

TRANSCRIPTOMIC PROFILING IN MILD COGNITIVE IMPAIRMENT AND
ALZHEIMER'S DISEASE USING NEUROIMAGING ENDOPHENOTYPES

Apoorva Bharthur Sanjay

Submitted to the faculty of the University Graduate School

in partial fulfillment of the requirements

for the degree

Doctor of Philosophy

in the Program of Medical Neuroscience,

Indiana University

December 2022

Accepted by the Graduate Faculty of Indiana University, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Doctoral Committee

Karmen K. Yoder, Ph.D., Chair

Liana G. Apostolova, M.D., MSc, FAAN

July 21, 2022

Shannon L. Risacher, Ph.D.

Sujuan Gao, Ph.D.

Kelly Nudelman Ph.D.

© 2022

Apoorva Bharthur Sanjay

ACKNOWLEDGMENTS

I would like to firstly thank my primary mentor, Dr. Liana Apostolova, without whom this work would not have been possible. She has fostered an environment of excellence and has been instrumental in me becoming a better scientist and critical thinker. She has always been supportive and encouraging, providing me guidance when I needed but giving me enough independence to be able to follow through on my ideas and projects. As a mentor, she has also nurtured my career development, enabling my involvement in writing grants, mentoring students, and leading successful collaborations, as well as encouraging me to present my work at national and international conferences. She is truly my role model, in both personal and professional capacity. I would like to thank Dr. Diana Svaldi who taught me so much and helped shape and formulate a lot of the ideas executed in this work. I also thank other Apostolova lab members for their help and guidance.

I would like to acknowledge my committee members, Drs. Karmen Yoder, Shannon Risacher, Sujuan Gao and Kelly Nudelman. Their guidance and expertise have been a great source of support throughout this journey, and their genuine interest in my well-being and growth has been a tremendous positive force in my graduate career. I am extremely grateful to them for all they have done.

I would like to thank my dear family, who have managed to always be there for me and support my dreams despite living on a different continent. They have been patient, kind and encouraging through good times and bad, and I cannot express how grateful I am to them and their pride in my work. My friends are my family away from home, and my PhD journey has been warm, happy, and fun filled

thanks to them. Their love and friendship are a source of so much joy and strength to me, and I'd like to sincerely thank them for that.

The analyses reported in this work 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 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). 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.

Apoorva Bharthur Sanjay

TRANSCRIPTOMIC PROFILING IN MILD COGNITIVE IMPAIRMENT AND
ALZHEIMER'S DISEASE USING NEUROIMAGING ENDOPHENOTYPES

Alzheimer's disease (AD) is a devastating neurodegenerative disease affecting more than 6 million Americans and 50 million people worldwide currently. It is an irreversible neurodegenerative disease which causes decline in memory, cognition, personality, and other functions which eventually lead to death due to complete brain failure.

Recently there has been a lot of research that has focused on enabling early intervention and disease prevention in AD which could have a significant impact on this disease, be crucial for life management, assessment of risk for future generations, and assistance in end-of-life preparation. For a late-life complex multifactorial disease, such as AD, where both genetic and environmental factors are involved, integrating multiple layers of genetic, imaging, and other biomarker data is a critical step for therapeutic discovery and building predictive risk assessment tools.

The multifactorial nature of AD suggests that multiple therapeutic targets need to be identified and tested together. Hence, we need a systems-level approach to build biomarker profiles which can be used for drug discovery and screening/risk assessment. The research presented in this dissertation focuses on utilizing a systems level approach to identify promising imaging genetics biomarkers that provide insight into dysregulated biological pathways in AD

pathogenesis and identify critical mRNA measures that can be investigated further within the scope of novel therapeutics, as well as input variables in predictive models for AD risk, screening, and diagnosis. The overall research goal was the development of systems level, imaging genetics biomarker signatures to serve as tools for risk analysis and therapeutic discovery in AD. The specific outcomes of the analyses were characterization of patterns in gene expression at systems level using neuroimaging endophenotypes, and identification of specific driver genes and genotypic variants, which can inform predictive modeling for diagnosis, risk, and pathogenic profiling in AD.

Karmen K. Yoder, Ph.D., Chair

Liana G. Apostolova, M.D., MSc, FAAN

Shannon L. Risacher, Ph.D.

Sujuan Gao, Ph.D.

Kelly Nudelman Ph.D.

TABLE OF CONTENTS

List of Tables.....	xi
List of Figures.....	xii
Introduction.....	1
Background and Justification.....	3
Neuroimaging endophenotypes.....	9
Chapter 1: Characterization of gene expression patterns in Mild Cognitive Impairment using a transcriptomics approach and neuroimaging endophenotypes.....	12
Introduction:.....	12
Methods:.....	14
Dataset.....	14
Preliminary data reduction.....	17
Univariate surface mapping.....	18
Persistent Homology.....	18
Unsupervised kernel clustering.....	21
Cluster Analysis.....	21
Gene Enrichment Analysis and Driver Genes Identification.....	22
Cluster Level Associations.....	23
Variant Analyses in Driver Genes.....	23
Validation in an external dataset.....	23
Results:.....	24
Subject demographics and identification of gene set.....	24
Cluster analysis and significant clusters.....	25
Biological processes associated with clusters and driver genes.....	25
Cluster Level Associations.....	32
Variant Analyses:.....	33
Validation in ADNI.....	35
Discussion:.....	36
Chapter 2: Transcriptomic profiling of brain amyloidosis and Alzheimer's disease phenotypes identifies important novel gene targets from peripheral blood gene expression data.....	51
Introduction:.....	51
Methods:.....	53

Datasets.....	53
Data availability.....	57
Gene expression analyses.....	57
Driver Genes analyses.....	58
Variant Analyses.....	59
Results:.....	60
Gene expression analyses.....	60
Gene enrichment and cluster level associations.....	61
Driver gene analyses.....	64
Variant Analyses:.....	70
3D SPM analyses:.....	74
Discussion:.....	75
Chapter 3: Transcriptomic profiling in Mild Cognitive Impairment using peripheral blood gene co-expression networks.....	82
Introduction:.....	82
Methods:.....	83
Dataset.....	83
Differential expression and gene co-expression matrix.....	85
Network backbone construction.....	85
Clustering and module Eigengene.....	86
Network Analyses:.....	87
Univariate surface mapping.....	88
Results:.....	88
Clusters with significance to amnesic MCI phenotype.....	90
Network Analyses on cluster 37.....	91
Hub genes associated with cortical atrophy:.....	93
Discussion:.....	94
Summary.....	99
References.....	105
Curriculum Vitae	

LIST OF TABLES

Table 1: ImaGene Demographics	15
Table 2: ADNI Demographics.....	19
Table 3: Driver genes and their fold change	30
Table 4: Variant analyses within driver genes	34
Table 5: Gene Ontology analyses for the differentially expressed genes within the significant clusters	43
Table 6: Gene Ontology analyses for all genes within the significant clusters	45
Table 7: AD relevant driver genes identified from significant clusters	47
Table 8: ADNI demographics table	55
Table 9: ImaGene demographics	57
Table 10: Differential peripheral blood expression levels by clinical diagnosis for driver genes.....	64
Table 11: Driver genes' correlations with amyloid SUVR in ADNI and ImaGene	66
Table 12: Variant analyses in differentially expressed genes and driver genes..	71
Table 13: Significant SNPs identified through GWAS analyses.....	72
Table 14: ImaGene sample demographics	84
Table 15: Genes in cluster 37.....	90

LIST OF FIGURES

Figure 1: Overview of methods.....	17
Figure 2: Persistent homology pipeline representation with one gene.	20
Figure 3: (A) Visualization of 20 cluster solution.....	27
Figure 4: Gene clusters visualizing nodes and relevant biological processes.....	29
Figure 5: Molecular signatures of driver genes across diagnostic groups in A) ImaGene and B) ADNI.....	32
Figure 6: Cluster association(T-statistic) with cortical thickness.	33
Figure 7: ROC using driver genes from each significant cluster for ADNI (blue) and ImaGene(red).....	36
Figure 8: Visualization of k-means clustering	61
Figure 9: Visualization of the associations of each cluster with relevant clinical and biomarker traits.....	63
Figure 10: ROC curves in the validation sample (ImaGene).....	69
Figure 11: Association pattern of driver genes with brain amyloidosis in the ADNI sample (n=356).....	75
Figure 12: Heatmap representation of co-expression networks.....	89
Figure 13: Network representation of cluster 37.....	92
Figure 14: Cluster 37 network representation (blue) with hub nodes in red	93
Figure 15: Association of hub genes with cortical thickness	94

Introduction

Alzheimer's disease (AD) is a devastating neurodegenerative disease affecting more than 6 million Americans and 50 million people worldwide currently. It is an irreversible neurodegenerative disease which causes decline in memory, cognition, personality, and other functions which eventually lead to death due to complete brain failure. Recently there has been a lot of research that has focused on enabling early intervention and disease prevention in AD which could be impactful, and would also be crucial for life management, to assessment of risk for future generations and assistance in end-of-life preparation. Biomarkers are essential to this effort, as they are a defined characteristic which can be measured as an indicator of normal biological process, pathogenic process, or response to an exposure or intervention. Development of therapeutically relevant biomarkers is a necessary line of research, as established cognitive instruments commonly used to diagnose and follow AD disease progression have limited usefulness in the latent AD stages[1].

The hallmark pathologies observed in AD pathology are the cerebral extracellular deposition of β -Amyloid ($A\beta$) and the intracellular formation of neurofibrillary tangles by hyperphosphorylated tau. There are several other conditions that are associated with these hallmark pathologies. Broadly, they are associated with vascular alterations, systemic inflammation, genetic/epigenetic status, mitochondrial dysfunction, and lipid metabolism. Valid and predictive re-symptomatic risk assessments (based in part on $A\beta$ and tau) are needed to define

essential windows in which disease-modifying and preventative interventions for AD) can be maximally effective. Currently established biomarkers such as positron emission tomography (PET) (Positron Emission Tomography) amyloid imaging and cerebrospinal fluid (CSF) levels of (Cerebrospinal Fluid) tau/A β are expensive and/or invasive. Structural imaging biomarkers, such as gray matter and white matter volume, while less expensive, can also be non-specific, and may be confounded by co-morbid medical conditions. Nonetheless, these imaging techniques are still important for characterizing AD phenotypes. Blood-based biomarkers such as transcriptomic, proteomic and metabolomics, and other omics data are important because they represent a more specific, less invasive, and potentially less expensive approach for aiding in AD detection and monitoring of disease progression. For a late-life complex multifactorial disease such as AD, in which both genetic and environmental factors are involved, integration of multiple layers of genetic, imaging, and other biomarker data is a critical step for the development of robust risk assessment platforms. Genetic data especially can provide better understanding of the underlying pathogenesis and be very useful for both biomarker and therapeutic development. The multifactorial nature of AD suggests that multiple therapeutic targets need to be identified and tested together. Hence, a systems-level approach is needed to build biomarker profiles that could be used for drug discovery and screening/risk assessment.

The purpose of this body of work was to utilize a systems level approach to identify promising integrative imaging/genetics biomarkers and gene-interaction networks which could provide insight into dysregulated biological pathways in AD

pathogenesis and to identify critical mRNA measures that could be investigated further within the scope of novel therapeutics as well as be used as input variables within predictive models for AD risk, screening, and diagnosis. The datasets used for the analyses were data from the Imaging and Genetic Biomarkers of Alzheimer's Disease study (ImaGene) and the Alzheimer's Disease Neuroimaging Initiative study (ADNI)[2]. The ImaGene dataset consists of longitudinal clinical, cognitive, imaging, epigenetic, genomic, and transcriptomic data on 160 subjects (50 Normal controls (NC) and 100 MCI (Mild Cognitive Impairment), 35 of which converted to dementia during the study.

Background and Justification

Evidence for a genetic component in neurodegenerative disorders such as AD is overwhelming. The primary risk alleles of apolipoprotein (APP) and presenilin (PSEN) have been known for several decades[3]. More recently, at least 21 additional genetic risk loci have been identified for the non mendelian form of Late Onset AD (LOAD) in genome-wide association studies (GWAS) and massive parallel resequencing (MPS) efforts[4, 5]. These studies reiterate the multifactorial nature of AD. Analyses of genetic data, such as full genome screens, require large sample sizes, which make it costly and logistically challenging. Moreover, the predictive and diagnostic screening for causal mutations accounts for only a small portion of AD patients. Many mutations in the amyloid precursor protein (APP) and presenilin 1 and 2 (PSEN1 and PSEN2) have been reported as the pathogenic causes of early-onset AD (EOAD), which accounts for up to 5% of all AD cases[6]. Pathogenic mutations are identified in only ~6% of patients through genetic

screening because they focus on *APP*, *PSEN1* and *PSEN2*[7]. Massive parallel sequencing has offered a new diagnostic tool to screen for multiple putative risk genes, but this requires a reliable, pre-defined disease spectrum gene panel to eliminate the need for solely clinical diagnostic parameters. However, no such panel currently exists in the absence of knowledge of clinical symptoms. This means that predictive diagnoses during the early latent stages of AD (when there is often no clinical manifestation) is exceptionally difficult.

Emerging data suggest that peripheral blood markers may provide a suitable surrogate for gene expression patterns identified in CNS disease. For example, 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[8]. In addition, systematic evaluation of gene expression in blood and brain has shown that whole-blood gene expression profiles share significant similarities with that of CNS tissues[9]. A significant amount of work has been conducted in analysis of blood-based markers within the context of predicting progression of AD, conversion from mild cognitive impairment (MCI) to AD, and risk for future AD. A set of ten lipids from peripheral blood were identified that predicted conversion to amnesic mild cognitive impairment (aMCI) or AD in 525 community dwelling older participants [10]; however, these findings have not been successfully cross- validated. Hye et al. analyzed plasma proteomics from three independent cohorts and found a set of 10 proteins that predicted progression from MCI to AD[11]. Another focus in recent years has been on plasma t-tau as a

potential blood-based biomarker for diagnostics and screening[12]. Total tau (t-tau) and tau phosphorylated at threonine 181 (p-tau-181) are biomarkers for the presence of Alzheimer's disease (AD) pathology in the brain[13]. However, there is no clear correlation between plasma and CSF t-tau concentrations[12, 14], and this finding of low correlation between CSF and blood biomarkers is reported frequently[15]. A study examined serum autoantibodies within a classification analysis and found that the top fifty differentially expressed autoantibodies had greater than 95% sensitivity and specificity for discriminating MCI from all other diagnostic categories within a cross-sectional cohort[16]. Taken together, these studies point towards the utility of using blood-based biomarkers, with an emphasis on quantitative analytes/metabolites for indexing disease state and progression. Additionally transcriptomic profiling or studying gene expression from blood can be a useful tool for risk analysis and therapeutic development, especially in prodromal AD.

Transcriptomic analyses on brain tissue are restricted to postmortem data, and thus represent primarily profiles of end-stage AD, which may not be generalizable to pre-symptomatic or early symptomatic stages of AD. Gene expression and genetic variant analyses may be a key bridge to the development of robust blood-based risk assessment platforms, as it has been observed that 80% of the genes expressed in human brain are also expressed in the peripheral blood lymphocytes[17]. Microarray and RNA-sequencing based transcriptome studies of postmortem brain in AD and healthy controls have identified differentially expressed genes and associated functional pathways, yielding a core set of

differentially expressed pathways including immune response, apoptosis, cell proliferation, energy metabolism, and synaptic transmission[18, 19]. RNA-Seq analyses of human AD postmortem brains have shown that there are novel isoforms which are dysregulated in parietal cortex associated with immune response and lipid metabolism[20].

Gene expression can also be studied as gene co-expression networks, which looks at relationship between genes, where a node represents a gene and an edge joining two nodes represents their relationship. Use of gene co-expression networks may help identify novel gene interactions and subnetworks. Gene-gene correlation networks offer the potential for identifying new 'hub genes' in disease susceptibility and progression. Using network analysis tools, Miller et al et al. published a systems-level analysis of transcriptional changes in AD and normal aging using gene expression data from the CA1 region of hippocampus. They identified conserved modules between AD and aging, as well as other biologically relevant driver genes[21]. Liang et al recently identified two modules of co-expressed genes from hippocampal tissue associated with AD clinical severity that were related to NF-KB signaling and cGMP-PKG signaling pathways. They also confirmed the validity of the hub genes in AD transgenic mice[22]. The studies mentioned above reinforce the validity of studying genetic data and gene expression networks in identification of novel targets or dysregulated pathways, with the caveat being that most have focused on postmortem brain tissue. We therefore aimed to address the utility of identifying promising genetic signatures

from blood at the prodromal AD stage with the intent of discovering novel therapeutic targets.

An accurate interpretation of the association between a genetic variant and pathogenesis in AD is difficult because of its pathophysiological complexity and pleiotropy of risk genes. A myriad of molecular factors interacts in multiple networks dynamically, and at different biological levels. One way to address this complexity is by combining genetic variant data with gene expression studies. A fundamental concept of genetics is that the genotype influences the phenotype. Given the low phenotypic variance explained by any single SNP, it is expected that groups of SNPs cluster within groups of functionally-related genes within biological systems[23]. Using eQTL (expression Quantitative Trait Loci) network analysis, Fagny et al. found communities (clusters of SNPs) associated with tissue specific processes, and these clusters elucidate the regulatory role of SNPs in gene expression networks[24]. In another study, genetic variants in hub genes (genes with higher connectivity to other genes in the network) are predicted to have a more profound effect on network dysfunction and disease than other genes in the network[25]. *TYROBP* (a receptor-activating subunit component of natural killer (NK) cells) was identified as a central regulator of the immune/microglia networks, as were *TREM2*, *MS4A4A*, *MS4A6A*, and *CD33*[26]. *TYROBP* is a binding partner of *TREM2*, and a genetic variant in *TYROBP* is reported to regulate *TREM2* expression levels[26]. *TYROBP* is the downstream adaptor and putative signaling partner for several receptors implicated in Alzheimer's disease (AD)[27]. Additionally, *PTK2B* has been reported as a hub regulator identified through

microarray meta-analysis in the frontal cortex[28]. *PTK2B* regulates important biological processes involved in neuronal differentiation and electrical maturation[29]. Given the potential utility of understanding interactions between gene expression and gene networks for biomarker development in AD, we investigated the genotype-phenotype relationship in AD pathogenesis within the context of genetic network analyses of peripheral blood-based gene expression.

