

## RESEARCH ARTICLE

# Characterization of gene expression patterns in mild cognitive impairment using a transcriptomics approach and neuroimaging endophenotypes

Apoorva Bharthur Sanjay<sup>1</sup>  | Alice Patania<sup>2</sup> | Xiaoran Yan<sup>2</sup> | Diana Svaldi<sup>1</sup> |  
Tugce Duran<sup>3</sup> | Niraj Shah<sup>1</sup> | Sara Nemes<sup>1</sup> | Eric Chen<sup>1</sup> | Liana G. Apostolova<sup>1</sup><sup>1</sup> Department of Neurology, Indiana University School of Medicine, Indianapolis, Indiana, USA<sup>2</sup> Indiana University Network Sciences Institute, Indiana University, Bloomington, Indiana, USA<sup>3</sup> Department of Internal Medicine, Section of Gerontology & Geriatric Medicine, Wake Forest School of Medicine, Winston-Salem, North Carolina, USA**Correspondence**Apoorva Bharthur Sanjay and Liana G. Apostolova, Department of Neurology, Indiana University School of Medicine, 355W 16th Street, Suite 4022, Indianapolis, IN 46202, USA.  
E-mail: [abharthu@iu.edu](mailto:abharthu@iu.edu); [lapostol@iu.edu](mailto:lapostol@iu.edu).**Funding information**

The Eli Lilly Stark Neurosciences Predoctoral Research fellowship; NIA, Grant/Award Numbers: R21 AG072101, NIA R01 AG040770, NIA K02 AG048240, NIA P30 AG010133, K01 AG049050, R56/U01 AG057195; Alzheimer's Disease Neuroimaging Initiative, Grant/Award Number: U01 AG024904; DOD ADNI, Grant/Award Number: W81XWH-12-2-0012

**Abstract****Introduction:** Identification of novel therapeutics and risk assessment in early stages of Alzheimer's disease (AD) is a crucial aspect of addressing this complex disease. We characterized gene-expression patterns at the mild cognitive impairment (MCI) stage to identify critical mRNA measures and gene clusters associated with AD pathogenesis.**Methods:** We used a transcriptomics approach, integrating magnetic resonance imaging (MRI) and peripheral blood-based gene expression data using persistent homology (PH) followed by kernel-based clustering.**Results:** We identified three clusters of genes significantly associated with diagnosis of amnesic MCI. The biological processes associated with each cluster were mitochondrial function, NF- $\kappa$ B signaling, and apoptosis. Cluster-level associations with cortical thickness displayed canonical AD-like patterns. Driver genes from clusters were also validated in an external dataset for prediction of amyloidosis and clinical diagnosis.**Discussion:** We found a disease-relevant transcriptomic signature sensitive to prodromal AD and identified a subset of potential therapeutic targets associated with AD pathogenesis.**KEYWORDS**Alzheimer's disease, apoptosis and cell proliferation, cortical thickness, gene expression, imaging genetics, NF- $\kappa$ B signaling, risk genes

## 1 | INTRODUCTION

The prodromal stages of Alzheimer's disease (AD) can provide an essential window in which disease-modifying and preventative interventions can be maximally effective. Currently established biomarkers like positron emission tomography (PET), amyloid imaging and cerebrospinal fluid (CSF) tau/amyloid beta ( $A\beta$ ) are expensive and/or invasive. Structural imaging biomarkers while less expensive are non-specific. Blood-based biomarkers are important because they represent a less invasive and potentially cheaper approach for aiding AD detection and therapeutic discovery.

For a late-life complex multifactorial disease such as AD, with both genetic and environmental factors involved, integrating multiple layers of genetic, imaging, and other biomarker data is a critical step in identifying distinct pathogenic profiles and uncovering novel dysregulated pathways/biological processes for therapeutics. Additionally, the benefit from biomarker-driven risk assessment tools lies in eliminating the need for decisions based solely on clinical parameters, especially during the latent stages of AD in which clinical manifestation is inconclusive. Currently, biomarkers are the only feasible approach for identifying and estimating disease-related traits in early stages of AD when a therapeutic intervention can achieve its greatest impact.

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2022 The Authors. *Alzheimer's & Dementia* published by Wiley Periodicals LLC on behalf of Alzheimer's Association

Evidence for a genetic component in neurodegenerative disorders like AD is overwhelming. Autosomal dominant pathogenic mutations in *APP*, *PSEN1*, and *PSEN2* are identified in only  $\approx 6\%$  of patients,<sup>1</sup> accounting for a small portion of AD patients. At least 21 additional genetic risk loci have been identified for the genetically complex sporadic form of AD in genome-wide association studies (GWAS) and massive parallel resequencing (MPS) efforts.<sup>2,3</sup> These studies re-emphasize the multifactorial nature of AD. Emerging data suggest that peripheral blood may provide a suitable surrogate for gene expression patterns identified in the central nervous system (CNS). Peripheral blood microarray studies have successfully identified candidate genes for Parkinson's disease—another common neurodegenerative disorder in the elderly that, like AD, is caused by misfolding and deposition of aberrant protein species.<sup>4</sup> In addition, systematic evaluation of comparability of gene expression in blood and brain has shown that whole-blood gene expression profile shares significant similarities with that of CNS tissues.<sup>4</sup> In contrast to DNA sequencing transcriptomics, the study of gene expression profiles informs not only about inherited but also non-inherited genomic signals.<sup>4</sup> Microarray and RNA-sequencing based transcriptome studies of *post mortem* brains of AD and control subjects have identified differentially expressed genes, yielding a core set of differentially expressed pathways including immune response, apoptosis, cell proliferation, energy metabolism, and synaptic transmission.<sup>5,6</sup> Pathways involved in inflammation, DNA damage response, cell cycle and neuronal homeostasis were found to be dysregulated in peripheral blood.<sup>7</sup> Several studies have shown that an AD-specific mRNA signature that differentiates AD from cognitively normal (CN) controls can be detected in peripheral blood.<sup>8,9</sup> There is still a wide gap of knowledge on transcriptomic profiling for risk analysis and how it may, in conjunction with other cognitive, imaging and blood biomarkers, inform predictive modeling and therapeutic development. In the following analyses, we aimed to identify relevant transcriptomic signatures sensitive to cortical atrophy and amnesic MCI diagnosis. These signatures consist of a limited number of critically important RNA measures that capture early disease specificity, while minimizing noise and tissue level variability.

We applied a transcriptomics approach integrating genetic and neuroimaging data using an applied mathematical tool known as persistent homology (PH) followed by a statistical kernel-based clustering. We further investigated the driver genes for variants or single nucleotide polymorphisms (SNPs), which have significant association with amnesic MCI diagnosis. The clustering solution and genes obtained from our novel pipeline were further validated in an external dataset derived from the Alzheimer's Disease Neuroimaging Initiative (ADNI).

## 2 | METHODS

### 2.1 | Dataset

Our discovery sample included 160 subjects from the Imaging and Genetic Biomarkers of Alzheimer's Disease (ImaGene) study who were clinically diagnosed as either (1) MCI (N = 108) or (2) CN (N = 52). The MCI group was further divided into those presenting with amnesic

### RESEARCH IN CONTEXT

**Systematic review:** The authors reviewed the literature using traditional sources. Blood biomarkers is an important area of Alzheimer's disease (AD) research. While there is a breadth of research on utility of blood and serum biomarkers, there are some aspects of transcriptomic data from blood that are yet to be addressed, specifically with respect to brain atrophy and mild cognitive impairment.

**Interpretation:** Our methods combined neuroimaging markers with transcriptomic data to identify clusters of genes and driver genes associated with prodromal AD using a novel persistent homology pipeline followed by kernel-based clustering. The findings from our results were validated in an external dataset.

**Future directions:** Peripheral blood transcriptomic data at earlier stages of AD can be used to identify and screen for multiple novel therapeutic targets and as a preliminary risk assessment/prognostic tool. It can also lead to better understanding of critical disease-related pathways and pathogenic mechanisms.

tic MCI (aMCI, N = 70) or those presenting with the non-amnesic phenotype (naMCI, N = 38; Tables 1).

#### 2.1.1 | Imaging

The detailed imaging protocol has been previously published.<sup>10</sup> All subjects received annual 1.5 T magnetic resonance imaging (MRI) scans following the University of California Los Angeles Alzheimer's Disease Research Center protocol consisting of coronal F13 D T1 MPRAGE: TR 28, TE 4.5, FOV 22 cm, matrix 256 × 192, slice/gap 1.5/0 mm. Measures of neurodegeneration were obtained from coronal T1-weighted MPRAGE sequences. Scans were processed using the FreeSurfer (version 6.0) longitudinal pipeline, to obtain region-specific and global measures of atrophy.<sup>10</sup>

#### 2.1.2 | Microarray-based gene expression

All subjects provided yearly peripheral blood RNA. Total RNA was extracted using the PAXgene blood RNA kit (Qiagen). Total RNA (200 ng) was amplified, labeled, and hybridized on Illumina Human BeadChips, querying the expression of  $\approx 24$  K RefSeq-curated gene targets. Slides were processed and scanned with Illumina BeadStation platform. Raw data were collected, loaded in the statistical software R, and log transformed. Poor-quality arrays were excluded from further analyses. Data were normalized using quantile normalization. mRNA levels were  $\log_2$ -transformed.

All subjects provided DNA at baseline. DNA was labeled, fragmented, and hybridized on Illumina IM chips according to Illumina instructions. The Illumina 1 M SNP array assays 1.2 mln markers

**TABLE 1** ImaGene demographics

| Variable                           | NC<br>(n = 52) | Amnesic MCI<br>(n = 70) | Non-amnesic MCI<br>(n = 38) | P-value |
|------------------------------------|----------------|-------------------------|-----------------------------|---------|
| Age, years                         | 69.0 (7.9)     | 69.8 (8.5)              | 69.8 (8.5)                  | .9      |
| Education, years                   | 17.6 (2.04)    | 15.5 (2.7)              | 16.5 (2.88)                 | .001*   |
| Sex (M/F)                          | 30/21          | 26/43                   | 20/18                       | n.s     |
| MMSE                               | 28.8 (1.2)     | 27.0 (2.5)              | 27.9 (1.9)                  | <.001*  |
| Hippocampal vol (mm <sup>3</sup> ) | 8602 (1092)    | 7990 (1324)             | 8718 (1023)                 | .002*   |

per sample, including >100,000 copy number variants, providing the highest genotyping density available on the market. Arrays were scanned using Illumina iScan equipment. Proprietary software performance was compared to new segmentation methods available within the Bioconductor project (“affyio” and “oligo” packages) showing improved calling performance over the Illumina software. Low-confidence calls and SNPs not in Hardy-Weinberg equilibrium were excluded from further analyses. We have already imputed all missing genotypes using MACH and minimac in a two-stage procedure using the 1000 Genomes project pilot data as a reference panel for inferring missing genotypes. Minimac yielded the posterior probabilities of the imputed genotypes at un-genotyped marker loci for everyone.  $r^2$  value equal to 0.30 was set as the threshold to accept each imputed genotype.

