzheimer disease,” *Cold Spring Harb. Perspect. Med.*, vol. 2, no. 8, 2012, doi: 10.1101/cshperspect.a006239.

[2] J. C. Lee, “Beyond disease susceptibility—Leveraging genome-wide association studies for new insights into complex disease biology,” *HLA*, vol. 90, no. 6. Blackwell Publishing Ltd, pp. 329–334, 01-Dec-2017, doi: 10.1111/tan.13170.

[3] I. E. Jansen *et al.*, “Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer’s disease risk,” *Nat. Genet.*, vol. 51, no. 3, pp. 404–413, Mar. 2019, doi: 10.1038/s41588-018-0311-9.

[4] J. Verheijen and K. Sleegers, “Understanding Alzheimer Disease at the Interface between Genetics and Transcriptomics,” *Trends in Genetics*, vol. 34, no. 6. Elsevier Ltd, pp. 434–447, 01-Jun-2018, doi: 10.1016/j.tig.2018.02.007.

[5] N. Milind *et al.*, “Transcriptomic Stratification of Late-Onset Alzheimer’s Cases Reveals Novel Genetic Modifiers of Disease Pathology,” *bioRxiv*, p. 51404, Sep. 2019, doi: 10.1101/763516.

[6] N. T. Seyfried *et al.*, “A Multi-network Approach Identifies Protein-Specific Co-expression in Asymptomatic and Symptomatic Alzheimer’s Disease,” *Cell Syst.*, vol. 4, no. 1, pp. 60-72.e4, Jan. 2017, doi: 10.1016/j.cels.2016.11.006.

[7] J. A. Miller, R. L. Woltjer, J. M. Goodenbour, S. Horvath, and D. H. Geschwind, “Genes and pathways underlying regional and cell type changes in Alzheimer’s disease,” *Genome Med.*, vol. 5, no. 5, p. 48, May 2013, doi: 10.1186/gm452.

[8] S. Mukherjee *et al.*, “Molecular estimation of neurodegeneration pseudotime in older brains,” *bioRxiv*, p. 686824, Jun. 2019, doi: 10.1101/686824.

[9] B. De Strooper and E. Karran, “The Cellular Phase of Alzheimer’s Disease,” *Cell*, vol. 164, no. 4. Cell Press, pp. 603–615, 11-Feb-2016, doi: 10.1016/j.cell.2015.12.056.

[10] S. Mostafavi *et al.*, “A molecular network of the aging human brain provides insights into the pathology and cognitive decline of Alzheimer’s disease,” *Nat. Neurosci.*, vol. 21, no. 6, pp. 811–819, Jun. 2018, doi: 10.1038/s41593-018-0154-9.

[11] B. A. Logsdon *et al.*, “Meta-analysis of the human brain transcriptome identifies heterogeneity across human AD coexpression modules robust to sample collection and methodological approach,” *bioRxiv*, p. 510420, Jan. 2019, doi: 10.1101/510420.

[12] P. L. De Jager *et al.*, “Data descriptor: A multi-omic atlas of the human frontal cortex for aging and Alzheimer’s disease research,” *Sci. Data*, vol. 5, no. 1, pp. 1–13, Aug. 2018, doi: 10.1038/sdata.2018.142.

[13] M. Mørup and L. K. Hansen, “Archetypal analysis for machine learning and data mining,” *Neurocomputing*, vol. 80, pp. 54–63, Mar. 2012, doi: 10.1016/J.NEUCOM.2011.06.033.

[14] V. Kleshchevnikov, “ParetoTI: R toolbox for Archetypal Analysis and Pareto Task Inference on single cell data.” 2019.

[15] Y. Guangchuang, L.-G. Wang, Y. Han, and Q.-Y. He, “clusterProfiler: an R package for comparing biological themes among gene clusters,” *Omi. A J. Integr. Biol.*, vol. 16, no. 5, pp. 284–287, 2012, doi: 10.1089/omi.2011.0118.

[16] A. T. McKenzie *et al.*, “Brain Cell Type Specific Gene Expression and Co-expression Network Architectures,” *Sci. Rep.*, vol. 8, no. 1, p. 8868, Dec. 2018, doi: 10.1038/s41598-018-27293-5.

[17] H. M. Kang *et al.*, “Variance component model to account for sample structure in genome-wide association studies,” *Nat. Genet.*, vol. 42, no. 4, pp. 348–354, 2010, doi: 10.1038/ng.548.

[18] R. Magno and A.-T. Maia, “gwasrapidd: an R package to query, download and wrangle GWAS Catalog data,” *Bioinformatics*, pp. 1–2, 2019.

[19] P. Shannon *et al.*, “Cytoscape: A Software Environment for Integrated Models,” *Genome Res.*, vol. 13, no. 22, pp. 2498–2504, 2003, doi: 10.1101/gr.1239303.metabolite.