The need for a cost-effective way that can serve as the first step in a multistage diagnostic framework in AD is an integral part of biomarker development in AD research. The use of sophisticated computational methods and statistical modeling for diagnostics and screening could minimize human error and ultimately would make results readily available to clinicians. Most importantly, it allows for integration of vast and diverse datasets to facilitate reliable classification and prognosis. In an effort towards this goal, multiple studies like the Alzheimer's Disease Neuroimaging Initiative (ADNI), AddNeuroMed, and ImaGene collect and store data ranging from clinical data; imaging data; genomic, transcriptomic, proteomic, and metabolomic data; and gene expression data[30]. Deep-learning based automated diagnostic frameworks [31] and hyper spectral imaging-based methods[32] are examples of developments in computational tools for diagnosis. Huang et al. recently identified candidate AD genes from human brain-specific gene network data using an support vector machine (SVM) machine learning approach[33]. In another study, using an eQTL framework, Hu et al. found several candidate SNPs that have a strong relationship with AD[34]. Integration of gene expression and genotype data for prediction using machine learning and statistical

modeling will be useful for development of the computation-aided diagnostic frameworks for AD, and we employed this methodology for assessment of our transcriptomic measures. External validation of a model's predictive performance is crucial for determining the reliability and accuracy of model predictions because variables found with data-driven in sample prediction algorithms do not always cross-validate very well [10], we validated our models and predictors found through analyses of genetic and gene expression data from the initial discovery dataset(the dataset our model was trained and built on) in a second, external dataset.

Neuroimaging endophenotypes

Genetic studies based on case-control analyses and gene expression makes identification of genetic risk factors and candidate genes implicated in AD difficult because they are based on clinical diagnosis alone. Putative AD diagnosis is based of cognitive impairment that is identified with standard neuropsychiatric tests. The outcomes of standard cognitive tests may be subject to effects unrelated to AD, such as fatigue, anxiety, general test-taking ability, and practice effects. Occasionally, well-educated people suffering from cognitive decline can appear normal in a clinical setting, whereas cognitively normal individuals with extreme anxiety may appear to be impaired. There are other late-life dementias that can be clinically misdiagnosed as AD based on neuropsychiatric test results. Therefore, the use of brain endophenotypes, which are objective and highly reproducible over time, may facilitate identification of AD genetic risk factors, more importantly, may permit understanding of these risk factors' impacts on processes like amyloid and tau pathology, hypometabolism, and neurodegeneration. Imaging metrics also

offer versatility in terms of pathophysiological mechanisms such as structural or (e.g.), functional changes (e.g., loss in volume, cortical thinning), functional decline (e.g., altered task-based functional magnetic resonance imaging (, diffusion anisotropy reduction, white matter pathology, fMRI) activity, altered resting state fMRI network connectivity), white matter decline (e.g. reductions in diffusion weighted imaging signal, including white matter volume and/or lesions, white matter tract integrity), and pathological protein aggregation (e.g. amyloid and tau, as visualized with PET positivity).

Anatomical information from imaging biomarkers provides crucial disease-staging information. Neuroimaging can help distinguish the different phases of AD disease both temporally and anatomically[35]. Volumetric changes can be observed through analyses of structural MRI data[36]. In AD, overall cortical thinning, and reduction of gray matter in the hippocampus, the entorhinal cortex and prefrontal cortex are observed, with hippocampal and entorhinal shrinkage occurring in preclinical and early stages of disease[37-40]. Amyloid PET provides information on amyloid aggregation [22]. Studies show that parietal cortices (posterior cingulate, retrosplenial cortex, precuneus) are key regions of early deposition[41], which are consistent with post-mortem Braak amyloid stages (REF). These regions have strong connections to medial temporal lobe, a site of early tau pathology[42]. Tau pathology correlates strongly with neurodegeneration and with cognitive impairment, and tau PET imaging can capture histology-based Braak and Braak staging of tau pathology[43, 44]. Given the importance of these

neuroimaging biomarkers, we have incorporated them in our predictive models with the transcriptomic data.

Chapter 1: Characterization of gene expression patterns in Mild Cognitive Impairment using a transcriptomics approach and neuroimaging endophenotypes

Introduction:

The prodromal stages of Alzheimer's Disease (AD) can provide an essential window where disease-modifying and preventative interventions can be maximally effective. Currently established biomarkers like Positron Emission Tomography (PET), amyloid imaging and cerebrospinal fluid (CSF) tau/A β 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, where both genetic and environmental factors are 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 where 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. 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 ~6% of

patients [7], 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 [4, 45]. These studies re-emphasize the multifactorial nature of AD.

Emerging data suggests that peripheral blood may provide a suitable surrogate for gene expression patterns identified in the 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[8] 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 [8]. In contrast to DNA sequencing transcriptomics, the study of gene expression profiles informs not only about inherited but also non-inherited genomic signals [4]. Microarray and RNA-sequencing based transcriptome studies of postmortem 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 [18, 19]. Pathways involved in inflammation, DNA damage response, cell cycle and neuronal homeostasis were found to be dysregulated in peripheral blood [46]. Several studies have shown that an AD-specific mRNA signature that differentiates AD from cognitively normal controls (CN) can be detected in peripheral blood [47, 48]. 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 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).

Methods:

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 (aMCI, N = 70) or those presenting with non-amnesic phenotype (naMCI, N = 38) (**Table1**).

Imaging: The detailed imaging protocol has been previously published [10]. All subjects received annual 1.5T MRI scans following the UCLA Alzheimer's Disease Research Center protocol consisting of coronal FI3D T1 MPRAGE: TR

28, TE 4.5, FOV 22 cm, matrix 256x192, slice/gap 1.5/0 mm. Measures of neurodegeneration were obtained from coronal T1-weighted MPRAGE sequences. Scans were processed using Freesurfer (version 6.0) longitudinal pipeline, to obtain region specific and global measures of atrophy [49].

Table 1: ImaGene Demographics

Variable Mean (std)	NC (n=52)	aMCI (n=70)	naMCI (n=38)	p- value
Age	69.03(7.9)	69.82(8.5)	69.75(8.5)	0.9
Education	17.6(2.04)	15.5(2.7)	16.5(2.88)	0.001*
Gender(M/F)	30/21	26/43	20/18	N.S.
MMSE	28.8(1.2)	27(2.5)	27.9(1.9)	<0.001*
Hippocampus vol (mm³)	8602 (1092)	7990 (1324)	8718 (1023)	0.002*

Microarray-based Gene Expression: All subjects provided yearly peripheral blood RNA. Total RNA was extracted using the PAXgene blood RNA kit (Qiagen). Total RNA (200ng) was amplified, labeled, and hybridized on Illumina Human BeadChips, querying the expression of ~24K 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 are log₂-transformed.

Microarray-based SNP genotyping: All subjects provided DNA at baseline. DNA was labeled, fragmented, and hybridized on Illumina IM chips according to Illumina instructions. The Illumina 1M SNP array assays 1.2 mln markers per sample, including over 100,000 copy number variants, providing the highest

genotyping density available on the market. Arrays were scanned using the Illumina iScan equipment. Proprietary software performance was compared with 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.

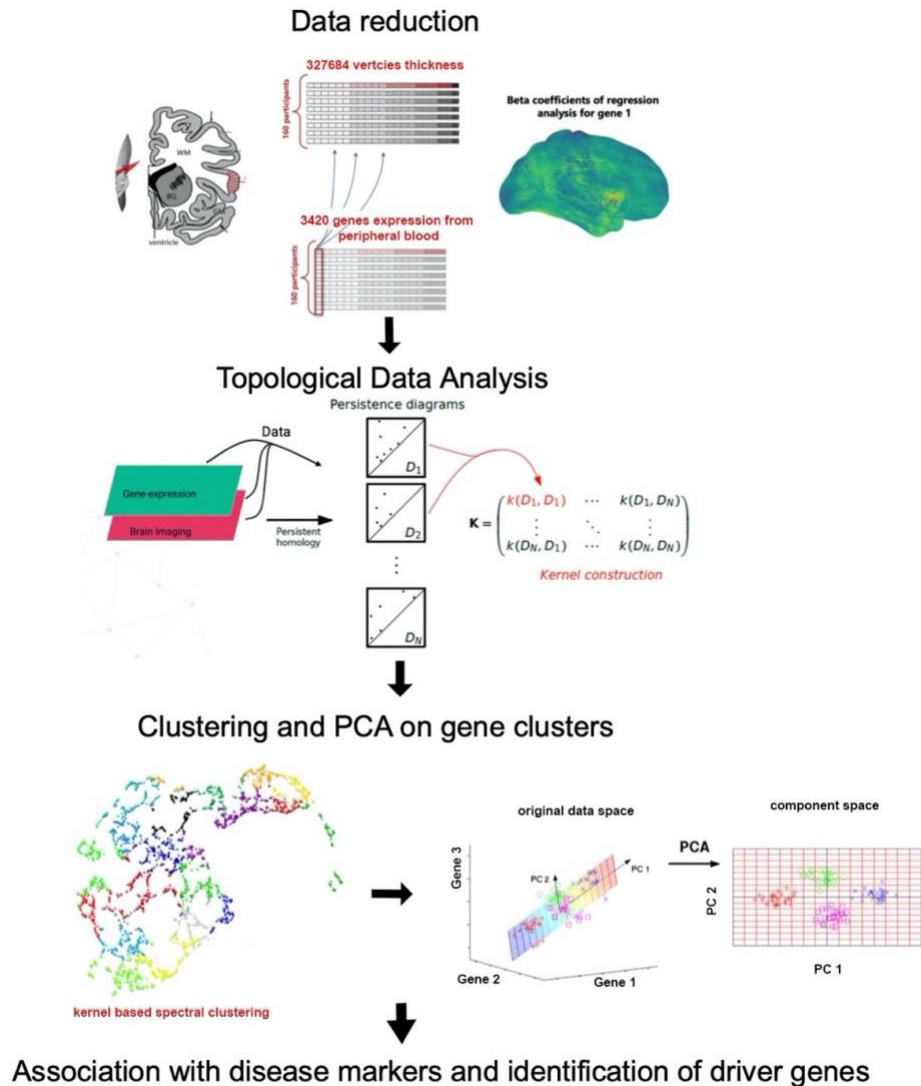


Figure 1: Overview of methods

Preliminary data reduction

To reduce noise and find endophenotype-specific transcriptomic data from blood, we performed an initial data reduction step by selecting transcripts which were 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 which were significantly ($p < 0.05$) associated to both neurodegeneration measures.

Univariate surface mapping

An average surface was constructed by computing the Talairach coordinates at each vertex for each subject [50]. The scans were then averaged together using FreeSurfer6.0. Vertex-wise regressions for each of the transcripts was performed using age, gender, and education as covariates to map the association of the gene expression value with average cortical thickness using a MATLAB toolbox SurfStat [51]. The beta coefficients were used for the persistent homology pipeline.

Persistent Homology

The topological method known as persistent homology (PH) builds a data-driven coarse descriptor of a weighted discretized surface while retaining meaningful geometric information [52]. We analyzed the SurfStat triangular meshes using PH to characterize each map through the evolution of homological features (Figure 2-A) 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[53] 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 2-B). 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.

Table 2: ADNI Demographics

Variable	CN	EMCI	LMCI	AD	p-value
Mean (std)	(n=154)	(n=215)	(n=104)	(n=42)	
Age	73.69(6.01)	70.95(7.44)	72.02(7.37)	75.31(9.37)	0.0001*
Gender(M/F)	72/82	115/100	57/47	25/17	N.S
MMSE	29.03(1.21)	28.32(1.49)	27.60(1.76)	22.88(1.92)	<0.0001*
AV45 SUVR	1.11(0.19)	1.16(0.21)	1.27(0.23)	1.39(0.22)	<0.0001*

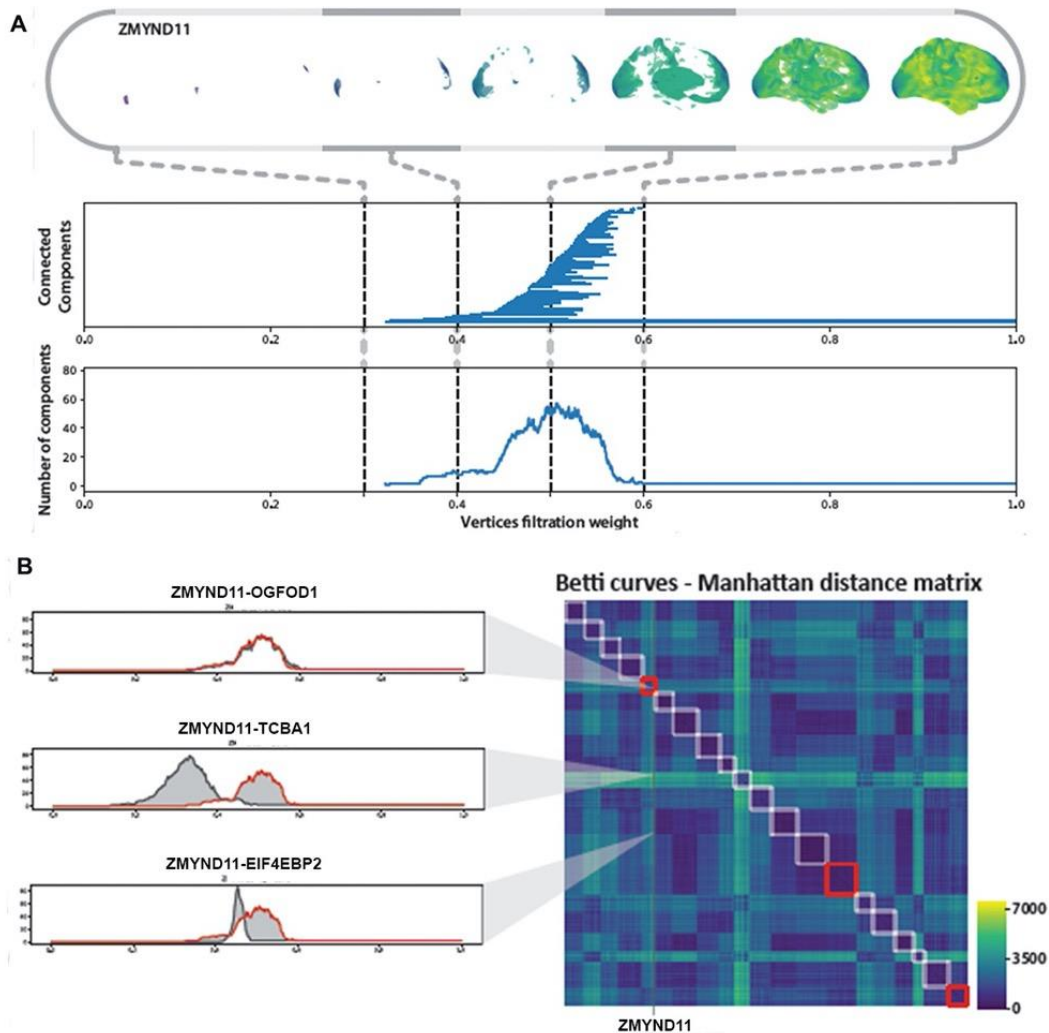


Figure 2: Persistent homology pipeline representation with one gene. (A) The filtration obtained from the cortical map relative to gene ZMYND11. For increasing threshold values (on the x-axis), we record the appearance of topological features of the surface (individual components and holes) and when 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

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 [54]. Persistent homology can be fit into the framework via the mathematical representation of kernel matrices [55]. 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 [56]. 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.

Cluster Analysis

To identify disease relevant clusters, we performed principal component decomposition on each cluster to represent the cluster by its first principal component or the “eigen-gene” representing the set of genes in that cluster [57]. The eigen-gene represent 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 which were differentially expressed in aMCI vs. 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 ANOVA, $p > 0.05$). 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, $p_{\text{fdr}} < 0.05$) for each iteration followed by FDR correction.

Gene Enrichment Analysis and Driver Genes Identification

Gene enrichment analysis was performed using the topGO package in R using the weight01 algorithm [58]. P-values computed by two-sample t-test comparing aMCI vs. the grouped CN and naMCI subjects were used as scores for the gene ontology 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. REVIGO tool was used to visualize the enriched biological processes. REVIGO summarizes long Gene Ontology 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[59]. The GO terms from the gene enrichment analysis were provided to REVIGO for visualization of enriched processes. The driver genes were further investigated for role and function in AD pathology.

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.

Variant Analyses in Driver Genes

We used the tool MAGMA (Multi-marker Analysis of GenoMic Annotation)[60] to analyze variants in the driver genes. The gene analysis 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 amnesic MCI phenotype.

Validation in an external dataset

The clusters or sets of genes obtained through our pipelines were validated in an external dataset which consisted of MRI, gene expression (specific transcripts used for the discovery analysis), demographic and amyloid PET data. We identified 515 subjects from the ADNI study [2] 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). Pre-processed data was downloaded wherein processing for mRNA values and neuroimaging analyses were performed according to ADNI

protocol [61]. The gene expression data was quantile normalized and log₂ transformed. We applied a logistic regression classification (with age and gender 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 SUVR > 1.11 in ADNI and Flutemetamol SUVR>1.17 in ImaGene). These cut-offs for Florbetapir and Flutemetamol [62] 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.