## 2.2 | Preliminary data reduction

To reduce noise and find endophenotype-specific transcriptomic data from blood, we performed an initial data-reduction step by selecting transcripts significantly associated with both hippocampal volume and average cortical thickness. A simple linear regression model was applied to each of the 25,000 transcripts with hippocampal volume and cortical thickness as the outcome variable to select the unique union of transcripts that were significantly ( $P < .05$ ) associated to both neurodegeneration measures.

## 2.3 | Univariate surface mapping

An average surface was constructed by computing the Talairach coordinates at each vertex for each subject.<sup>11</sup> The scans were then averaged using FreeSurfer6.0. Vertex-wise regressions for each of the transcripts was performed using age, sex, and education as covariates to map the association of the gene expression value with average cortical thickness using a MATLAB toolbox SurfStat.<sup>12</sup> The beta coefficients were used for the persistent homology pipeline.

## 2.4 | Persistent homology

The topological method known as PH builds a data-driven coarse descriptor of a weighted discretized surface while retaining meaningful geometric information.<sup>13</sup> We analyzed the SurfStat triangular meshes

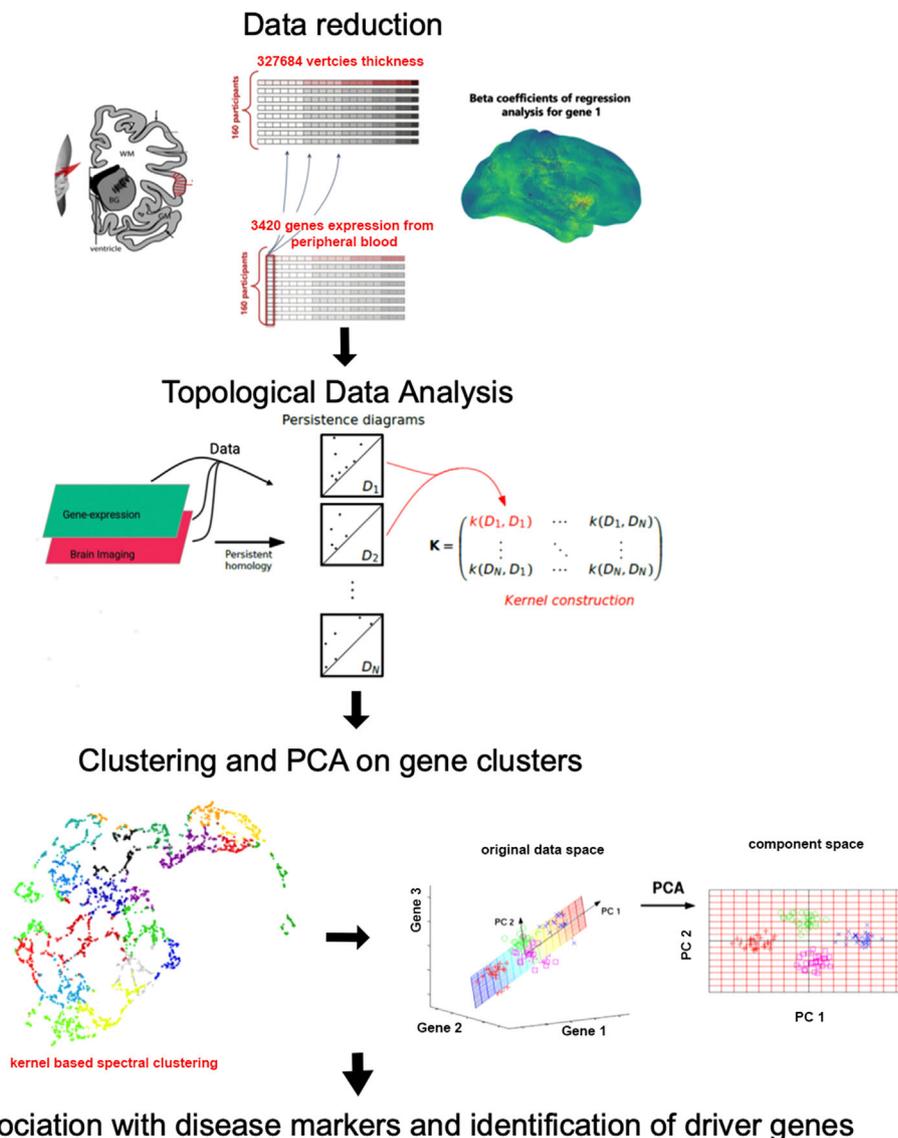
using PH to characterize each map through the evolution of homological features (Figure 1A) across increasing vertex-wise thresholds. Vertices were normalized between [0,1] across all 3420 cortical genes maps. The PH algorithm analyzes the mesh and adds a triangle when all tree vertices have weight below or equal to the threshold. As the thresholds increase, topological features of the surface (components and holes) appear. These features can disappear either by merging with older ones in the case of components, or by being filled up in the case of holes. The birth and death of each feature and the number of features characterize the shape and intensity of the map. This information can be summarized in Betti curves, which represent the number of features present at different threshold values. The Betti curves are indicative of the global distribution of the beta coefficients along the cortical surface. To compare and cluster the maps, we computed the Manhattan distance<sup>14</sup> between each pair of Betti curves. The distance between two Betti curves is then the difference in number of features (components or cycles) across all thresholds (Figure 2B). Two distance matrices were built, one for each dimension of the features under study (components and holes), representing pairwise similarity or dissimilarity in homological features between the cortical gene maps.

## 2.5 | Unsupervised kernel clustering

Multiple kernel learning is an established framework for representation and integration of different modalities of data, including vectors, strings, graphs, and topological features.<sup>15</sup> PH can be fit into the framework via the mathematical representation of kernel matrices.<sup>16</sup> Based on the Betti curve differences, we built Laplace radial basis kernels representing pairwise similarity in homological features between the association pattern of genes and cortical thickness. We then applied a kernel spectral clustering algorithm from the kernelab package in R.<sup>17</sup> This clustering algorithm was applied to the components distance matrix only. To choose the optimum number of clusters, we applied two methods—the elbow method and the silhouette analysis—for choosing the number of clusters.

## 2.6 | Cluster analysis

To identify disease-relevant clusters, we performed principal component decomposition on each cluster to represent the cluster by its first

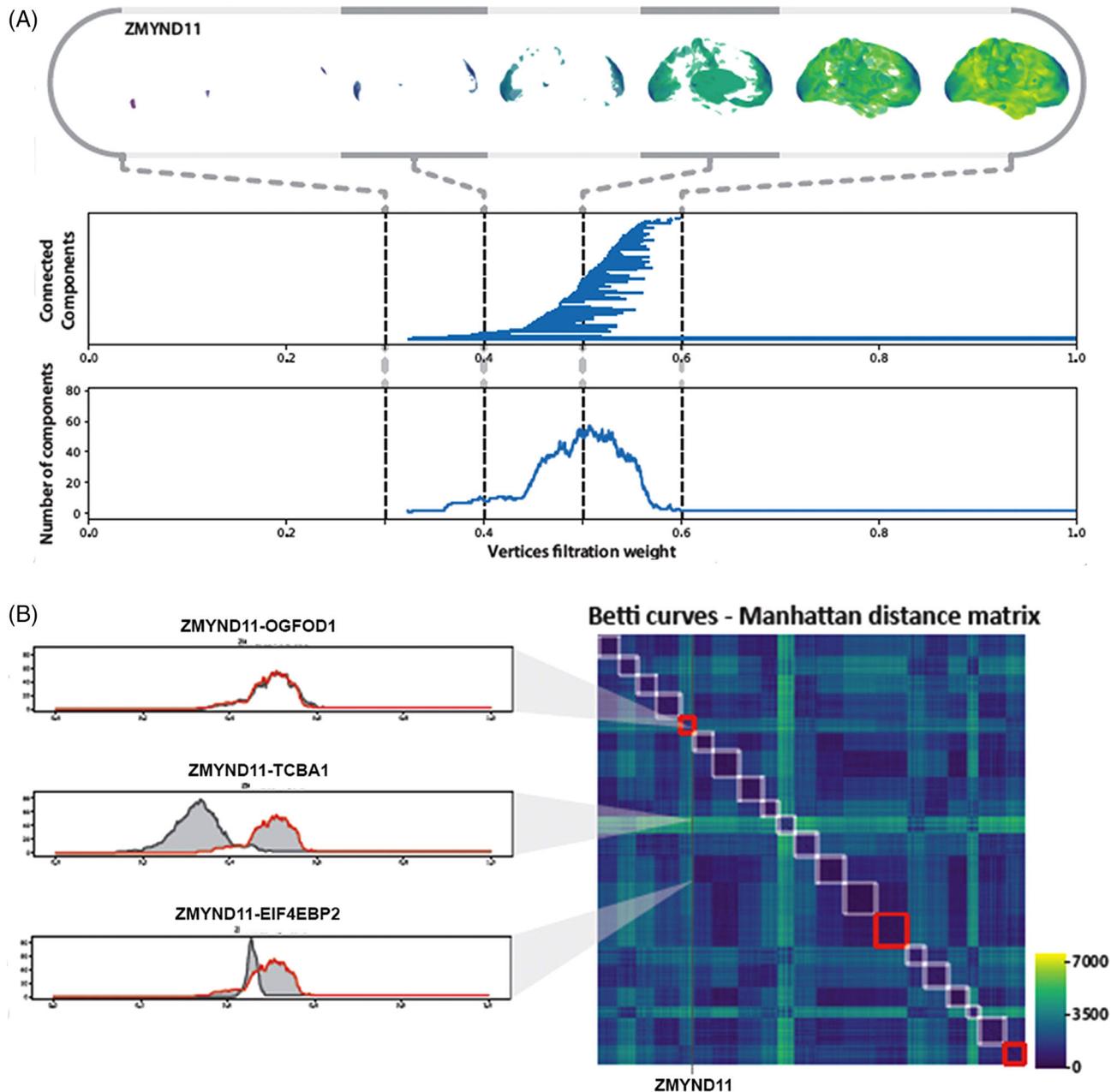


**FIGURE 1** Overview of analyses pipeline. PCA, principal component analysis

principal component or the “eigen-gene” representing the set of genes in that cluster.<sup>18</sup> The eigen-gene represents a data vector of values that summarizes the gene expression values within a given cluster for a given participant. These individual eigen-gene values were used to identify clusters that were differentially expressed in aMCI versus CN pooled with naMCI (two-sample *t*-test). Assumption for normality was verified for the *t*-tests. We compared the aMCI group to the combined naMCI and CN group as there were no significant differences between naMCI and CN (one-way analysis of variance,  $P > .05$ ). False discovery rate (FDR) correction was applied to identify significant clusters that survived multiple testing correction and had differential expression at the cluster level between diagnostic groups. For validation, we permuted a null model with random assignment of cluster number while maintaining cluster size and distribution for 1000 iterations and calculated the number of significant clusters (two-sample *t*-test,  $P_{\text{fdr}} < .05$ ) for each iteration followed by FDR correction.