Results:

Subject demographics and identification of gene set

The three diagnostic groups had no significant difference in mean age and gender distribution. There was a significant difference in mean number of years of education, MMSE and hippocampal volume ($p < 0.05$) (Table1). Using a simple linear model, we identified 3420 genes from ~25,000 transcripts which were significantly associated with both hippocampal volume and cortical thickness which were mapped onto the group average cortical thickness using vertex wise regression with age, gender and education as covariates to obtain betti distances from Persistent Homology.

Cluster analysis and significant clusters

We obtained an optimal clustering solution of 20 clusters from the elbow and silhouette analysis. Figure 3-A shows the representation of the twenty-cluster solution. We identified three clusters which were significantly associated with disease diagnosis post FDR correction (two sample ttest, $p < 0.05$)- cluster-5, cluster-14, and cluster-20. For the null model validation, 1000 iterations yielded our 3-cluster solution outside the 95% confidence interval post FDR correction (two sample t-test, $p_{\text{fdr}} < 0.05$) 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 ($p_{\text{fdr}} < 0.05$). Cluster 14 consisted of 255 genes of which 53 were differentially expressed ($p_{\text{fdr}} < 0.05$) and cluster 20 consisted of 157 genes of which two were differentially expressed ($p_{\text{fdr}} < 0.05$). All 3,420 genes along with their clustering solution are summarized supplementary file S1.

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 supplementary Table 5. **Positive regulation of apoptotic processes and cell proliferation** were the most significant biological processes associated with cluster 5 (Figure 3-B). Driver gene analysis yielded 2 differentially expressed cluster 5 genes SPINK6 and ZMYND11 ($p_{\text{fdr}} < 0.05$). Cluster 14 was associated with **NF-kappa B signaling pathway** (Figure 3-C). There were 53 differentially

expressed genes in the cluster. For cluster 20, the main overrepresented pathway was **cellular response to retinoic acid and mitochondrial respiratory chain complex processes** (Figure 3-D).

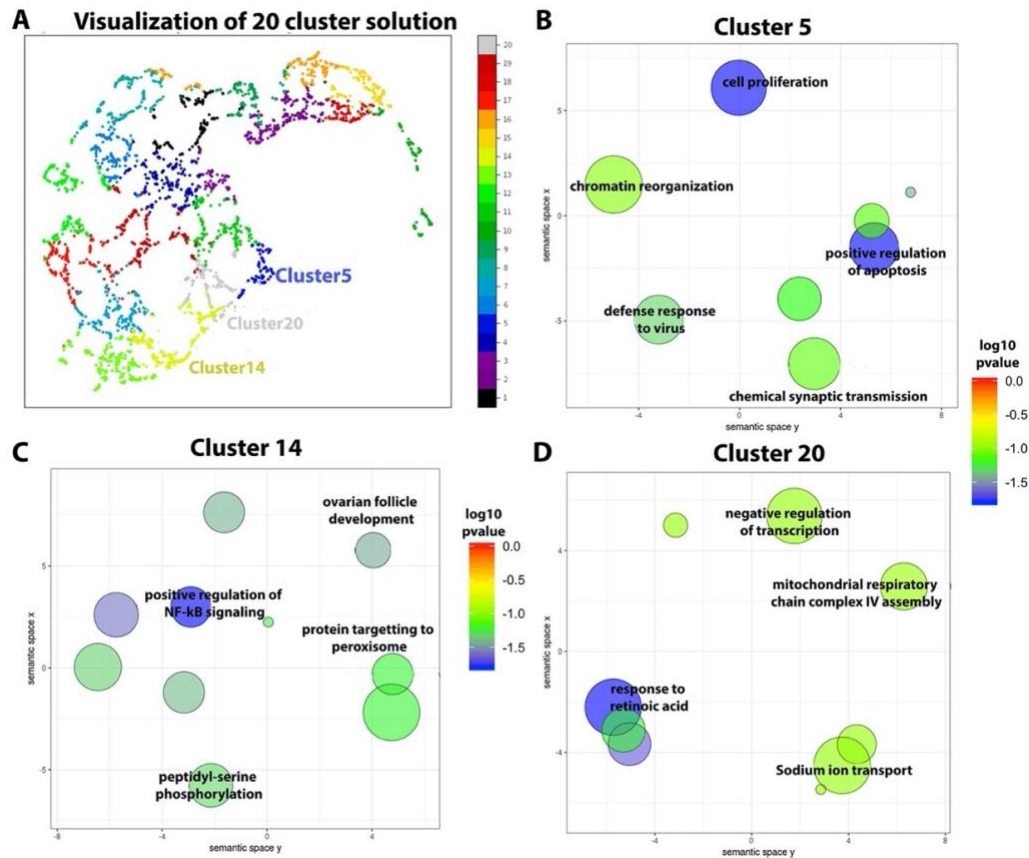


Figure 3: (A) Visualization of 20 cluster solution. Three clusters were significantly associated with amnesic MCI diagnosis post FDR correction; cluster5,14 and 20. Gene enrichment analysis with significant over-represented pathways in (B) Cluster 5 (C) Cluster 14 and (D) Cluster 20. Blue and green bubbles are 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 i.e., the semantic similarity.

Gene enrichment results with the top ten enriched biological pathways using all genes within the significant clusters are summarized in supplementary Table S2. Figure 4 represents gene networks of the significant clusters with nodes grouped by top overrepresented biological processes based on the gene ontology 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' (Supplementary Table 6). All 57 driver genes were downregulated in aMCI. The driver genes and their fold change are summarized in Table 3.

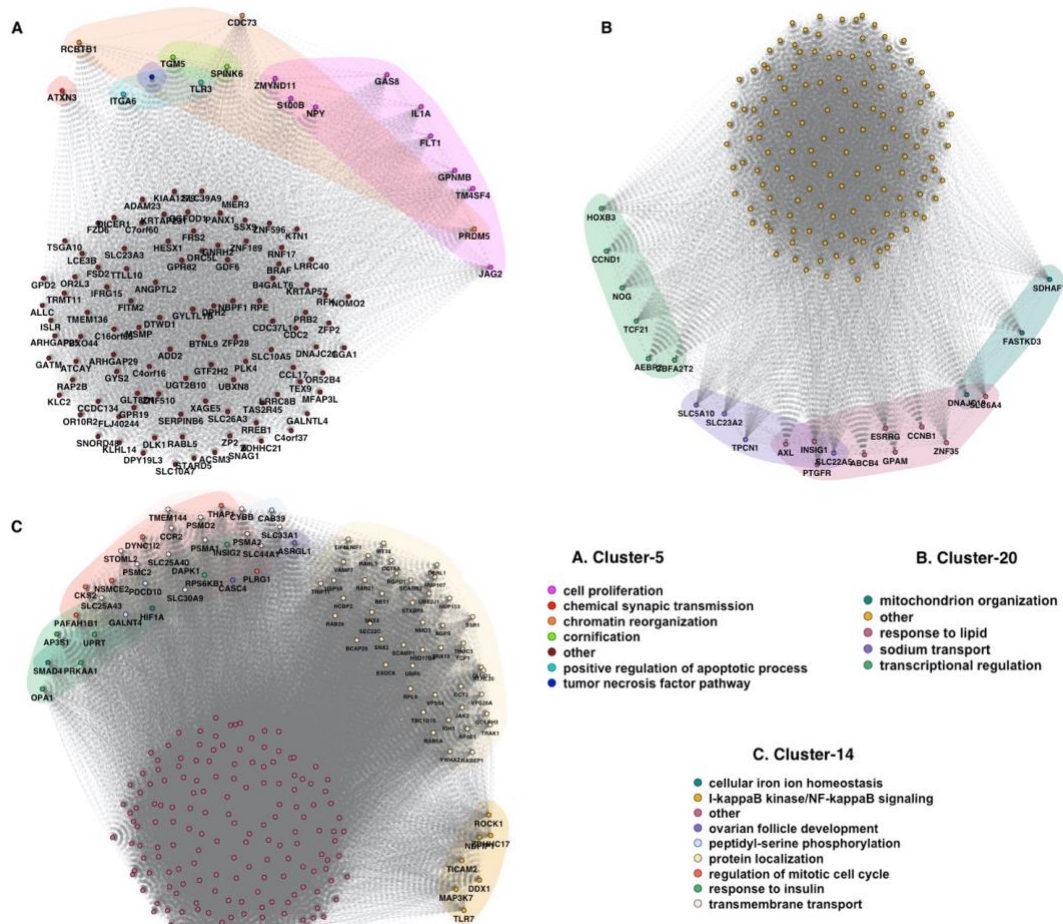


Figure 4: Gene clusters visualizing nodes and relevant biological processes

Table 3: Driver genes and their fold change

Downregulated in amnesic MCI sample		
Driver-Gene	fold change	p-value (p_{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	<0.001
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
SLC44A1	-0.016	0.006
SMAD4	-0.017	0.002
SNX4	-0.019	0.001
SPINK6	-0.007	0.001
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
VPS26A	-0.014	0.007
ZDHHC17	-0.02	0.001
ZMYND11	-0.008	0.001
ZNF654	-0.016	0.001

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 5-A). In the ADNI sample, the LMCI and AD groups show reasonable under expression of the driver genes compared to the other two groups; CN and EMCI, but this difference was not as pronounced as it was in the ImaGene sample (Figure 5-B).

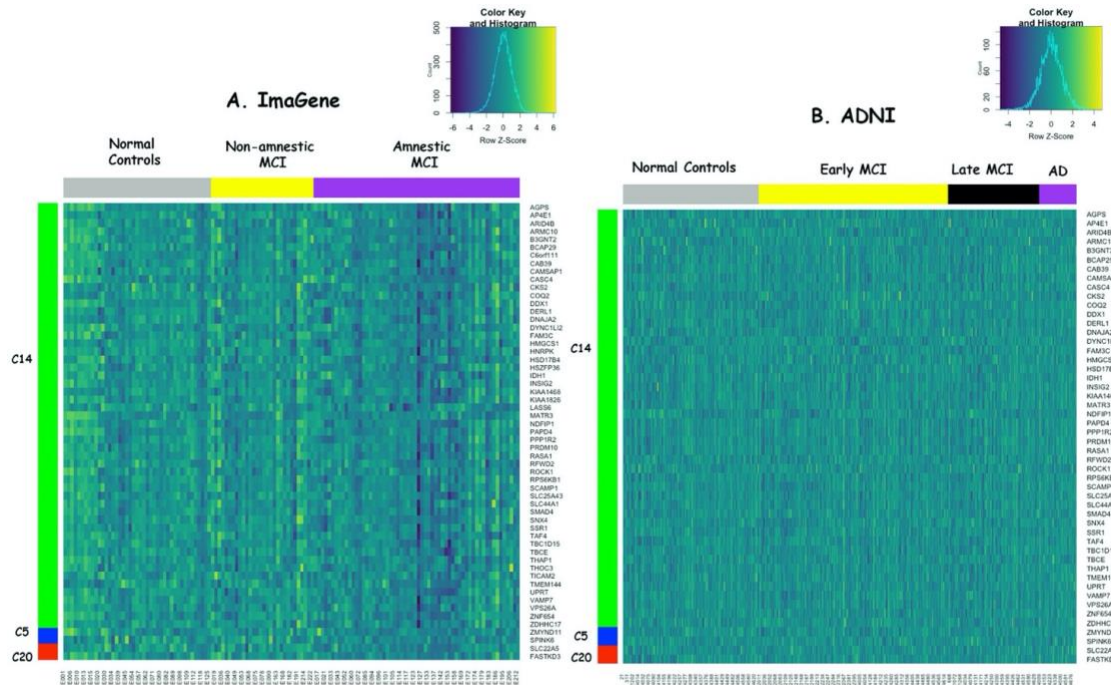


Figure 5: Molecular signatures of driver genes across diagnostic groups in A) ImaGene and B) ADNI. Rows represent driver genes and columns represent subjects. Color blocks represent cluster data(row) and diagnosis(column)

Cluster Level Associations

SurfStat mapping of cluster-level associations with cortical thickness showed AD-like patterns (Figure 6) [63]. All three clusters showed negative association with cortical thickness in the medial, inferior, and lateral temporal, the precuneus, posterior cingulate, the 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-kappa B signaling pathway**).

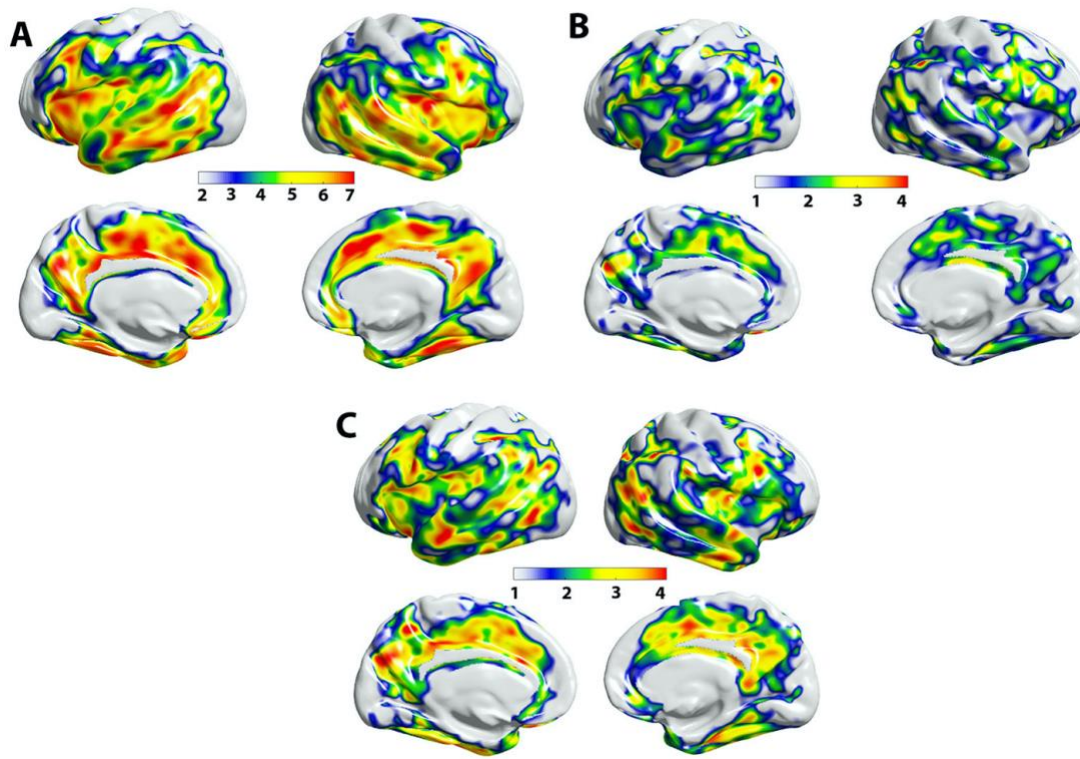


Figure 6: Cluster association(T-statistic) with cortical thickness($T > 1.96$, $p < 0.05$).

(A) Cluster 5 (B) Cluster 14 and (C) Cluster 20

Variant Analyses:

We identified two driver genes ARMC10 and KIAA1468 from the 57 driver genes with SNPs significantly associated with aMCI in our sample. 47 SNPs were annotated to the ARMC10 gene on chromosome 7, which had a significant association to aMCI phenotype with p-value of 0.02 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 0.04 and Z statistic 1.697. The results from the variant the MAGMA variant analyses are summarized in Table 4. We further investigated the association of individual SNPs within the genes to

disease phenotype using PLINK[64] GWAS tools but none of the individual SNPs had a significant genome wide association post FDR correction.

Table 4: Variant analyses within driver genes

Gene Name	Chromosome	Start (bp)	Stop (bp)	ZSTAT	P-value
ARMC10	7	102715174	102740210	2.00	0.02
KIAA1468	18	59854506	59974355	1.70	0.04
SNX4	3	125165488	125239058	1.46	0.07
PAPD4	5	78908243	78982471	1.37	0.09
ROCK1	18	18529701	18691812	1.36	0.09
AP4E1	15	51200780	51298097	1.27	0.10
ZMYND11	10	180405	300577	1.09	0.14
TICAM2	5	114914339	114952142	0.97	0.17
DNAJA2	16	46989274	47007625	0.95	0.17
PPP1R2	3	195241221	195270224	0.92	0.18
TMEM144	4	159122749	159176439	0.81	0.21
LASS6	2	169312759	169631644	0.77	0.22
IDH1	2	209100951	209120478	0.77	0.22
DERL1	8	124025404	124054663	0.77	0.22
CAMSAP1	9	138700333	138799060	0.73	0.23
PRDM10	11	129769601	129872730	0.68	0.25
C6orf111	6	99846534	99873263	0.67	0.25
DDX1	2	15731745	15771235	0.65	0.26
CKS2	9	91926113	91931618	0.52	0.30
ZNF654	3	88108394	88193814	0.51	0.30
ZDHHC17	12	77157854	77247481	0.47	0.32
NDFIP1	5	141488324	141534008	0.47	0.32
VPS26A	10	70883908	70932617	0.32	0.38
RPS6KB1	17	57970407	58027787	0.23	0.41
SPINK6	5	147582357	147594700	0.20	0.42
RFWD2	1	175913967	176176386	0.20	0.42
CAB39	2	231577557	231685790	0.19	0.43
THAP1	8	42691817	42698474	0.16	0.44
SLC22A5	5	131705396	131731306	0.10	0.46
SMAD4	18	48556583	48611412	0.04	0.48
FASTKD3	5	7859272	7869150	0.01	0.49
B3GNT2	2	62423262	62451866	0.01	0.50
ARID4B	1	235330210	235491532	-0.05	0.52
THOC3	5	175386534	175395318	-0.05	0.52
TAF4	20	60549854	60640866	-0.13	0.55
AGPS	2	178257471	178408564	-0.15	0.56
MATR3	5	138609441	138667366	-0.18	0.57
SSR1	6	7281283	7313541	-0.20	0.58
TBC1D15	12	72233487	72320629	-0.35	0.64
CASC4	15	44580909	44707959	-0.51	0.70
INSIG2	2	118846033	118867604	-0.52	0.70
COQ2	4	84184972	84206067	-0.57	0.72
TBCE	1	235530728	235612280	-0.58	0.72
HSD17B4	5	118788138	118878030	-0.59	0.72
RASA1	5	86564070	86687743	-0.59	0.72
BCAP29	7	107220422	107263762	-0.59	0.72
HSZFP36	19	11832080	11849824	-0.67	0.75
FAM3C	7	120988905	121036422	-0.82	0.79
SCAMP1	5	77656339	77776562	-0.92	0.82
DYNC1L1I2	16	66754796	66785526	-1.03	0.85
SLC44A1	9	108006906	108200785	-1.14	0.87
KIAA1826	11	105878629	105892981	-1.30	0.90
HNRPK	9	86582998	86595692	-1.48	0.93
HMGCS1	5	43287572	43313614	-2.00	0.98

Validation in ADNI

Our logistic regression model (with age and gender as covariates) to predict amyloid positivity using the driver gene transcripts yielded an AUC (Area Under Curve) of 0.74 in ADNI and 0.73 in ImaGene (Figure 7A). Our logistic regression model to predict amnesic MCI 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 (Figure 7 C-D) and inferior parietal thickness in the ADNI dataset ($p < 0.05$, data not shown).

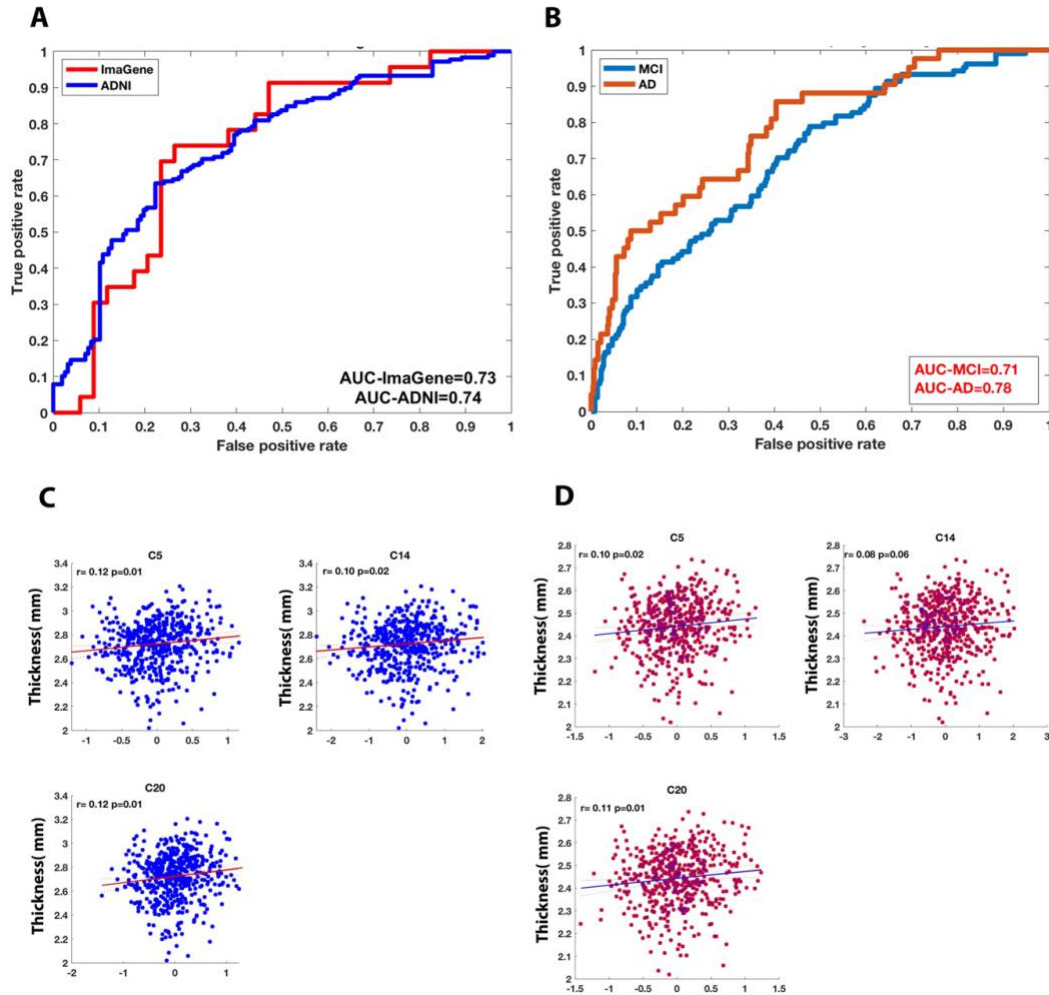


Figure 7: ROC using driver genes from each significant cluster for ADNI (blue) and ImaGene (red). (A) amyloidosis and (B) MCI and AD diagnosis in ADNI cohort. (C) Cluster level association with average inferior temporal thickness and (D) average cortical thickness (x-axis: eigen gene value, y-axis: thickness measure).

Discussion:

The need for a cost-effective way that can serve as the 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 together with respect to the various phenotypic manifestations of the disease. While there are many highly sensitive plasma and fluid biological markers which 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 multiple downstream and upstream processes which 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 amnesic MCI phenotype and can also help in risk prediction for conversion to AD. Persistent homology 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* (Zinc Finger MYND-Type Containing 11) and *SPINK6* (Serine Peptidase Inhibitor Kazal Type 6). *ZMYND11* is a protein coding gene that has been associated with mental retardation, and autosomal dominant non-

syndromic intellectual disability [65] . *SPINK6* is a kazal-type serine protease inhibitor that acts on kallikrein-peptidases in the skin. It is a gene associated with keratinization [66]. To date, *SPINK6* has not been linked to central nervous system processes or diseases.

Cluster 14 which was mainly associated with NF- κ B signaling had 53 differentially expressed driver genes. 18 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 TGF-beta signaling pathway (Table 7). The neuroprotective cytokine TGF-beta is increased in AD and is associated with chronic neuroinflammation which is hypothesized to lead to neurodegeneration[67]. *ROCK1*(Rho-associated protein kinase 1) is increased in AD and *ROCK1* depletion reduces amyloid beta levels in the brain[68]. *SMAD4* regulates TGF-Beta signaling pathway through feedback mechanism[69]. Reducing *RPS6KB1* expression has shown to improve spatial memory and synaptic plasticity in a mouse model of AD[70]. *NDFIP1*, *TICAM2* and *ZDHH17* are associated with positive regulation of I-kappaB kinase/NF-kappaB 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[71]. *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[72]. *ZDHH17* is a protein coding gene involved in palmitoylation. Disruption of protein

palmitoylation has been implicated in pathogenesis of neurodegenerative diseases, including AD[73], Huntington's disease, schizophrenia and intellectual disability. Detailed description of driver genes relevant to neurodegeneration and AD along with their molecular function are summarized in supplementary 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, folate-sensitive and disorders of intracellular cobalamin metabolism [74]. *FASTKD3* interacts with components of mitochondrial respiratory and translation machineries [75]. 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 [76]. *SLC22A5* (Solute Carrier Family 22 Member 5) is a gene involved with fatty acid metabolism in mitochondria [77]. 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 [78].

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 abeta induced mitochondrial fission and neuronal death[79]. *KIAA1468*(RELCH) regulates intracellular cholesterol distribution from recycling endosomes[80]. Although individual SNPs did not survive genome wide association($p_{\text{fdr}} > 0.05$), further studies are warranted in a larger sample size for identification of disease

relevant variants or eQTLs 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 since external validation of a model's predictive performance is extremely crucial to examine how reliable and accurate model predictions are. It is important to note that historically many variables found using prediction algorithms do not cross-validate very well [81] 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 which are potentially biologically relevant despite acquisition methods. In efforts to standardize and improve cross-platform analyses, data harmonization methods which 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 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, ANMI, 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 [82]. 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 5 gene-set signature for classifying normal controls vs MCI and normal controls vs 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 vs control using PBMCs as the RNA source[83]. 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 identified 207 differentially expressed genes using different AD tissue microarray datasets in the NCBI_GEO database[84]. 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 β and p-tau plasma assays [85, 86], it is less likely that gene expression data will be clinically useful, but blood gene expression does help 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 role in diagnostic prediction algorithms and as therapeutic targets through experimental validation.

Overall, our analysis has contributed towards 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.

Table 5: Gene Ontology analyses for the differentially expressed genes within the significant clusters

Cluster 5						
	GO.ID	Term	Annot ated	Rank in classic Fisher	Classic Fisher	weigh t01
2	GO:0043065	positive regulation of apoptotic process	3	198	1	0.015
1	GO:0008283	cell proliferation	15	125	0.307	0.015
3	GO:2001237	negative regulation of extrinsic apoptosis	3	43	0.066	0.028
4	GO:0070268	cornification	2	27	0.044	0.034
5	GO:0042981	regulation of apoptotic process	9	101	0.191	0.046
6	GO:0051607	defense response to virus	2	28	0.044	0.046
7	GO:0043123	positive regulation of I-kappaB kinase/N.F kappaB	2	199	1	0.063
8	GO:0032760	positive regulation of tumor necrosis factor production	2	200	1	0.073
9	GO:0007268	chemical synaptic transmission	4	201	1	0.075
10	GO:0006325	chromatin organization	5	71	0.109	0.095
Cluster 14						
	GO.ID	Term	Annot ated	Rank in classic Fisher	Classic Fisher	weigh t01
1	GO:0043123	positive regulation of I-kappaB kinase/NF-kappaB	5	524	1	0.016
2	GO:0006879	cellular iron ion homeostasis	3	525	1	0.026
3	GO:0001541	ovarian follicle development	2	77	0.084	0.036
4	GO:0032868	response to insulin	4	166	0.163	0.038
5	GO:0031398	positive regulation of protein ubiquitination	3	526	1	0.04
6	GO:0007346	regulation of mitotic cell cycle	8	527	1	0.046
7	GO:0018105	peptidyl-serine phosphorylation	4	528	1	0.049
8	GO:1903595	positive regulation of histamine secretion	2	78	0.084	0.053
9	GO:0055085	transmembrane transport	14	529	1	0.059
10	GO:0006625	protein targeting to peroxisome	4	530	1	0.059

Cluster 20						
	GO.ID	Term	Annot ated	Significant	classic Fisher	weigh t01
1	GO:004 2493	response to drug	11	1	0.18	0.025
2	GO:007 1300	cellular response to retinoic acid	2	0	1	0.033
3	GO:003 2355	response to estradiol	3	0	1	0.08
4	GO:000 0122	negative regulation of transcription by ...	6	0	1	0.136
5	GO:000 6814	sodium ion transport	3	1	0.051	0.137
6	GO:005 2106	quorum sensing	1	1	0.017	0.138
7	GO:006 0731	positive regulation of intestinal epithelial structure maintenance	1	1	0.017	0.138
8	GO:007 0715	sodium-dependent organic cation transport	1	1	0.017	0.138
9	GO:190 2603	carnitine transmembrane transport	1	1	0.017	0.138
1 0	GO:003 3617	mitochondrial respiratory chain complex assembly	1	1	0.017	0.142

Table 6: Gene Ontology analyses for all genes within the significant clusters

Cluster 5						
	GO.ID	Term	Annota ted	Rank in classic Fisher	classic Fisher	weight 01
1	GO:0090 630	activation of GTPase activity	2	239	1	0.033
2	GO:0001 525	angiogenesis	6	240	1	0.055
3	GO:0043 066	negative regulation of apoptotic process	5	75	0.25	0.068
4	GO:0006 457	protein folding	2	241	1	0.081
5	GO:0070 374	positive regulation of ERK1 and ERK2 cas...	3	242	1	0.082
6	GO:0000 186	activation of MAPKK activity	2	243	1	0.128
7	GO:0001 676	long-chain fatty acid metabolic process	1	244	1	0.138
8	GO:0006 633	fatty acid biosynthetic process	1	245	1	0.138
9	GO:0042 632	cholesterol homeostasis	1	246	1	0.138
1 0	GO:0006 637	acyl-CoA metabolic process	2	247	1	0.138
Cluster 14						
	GO.ID	Term	Annota ted	Rank in classic Fisher	classic Fisher	weight 01
1	GO:0045 727	positive regulation of translation	4	658	1	0.047
2	GO:0051 131	chaperone-mediated protein complex assembly.	2	659	1	0.048
3	GO:0045 737	positive regulation of cyclin- dependent ...	2	660	1	0.053
4	GO:0042 795	snRNA transcription by RNA polymerase II	6	661	1	0.062
5	GO:0000 186	activation of MAPKK activity	4	662	1	0.067
6	GO:0035 556	intracellular signal transduction	41	608	0.8	0.077
7	GO:0032 781	positive regulation of ATPase activity	3	663	1	0.081
8	GO:0006 457	protein folding	9	664	1	0.085
9	GO:0019 886	antigen processing and presentation of e...	2	665	1	0.09
1 0	GO:0070 125	mitochondrial translational elongation	4	666	1	0.098

Cluster 20						
	GO.ID	Term	Annot ted	Rank in classic Fisher	classic Fisher	weight 01
1	GO:0055 085	transmembrane transport	14	287	1	0.02
2	GO:0071 347	cellular response to interleukin-1	2	288	1	0.041
3	GO:0071 356	cellular response to tumor necrosis fact...	2	289	1	0.041
4	GO:0010 971	positive regulation of G2/M transition o...	2	290	1	0.046
5	GO:0045 737	positive regulation of cyclin- dependent ...	2	291	1	0.046
6	GO:0006 367	transcription initiation from RNA polymerase	3	292	1	0.048
7	GO:0007 080	mitotic metaphase plate congression	2	293	1	0.063
8	GO:0030 198	extracellular matrix organization	2	294	1	0.079
9	GO:0001 974	blood vessel remodeling	2	295	1	0.084
1 0	GO:0010 976	positive regulation of neuron projection...	2	296	1	0.11

Table 7: AD relevant driver genes identified from significant clusters

Tissues Expressed (RNA expression)	Function	Alzheimer's Disease/Neurodegeneration Relevance
Detected in all; low tissue specificity	Catalyzes the conversion of acyl-DHAP to alkyl-DHAP through the exchange of the acyl chain in acyl-DHAP for a long chain fatty alcohol. Important enzyme for plasmalogen synthesis.	Processing of APP results in the formation of APP intracellular domain (AICD) and Amyloid Beta (A β). AICD increases AGPS transcription and protein level leading to increased synthesis of plasmalogens under normal physiological conditions. Pathological levels of A β increases ROS production which oxidize plasmalogens and reduce their levels [87]
Detected in all; low tissue specificity	Armc10 protein shown to regulate mitochondrial trafficking via interactions with KIF5/Miro/Trak2 complex.	Overexpression of Armc10 in hippocampal neurons plays a protective role against A β induced mitochondrial fragmentation [79]
Detected in all; low tissue specificity	Catalyzes the final steps in the synthesis of CoQ10	COQ2 mutations result in COQ10 deficiencies and are associated with increased risk for multiple system atrophy [88]
Detected in all; low tissue specificity	ATP dependent RNA helicase	Biomarker for AD progression [89]
Detected in all; low tissue specificity	Component of ER-associated degradation (ERAD). Recognizes substrates in the ER and works in a complex to translocate it across the ER membrane into the cytosol for degradation	Involved in the translocation of APP for degradation. HtrA2 (protease) get recruited to complexes containing Derlin-1 and assist with ERAD. HtrA2 may regulate cellular levels of APP via ERAD [90]
Detected in all; low tissue specificity	Co-chaperone of Hsp70s in protein folding	May suppress/limit tau aggregation. It was also found to be significantly effective at suppressing tau prion activity (an ability of tau aggregates that can induce aggregation of soluble tau). It was also found to be

		upregulated in brain tissue samples of patients with MCI and AD. The levels were diminished in AD patients compared to MCI suggesting that DNAJA2 could have a protective role in early neurodegeneration progression [91]
Detected in all; low tissue specificity	Acts on peroxisomal beta-oxidation pathway for fatty acids. Catalyzes the formation of 3-ketoacyl-CoA intermediates from both straight-chain and 2-methyl-brnched-chain fatty acids	Docosaehaenoic acid (DHA) (omega-3-FA) has been shown to decrease the risk of developing AD and levels of DHA are reduced in the brains of AD patients. The liver plays an important role in supplying DHA to the brain. Liver tissues of AD patients have been shown to have reduced DHA levels but increased levels of tetracoashexanoic acid (DHA precursor) suggesting a defect in the beta-oxidation of this precursor to DHA by D-bifunctional protein (DBP). DBP is encoded by <i>HSD17B4</i> , and the expression of this gene is reduced in AD [92]
Detected in all; low tissue specificity	Ceramide synthase that catalyzes formation of ceramide from sphingosine and acyl-CoA substrates, with high selectivity toward palmitoyl-CoA as acyl donor. Ceramides generated play a role in inflammatory response	Downregulated in AD [93]
Detected in all; low tissue specificity	regulation of DNA virus-mediated innate immune response by assembling into the HDP-RNP complex	Activities of MATR3 decreased in AD [93]
Detected in all; low tissue specificity	Activates Nedd4 family E3 ubiquitin ligases and is essential for T regulatory cell stability and function	Increased expression of divalent metal transporter 1 (DMT1) has been shown to play a role in iron dyshomeostasis and A β generation in the brain of AD patients. Ndfip1

		can degrade DMT1 via ubiquitination and reduce intracellular iron accumulation. Ndfip1 immunoreactivity was shown to be decreased in the cortex and hippocampus of APP/PS1 transgenic mice compared to WT [71]
Detected in all; low tissue specificity	Key regulator of actin cytoskeleton and cell polarity	ROCK1 levels are significantly elevated in MCI and AD patient brains. ROCK1 knockout mice showed decreased levels of APP. ROCK1 may be a therapeutic target for decreasing A β production[68]
Detected in all; low tissue specificity	Acts downstream of mTOR signaling in response to growth factors and nutrients to promote cell proliferation, cell growth and cell cycle progression. Regulates protein synthesis via phosphorylation of EIF4B, RPS6 and EEF2K and contributes to cell survival by repressing the pro-apoptotic function of BAD	Deletion of 22 amino acid residues in the RPS6KB1 sequence associated with AD [94]
Detected in all; low tissue specificity	binds to smad-binding elements (SBE) regions on DNA and regulate transcription of certain genes involved in cell growth and proliferation	Smad4 expression shown to be increased in AD patient brain tissue. This correlated with increased levels of APP and TGF- β [69]
Detected in all; low tissue specificity	Involved in regulation of endocytosis and intracellular trafficking. Involved in the recycling of endocytosed transferrin receptors and prevent their degradation.	Significantly decreased levels of SNX4 protein in late-stage AD patient brains compared to controls. In mouse studies, it was found that SNX4 was increased in the brains of young APP/PS1 mice and decreased in older mice. Overexpression of SNX4 resulted in increased BACE-1 levels and led to

		increased BACE-1 mediated APP processing, whereas knockdown of SNX4 resulted in the opposite effects [95]
Detected in all; low tissue specificity	Involved in beta-tubulin folding. Needed for the proper maintenance of neuronal microtubule network	TBCE knockdown mouse neurons resulted in an accumulation of misfolded tubulins that were not incorporated in microtubules. Knockdown also resulted in the accumulation of phosphorylated tau in the neuronal cell bodies. This suggests that disruption of normal tubulin/microtubule biosynthesis may be associated with tau accumulation [96]
Detected in all; enriched in Brain	carbohydrate transmembrane transporter	Downregulated in the ACC in Dementia with Lewy Bodies patients. Regulated by miR-25 [97]
Detected in all; low tissue specificity	Palmitoyltransferase specific for certain neuronal proteins. Might be involved in the targeting of important proteins involved in the initiation of endocytosis.	Nicotinamide mononucleotide adenylyltransferase (NMNAT2) is an NAD-synthesizing enzyme that is supplied to axons from neuronal cell bodies and is crucial for their survival. Decreases in supply of NMNAT2 may contribute to axon degeneration. Palmitoylation of NMNAT2 is necessary for it to be targeted to post-Golgi vesicles. Several ZDHHC palmitoyltransferases, in particular zDHHC17, play a role in in NMNAT2 palmitoylation and localization [98]