## 2.7 | Gene enrichment analysis and driver genes identification

Gene enrichment analysis was performed using the topGO package in R using the weight01 algorithm.<sup>19</sup> *P*-values computed by two-sample *t*-test comparing aMCI versus the grouped CN and naMCI subjects were used as scores for the Gene Ontology (GO) analysis to select the most disease relevant biological processes associated with the cluster of genes. From the disease-relevant clusters, driver genes were identified by performing differential expression analysis within clusters followed by FDR correction. The REVIGO tool was used to visualize the enriched biological processes. REVIGO summarizes long GO lists by reducing functional redundancies and visualizes the remaining GO terms in two-dimensional plots and semantic similarity measures between GO terms are calculated based on pre-established methods.<sup>20</sup> The GO terms from the gene enrichment analysis were provided to REVIGO for



**FIGURE 2** Persistent homology pipeline representation with one gene. A, The filtration obtained from the cortical map relative to gene *ZMYND11*. For increasing threshold values (on the x-axis), we record the appearance of topological features of the surface (individual components and holes) and when these features disappear either by merging with older ones in the case of individual components or being filled up in the case of holes. We can represent each feature as a horizontal bar whose extreme points are the birth and death of each filtration value a-b. Moreover, we can count the number of features present as we increase the threshold value. This gives us a Betti curve indicative of the global distribution of Beta values along the cortical surface. B, Distance matrices: Manhattan distance matrix for the Betti curves with genes sorted according to cluster labels for genes

visualization of enriched processes. The driver genes were further investigated for role and function in AD pathology.

## 2.8 | Cluster-level associations

Using the “eigen-gene” approach and vertex-wise regressions with cortical thickness in SurfStat, we investigated cluster-level associations

with cortical thickness to identify patterns or region-specific presentation of cortical atrophy.

## 2.9 | Variant analyses in driver genes

We used the tool MAGMA (multi-marker analysis of genomic annotation)<sup>21</sup> to analyze variants in the driver genes. The gene analysis

**TABLE 2** ADNI demographics

| VariableMean (std) | CN (n = 154) | EMCI (n = 215) | LMCI (n = 104) | AD (n = 42) | P-value  |
|--------------------|--------------|----------------|----------------|-------------|----------|
| Age, years         | 73.7 (6.01)  | 71.0 (7.44)    | 72.0 (7.37)    | 75.3 (9.37) | .0001*   |
| Sex (M/F)          | 72/82        | 115/100        | 57/47          | 25/17       | n.s      |
| MMSE               | 29.0 (1.21)  | 28.3 (1.49)    | 27.6 (1.76)    | 22.9 (1.92) | < .0001* |
| AV45 SUVR          | 1.11 (0.19)  | 1.16 (0.21)    | 1.27 (0.23)    | 1.39 (0.22) | < .0001* |

in MAGMA is based on a multiple linear principal components regression model, using an F-test to compute the gene-variant-phenotype *P*-value. This model first projects the SNP matrix for a gene onto its principal components (PC), pruning away PCs with very small eigenvalues, and then uses those PCs as predictors for the phenotype in the linear regression model. We first annotated the SNPs onto genes using the imputed raw genotype data for our cohort and conducted a gene-level analysis step to compute associations between SNPs in the driver genes and the aMCI phenotype.

## 2.10 | Validation in an external dataset

The clusters or sets of genes obtained through our pipelines were validated in an external dataset consisting of MRI, gene expression (specific transcripts used for the discovery analysis), demographic, and amyloid PET data. We identified 515 subjects from the ADNI study<sup>22</sup> with the required overlapping data types used in our ImaGene discovery datasets (Table 2). Data used in the preparation of this article were obtained from the ADNI database ([adni.loni.usc.edu](http://adni.loni.usc.edu)). Pre-processed data were downloaded wherein processing for mRNA values and neuroimaging analyses were performed according to ADNI protocol.<sup>23</sup> The gene expression data was quantile normalized and log<sub>2</sub> transformed. We applied a logistic regression classification (with age and sex as covariates) for amyloid positivity and clinical diagnosis using the driver genes identified from our analyses in the ADNI dataset. Amyloid positivity was predicted using a logistic regression model positivity (florbetapir standardized uptake value ratio [SUVR] > 1.11 in ADNI and flutemetamol SUVR > 1.17 in ImaGENE). These cut-offs for florbetapir and flutemetamol<sup>24</sup> have been established according to previously published data. We also analyzed the FreeSurfer-derived values for average cortical thickness, average inferior temporal thickness, and average parietal thickness available in the ADNI dataset and their association with the significant gene clusters identified in our discovery dataset.

## 3 | RESULTS

### 3.1 | Subject demographics and identification of gene set

The three diagnostic groups had no significant difference in mean age and sex distribution. There was a significant difference in mean number

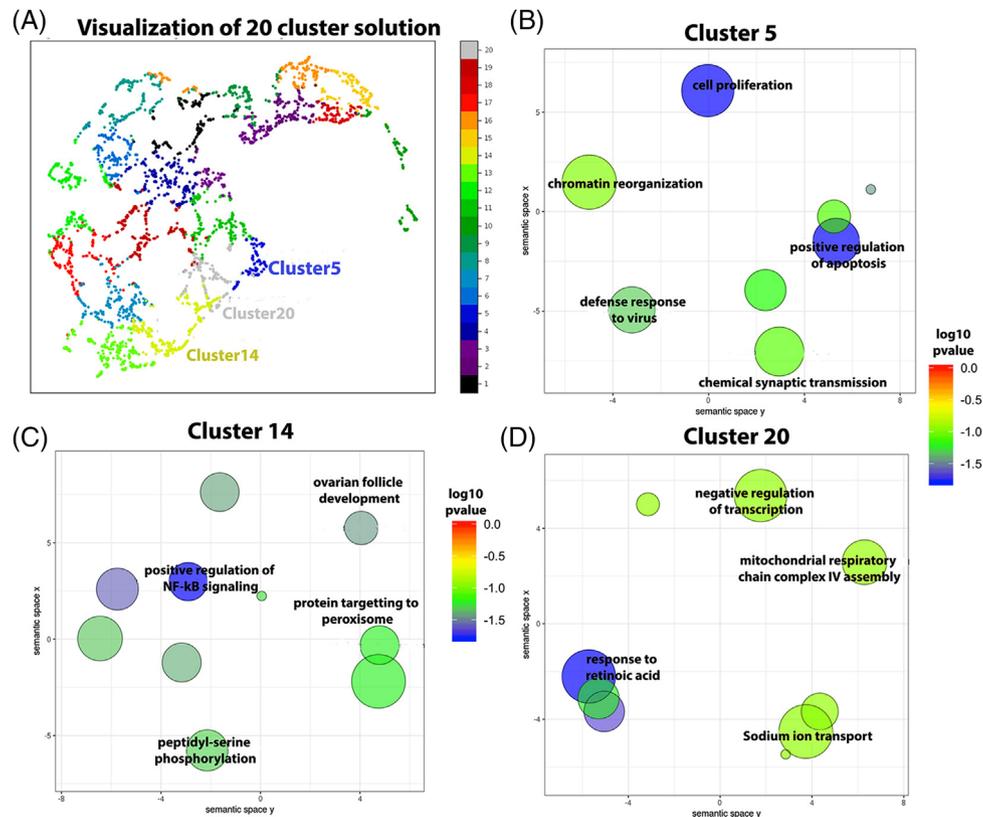
of years of education, Mini-Mental State Examination and hippocampal volume ( $P < .05$ ; Table 1). Using a simple linear model, we identified 3420 genes from  $\approx 25,000$  transcripts that were significantly associated with both hippocampal volume and cortical thickness mapped onto the group average cortical thickness using vertex-wise regression with age, sex, and education as covariates to obtain Betti distances from PH.

### 3.2 | Cluster analysis and significant clusters

We obtained an optimal clustering solution of 20 clusters from the elbow and silhouette analysis. Figure 3A shows the representation of the 20-cluster solution. We identified three clusters significantly associated with disease diagnosis post-FDR correction (two-sample *t*-test,  $P < .05$ )—cluster-5, cluster-14, and cluster-20. For the null model validation, 1000 iterations yielded our three-cluster solution outside the 95% confidence interval post-FDR correction (two-sample *t*-test,  $P_{\text{fdr}} < .05$ ), which provides convincing evidence of the presence of the three clusters having significant biological relevance and being sensitive to disease diagnosis. Cluster 5 consisted of 118 genes, of which two genes were differentially expressed ( $P_{\text{fdr}} < .05$ ). Cluster 14 consisted of 255 genes, of which 53 were differentially expressed ( $P_{\text{fdr}} < .05$ ) and cluster 20 consisted of 157 genes, of which two were differentially expressed ( $P_{\text{fdr}} < .05$ ). All 3420 genes along with their clustering solution are summarized File S1 in supporting information.