Chapter 2: Transcriptomic profiling of brain amyloidosis and Alzheimer's disease phenotypes identifies important novel gene targets from peripheral blood gene expression data

Introduction:

Alzheimer's disease (AD) is a devastating neurodegenerative disorder and the only top 10 cause of death in the United States that cannot be cured, prevented or even slowed [99]. AD places a huge burden on the healthcare system, with annual costs exceeding a quarter of a trillion dollars [99]. A fundamental approach to address this multifaceted disease is biomarker development and identification of novel therapeutics for the early disease stages. For a late-life complex multifactorial disease, such as AD, where both genetic and environmental factors are involved, integrating multiple layers of genetic, imaging, and other biomarker data is a critical step for the development of robust risk assessment platforms and for identifying important therapeutic targets for disease intervention. There are many factors that contribute to AD pathology like amyloid deposition, such as vascular alterations, systemic inflammation, genetic/epigenetic status, tau phosphorylation, mitochondrial dysfunction and more; however, the hallmark pathologies observed in AD are the cerebral extracellular deposition of β -amyloid ($A\beta$) and the intracellular formation of neurofibrillary tangles (NFT) by hyperphosphorylated tau [100-102].

Although $A\beta$ deposition as a pharmacological target has been supported by genetic and experimental data, therapeutic approaches focused on slowing $A\beta$ production or removing $A\beta$ deposition have not had much success to date [101]. Recent advances have identified numerous factors including mitochondrial failure,

oxidative stress, inflammation, iron accumulation, cellular trafficking impairment and lipid metabolism alterations to contribute to the neurodegenerative pathological process in AD [103]. The multifactorial nature of AD suggests that multiple therapeutic targets need to be identified and tested together. Hence, we need a systems-level approach to build biomarker profiles which can be used for drug discovery and screening/risk assessment. While there are many highly sensitive plasma and fluid biomarkers which have shown promising results in predictive and prognostic models, mRNA measures add to this paradigm because they are directly correlated to physiological and pathophysiological changes and can shed light on multiple downstream and upstream processes. Transcriptomic analyses in AD are largely restricted to postmortem brain tissue and thus reflect only end-stage AD and not the therapeutically desirable pre-symptomatic and early symptomatic stages. Given that that 80% of the genes expressed in human brain are also expressed in peripheral blood lymphocytes [8], peripheral blood gene expression analyses can be fundamental to the development of robust blood-based risk assessment platforms. In an effort towards pre-symptomatic risk assessment and precision medicine and biomarker development for AD, we used a transcriptomic data analysis approach to evaluate the efficacy of characterizing and clustering peripheral blood gene expression data with respect to brain amyloidosis.

Methods:

Datasets

Discovery Dataset: ADNI: The discovery dataset used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). ADNI recruited participants at 57 sites in the US and Canada. 356 participants ADNI who had corresponding blood gene expression, Florbetapir SUVR, neurodegeneration measures and CSF measures (CN (Cognitively Normal) =120, EMCI (Early Mild Cognitive Impairment) =130, LMCI (Late Mild Cognitive Impairment) =72 and AD=34) were identified from the ADNI study (**Table 8**).

Microarray-based Gene Expression: ADNI Gene expression data were produced using the Affymetrix Human Genome U 219 array (Affymetrix, Santa Clara, CA). Peripheral blood samples were collected using PAXgene tubes for RNA analysis. Total RNA was extracted using the PAXgene Blood RNA Kit, following the protocol provided by the manufacturer. The quantity and quality of extracted RNA were assessed using the NanoDrop and the Agilent Bioanalyzer, respectively. Samples were randomized to plates, with checks to ensure sex and diagnosis balance, and hybridized to Affymetrix Human Genome U219 array plate. Array hybridization, washing, staining, and scanning were carried out in an Affymetrix GeneTitan system. The quality of gene expression data, including sample quality and hybridization, and overall signal quality, were analyzed using Affymetrix Expression Console software and Partek Genomic Suite 6.6, according to standard QC criteria provided by each software package. Raw expression values

obtained directly from CEL files were pre-processed using the Robust Multi-chip Average normalization method. The Affymetrix HG U219 Array contains 530,467 probes for 49,293 transcripts. All Affymetrix U219 probe sets were mapped and annotated with reference to the human genome (hg19). The ADNI Genetics Core performed several additional QC steps using the RMA(Robust Multichip Average) normalized expression array data [61, 104].

Amyloid PET imaging data: Preprocessed AV45 (florbetapir) PET data with summary and composite measures of standardized uptake value ratio (SUVR; standardized uptake values in cerebrum, intensity-normalized to cerebellum) were downloaded from the ADNI database and matched to gene expression data cross-sectionally. The image acquisition and processing pipelines are as previously established [105, 106]. The [¹⁸F]-florbetapir PET acquisition and preprocessing protocols are available at <http://www.adni-info.org>. In our main analyses, we used a subset of SUVR values from University of California, Berkeley downloaded from the ADNI database (<http://adni.loni.usc.edu>). The composite variable was derived by averaging the SUVRs across the frontal, anterior and posterior cingulate, lateral parietal, and lateral temporal gray matter regions. The University of California, Berkeley, protocols for [¹⁸F]-florbetapir preprocessing, coregistration, and normalization have been previously described [107].

T1-weighted structural MRI data and analyses: Processed FreeSurfer data with cortical volumes and thickness measures were downloaded and matched to gene expression data using timepoint/visit and subject ID as unique identifiers. Cortical reconstruction and volumetric segmentation were performed with the FreeSurfer

image analysis suite, freely available for download online (<http://surfer.nmr.mgh.harvard.edu/>). The technical details of these procedures are described in prior publications [108, 109].

Biomarker data: CSF A β 1-42, total tau (t-tau) and phosphorylated tau at position 181 (p-tau181) concentration data generated using the Research Use Only (RUO) INNOBIA AlzBio3 immunoassay (Fujirebio, Belgium) were downloaded from the ADNI database and matched to the visit at which gene expression data were obtained.

Genome-wide array genotyping data: ADNI1/GO/2 sample data whole-genome sequenced (WGS) at high coverage and genotyped using the Illumina Omni 2.5M BeadChip were downloaded for 812 subjects from the ADNI database (<http://adni.loni.usc.edu>). Detailed genotyping protocols were described previously [61, 104]

Table 8: ADNI demographics table

DX	CN (n=120)	EMCI (n=130)	LMCI (n=72)	AD (n=30)	p-value
Age, years Mean (SD)	73.5(6.1)	70.3(7.1)	71.5(6.9)	74.5(9.5)	<0.01
Sex (% male)	46%	51%	52%	60%	0.8
Education, years Mean (SD)	16.6(2.6)	16.1(2.5)	17.0(2.4)	15.3(2.7)	0.006
MMSE Mean (SD)	28.3(2.1)	28.1(1.9)	27.6(2.5)	26.9(3.1)	<0.01
Amyloid SUVR Mean (SD)	1.11(0.18)	1.17(0.2)	1.28(0.2)	1.40(0.2)	<0.01

Validation dataset: ImaGene: Our validation sample included 160 subjects from the Imaging and Genetic Biomarkers of Alzheimer's Disease (ImaGene) study who were clinically diagnosed as either (1) MCI (Mild Cognitive Impairment) (N = 108) or (2) CN (Cognitively Normal) (N = 52). The MCI group was further divided into those presenting with predominantly amnesic (aMCI, N = 70) or those presenting with non-amnesic phenotype (naMCI, N = 38) (**Table 9**). All subjects had gene expression and 1.5T MRI data available. In addition, a subset of the sample (n=57) had [¹⁸F]-futemetamol amyloid PET SUVR data available.

Microarray-based Gene Expression: Total RNA was extracted using the PAXgene blood RNA kit (Qiagen). RNA was extracted in batches using a semi-automated extraction system (Qiagen Qiacube). RNA was stored at -80C. RNA quantity was assessed with Nanodrop spectrophotometer, and quality was checked with Agilent Bioanalyzer Nanochips. Total RNA (200ng) was amplified, labeled, and hybridized on Illumina Human BeadChips, querying the expression of ~24K 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. Low-level quality-control analysis was performed using several indices, including inter-array Pearson correlation, and clustering based on variance. Poor quality arrays were excluded from further analyses. Data clustering allowed for detection and correction of batch effects, a potential confounder in this kind of studies. Data were normalized using quantile normalization. mRNA levels are log₂-transformed.

Table 9: ImaGene demographics

Variable	NC	aMCI	naMCI	p-value
Mean (std)	(n=52)	(n=70)	(n=38)	
Age	69.0(7.9)	69.8(8.5)	69.8(8.5)	0.9
Education	17.6(2.04)	15.5(2.7)	16.5(2.88)	0.001
Sex (%male)	57%	37%	52%	0.7
MMSE	28.8(1.2)	27.0(2.5)	27.9(1.9)	<0.001
Hippocampal vol (mm³)	8602 (1092)	7990 (1324)	8718 (1023)	0.002

Data availability

Data used in preparation of this article were obtained from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). Data from the ImaGene dataset will be made available upon request.

Gene expression analyses

Transcripts which were differentially expressed between the diagnostic groups in the ADNI study were identified (p-value corrected for false discovery rate $p_{fdr} < 0.05$, ANOVA(ANOVA). When multiple probes mapped to same gene, the median value of transcripts was used. A distance matrix using the pairwise Euclidian distance between the expression values was built and followed by k-means clustering. The optimal number of clusters were determined by silhouette analysis and elbow plot (**Supplementary Figure 1**). Principal component (PC) decomposition with rows representing genes and subjects as columns was performed to analyze the cluster level expression per subject in our sample, and identify the first PC data vector representing each gene cluster also known as the “eigengene” [57]. The variance

explained by the first PC for each cluster was >40%. This method of eigengene representation has been adapted from previously published methods [57]. All the above analyses were done using MATLAB (2019a).

Gene Enrichment and cluster-level associations

Gene enrichment analysis was performed using the topGO package in R using the weight01 algorithm [58]. The p-values from the differential expression analyses described above were used in the topGO analysis to select the most significant disease-relevant biological processes associated with the given cluster. The eigengene scores for each subject were used to perform cluster-level correlations with relevant AD phenotypes - age, education, hippocampal volume, amyloid SUVR, CSF p-tau181, CSF A β 1-42, CSF t-tau (Pearson correlation, $p < 0.05$).

Driver Genes analyses

To identify specific mRNA measures or genes that are critical to AD pathogenesis, such as those driving the association with amyloidosis in the brain, we determined the driver genes for each cluster. We defined driver genes as genes with significantly higher Pearson correlation with amyloid PET SUVR than the eigengene's correlation at the $p < 0.05$ significance level. We next investigated the between-group expression differences for these driver genes (ANOVAANOVA, $p < 0.01$) as well as their function and role in AD pathogenesis. To externally validate our results, we tested the association of the driver gene variables discovered in the ADNI dataset with brain amyloidosis and MCI diagnosis in the ImaGene dataset and validated their predictive accuracy using support vector machine (SVM) and logistic regression classifiers.

Variant Analyses

We used the tool Multi-marker Analysis of GenoMic Annotation (MAGMA) to analyze variants in the driver genes[60]. MAGMA uses a multiple linear principal components' regression model and linear regression statistics 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 the remaining PCs as predictors of the phenotype of choice 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 AD diagnosis phenotype.

Imaging and Spatial Parametric Mapping analyses

To visualize the regional pattern of associations in 3 dimensions, we downloaded all preprocessed [¹⁸F]-florbetapir PET data from the Laboratory of Neuroimaging Image Data Archive (<https://ida.loni.usc.edu>). We aligned the PET images to the corresponding MRI from the same visit, normalized them to MNI space using the transformation matrices obtained from the MNI spatial transformation and intensity-normalized the voxel values in the cerebrum to the average intensity value of the whole cerebellum, which resulted in standardized uptake value ratio (SUVR) images. To explore the spatial distribution of the associations, we reproduced the final linear regression models using voxel-wise regression in SPM12. Because of the exploratory nature of our secondary results, we allowed a less stringent visualization threshold: voxelwise threshold of $p < 0.01$ (uncorrected)

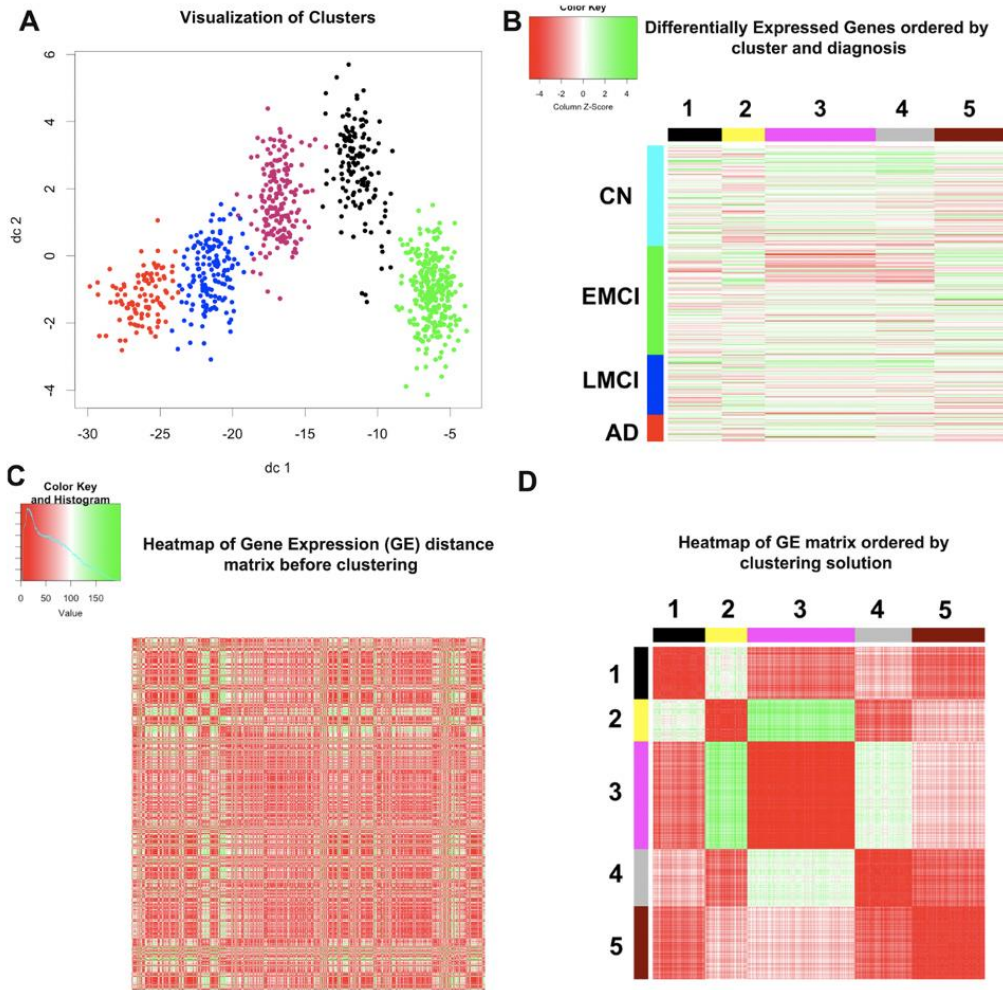
with the minimum cluster size (k) dependent on the minimum threshold level with p_{FWE} (Family Wise Error) <0.05 .