### 3.3 | Biological processes associated with clusters and driver genes

The gene enrichment analysis identified the overrepresented biological pathways in the significant clusters based on the *P*-values from the differential expression. Gene enrichment results using differentially expressed genes are summarized in Table S1 in supporting information. Positive regulation of apoptotic processes and cell proliferation were the most significant biological processes associated with cluster 5 (Figure 3B). Driver gene analysis yielded two differentially expressed cluster-5 genes, *SPINK6* (serine peptidase inhibitor Kazal type 6) and *ZMYND11* (zinc finger MYND-type containing 11;  $P_{\text{fdr}} < .05$ ). Cluster-14 was associated with NF- $\kappa$ B signaling pathway (Figure 3C). There were 53 differentially expressed genes in the cluster. For cluster-20, the main overrepresented pathway was cellular response to retinoic



**FIGURE 3** A, Visualization of 20 cluster solution. Three clusters were significantly associated with amnesic mild cognitive impairment diagnosis post–false discovery rate correction; cluster-5,-14, and -20. Gene enrichment analysis with significant over-represented pathways in (B) cluster-5, (C) cluster-14, and (D) cluster-20. Blue and green bubbles are Gene Ontology (GO) terms with more significant  $P$ -values than the orange and red bubbles. The bubbles'  $x$  and  $y$  coordinates were derived by applying multidimensional scaling to a matrix of the GO terms' semantic similarities; consequently, their closeness on the plot should closely reflect their closeness in the GO graph structure that is, the semantic similarity

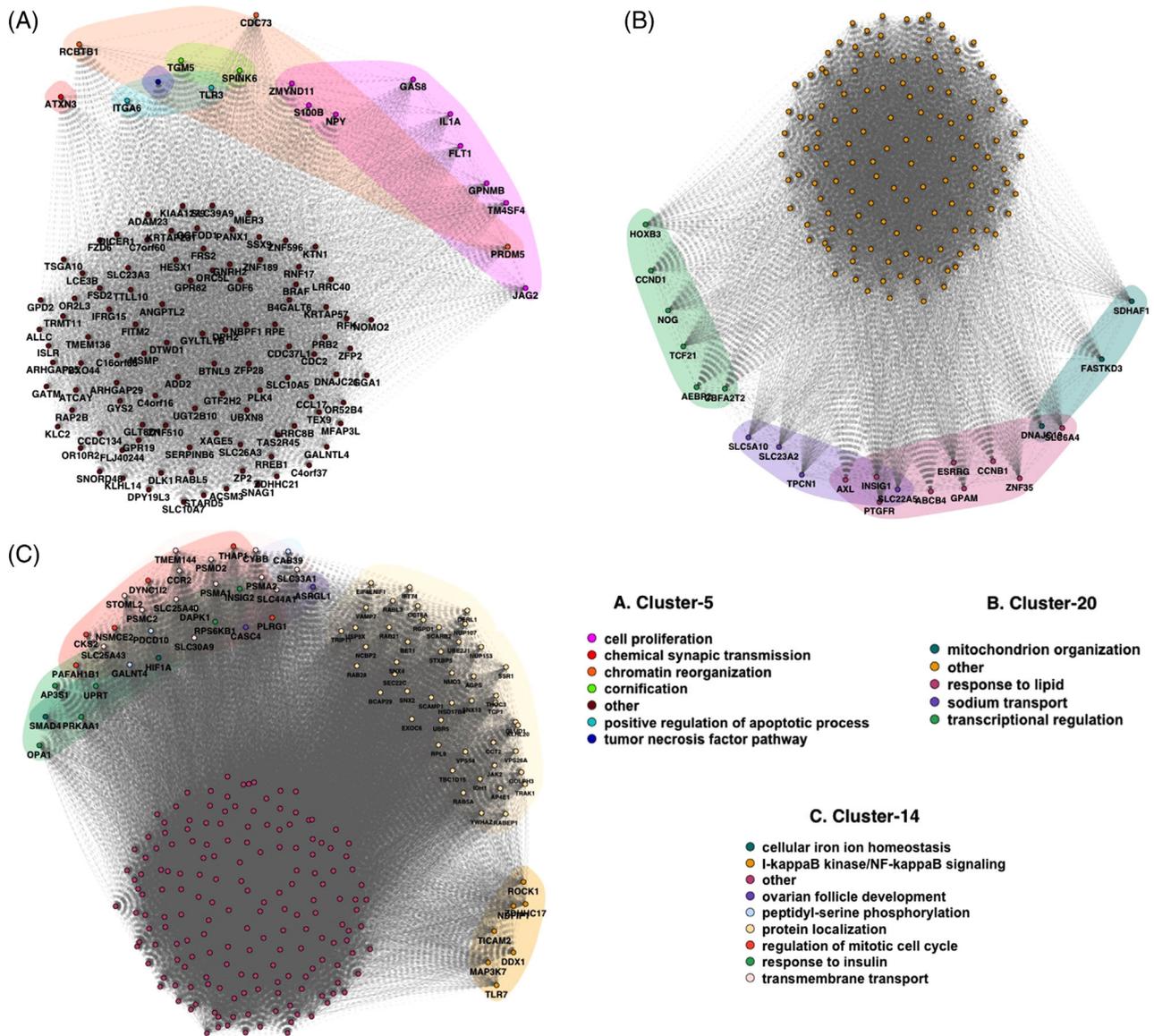
acid and mitochondrial respiratory chain complex processes (Figure 3D). Gene enrichment results with the top 10 enriched biological pathways using all genes within the significant clusters are summarized in Table S2 in supporting information. Figure 4 represents gene networks of the significant clusters with nodes grouped by top over-represented biological processes based on the GO analyses. GO analyses using all genes in the cluster yielded “activation of GTPase activity,” as the significantly overrepresented pathway for cluster-5. For cluster-14, significant biological processes were “positive regulation of translation” and “chaperone-mediated protein complex assembly” and for cluster-20, “transmembrane transport,” “cellular response to interleukin-1,” and “cellular response to tumor necrosis factor” (Table S2). All 57 driver genes were downregulated in aMCI. The driver genes and their fold change are summarized in Table 3. The molecular signature for the driver gene expression data is shown in Figure 5. In the ImaGene sample, the aMCI group shows a marked downregulation of the driver genes compared to normal controls and naMCI (Figure 5A). In the ADNI sample, the late MCI and AD groups show reasonable under expression of the driver genes compared to the other two groups—CN and early MCI—but this difference was not as pronounced as it was in the ImaGene sample (Figure 5B).

### 3.4 | Cluster-level associations

SurfStat mapping of cluster-level associations with cortical thickness showed AD-like patterns (Figure 6).<sup>25</sup> All three clusters showed negative association with cortical thickness in the medial, inferior, and lateral temporal; the precuneus; posterior cingulate; lateral parietal; and the frontal lobes. These were most significant for cluster-5 (regulation of apoptotic processes and cell proliferation) and least significant for cluster-14 (NF-kB signaling pathway).

### 3.5 | Variant analyses

We identified two driver genes, *ARMC10* and *KIAA1468*, from the 57 driver genes with SNPs significantly associated with aMCI in our sample. Forty-seven SNPs were annotated to the *ARMC10* gene on chromosome 7, which had a significant association to aMCI phenotype with  $P$ -value of .02 and  $Z$ -statistic 1.998. For *KIAA1468* on chromosome 18, 537 SNPs were annotated to the gene, which had a significant association to aMCI phenotype with  $P$ -value .04 and  $Z$  statistic 1.697. The results from the MAGMA variant analyses are summarized in Table 4.



**FIGURE 4** Gene clusters visualizing nodes and relevant biological processes. The top enriched biological processes obtained through gene ontology analyses and the gene nodes in the cluster represented for A) Cluster-5 B) Cluster-20 and C) Cluster-14

We further investigated the association of individual SNPs within the genes to disease phenotype using PLINK<sup>26</sup> GWAS tools but none of the individual SNPs had a significant genome-wide association post-FDR correction.

### 3.6 | Validation in ADNI

Our logistic regression model (with age and sex as covariates) to predict amyloid positivity using the driver gene transcripts yielded an area under the curve (AUC) of 0.74 in ADNI and 0.73 in ImaGene (Figure 7A). Our logistic regression model to predict aMCI or AD diagnosis in ADNI produced an AUC of 0.71 and 0.78, respectively (Figure 7B). Cluster-5 and cluster-20 were significantly associated with average cortical thickness, inferior temporal thickness (Figures 7C–7D), and

inferior parietal thickness in the ADNI dataset ( $P < .05$ , data not shown).

## 4 | DISCUSSION

The need for a cost-effective first step in a multistage diagnostic framework in AD is an integral part of biomarker development in AD research. With the evidence being what it is about the multifactorial nature of AD, multiple therapeutic targets need to be identified and tested with respect to the various phenotypic manifestations of the disease. While there are many highly sensitive plasma and fluid biological markers that have shown promising results in predictive and prognostic models, mRNA measures can be critical and add to this paradigm because they are directly correlated to physiological changes and have

**TABLE 3** Identified driver genes and their foldchange (controls vs. aMCI)

| Downregulated in amnesic MCI |            |                  |
|------------------------------|------------|------------------|
| Driver-gene                  | foldchange | $P_{\text{fdr}}$ |
| AGPS                         | -0.012     | 0.009            |
| AP4E1                        | -0.013     | < 0.001          |
| ARID4B                       | -0.013     | 0.007            |
| ARMC10                       | -0.018     | 0.001            |
| B3GNT2                       | -0.017     | < 0.001          |
| BCAP29                       | -0.017     | 0                |
| C6orf111                     | -0.021     | < 0.001          |
| CAB39                        | -0.016     | 0.004            |
| CAMSAP1                      | -0.013     | < 0.001          |
| CASC4                        | -0.013     | 0.001            |
| CKS2                         | -0.016     | 0.005            |
| COQ2                         | -0.015     | 0.005            |
| DDX1                         | -0.014     | 0.007            |
| DERL1                        | -0.015     | 0.002            |
| DNAJA2                       | -0.014     | 0.006            |
| DYNC1LI2                     | -0.009     | 0.009            |
| FAM3C                        | -0.017     | < 0.001          |
| FASTKD3                      | -0.014     | < 0.001          |
| HMGCS1                       | -0.016     | 0.001            |
| HNRPK                        | -0.021     | < 0.001          |
| HSD17B4                      | -0.011     | 0.004            |
| HSZFP36                      | -0.013     | 0.004            |
| IDH1                         | -0.012     | 0.004            |
| INSIG2                       | -0.016     | 0.003            |
| KIAA1468                     | -0.02      | 0.004            |
| KIAA1826                     | -0.022     | < 0.001          |
| LASS6                        | -0.015     | 0.006            |
| MATR3                        | -0.011     | 0.011            |
| NDFIP1                       | -0.017     | 0.004            |
| PAPD4                        | -0.016     | 0.009            |
| PPP1R2                       | -0.017     | 0.007            |
| PRDM10                       | -0.014     | 0.001            |
| RASA1                        | -0.019     | < 0.001          |
| RFWD2                        | -0.013     | 0.002            |
| ROCK1                        | -0.016     | 0.012            |
| RPS6KB1                      | -0.017     | 0.005            |
| SCAMP1                       | -0.02      | 0.001            |
| SLC22A5                      | -0.014     | < 0.001          |
| SLC25A43                     | -0.013     | 0.006            |
| SLC44A1                      | -0.016     | 0.006            |
| SMAD4                        | -0.017     | 0.002            |
| SNX4                         | -0.019     | < 0.001          |
| SPINK6                       | -0.007     | < 0.001          |

(Continues)

**TABLE 3** (Continued)

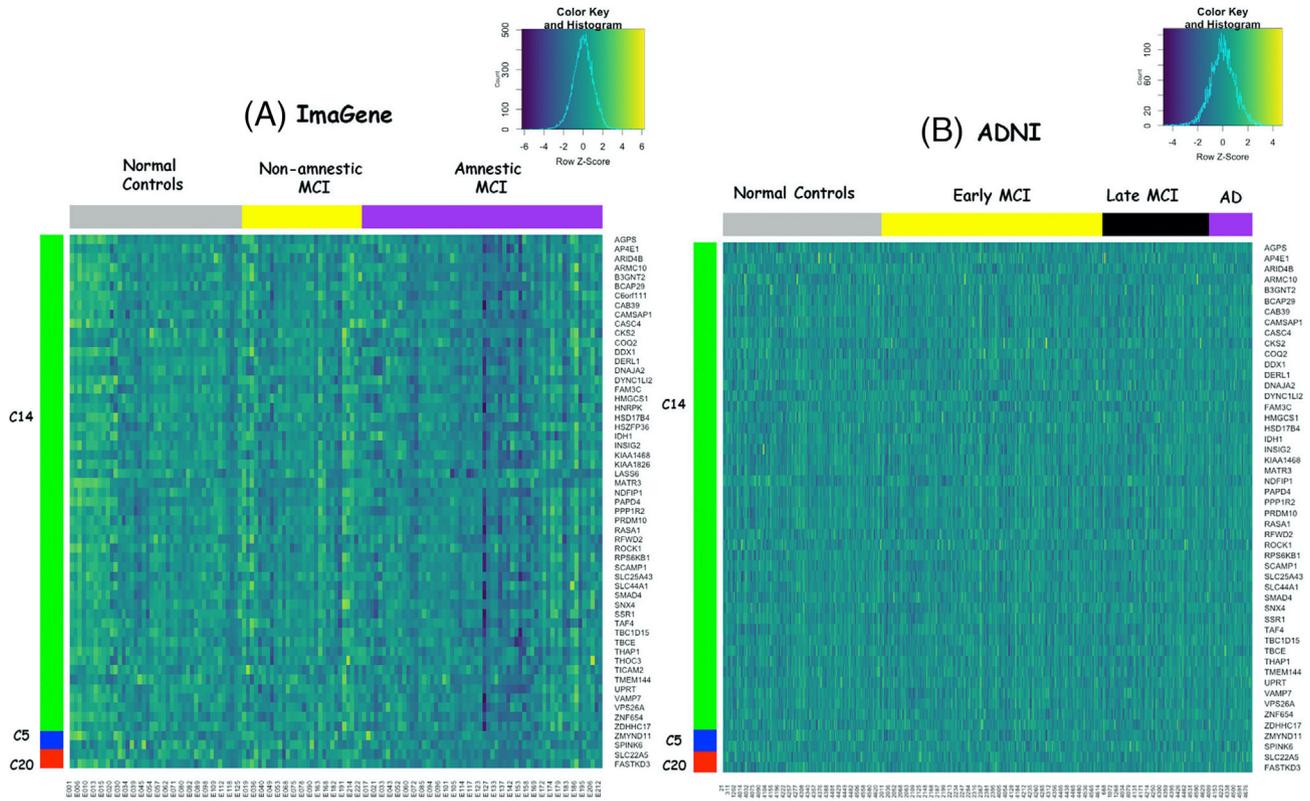
| Downregulated in amnesic MCI |            |                  |
|------------------------------|------------|------------------|
| Driver-gene                  | foldchange | $P_{\text{fdr}}$ |
| SSR1                         | -0.013     | 0.006            |
| TAF4                         | -0.019     | < 0.001          |
| TBC1D15                      | -0.02      | 0.001            |
| TBCE                         | -0.015     | 0.002            |
| THAP1                        | -0.015     | 0.003            |
| THOC3                        | -0.016     | 0.006            |
| TICAM2                       | -0.013     | 0.005            |
| TMEM144                      | -0.012     | 0.006            |
| UPRT                         | -0.018     | < 0.001          |
| VAMP7                        | -0.015     | 0.005            |
| VPS26A                       | -0.014     | 0.007            |
| ZDHHC17                      | -0.02      | 0.001            |
| ZMYND11                      | -0.008     | < 0.001          |
| ZNF654                       | -0.016     | < 0.001          |

multiple downstream and upstream processes that can be used for therapeutic intervention and risk assessment/screening.

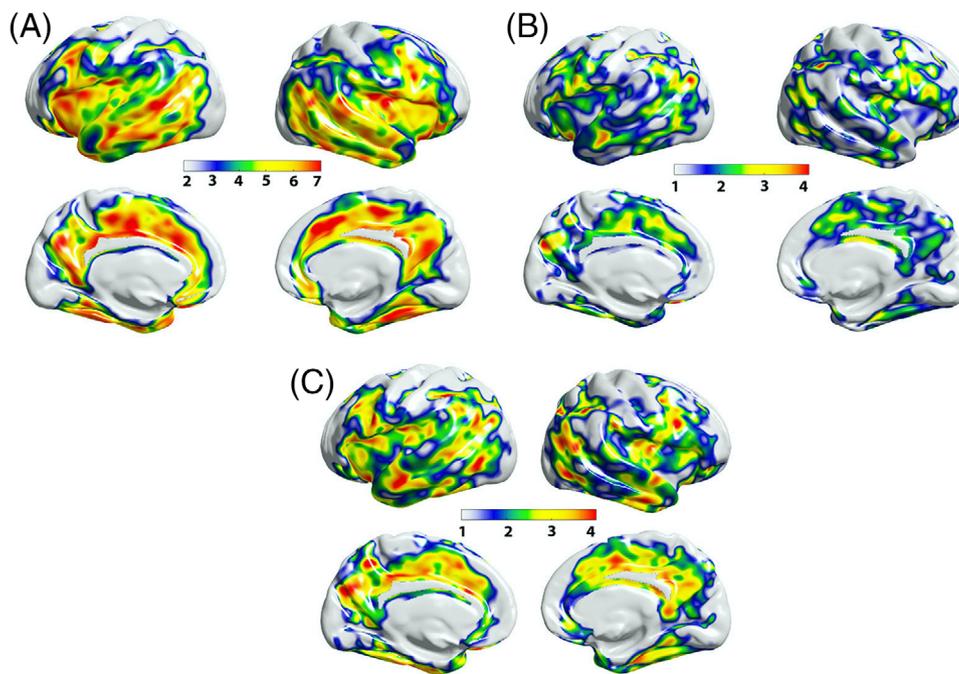
Using a data-driven approach, we successfully reduced large and noisy transcriptomic data from peripheral blood to a significantly smaller gene set specific to neurodegeneration and sensitive to disease diagnosis. Through our novel pipeline, we identified transcripts from peripheral blood that are associated with the aMCI phenotype and can also help in risk prediction for conversion to AD. PH allowed us to compress in single summary the shape and intensity of each brain map giving less weight to small fluctuations in the spatial distribution of the beta coefficient values across the cortex. Moreover, the topological summaries are an optimal tool to reduce the redundancies in the weighted triangular meshes in a way that allowed us to use sophisticated and memory expensive algorithms in a large gene pool.

We identified many AD-relevant genes and few novel genes that can potentially be important therapeutic targets. For cluster-5, the main driver genes identified were *ZMYND11* and *SPINK6*. *ZMYND11* is a protein-coding gene that has been associated with mental retardation, and autosomal dominant non-syndromic intellectual disability.<sup>27</sup> *SPINK6* is a Kazal-type serine protease inhibitor that acts on kallikrein peptidases in the skin. It is a gene associated with keratinization.<sup>28</sup> To date, *SPINK6* has not been linked to CNS processes or diseases.

Cluster-14, which was mainly associated with NF- $\kappa$ B signaling, had 53 differentially expressed driver genes. Eighteen of these genes were found to have a direct relevance in AD pathophysiology. Three of the driver genes (*ROCK1*, *SMAD4*, and *RPS6KB1*) are associated with the transforming growth factor beta (TGF- $\beta$ ) signaling pathway (Table S3 in supporting information). The neuroprotective cytokine TGF- $\beta$  is increased in AD and is associated with chronic neuroinflammation, which is hypothesized to lead to neurodegeneration.<sup>29</sup> *ROCK1* (Rho-associated protein kinase 1) is increased in AD and *ROCK1* depletion reduces A $\beta$  levels in the brain.<sup>30</sup> *SMAD4* regulates TGF- $\beta$  signaling



**FIGURE 5** Molecular signatures of driver genes across diagnostic groups in (A) ImaGene and (B) Alzheimer's Disease Neuroimaging Initiative. Rows represent driver genes and columns represent subjects. Color blocks represent cluster data (row) and diagnosis (column)



**FIGURE 6** Cluster association (T-statistic) with cortical thickness ( $T > 1.96, P < .05$ ): (A) cluster-5, (B) cluster-14, and (C) cluster-20