Results:

Gene expression analyses

Our differential expression analysis yielded 864 genes from ~49,000 transcripts which were differentially expressed across diagnostic groups ($p_{FDR}<0.05$). K-means clustering based on Euclidian distance with optimized clustering solution provided 5 distinct clusters of genes (**Figure 8A**). The raw gene expression values once ordered by clustering labels and diagnosis showed distinct expression patterns between groups and across clusters (**Figure 8B**). The gene expression matrix heatmap similarly showed distinct clusters of genes sharing similar expression profiles when ordered by the clustering labels obtained from our analysis (**Figure 8C-D**).

Figure 8: Visualization of k-means clustering (A) and raw gene expression data ordered by clustering solution and diagnosis (B). Heatmap of the gene-gene network before (C) and after clustering (D)



Gene enrichment and cluster level associations

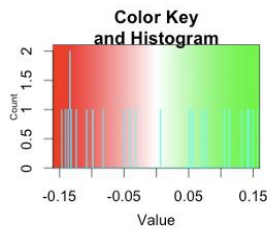
Gene enrichment analysis using topGO and the weight01 algorithm provided five main broad biological pathways that were overrepresented for each cluster with reference to Alzheimer's Disease phenotype. The main overrepresented pathway in cluster 1 was mitochondrial function with the top significant node in the

directed acyclic graph (DAG) representing mitochondrial respiratory chain complex assembly (GO: 0022108). The main enriched pathway in cluster 2 was protein complex oligomerization with the top significant nodes being protein heterooligomerization and oxygen transport (GO:0051291 and GO:0015671, respectively). In cluster 3 the main significant node in the DAG was, Transforming Growth Factor β (TGF- β) receptor signaling pathway (GO: 0007179). Cluster 4 was significantly associated with protein autophosphorylation (GO:0046777) and cluster 5 with activation of MAP (Mitogen-activated Protein) kinase kinase (MAPKK; GO :0000186).

We found that all five clusters were significantly associated with amyloid SUVR with cluster 1 (mitochondrial function), cluster 2 (Protein oligomerization), and cluster 5 (MAPKK activation) having negative association while cluster 3 (TGF- β receptor signaling) and cluster 4 (protein autophosphorylation) having positive association (all $p < 0.05$; **Figure 9**). CSF levels of A β 1-42 were significantly associated with all five clusters (all $p < 0.05$), and the direction of association was always opposite the one observed with amyloid PET SUVR. The TGF- β signaling cluster was positively associated with CSF p-tau levels ($p = 0.004$), while the mitochondrial function and MAPKK clusters were negatively associated with CSF p-tau levels ($p = 0.025$ and, $p = 0.015$, respectively) the TGF- β . The associations with CSF tau species were always opposite of those seen with A β 1-42. The protein autophosphorylation cluster was negatively associated with hippocampal volume ($p = 0.018$) while the mitochondrial function, protein oligomerization and MAPKK activation clusters were positively associated with hippocampal volume in the

sample ($p=0.011$, $p=0.004$ and $p=0.041$, respectively). These associations were in the opposite direction to those observed with amyloid PET SUVR. The TGF- β signaling cluster was negatively associated with education ($p=0.015$). None of the clusters were associated with age.

Figure 9: Visualization of the associations of each cluster with relevant clinical and biomarker traits. Positive correlations are in green; negative correlations are in red. The color intensity depicts the strength of the correlations. The numeric values are p-values for the correlation.



0.91	0.466	0.05	0.561	0.166	Age
0.35	0.201	0.015	0.367	0.314	Education
0.011	0.004	0.138	0.018	0.041	Hippocampal Volume
0.011	0.011	0.004	0.016	0.008	Amyloid SUVR (AV45)
0.01	0.006	0.013	0.015	0.017	CSF Ab42
0.025	0.077	0.004	0.057	0.015	CSF pTau
Mitochondrial fuction	Protein Oligomerization	TGF Beta signaling	Protein auto-phosphorylation	MAPKK activation	

Driver gene analyses

Across the five clusters we identified 28 driver genes defined as having stronger correlation with [¹⁸F]-florbetapir SUVR than the cluster-level correlation of the eigengene (i.e., first PC). We used one-way between-subject ANOVA to analyze the differential expression of the driver genes between diagnostic groups followed by Tukey's *post hoc* test to identify pairwise differences. The statistical analyses for the driver genes in the ADNI dataset are summarized in **Table 10**.

Table 10: Differential peripheral blood expression levels by clinical diagnosis for driver genes

Gene-Mean (SD)	CN (n=120)	EMCI (n=130)	LMCI (n=72)	AD (n=34)	p-value
BAIAP3	4.02(0.26)	4.04(0.22)	4.07(0.22)	3.87(0.29)	<0.0001** *
E2F2	9.37(0.72)	9.4(0.65)	9.34(0.62)	8.82(0.67)	<0.0001** *
PSMF1	9.44(0.43)	9.47(0.42)	9.44(0.42)	9.13(0.37)	<0.0001** *
SMOX	7.78(0.78)	7.8(0.75)	7.81(0.73)	7.22(0.71)	<0.0001** *
UBE2O	5.68(0.75)	5.76(0.66)	5.69(0.63)	5.22(0.56)	<0.0001** *
TNS1	7.21(0.93)	7.26(0.96)	7.16(0.88)	6.57(0.94)	0.001***
MCMBP	7.4(0.26)	7.48(0.23)	7.46(0.27)	7.57(0.18)	0.002*
NMI	9.94(0.27)	9.91(0.27)	9.92(0.25)	10.1(0.24)	0.002***
PRDM16	2.02(0.16)	1.98(0.12)	2.01(0.14)	1.92(0.13)	0.002***
RNF11	4.57(0.83)	4.66(0.77)	4.61(0.75)	4.09(0.73)	0.002***
TRIM58	9.97(0.56)	10.05(0.57)	10.01(0.52)	9.63(0.64)	0.002***

TCF3	5.03(0.33)	5.08(0.32)	5.03(0.3)	4.86(0.27)	0.004***
CD99L2	3.45(0.24)	3.37(0.2)	3.43(0.2)	3.35(0.18)	0.005**
ST6GALNAC4	4.66(0.74)	4.65(0.64)	4.68(0.61)	4.24(0.51)	0.006***
DECR1	10.39(0.16)	10.42(0.15)	10.39(0.16)	10.48(0.11)	0.008***
CLIC1	11.58(0.11)	11.57(0.11)	11.59(0.09)	11.63(0.09)	0.01***
FAM46C	8.67(1.02)	8.73(0.98)	8.62(0.93)	8.11(0.99)	0.01***
FAM92B	2.63(0.24)	2.54(0.18)	2.59(0.19)	2.56(0.19)	0.01**
KANK2	4.9(1.16)	4.92(0.98)	4.83(1.06)	4.27(0.96)	0.01***
PAX3	2.29(0.17)	2.26(0.15)	2.24(0.16)	2.2(0.14)	0.01*
PDDC1	4.45(0.29)	4.4(0.29)	4.41(0.31)	4.26(0.29)	0.01*
SH3GLB2	6.19(0.46)	6.18(0.42)	6.14(0.38)	5.93(0.37)	0.01***
TAS2R40	2.44(0.13)	2.41(0.12)	2.41(0.1)	2.37(0.1)	0.01*
WNK3	2.09(0.22)	2.01(0.18)	2.09(0.26)	2.05(0.18)	0.01**
EIF4G3	6.66(0.33)	6.7(0.28)	6.68(0.26)	6.83(0.22)	0.02*
ITM2B	11.47(0.11)	11.48(0.12)	11.48(0.12)	11.54(0.08)	0.02***
MMP12	2.71(0.15)	2.66(0.13)	2.67(0.16)	2.66(0.14)	0.02**
RFWD2	9.62(0.2)	9.64(0.19)	9.64(0.19)	9.73(0.19)	0.02*
* MCI vs AD ** CN vs MCI *** CN vs AD					

Next, we identified the transcripts corresponding to the driver genes in our validation dataset (ImaGene) and tested for correlations with amyloidosis ([¹⁸F]-flutemetamol (Flutemetamol SUVR, *N*=57). The correlations of the driver genes with amyloid SUVR for both datasets are summarized in **Table 11**.

Table 11: Driver genes' correlations with amyloid SUVR in ADNI and ImaGene

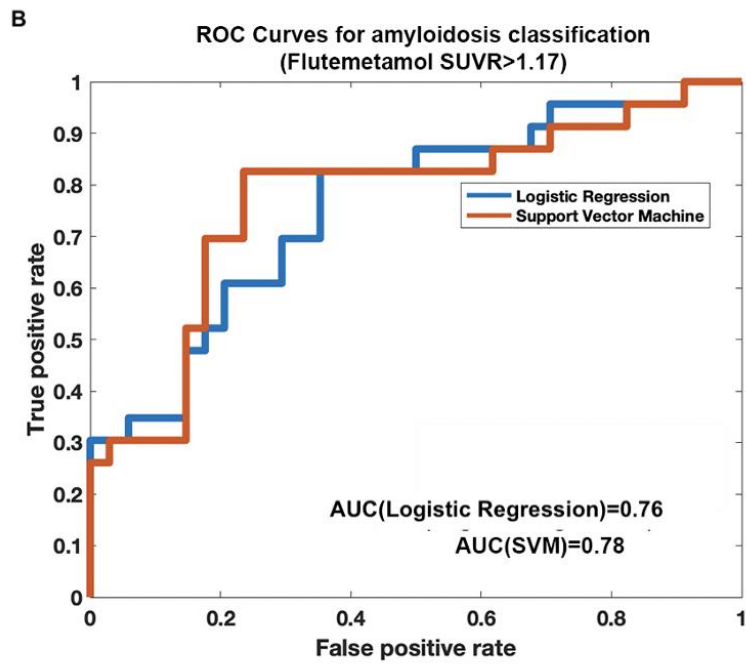
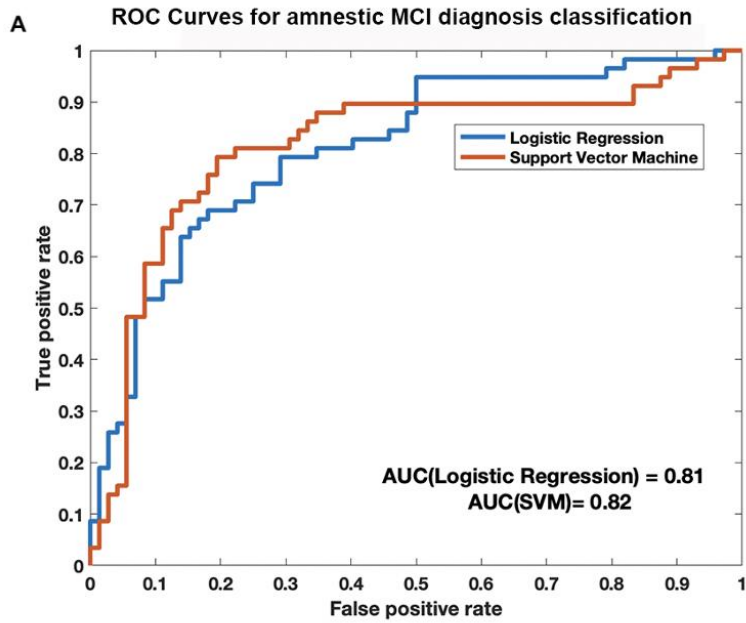
Gene	Function	Cluster	ADNI		ImaGene	
			(n=356)		(n=57)	
			Correlation with Florbetapir SUVR	p-value	Correlation with Flutemetamol SUVR	p-value
<i>E2F2</i>	apoptotic signaling pathway	1-Mitochondrial function	-0.16	0.003	0.08	0.53
<i>PSMF1</i>	proteasome function		-0.13	0.025	0	0.97
<i>TNS1</i>	actin binding/signal transduction		-0.11	0.035	0.09	0.52
<i>TRIM58</i>	erythropoiesis		-0.13	0.015	-0.02	0.87
<i>CLIC1</i>	activation of cAMP dependent PKA	2-Protein oligomerization	0.14	0.008	0.07	0.62
<i>DEC R1</i>	mitochondrial fatty acid beta oxidation		0.15	0.005	0.25	0.06
<i>ITM2B</i>	amyloid beta A4 precursor protein processing		0.15	0.006	0.11	0.43
<i>NMI</i>	IL2 signaling pathway		0.13	0.013	0.2	0.13
<i>RFWD2</i>	proteasomal degradation		0.15	0.006	-0.25	0.06
<i>CD99L2</i>	cell adhesion process	3-TGF- β signaling pathway	-0.15	0.005	-0.12	0.37
<i>FAM92B</i>	cilium biogenesis		0.15	0.005	0.37	0.005
<i>MMP12</i>	TGF beta signaling pathway		-0.15	0.006	-0.06	0.67
<i>PAX3</i>	neural development and myogenesis		-0.15	0.004	0	1

<i>PRD</i> <i>M16</i>	repressor of TGF beta signaling		0.16	0.003	-0.04	0.74
<i>TAS</i> <i>2R40</i>	Taste receptor		-0.17	0.001	-0.18	0.18
<i>WNK</i> <i>3</i>	Ion channel transport		-0.15	0.004	-0.07	0.59
<i>EIF4</i> <i>G3</i>	cytokine signaling in immune system	4-Protein autophosphorylation	0.15	0.003	0.11	0.42
<i>BAIA</i> <i>P3</i>	recycling of secretory vesicle transmembrane proteins	5-MAPKK activation	-0.12	0.003	-0.22	0.1
<i>KAN</i> <i>K2</i>	proapoptotic proteasomal degradation		-0.13	0.017	0.11	0.42
<i>RNF</i> <i>11</i>	inflammatory signaling pathway		-0.12	0.027	-0.3	0.02
<i>SH3</i> <i>GLB</i> <i>2</i>	endocytosis		-0.16	0.003	-0.04	0.77
<i>SMO</i> <i>X</i>	role in neurotransmission through regulation of cell surface receptor activity		-0.12	0.018	-0.16	0.23
<i>ST6</i> <i>GAL</i> <i>NAC</i> <i>4</i>	biosynthesis of ganglioside GD1A from GM1B		-0.12	0.021	0.12	0.36
<i>TCF</i> <i>3</i>	transcriptional regulator of neuronal differentiation		-0.16	0.003	-0.16	0.23
<i>UBE</i> <i>20</i>	NF kappaB activation negative regulation		-0.14	0.01	-0.18	0.18

While the driver genes did not achieve significant association in the ImaGene dataset due to the much smaller sample size, we observed that the direction of the correlations was largely in agreement and that the coefficient values were comparable for most of the transcripts across both datasets (**Table 11**). *FAM92B*

and *RNF11* were two driver genes that were significantly associated with amyloidosis in the ImaGene dataset ($r=-0.37$, $p=0.005$ and $r=-0.3$, $p=0.02$, respectively). A logistic regression model which included the 28 driver genes (**Table 11**) and age and sex as covariates predicted amnesic MCI diagnosis in ImaGene ($n=160$) with an AUC of 0.82. A radial basis function SVM classifier provided an identical AUC of 0.82 for diagnosing amnesic MCI (**Figure 10A**). Next, we used the four gene transcripts with significant or trend-level ($0.05 < p < 0.1$) [¹⁸] Flutametamol SUVR correlation values in ImaGene - *RNF11*, *RFWD2*, *FAM92B* and *DECR1*, along with age and sex as covariates in a logistic regression and a SVM classifier tasked with predicting amnesic MCI diagnosis. These models yielded an AUC of 0.75 and 0.77, respectively (**Figure 10B**).

Figure 10: ROC curves in the validation sample (ImaGene) for driver gene prediction of amnesic MCI diagnosis (n=160, panel A) and amyloid positivity (n=57, panel B)



Variant Analyses:

One of the driver genes, KANK2, and overall, 28 of the differentially expressed genes had SNPs significantly associated with AD and MCI phenotype (**Table 12**). To identify potentially important SNPs within these genes, we performed a standard association analysis using Fishers exact test to generate significance with the GWAS toolset plink[64] and extracted a gene-based report for the 29 genes. We found that 14 genes contained SNPs with significance of $p < 0.05$ and 6 of those genes contained SNPs with significance of $p < 0.01$ (**Table 13**).