**TABLE 4** Variant analyses within driver genes

| Gene Name | Chromosome | Start (bp) | Stop (bp) | N-snps | Zstat  | P-value |
|-----------|------------|------------|-----------|--------|--------|---------|
| ARMC10    | 7          | 102715174  | 102740210 | 46     | 1.998  | .022    |
| KIAA1468  | 18         | 59854506   | 59974355  | 532    | 1.697  | .044    |
| SNX4      | 3          | 125165488  | 125239058 | 284    | 1.464  | .071    |
| PAPD4     | 5          | 78908243   | 78982471  | 262    | 1.367  | .085    |
| ROCK1     | 18         | 18529701   | 18691812  | 156    | 1.358  | .087    |
| AP4E1     | 15         | 51200780   | 51298097  | 285    | 1.265  | .102    |
| ZMYND11   | 10         | 180405     | 300577    | 297    | 1.093  | .137    |
| TICAM2    | 5          | 114914339  | 114952142 | 95     | 0.965  | .167    |
| DNAJA2    | 16         | 46989274   | 47007625  | 21     | 0.946  | .172    |
| PPP1R2    | 3          | 195241221  | 195270224 | 91     | 0.918  | .179    |
| TMEM144   | 4          | 159122749  | 159176439 | 130    | 0.811  | .208    |
| LASS6     | 2          | 169312759  | 169631644 | 1210   | 0.775  | .219    |
| IDH1      | 2          | 209100951  | 209120478 | 54     | 0.772  | .220    |
| DERL1     | 8          | 124025404  | 124054663 | 94     | 0.766  | .221    |
| CAMSAP1   | 9          | 138700333  | 138799060 | 220    | 0.725  | .234    |
| PRDM10    | 11         | 129769601  | 129872730 | 411    | 0.677  | .249    |
| C6orf111  | 6          | 99846534   | 99873263  | 69     | 0.667  | .252    |
| DDX1      | 2          | 15731745   | 15771235  | 139    | 0.646  | .259    |
| CKS2      | 9          | 91926113   | 91931618  | 24     | 0.516  | .302    |
| ZNF654    | 3          | 88108394   | 88193814  | 257    | 0.510  | .305    |
| ZDHC17    | 12         | 77157854   | 77247481  | 379    | 0.470  | .319    |
| NDFIP1    | 5          | 141488324  | 141534008 | 206    | 0.465  | .321    |
| VPS26 A   | 10         | 70883908   | 70932617  | 159    | 0.318  | .371    |
| RPS6KB1   | 17         | 57970407   | 58027787  | 120    | 0.232  | .408    |
| SPINK6    | 5          | 147582357  | 147594700 | 47     | 0.198  | .421    |
| RFWD2     | 1          | 175913967  | 176176386 | 804    | 0.196  | .422    |
| CAB39     | 2          | 231577557  | 231685790 | 335    | 0.188  | .425    |
| THAP1     | 8          | 42691817   | 42698474  | 13     | 0.159  | .436    |
| SLC22A5   | 5          | 131705396  | 131731306 | 128    | 0.102  | .459    |
| SMAD4     | 18         | 48556583   | 48611412  | 123    | 0.040  | .484    |
| FASTKD3   | 5          | 7859272    | 7869150   | 38     | 0.014  | .494    |
| B3GNT2    | 2          | 62423262   | 62451866  | 77     | 0.009  | .496    |
| ARID4B    | 1          | 235330210  | 235491532 | 535    | -0.047 | .518    |
| THOC3     | 5          | 175386534  | 175395318 | 1      | -0.052 | .520    |
| TAF4      | 20         | 60549854   | 60640866  | 397    | -0.133 | .553    |
| AGPS      | 2          | 178257471  | 178408564 | 503    | -0.146 | .558    |
| MATR3     | 5          | 138609441  | 138667366 | 95     | -0.181 | .572    |
| SSR1      | 6          | 7281283    | 7313541   | 202    | -0.204 | .581    |
| TBC1D15   | 12         | 72233487   | 72320629  | 280    | -0.352 | .637    |
| CASC4     | 15         | 44580909   | 44707959  | 237    | -0.513 | .696    |
| INSIG2    | 2          | 118846033  | 118867604 | 62     | -0.517 | .694    |
| COQ2      | 4          | 84184972   | 84206067  | 58     | -0.572 | .716    |
| TBCE      | 1          | 235530728  | 235612280 | 333    | -0.580 | .719    |
| HSD17B4   | 5          | 118788138  | 118878030 | 419    | -0.591 | .722    |
| RASA1     | 5          | 86564070   | 86687743  | 263    | -0.591 | .722    |

(Continues)

**TABLE 4** (Continued)

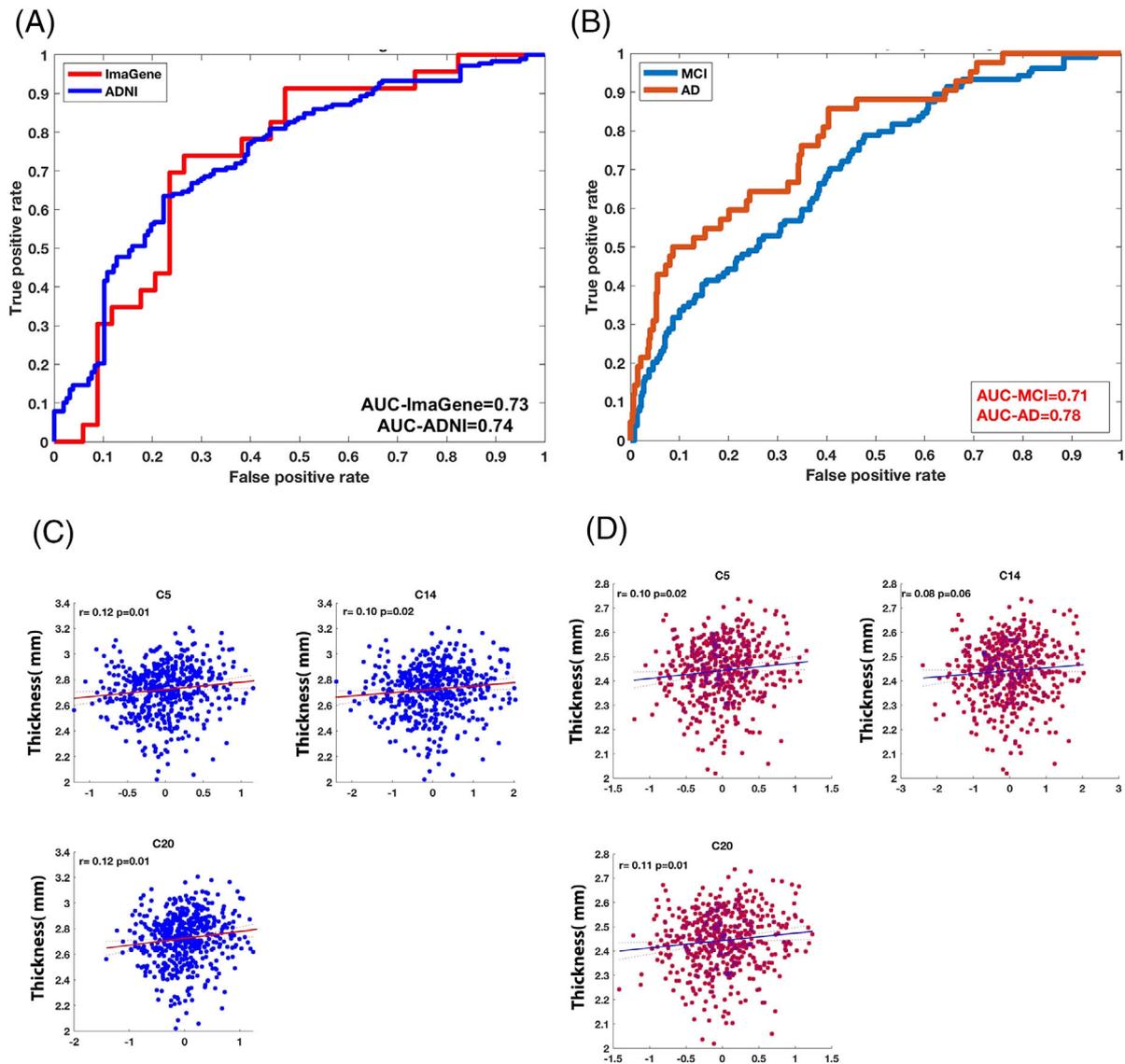
| Gene Name | Chromosome | Start (bp) | Stop (bp) | N-snps | Zstat  | P-value |
|-----------|------------|------------|-----------|--------|--------|---------|
| BCAP29    | 7          | 107220422  | 107263762 | 111    | -0.593 | .723    |
| HSZFP36   | 19         | 11832080   | 11849824  | 61     | -0.667 | .747    |
| FAM3C     | 7          | 120988905  | 121036422 | 150    | -0.822 | .794    |
| SCAMP1    | 5          | 77656339   | 77776562  | 395    | -0.924 | .822    |
| DYNC1LI2  | 16         | 66754796   | 66785526  | 86     | -1.033 | .849    |
| SLC44A1   | 9          | 108006906  | 108200785 | 572    | -1.142 | .873    |
| KIAA1826  | 11         | 105878629  | 105892981 | 34     | -1.302 | .903    |
| HNRPK     | 9          | 86582998   | 86595692  | 37     | -1.481 | .930    |
| HMGCS1    | 5          | 43287572   | 43313614  | 79     | -2.001 | .977    |

pathway through feedback mechanism.<sup>31</sup> Reducing *RPS6KB1* expression has been shown to improve spatial memory and synaptic plasticity in a mouse model of AD.<sup>32</sup> *NDFIP1*, *TICAM2*, and *ZDHC17* are associated with positive regulation of I-kB kinase/NF-kB signaling, which is a key regulatory mechanism in innate immunity and known to be associated with AD pathogenesis. Lower expression of *NDFIP1* has been reported to be associated with AD pathogenesis through decreasing DMT1 degradation and increasing iron influx.<sup>33</sup> *TICAM2* is involved in Toll receptor signaling (TLR4 signaling). TLR4-mediated signaling has been reported to contribute to the pathogenesis of age-related neurodegenerative diseases, including AD.<sup>34</sup> *ZDHC17* is a protein coding gene involved in palmitoylation. Disruption of protein palmitoylation has been implicated in pathogenesis of neurodegenerative diseases, including AD,<sup>35</sup> Huntington's disease, schizophrenia, and intellectual disability. A detailed description of driver genes relevant to neurodegeneration and AD along with their molecular function are summarized in Table S3.