Table 12: Variant analyses in differentially expressed genes and driver genes

Gene	Chromosome	NSNPS	Z-stat	P-value
<i>KANK2</i>	19.00	45.00	1.74	0.04
<i>BCL2L1</i>	20.00	23.00	1.80	0.04
<i>CSNK1E</i>	22.00	21.00	2.20	0.01
<i>CYP4A11</i>	1.00	14.00	3.25	<0.01
<i>GOLGA3</i>	12.00	60.00	2.33	0.01
<i>GPR3</i>	1.00	3.00	1.87	0.03
<i>LY75</i>	2.00	93.00	1.77	0.04
<i>MT1X</i>	16.00	2.00	1.90	0.03
<i>NDUFA5</i>	7.00	6.00	1.91	0.03
<i>NEFM</i>	8.00	5.00	3.61	0.00
<i>PLXNB3</i>	X	1.00	2.13	0.02
<i>RPL34</i>	4.00	4.00	1.80	0.04
<i>SELE</i>	1.00	23.00	2.67	<0.01
<i>TRAPPC10</i>	21.00	48.00	1.78	0.04
<i>UCP2</i>	11.00	4.00	1.82	0.03
<i>YWHAG</i>	7.00	22.00	1.67	0.05
<i>YBX3</i>	12.00	9.00	1.84	0.03
<i>PRKRA</i>	2.00	7.00	2.00	0.02
<i>EIF5B</i>	2.00	15.00	1.74	0.04
<i>PHTF1</i>	1.00	23.00	2.82	<0.01
<i>VAX1</i>	10.00	3.00	1.85	0.03
<i>COPZ1</i>	12.00	13.00	2.70	<0.01
<i>CARD8</i>	19.00	64.00	2.23	0.01
<i>KANK2</i>	19.00	45.00	1.74	0.04
<i>RSBN1</i>	1.00	29.00	2.10	0.02
<i>SARNP</i>	12.00	29.00	1.95	0.03
<i>ZNF518B</i>	4.00	27.00	1.75	0.04
<i>OR9A2</i>	7.00	2.00	1.84	0.03
<i>DEFB124</i>	20.00	3.00	1.93	0.03
<i>MIA3</i>	1.00	31.00	2.01	0.02

Table 13: Significant SNPs identified through GWAS analyses

Gene	CH R	SNP	BP	BETA	SE	t	p
BCL2L1	12	kgp1109473 3	12233244	0.2177	0.0840 5	2.59	0.00976 2
	12	rs885720	12248099	0.1024	0.0366 2	2.797	0.00528 7
LY75	2	rs11690478	16038690 3	0.09657	0.0363 3	2.658	0.00801 2
	2	kgp1452043 1	16042560 5	-0.3617	0.1214	-2.98	0.00296 8
HTF4	1	rs34258372	11409575 0	0.07979	0.0304 5	2.62	0.00895 8
COP1	12	kgp1895034 6	53008372	-0.4451	0.1715	- 2.596	0.00960 3
	12	kgp1132949 3	53015076	0.09267	0.0327 5	2.83	0.00477 1
CARD8	19	kgp1151166 6	53409877	-0.2934	0.1086	- 2.701	0.00704 7
	19	rs1978616	53413825	0.08562	0.0329 7	2.597	0.00958
MIA3	1	kgp3780310	22087246 9	-0.2898	0.0925 2	- 3.133	0.00179 5
KANK2	19	kgp574740	11166791	-0.1274	0.0560 6	- 2.273	0.02328
BCL2L1	12	rs2909009	12218529	-0.1022	0.0516 8	- 1.977	0.04835
	12	kgp4814914	12219367	-0.1018	0.0517 1	- 1.969	0.04929
	12	kgp1896842 3	13190565 8	-0.7145	0.3271	- 2.184	0.02923
LY75	2	rs1863218	16036878 7	- 0.06258	0.0288 3	- 2.171	0.03023
	2	rs11693108	16042858 7	-0.0572	0.0283 4	- 2.018	0.04388
	2	kgp1493271 7	16045111 1	-0.7155	0.2833	- 2.525	0.01175
	2	kgp1442835 0	16045250 6	- 0.05808	0.0283 7	- 2.047	0.04099
NEFM	8	kgp1684425	24829265	-0.3827	0.1896	- 2.018	0.0439

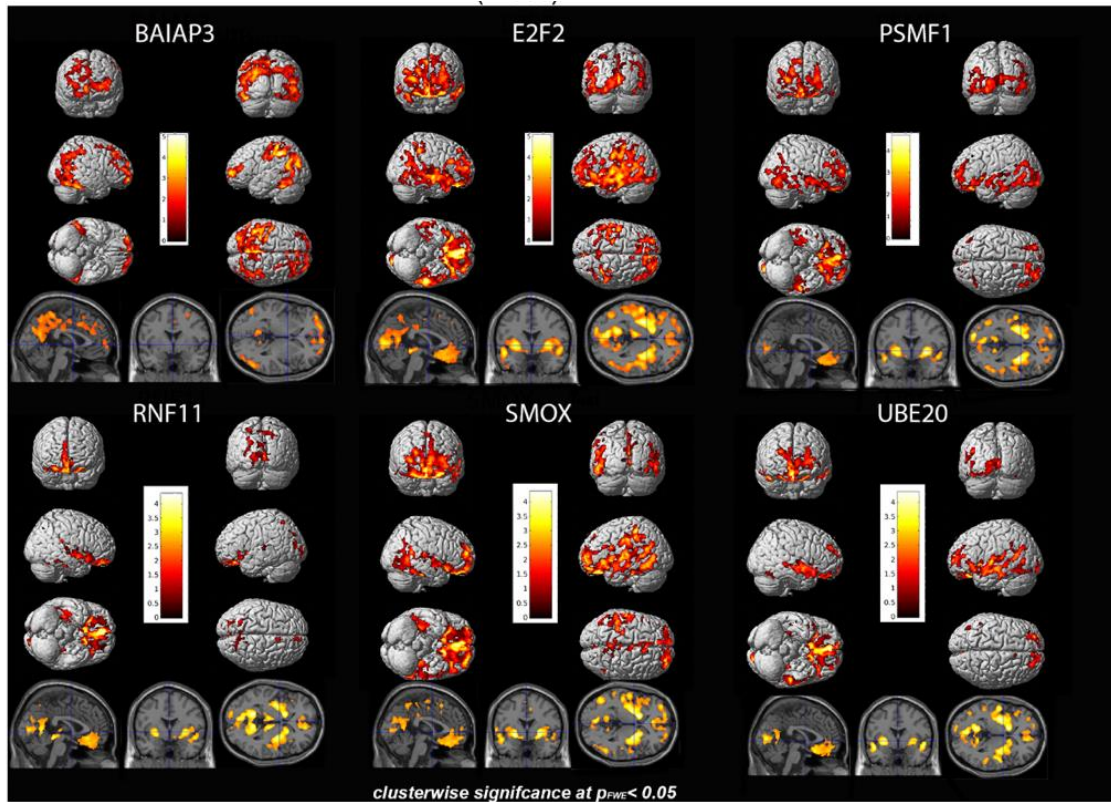
	8	kgp1722736	24830332	0.1144	0.0520 8	2.196	0.02841
RPL34	4	kgp2910324	10977028 3	0.07488	0.0374 7	1.998	0.04602
YWHAG	7	kgp1359574 4	75812297	1.29	0.5657	2.28	0.02287
	7	kgp1365483 2	75821885	-0.428	0.2153	- 1.988	0.04711
HTF1	1	kgp2282782 1	11405164 9	-0.353	0.1717	- 2.055	0.04017
	1	kgp937647	11405799 6	- 0.07298	0.0330 7	- 2.207	0.0276
	1	rs7554019	11408572 3	- 0.05937	0.0278 1	- 2.135	0.03309
VAX1	10	kgp7774102	11888536 4	-0.4166	0.18	- 2.315	0.02086
	10	kgp4602200	11888552 3	- 0.06687	0.0278 3	- 2.403	0.01649
	10	kgp2167921 7	11888655 1	-0.4298	0.2148	- 2.001	0.0457
COPZ1	12	rs665089	53015294	-0.3325	0.1581	- 2.103	0.03581
	12	rs583766	53023161	- 0.06997	0.0285 9	- 2.447	0.01461
	12	kgp1876701 3	53023745	-0.515	0.2538	- 2.029	0.04276
CARD8	19	kgp3239623	53405497	- 0.06212	0.0274 1	- 2.266	0.02373
	19	kgp4183224	53408899	0.08221	0.0409	2.01	0.04475
	19	rs7256400	53411448	0.06866	0.0320 2	2.144	0.03231
	19	kgp2154172 2	53411790	-0.7145	0.3271	- 2.184	0.02923
CARD8	19	kgp302691	53415028	- 0.05426	0.0273 9	- 1.981	0.04792
RSBN1	1	kgp6226105	11413661 8	0.2683	0.122	2.199	0.02813
RSBN1	1	rs1217225	11413999 1	- 0.05964	0.0271 7	- 2.195	0.02845
RSBN1	1	rs4838994	11415160 6	- 0.09573	0.0432 3	- 2.215	0.02706
ZNF518 B	4	kgp2648435	10053431	-0.0635	0.0266 4	- 2.383	0.01738

<i>MIA3</i>	1	rs437078	22086862 7	-0.1109	0.0468	-2.37	0.01801
<i>MIA3</i>	1	rs439513	22088449 8	- 0.09497	0.0396 7	- 2.394	0.01689
<i>MIA3</i>	1	rs368664	22088703 5	-0.1424	0.0717 9	- 1.984	0.04765

3D SPM analyses:

The top driver genes that showed maximum pairwise differences in CN vs AD and MCI vs AD in ADNI were *BAIAP3*, *E2F2*, *PSMF1*, *SMOX*, *UBE20* and *RNF11*. We used voxel-wise multiple regression in SPM12 with age and sex as covariates to visualize the pattern of association of the expression of these genes with brain amyloidosis. All six genes had negative association with amyloidosis, meaning lower expression of these genes was associated with higher amyloid deposition in agreement with the direction of the correlations with mean SUVR shown in **Table 11**. The transcripts showed strong association with brain amyloid deposition in the lateral temporal, lateral parietal and temporo- and parieto-occipital areas, in addition to regions of the frontal lobe (**Figure 11**).

Figure 11: Association pattern of driver genes with brain amyloidosis in the ADNI sample (n=356)



Discussion:

A multifactorial complex disease like AD, with multiple pathogenic mechanisms working together, calls for robust and reliable gene panels which can be used to uncover novel therapeutic targets and build models for risk assessment and screening. Targeted personalized therapies based on neurodegenerative and pathogenic profiles should be based on a broad, multitarget systematic approach. We derived predictive information from a large and inherently noisy peripheral blood transcriptomic dataset and analyzed it with respect to brain amyloidosis and clinical disease manifestations. Using a data-driven method and employing prior

information, we identified and characterized networks associated with important pathogenic mechanisms in AD and further reduced the transcriptomic data to a few critical mRNA measures. We observed clean separation of differentially expressed genes into distinct clusters, which may suggest a shared expression profile for certain pathogenic mechanisms. A systems approach like this can help provide novel targets which can be screened for potential therapeutic interventions.

The gene enrichment analyses identified several significant clusters of biological pathways that were overrepresented in AD. Mitochondrial dysfunction, represented within cluster X, is a highly relevant biological process in AD pathogenesis [110]. The driver genes in this cluster were *E2F2*, *TNS1* and *TRIM58*. *E2F2* plays a crucial role in control of cell cycle and regulates the activity of tumor suppression proteins. It is transcriptionally induced in neuronal cells after DNA damage and apoptotic response [111]. *E2F2* showed strong associations with disease status and brain amyloidosis. Altered expression of *TNS1* in astrocytes in the immune response pathway has been associated with Alzheimer's Disease[112]. *TRIM58* was significantly downregulated in whole blood mRNA of patients with Parkinson's Disease[113].

The protein oligomerization cluster (cluster Y) likely represents the pathogenic mechanisms involved in the formation of neurotoxic A β oligomers and plaques, and neurofibrillary tangles. *ITM2B*, *DECR1* and *CLIC1* were the main genes identified in this cluster. *ITM2B* gene has been associated with accumulation of amyloidogenic BRI2-derived peptides impairing synaptic long-

term potentiation [114]. *CLIC1* plays a role in A β -induced generation of reactive oxygen species by microglia [115] and can be a potential target for neuroinflammatory management. *DECR1* encodes an enzyme engaged in beta-oxidation and metabolism of unsaturated fatty enoyl-CoA esters. This gene is dysregulated in mice fed with cholesterol rich diet and associates with disruption of liver-brain axis signaling and metabolic stress [116].

The TGF- β signaling cluster is linked with the chronic inflammation and plays a role in both A β clearance and deposition. It has been shown to be dysregulated in AD, and hence, targeting the TGF- β network may lead to novel therapies [67, 117]. *FAM92B*, *CD99L2*, *WNK3*, *PAX3*, *PRDM16* and *TAS2R40* were the main driver genes in this cluster. *CD99L2* (CD99 Molecule Like 2) is protein coding gene which plays a role in leukocyte extravasation and has been associated with chorioangioma and cerebral palsy. It plays a role in blood-brain barrier and Immune cell transmigration signaling pathway [118]. *WNK3* (WNK Lysine Deficient Protein Kinase 3) is involved in an important signaling cascade that regulates the cation-chloride cotransporters. Inhibition of WNK3 Kinase signaling has been reported to reduce brain damage in ischemic neuroglial injury and suggests its potential role as a therapeutic target for neuroprotection in mouse models [119]. *PAX3* (paired box gene 3) is a transcription factor which plays a critical role during fetal growth and is active in neural crest cells for neural development and neurogenesis. It has been found to be one of the differentially methylated genes in hippocampal samples from a cohort of pure AD patients and controls [120]. The putative role of *PAX3* in adult neurogenesis has been linked to

hippocampal-dependent learning and memory tasks [121-123] and has been shown to be reduced with aging [124, 125]. *PRDM16* (PR/SET Domain 16) is a transcriptional regulator gene which plays a crucial role in neurovascular signaling and angiogenesis during cortical development and formation of neurovascular network [126]. *TAS2R40* (Taste 2 Receptor Member 40) is a taste receptor gene. Dysregulation of selected olfactory and frontal cortex in a gradient comparable to Braak staging has been reported in AD mouse models [127].

Numerous studies have reported that altered protein phosphorylation states of several proteins, like A β precursor protein and tau, are closely linked with AD [128]. Regulating signaling cascades in phosphorylation provide promising avenues for therapeutics [128, 129]. The driver genes identified from the protein autophosphorylation cluster were *EIF4G3* and *MCMBP*. *EIF4G3* is a translation initiation factor, the splicing of which is frequently disrupted in autism [130]. *EIF4G3* microexons suppress the expression of critical synaptic proteins and mice deficient in the gene display social behavior and memory deficits [130]. *MCMBP* (Minichromosome Maintenance Binding Protein) has been found to play a major role in Colorectal Carcinoma [131], but it has not yet been implicated in AD pathophysiology.

Cluster 5 was enriched in genes associated with MAPKK activation. MAPK pathways are involved in the pathophysiology and pathogenesis of AD. All MAPK pathways, i.e., the extracellular signal-regulated kinase (ERK), c-Jun N-terminal kinase (JNK) and p38 pathways, are activated in vulnerable neurons in patients with AD [132-134]. *BAIAP3*, *SMOX*, *UBE2O* and *RNF11* were the main driver

genes in this cluster and were also strongly associated with brain amyloidosis. The roles and utility of these genes as therapeutic targets for AD need to be further investigated. *BAIAP3* (BAI1 Associated Protein 3) gene encodes a brain specific angiogenesis inhibitor and is associated with optic atrophy and Alpha Thalassemia-Intellectual Disability Syndrome Type1 [135]. *SMOX* (Spermine oxidase) is a member of polyamine oxidases and has been reported to be involved in ischemic brain damage. It has been found to be an important mediator of cerebral ischemia injury. *SMOX* downregulation decreased neuronal apoptosis and inflammatory reactions in rat studies [136]. *UBE2O* (Ubiquitin Conjugating Enzyme E2 O) mediates monoubiquitination of target proteins. *RNF11* (RING finger protein 11) is also a member of ubiquitin-editing protein complex. Both *UBE2O* and *RNF11* negatively regulate NF- κ B activation [137, 138], which results in tissue specific brain inflammation and is implicated in AD and aging [139].

The variant analyses of the driver genes and differentially expressed genes identified 29 genes containing SNPs which have significant association with disease phenotype. Six of these genes also had individual SNPs with significant associations with (disease phenotype, amyloidosis). *BCL2L1* is a protein coding gene. Higher levels of Bcl-xl, a gene product of *BCL2L1*, have been found in reactive microglia from patients with AD and other neurological diseases[140]. Lymphocyte antigen 75 gene (LY75) codes for CD205, a dendritic cell surface receptor that interacts with MHC class I molecules²⁰ and plays an important role in T cell function[141, 142]. *COPZ1* is involved in coat protein complex I (COPI)-dependent trafficking. It has been shown that a reduced COPI-dependent

trafficking in vivo leads to a decrease in the amyloid plaque burden in an AD mouse model, and an improvement of the memory impairment observed in those mice . Twelve SNPs in COPI genes have been previously associated with AD risk[143]. Other genes with significant SNPs were *CARD8* and *MIA3*. *CARD8* has been associated with increased AD risk in women[144]. *MIA3* encodes the protein TANGO1 (Transport and Golgi organization protein 1) and has been linked to diabetes and alcoholic neuropathy; it has not yet been implicated in AD[145].

A classifier built on these driver genes showed a promising AUC of 0.82 in predicting amnesic MCI and an AUC of 0.77 for detecting brain amyloidosis in an external validation sample (ImaGene). Given that the relationship of these genes to AD pathophysiology validates in an external sample, these genes might be good candidates for further investigation as potential targets for early-stage screening and/or intervention.

To the best of our knowledge, this is one of the first studies to systematically evaluate transcriptomic data and gene networks from peripheral blood with respect to amyloidosis and clinical phenotypes. We identified several genes which warrant further investigation in experimental setting and animal models. Some limitations of the study merit consideration. In light of recent success of ultrasensitive A β and p-tau plasma assays [85, 86, 146], it is less likely that gene expression data will be clinically useful; however, whole blood gene expression does aid in the understanding of critical biological pathways associated with AD risk in a way that single analyte plasma protein assays cannot. Whereas polygenic risk scores and genome-wide association studies have uncovered the contribution of single

nucleotide polymorphic variants to AD, there is still a wide gap in knowledge about the pleiotropy of these risk genes and the potential myriad downstream processes associated with them. Gene expression data can be critical in bridging this gap. In summary, our study has made use of openly available transcriptomic data to systematically evaluate and identify specific mRNA measures associated with brain amyloidosis and clinical diagnosis from pathways that have been associated with AD.

Chapter 3: Transcriptomic profiling in Mild Cognitive Impairment using peripheral blood gene co-expression networks

Introduction:

Alzheimer's disease (AD), the most common neurodegenerative disease worldwide, is estimated to affect 6.2 million people over 65 years of age in the United States and represents the 5th cause of death in this same age group. It is expected to affect 13.8 million people by 2060[147].