The driver genes in cluster-20 were *FASTKD3* and *SLC22A5*. *FASTKD3* (Fas-activated serine/threonine kinase domain 3) has been associated with neural tube defects and disorders of intracellular cobalamin metabolism.<sup>36</sup> *FASTKD3* interacts with components of mitochondrial respiratory and translation machineries.<sup>37</sup> A polymorphism in the pro-apoptotic gene *FASTKD2* (Fas-activated serine/threonine kinase domains 2; rs7594645-G), a member of the same family of proteins, has been associated with better memory performance and hippocampus structure in older adults.<sup>38</sup> *SLC22A5* (solute carrier family 22 member 5) is a gene involved with fatty acid metabolism in mitochondria.<sup>39</sup> It is a well-studied solute carrier in the blood-brain barrier that acts upon carnitine, stimulates the synthesis of acetylcholine, decreases oxidative stress, and prevents neurodegeneration.<sup>40</sup> Variant analyses using the driver genes identified that SNPs within genes *ARMC10* and *KIAA1468* had significant association with disease diagnosis. Overexpression of *ARMC10* in neurons has been reported to prevent A $\beta$ -induced mitochondrial fission and neuronal death.<sup>41</sup> *KIAA1468*(*RELCH*) regulates intracellular cholesterol distribution from recycling endosomes.<sup>42</sup> Although individual SNPs did not survive genome-wide association ( $P_{\text{fdr}} > .05$ ), further studies are warranted in a larger sample size for identification of

disease-relevant variants or expression quantitative trait loci in these genes. When applied to an external dataset with the same driver gene data available, we found that our set of genes validated reasonably well in predicting MCI and AD, and in predicting amyloid positivity. This is an important finding because external validation of a model's predictive performance is extremely crucial to examine reliability and accuracy of model predictions. It is important to note that historically many variables found using prediction algorithms do not cross-validate very well,<sup>43</sup> which was not the case here. The ambiguity of identifying gene transcripts spotted on microarrays based on annotation makes it harder to cross-reference genes based on gene accession number, clone identifier, or even the sequence of a complete gene, and probes on both microarray platforms may hybridize to different gene regions with different GC content. Therefore, we have limited our validation in ADNI to a smaller set of transcripts identified from the discovery dataset that are potentially biologically relevant despite acquisition methods. In efforts to standardize and improve cross-platform analyses, data harmonization methods that enable meta-analyses of gene expression data and cross-validation of the entire pipeline in external cohorts are warranted as an important future direction of the current analyses. While there are obvious challenges when it comes to data harmonization due to multiple processing platforms and standardization issues, important biological markers should validate in external datasets. In a recent study using ADNI, AddNeuroMed (ANM1), and ANM2 data sets, classifiers trained on blood gene expression only were able to classify AD with AUC of 0.657, 0.874, and 0.804 for ADNI, ANM1, and ANM2, respectively. In the external validation, the best AUCs were 0.697 (training: ADNI vs. testing: ANM1), 0.764 (training: ADNI vs. testing: ANM2), 0.619 (training: ANM1 vs. testing: ADNI), 0.79 (training: ANM1 vs. testing: ANM2), 0.655 (training: ANM2 vs. testing: ADNI), and 0.859 (training: ANM2 vs. testing: ANM1), respectively.<sup>44</sup> Our analysis incorporating neuroimaging data yielded more accurate prediction (AUC = 0.73 in ImaGENE and AUC = 0.74 in ADNI for amyloidosis; AUC = 0.71 and AUC = 0.78 for MCI and AD prediction in ADNI, respectively).

There have been several other studies that have addressed the utility of using transcriptomic data from blood. One study aimed at establishing a five-gene-set signature for classifying normal controls



**FIGURE 7** ROC using driver genes from each significant cluster for ADNI (blue) and ImaGENE (red) for (A) amyloidosis. B, ROC for MCI and AD diagnosis in ADNI cohort. C, Cluster-level association with average inferior temporal thickness, and (D) average cortical thickness (x-axis: eigen gene value, y-axis: thickness measure). AD, Alzheimer's disease; ADNI, Alzheimer's Disease Neuroimaging Initiative; MCI, mild cognitive impairment; ROC, receiver operating characteristic

versus MCI and normal controls versus AD based on their differential expression pattern, using publicly available datasets in Gene Expression Omnibus. They reported AUCs of 0.47 and 0.5, respectively, for GSE4229 and GSE85426 in classifying AD versus control using peripheral blood mononuclear cells as the RNA source.<sup>45</sup> Our gene set has better classification performance and is also associated with amyloid and neurodegeneration biomarkers. A recent study conducted a meta-analysis of gene expression in AD and identified 207 differentially expressed genes using different AD tissue microarray datasets in the NCBI\_GEO database.<sup>46</sup> We found that 22 of the driver genes from our analyses were also downregulated in the MCI and AD cohorts from their gene signature set.

Some limitations of the study merit consideration. Considering recent success of ultrasensitive  $A\beta$  and phosphorylated tau plasma assays,<sup>47,48</sup> it is less likely that gene expression data will be clinically

useful, but blood gene expression does help us understand critical peripheral biological pathways associated with AD risk. Given the smaller sample size ( $N = 160$ ) of our discovery dataset, we have validated the driver genes and clusters in a larger external dataset (ADNI,  $N = 515$ ) with reasonable success, but to build more robust predictive models and screen for targets, the genes identified from our analysis warrant further investigation for potential roles in diagnostic prediction algorithms and as therapeutic targets through experimental validation.

Overall, our analysis has contributed to identification of gene expression biomarkers associated with baseline diagnosis of MCI and future conversion to AD dementia and aided in improvement of our understanding of critical disease-related pathways and systematic changes that occur in prodromal AD using blood-based biomarkers.

## ACKNOWLEDGMENTS

The analyses reported in this manuscript were funded by the Eli Lilly Stark Neurosciences Predoctoral Research fellowship, NIA R21 AG072101, NIA R01 AG040770, NIA K02 AG048240, NIA P30 AG010133, K01 AG049050, R56/U01 AG057195 and the Easton Consortium for Alzheimer's Drug Discovery and Biomarker Development. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. ADNI data collection and sharing 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: Alzheimer's Association; Alzheimer's Drug Discovery Foundation; BioClinica, Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; GE Healthcare; Innogenetics, N.V.; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC; Johnson & Johnson Pharmaceutical Research & Development LLC; Medpace, Inc.; Merck & Co., Inc.; Meso Scale Diagnostics, LLC; NeuroRx Research; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Synarc Inc.; and Takeda Pharmaceutical Company. 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 Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

## CONFLICT OF INTERESTS

Liana G. Apostolova receives funding from NIA U01 AG057195; NIA R01 AG057739; NIA P30 AG010133; Alzheimer Association LEADS GENETICS 19-639372; Roche Diagnostics RD005665. Research support: AVID Pharmaceuticals and Life Molecular Imaging. Consulting: Biogen, Two Labs, Eli Lilly, NIH, Florida Dept. Health, NIH Biobank. Data Safety and Monitoring Board: IQVIA, NIA R01 AG061111, UAB Nathan Schick Center. Stock: Semiring, Inc., Cassava Neurosciences. Honoraria: AAN, MillerMed, ASiM, Health and Hospitality Corporation, Mayo Clinic. Editor-in-chief for *Disease Assessment and Disease Monitoring*. Xiaoran Yan: received NLG-Libraries-FY18-2 funding in the last 36 months. Diana Svaldi: Currently an employee of Eli Lilly and received NIH- F32AG062157 funding in the last 36 months. Niraj Shah: T32 training grant in Alzheimer's Disease, Stark Neurosciences Research Institute, IUSM. All the other authors declare no other competing interests/disclosures.

## AUTHOR CONTRIBUTIONS

Liana G. Apostolova had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the

data analysis. Study concept and design: Bharthur Sanjay, Patania, Yan, Svaldi, Apostolova. Acquisition, analysis, or interpretation of data: sll authors. Drafting of the manuscript: Bharthur Sanjay, Apostolova. Critical revision of the manuscript for important intellectual content: Apostolova, Patania, Yan, Svaldi. Statistical analysis: Bharthur Sanjay, Yan, Patania, Svaldi. Administrative, technical, or material support: Apostolova, Patania

## ORCID

Apoorva Bharthur Sanjay  <https://orcid.org/0000-0001-9082-0587>

## REFERENCES

1. Van Cauwenberghe C, Van Broeckhoven C, Sleegers K. The genetic landscape of Alzheimer disease: clinical implications and perspectives. *Genet Med*. 2016;18(5):421-430.
2. Lambert J-C, Ibrahim-Verbaas CA, Harold D, et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat Genet*. 2013;45(12):1452-1458.
3. Jansen IE, Savage JE, Watanabe K, et al. Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. *Nat Genet*. 2019;51(3):404-413.
4. Scherzer CR, Eklund AC, Morse LJ, et al. Molecular markers of early Parkinson's disease based on gene expression in blood. *Proc Natl Acad Sci U S A*. 2007;104(3):955-960.
5. Li X, Long J, He T, Belshaw R, Scott J. Integrated genomic approaches identify major pathways and upstream regulators in late onset Alzheimer's disease. *Scientific Reports*, 2015;5(1). <https://doi.org/10.1038/srep12393>
6. Ciryam P, Kundra R, Freer R, Morimoto RI, Dobson CM, Vendruscolo M. A transcriptional signature of Alzheimer's disease is associated with a metastable subproteome at risk for aggregation. *Proc Natl Acad Sci U S A*. 2016;113(17):4753-4758.
7. Leandro GS, Evangelista AF, Lobo RR, Xavier DJ, Moriguti JC, Sakamoto-Hojo ET. Changes in expression profiles revealed by transcriptomic analysis in peripheral blood mononuclear cells of Alzheimer's disease patients. *J Alzheimers Dis*. 2018;66(4):1483-1495.
8. Rye PD, Booi BB, Grave G, et al. A novel blood test for the early detection of Alzheimer's disease. *J Alzheimers Dis*. 2011;23(1):121-129.
9. Booi BB, Lindahl T, Wetterberg P, et al. A gene expression pattern in blood for the early detection of Alzheimer's disease. *J Alzheimers Dis*. 2010;23(1):109-119.
10. Reuter M, Schmansky NJ, Rosas HD, Fischl B. Within-subject template estimation for unbiased longitudinal image analysis. *Neuroimage*. 2012;61(4):1402-1418.
11. Chau W, Mcintosh AR. The Talairach coordinate of a point in the MNI space: how to interpret it. *Neuroimage*. 2005;25(2):408-416.
12. Chung MK, Worsley KJ, Nacewicz BM, Dalton KM, Davidson RJ. General multivariate linear modeling of surface shapes using SurfStat. *Neuroimage*. 2010;53(2):491-505.
13. Edelsbrunner H, Letscher D, Zomorodian A. Topological persistence and simplification. *Discrete Comput Geometry*, 2002;28(4), 511-533. <https://doi.org/10.1007/s00454-002-2885-2>
14. Craw S. Manhattan Distance. In: Sammut C, Webb GI, eds. *Encyclopedia of Machine Learning and Data Mining*. Springer; 2017:790-791. Editors.
15. Gönen M, Alpaydin E. Multiple kernel learning algorithms. *J Mach Learn Res* 2011;12:2211-2268.
16. Reininghaus J, Huber S, Bauer U & Kwitt R. A stable multi-scale kernel for topological machine learning. In Proceedings of the IEEE conference on computer vision and pattern recognition, 2015:4741-4748.
17. Karatzoglou A, Smola A, Hornik K. kernlab - An S4 package for kernel methods in R. *J Stat Softw Articles*. 2004;11(9):1-20.

18. Langfelder P, Horvath S. Eigengene networks for studying the relationships between co-expression modules. *BMC Syst Biol*. 2007;1(1). <https://doi.org/10.1186/1752-0509-1-54>
19. Alexa A & Rahnenfuehr J topGO: Enrichment Analysis for Gene Ontology. R package version 2.40.0. 2020.
20. Supek F, Bošnjak M, Škunca N, Šmuc T. REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS ONE*. 2011;6(7):e21800. <https://doi.org/10.1371/journal.pone.0021800>
21. de Leeuw CA, Mooij JM, Heskes T, Posthuma D. MAGMA: generalized gene-set analysis of GWAS data. *PLoS Comput Biol*. 2015;11(4):e1004219. <https://doi.org/10.1371/journal.pcbi.1004219>
22. Petersen RC, Aisen PS, Beckett LA, et al. Alzheimer's Disease Neuroimaging Initiative (ADNI): clinical characterization. *Neurology*. 2010;74(3):201-209.
23. Saykin AJ, Shen L, Foroud TM, et al. Alzheimer's disease neuroimaging initiative biomarkers as quantitative phenotypes: genetics core aims, progress, and plans. *Alzheimers Dement*. 2010;6(3):265-273.
24. Thurfjell L, Lilja J, Lundqvist R, et al. Automated quantification of 18F-flutemetamol PET activity for categorizing scans as negative or positive for brain amyloid: concordance with visual image reads. *J Nucl Med*. 2014;55(10):1623-1628.
25. Du A-T, Schuff N, Kramer JH, et al. Different regional patterns of cortical thinning in Alzheimer's disease and frontotemporal dementia. *Brain*. 2007;130:1159-1166. Pt 4.
26. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*. 2007;81(3):559-575.
27. Yates TM, Drucker M, Barnicoat A, et al. ZMYND11-related syndromic intellectual disability: 16 patients delineating and expanding the phenotypic spectrum. *Hum Mutat*. 2020;41(5):1042-1050.
28. Meyer-Hoffert U, Wu Z, Kantyka T, et al. Isolation of SPINK6 in human skin: selective inhibitor of kallikrein-related peptidases. *J Biol Chem*. 2010;285(42):32174-32181.
29. Von Bernhardi R, Cornejo F, Parada GE, Eugenin J. Role of TGFβ signaling in the pathogenesis of Alzheimer's disease. *Front Cell Neurosci*. 2015;9:426-426.
30. Henderson BW, Gentry EG, Rush T, et al. Rho-associated protein kinase 1 (ROCK1) is increased in Alzheimer's disease and ROCK1 depletion reduces amyloid-β levels in brain. *J Neurochem*. 2016;138(4):525-531.
31. Du X, Pan Z, Li Q, Liu H, Li Q. SMAD4 feedback regulates the canonical TGF-β signaling pathway to control granulosa cell apoptosis. *Cell Death Dis*. 2018;9(2). <https://doi.org/10.1038/s41419-017-0205-2>
32. Caccamo A, Branca C, Talboom JS, et al. Reducing ribosomal protein s6 kinase 1 expression improves spatial memory and synaptic plasticity in a mouse model of Alzheimer's disease. *J Neurosci*. 2015;35(41):14042-14056.
33. Tian J, Zheng W, Li X-L, Cui Y-H, Wang Z-Y. Lower expression of Ndfip1 is associated with Alzheimer disease pathogenesis through decreasing DMT1 degradation and increasing iron influx. *Front Aging Neurosci*. 2018;10:165-165.
34. Calvo-Rodríguez M, García-Rodríguez C, Villalobos C, Núñez L. Role of toll like receptor 4 in Alzheimer's disease. *Front Immunol*. 2020;11. <https://doi.org/10.3389/fimmu.2020.01588>
35. Cho E, Park M. Palmitoylation in Alzheimer's disease and other neurodegenerative diseases. *Pharmacol Res*. 2016;111:133-151.
36. Boehm E, Zornoza M, Jourdain AA, et al. Role of FAST kinase domains 3 (FASTKD3) in post-transcriptional regulation of mitochondrial gene expression. *J Biol Chem*. 2016;291(50):25877-25887.
37. Simarro M, Gimenez-Cassina A, Kedersha N, et al. Fast kinase domain-containing protein 3 is a mitochondrial protein essential for cellular respiration. *Biochem Biophys Res Commun*. 2010;401:440-446.
38. Ramanan VK, Nho K, Shen L, et al. FASTKD2 is associated with memory and hippocampal structure in older adults. *Mol Psychiatry*. 2015;20(10):1197-1204.
39. Juraszek B, Nałęcz KA. SLC22A5 (OCTN2) Carnitine transporter—indispensable for cell metabolism, a Jekyll and Hyde of human cancer. *Molecules*. 2019;25(1):14. <https://doi.org/10.3390/molecules25010014>
40. Ferreira GC, Mckenna MC. L-Carnitine and Acetyl-L-carnitine roles and neuroprotection in developing brain. *Neurochem Res*. 2017;42(6):1661-1675.
41. Serrat R, Figueiro-Silva J, Navas-Pérez E, et al. The Armc10/SVH gene: genome context, regulation of mitochondrial dynamics and protection against Aβ-induced mitochondrial fragmentation. *Cell Death Dis*. 2014;5(4):e1163-e1163.
42. Sobajima T, Yoshimura S-I, Maeda T, Miyata H, Miyoshi E, Harada A. The Rab11-binding protein RELCH/KIAA1468 controls intracellular cholesterol distribution. *J Cell Biol*. 2018;217(5):1777-1796.
43. Casanova R, Varma S, Simpson B, et al. Blood metabolite markers of preclinical Alzheimer's disease in two longitudinally followed cohorts of older individuals. *Alzheimer's Dement: J Alzheimer's Association*. 2016;12(7):815-822.
44. Lee T, Lee H. Prediction of Alzheimer's disease using blood gene expression data. *Sci Rep*. 2020;10(1). <https://doi.org/10.1038/s41598-020-60595-1>
45. Hadar A, Gurwitz D. Peripheral transcriptomic biomarkers for early detection of sporadic Alzheimer disease?. *Dialogues in Clinical Neuroscience*. 2018;20(4):293-300.
46. Su L, Chen S, Zheng C, Wei H, Song X. Meta-analysis of gene expression and identification of biological regulatory mechanisms in Alzheimer's Disease. *Front Neurosci*. 2019;13. <https://doi.org/10.3389/fnins.2019.00633>
47. Schindler SE, Bollinger JG, Ovod V, et al. High-precision plasma β-amyloid 42/40 predicts current and future brain amyloidosis. *Neurology*. 2019;93(17):e1647-e1659. <https://doi.org/10.1212/wnl.0000000000008081>
48. Yang C-C, Chiu M-J, Chen T-F, Chang H-L, Liu B-H, Yang S-Y. Assay of plasma phosphorylated tau protein (Threonine 181) and total tau protein in early-stage Alzheimer's disease. *J Alzheimers Dis*. 2018;61(4):1323-1332.
49. Grimm MOW, Kuchenbecker J, Rothhaar TL, et al. Plasmalogen synthesis is regulated via alkyl-dihydroxyacetonephosphate-synthase by amyloid precursor protein processing and is affected in Alzheimer's disease. *J Neurochem*. 2011;116(5):916-925.
50. Ogaki K, Fujioka S, Heckman MG, et al. Analysis of COQ2 gene in multiple system atrophy. *Mol Neurodegener*. 2014;9:44-44.
51. Arefin AS, Mathieson L, Johnstone D, Berretta R, Moscato P. Unveiling clusters of RNA transcript pairs associated with markers of Alzheimer's disease progression. *PLoS One*. 2012;7(9):e45535-e45535.
52. Huttunen HJ, Guénette SY, Peach C, et al. HtrA2 regulates beta-amyloid precursor protein (APP) metabolism through endoplasmic reticulum-associated degradation. *J Biol Chem*. 2007;282(38):28285-28295.
53. Mok S-A, Condello C, Freilich R, et al. Mapping interactions with the chaperone network reveals factors that protect against tau aggregation. *Nat Struct Mol Biol*. 2018;25(5):384-393.
54. Astarita G, Jung K-M, Berchtold NC, et al. Deficient liver biosynthesis of docosahexaenoic acid correlates with cognitive impairment in Alzheimer's disease. *PLoS One*. 2010;5(9):e12538-e12538.
55. Katsel P, Li C, Haroutunian V. Gene expression alterations in the sphingolipid metabolism pathways during progression of dementia and Alzheimer's disease: a shift toward ceramide accumulation at the earliest recognizable stages of Alzheimer's disease?. *Neurochem Res*. 2007;32(4-5):845-856.
56. Jayapalan S, Subramanian D, Natarajan J. Computational identification and analysis of neurodegenerative disease associated protein kinases in hominid genomes. *Genes*. 2016;3(3):228-237.

57. Kim N-Y, Cho M-H, Won S-H, Kang H-J, Yoon S-Y, Kim D-H. Sorting nexin-4 regulates  $\beta$ -amyloid production by modulating  $\beta$ -site-activating cleavage enzyme-1. *Alzheimer's Res Ther.* 2017;9(1):4-4.
58. Fujiwara H, Watanabe S, Iwata M, et al. Inhibition of microtubule assembly competent tubulin synthesis leads to accumulation of phosphorylated tau in neuronal cell bodies. *Biochem Biophys Res Commun.* 2020;521(3):779-785.
59. Pietrzak M, Papp A, Curtis A, et al. Gene expression profiling of brain samples from patients with Lewy body dementia. *Biochem Biophys Res Commun.* 2016;479(4):875-880.
60. Milde S, Coleman MP. Identification of palmitoyltransferase and thioesterase enzymes that control the subcellular localization of axon survival factor nicotinamide mononucleotide adenylyltransferase 2 (NMNAT2). *J Biol Chem.* 2014;289(47):32858-32870.

## SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

**How to cite this article:** Bharthur Sanjay A, Patania A, Yan X, et al. Characterization of gene expression patterns in mild cognitive impairment using a transcriptomics approach and neuroimaging endophenotypes. *Alzheimer's Dement.* 2021;1-16. <https://doi.org/10.1002/alz.12587>