AD is a progressive disease that causes a decline in cognitive performance, that ultimately leads to inability to independently function in society. This disease is characterized by three main clinical stages: a preclinical state, followed by mild cognitive impairment and ultimately the AD dementia stage[35]. Given this clinical continuum, it is important to develop biomarkers which can help in the distinction among these states and the determination of risk factors for progression. With the multifactorial nature and complex pathogenesis of the disease, the exact pathophysiological mechanisms remain largely unknown with multiple theories being formulated throughout the years. The hallmark findings are the accumulation of neurofibrillary tangles and neuritic plaques and neuronal cell death[148]. The study of gene expression patterns as networks in the prodromal stages of the disease can be useful in determining novel therapeutics and provide better understanding of systemic dysregulated processes, and the use of blood based genetic data is useful to this effort. [149]

Co-expression network analysis is a methodology that allows the analysis of complex nonlinear gene-gene interactions in biological systems and can help to

create a robust risk assessment platform and identify important therapeutic targets[150]. Gene co-expression networks, while incredibly useful for assessing disease relevant genes and their interactions, are also inherently noisy due to the vast the complexity of the biological systems and it can be tricky to tease apart specific disease relevant networks and genes from regular although essential networks for biological processes. It is therefore important to remove redundant edges and identify a backbone of the network, to find driver nodes pathogenic interactions [151]. This approach allows to identify the genes that can inform therapeutic and biomarker development in MCI and AD. In the present analyses, we modeled transcriptomics data in the MCI stage as a gene co-expression network while applying a backbone method. Our goal was to identify sensitive modules and hub genes associated with prodromal AD.

Methods:

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 (Mild Cognitive Impairment) (N = 108) or (2) CN (Normal Controls) (N = 52). The MCI group was further divided into those presenting with amnesic (aMCI, N = 70) or those presenting with non-amnesic phenotype (naMCI, N = 38) (**Table 14**).

Table 14: ImaGene sample demographics

Variable Mean (std)	NC (n=52)	aMCI (n=70)	naMCI (n=38)	p- value
Age	69.03(7.9)	69.82(8.5)	69.75(8.5)	0.9
Education	17.6(2.04)	15.5(2.7)	16.5(2.88)	0.001*
Gender(M/F)	30/21	26/43	20/18	N.S.
MMSE	28.8(1.2)	27(2.5)	27.9(1.9)	<0.001*
Hippocampus vol (mm³)	8602 (1092)	7990 (1324)	8718 (1023)	0.002*

Microarray-based Gene Expression: All subjects provided yearly peripheral blood RNA. Total RNA was extracted using the PAXgene blood RNA kit (Qiagen). Total RNA (200ng) was amplified, labeled, and hybridized on Illumina Human BeadChips, querying the expression of ~24K 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 are log₂-transformed.

Imaging: The detailed imaging protocol has been previously published [10]. All subjects received annual 1.5T MRI scans following the UCLA Alzheimer's Disease Research Center protocol consisting of coronal FI3D T1 MPRAGE (magnetization-prepared rapid gradient-echo imaging). Measures of neurodegeneration were obtained from coronal T1-weighted MPRAGE sequences. Scans were processed using Freesurfer (version 6.0) longitudinal pipeline, to obtain region specific and global measures of atrophy [49].

Differential expression and gene co-expression matrix

We identified transcripts which were differentially expressed between the diagnostic groups in the lmaGene study (p-value corrected for false discovery rate $p_{\text{fdr}} < 0.05$, ANOVA). When multiple probes mapped to same gene, the median value of transcripts was used. A co-expression profile for the differentially expressed genes was derived by building an adjacency matrix using pairwise Pearson correlation coefficients for the gene expression values. A network is fully specified by its adjacency matrix a_{ij} , a symmetric $n \times n$ matrix with entries in $[0, 1]$ whose component a_{ij} encodes the network connection strength between nodes i and j . To calculate the adjacency matrix, an intermediate quantity called the co-expression similarity s_{ij} is first defined. The default method defines the co-expression similarity s_{ij} as the absolute value of the correlation coefficient between the profiles of nodes i and j : $s_{ij} = |\text{cor}(x_i, x_j)|$ [150]. These weighted undirected gene co-expression networks represent complex biological systems which can uncover important information on disease pathology and underlying biological pathways.

Network backbone construction

Weighted graphs are often used to capture distance associations among a set of node variables: every edge is a positive real number denoting a distance between the linked nodes. Most real-world network data like biological networks are semi metric, this means that the triangle inequality of Euclidean geometry where the shortest distance between at least two nodes in the graph is the direct edge between them is not observed for every edge in the graph and an indirect path via other nodes can be the shortest distance. For a full network, there is an invariant

sub-graph of the original graph that does not change with the closure computation, and which is sufficient to compute all shortest paths. We refer to this subgraph as the metric backbone of a complex network. The size of the backbone subgraph, in relation to the size of the original graph, defines the amount of redundancy in the network: edges not on this backbone are superfluous in the computation of shortest paths, as well as all network measures derived from shortest paths (e.g. efficiency, betweenness, centrality)[152]. We constructed network backbones for the gene co-expression networks using the shortest path method.

Clustering and module Eigengene

To identify communities of functionally related genes, we used the mathematical technique hierarchical clustering which groups genes into smaller clusters and then orders the clusters into higher-level systems. The correlations are ordered, and a node is created between the highest-scoring (geometrically closest) pair of rows—the two gene sequences that were most nearly coregulated. The matrix is then modified to represent the joined elements as a single node, and all distances between the newly formed node and other gene sequences (rows) in the matrix are calculated. It is not necessary to recalculate all correlations because only those involving the two rows joined in the new node have changed. Typically, the node is represented by a link in the dendrogram, the height of the link being directly proportional to the strength of the correlation. The process of creating proportional links and joining genes into clusters continues until all genes in the experiment have been joined into a single hierarchical cluster through links of appropriate length. We employed hierarchical clustering using the standard R function

hclust[153]. The number of clusters were defined using the Dynamic Tree Cut package in R[154]. We performed principal component decomposition on each cluster to represent the cluster by its first principal component or the “eigen-gene” representing the set of genes in that cluster [57]. The eigen-gene represent a data vector of values that summarizes the gene expression values within a given cluster for a given participant.

Network Analyses:

We further investigated clusters assigned through hierarchical clustering were as individual networks/ subnetworks. The top cluster showing most significant association with amnesic MCI diagnosis was identified ($p_{\text{fdr}} < 0.0001$). Network visualization tools were used to visualize the clusters and the backbones. The hub nodes were identified using degree centrality measure. Degree centrality assigns an importance score based on the number of links held by each node. In gene co-expression networks, the highest degree nodes or hubs can be useful regulators or drivers of their specific biological networks. We interpreted these hub nodes and backbone edges within the context of biological processes and relevance to AD pathology. We analyzed the biological processes associated with the cluster using the topGO package in R applying the weight01 algorithm[58]. The weight01 algorithm in the topGO package is a combination of weight and elim algorithms. The elim method investigates the nodes in the GO graph bottom-up and the weight method uses significance scores of connected nodes to compare and detect the locally most significant terms in the GO graph[58].

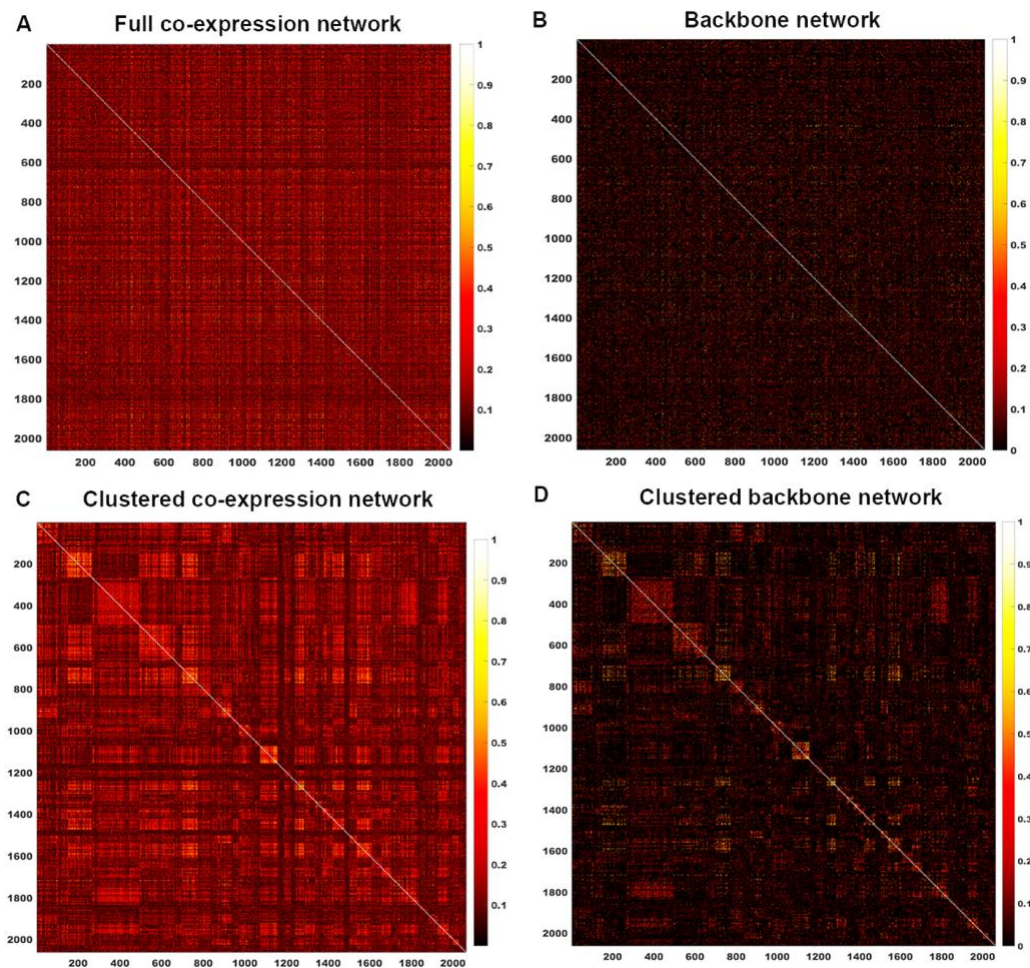
Univariate surface mapping

An average surface was constructed by computing and then averaging the Talairach coordinates at each vertex for each subject [50] using FreeSurfer6.0. Vertex-wise regressions for each of the hub gene transcripts was performed using age, gender, and education as covariates to map the association of the gene expression value with average cortical thickness using a MATLAB toolbox SurfStat [51].

Results:

We found 2062 genes differentially expressed between the diagnostic groups from 24,500+ transcripts ($p < 0.05$). The adjacency matrix was built for these differentially expressed genes, followed by network backbone construction using the shortest path method. Hierarchical clustering with dynamic tree cut algorithm gave us 42 clusters of genes. We can visualize the heatmap of the co-expression profile with the full network and network backbone in Figure 12. The backbone (Figure 12B) is much cleaner and has less noise, defined as redundant network edges compared to the original full network (Figure 12A). The heatmaps are organized by cluster assignments and there is a visible separation of clear clusters with significantly less noisy, potentially non-specific clusters in the backbone compared to the original network.

Figure 12: Heatmap representation of co-expression networks. A) Full co-expression network without clustering. B) Backbone network without clustering. C) Full co-expression network, ordered by cluster. D) Backbone network ordered by cluster.



Clusters with significance to amnesic MCI phenotype

To find most disease specific clusters, we used the eigengene method to find the cluster with the strongest association with aMCI phenotype. Of the 42 clusters, one cluster, cluster 37 had the strongest correlation to aMCI phenotype post FDR correction ($r=0.45$, $p_{\text{fdr}} < 0.0001$). There are 46 genes in this cluster (Table 15).

Table 15: Genes in cluster 37

<i>MANEAL</i>	<i>LOC390748</i>
<i>ADRA1A</i>	<i>LAMB3</i>
<i>LMX1A</i>	<i>GABRA4</i>
<i>PPFIBP2</i>	<i>ACSL5</i>
<i>SYNE2</i>	<i>CERKL</i>
<i>GPR133</i>	<i>FNBP1L</i>
<i>CCDC106</i>	<i>FBXW8</i>
<i>CHI3L2</i>	<i>C1orf141</i>
<i>NAT5</i>	<i>LOC439985</i>
<i>NEGR1</i>	<i>C21orf129</i>
<i>TSPAN11</i>	<i>COCH</i>
<i>C4orf6</i>	<i>HIST1H2AH</i>
<i>RAG2</i>	<i>NTS</i>
<i>ETV3L</i>	<i>ERCC4</i>
<i>CALCR</i>	<i>OR51F1</i>
<i>SNTG1</i>	<i>SPRR2E</i>
<i>PAX2</i>	<i>MAEL</i>
<i>NLRP13</i>	<i>NETO1</i>
<i>IPF1</i>	<i>STEAP2</i>
<i>OR2G6</i>	<i>ARHGAP4</i>
<i>FOXA1</i>	<i>SLN</i>
<i>PCDHB13</i>	<i>KAT2A</i>
<i>OR13A1</i>	<i>CNGA1</i>

Network Analyses on cluster 37

Using the metric backbone method, the number of network edges in cluster 37 were pruned from 1035 edges to 210 edges. The co-expression networks for cluster 37 can be visualized in Figure 13. The overlaying backbone represents all nodes in the network with a significantly lower number of edges. Due to cluster significance, there are potential gene-gene interactions that are involved in the disease pathology or can represent specific systemic changes occurring at a group level in our cohort, who represent early stages of AD. There were 10 hub nodes identified based on degree centrality, *CERKL*, *NLRP13*, *OR13A1*, *SNTG1*, *ERCC4*, *C4orf16(AP1AR)*, *HIST1H2AH*, *RAG2*, *CCDC106* and *C21orf129* (Figure 14). Gene ontology analyses using the degree centrality measures as scores gave us enriched biological processes associated with the cluster. The top significant nodes identified through topGO analyses were GO:0007568, biological processes associated with aging, GO:0001843, biological processes associated with neural tube closure, GO:0007608, biological processes associated with sensory perception of smell, GO:0006310, biological processes associated with DNA recombination” and GO:0050906, biological processes associated with Detection of chemical stimulus involved in sensory perception.

Figure 13: Network representation of cluster 37. A) Full network with edges represented as blue dashed lines. B) Full network with backbone, full network edges represented in dashed blue lines and backbone edges as solid red line. C- D) Magnified view of A and B, resp., to visualize edges of the network.

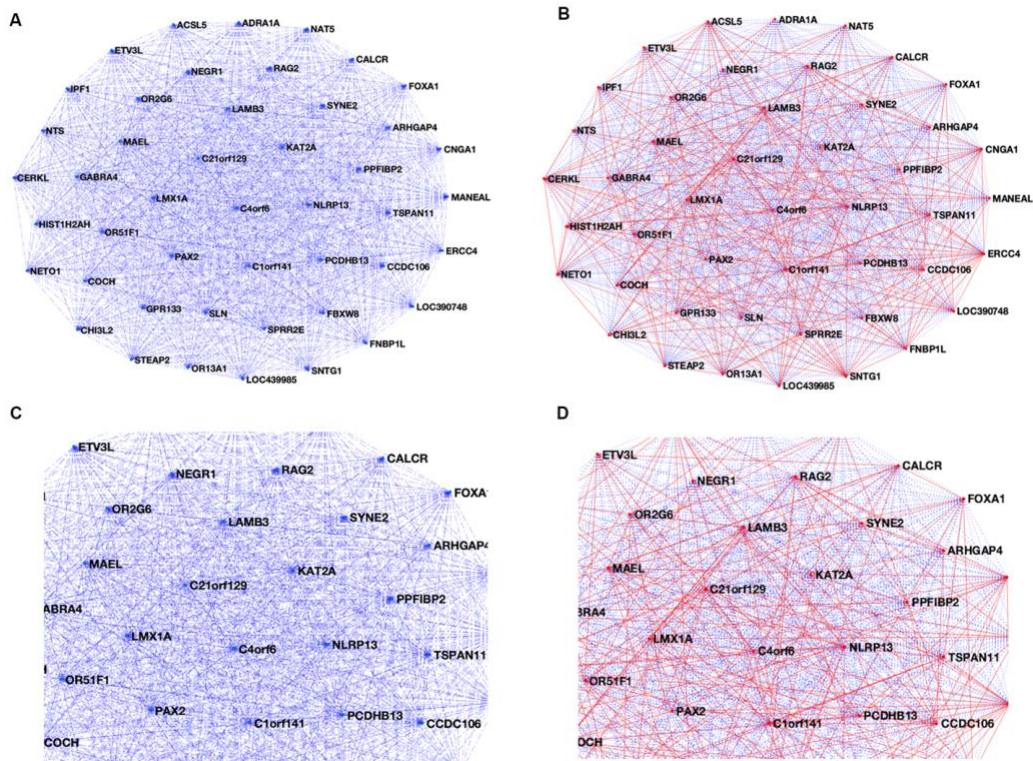
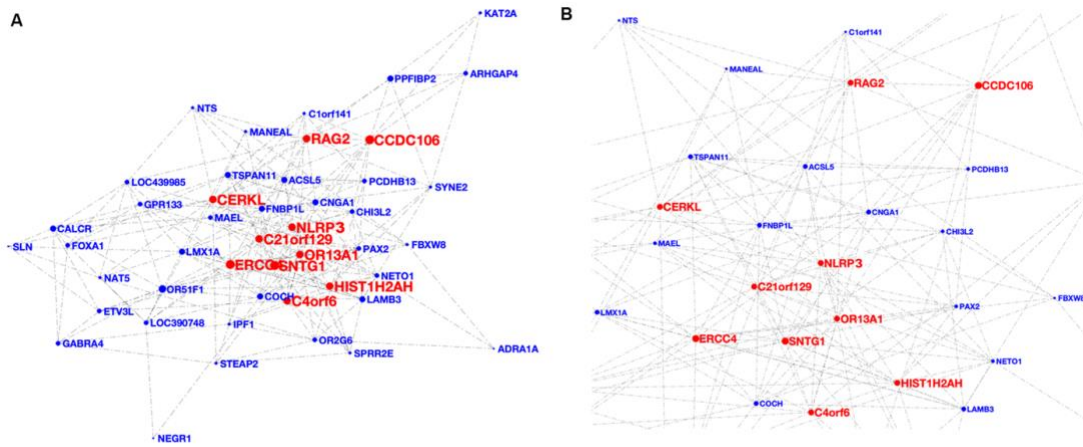


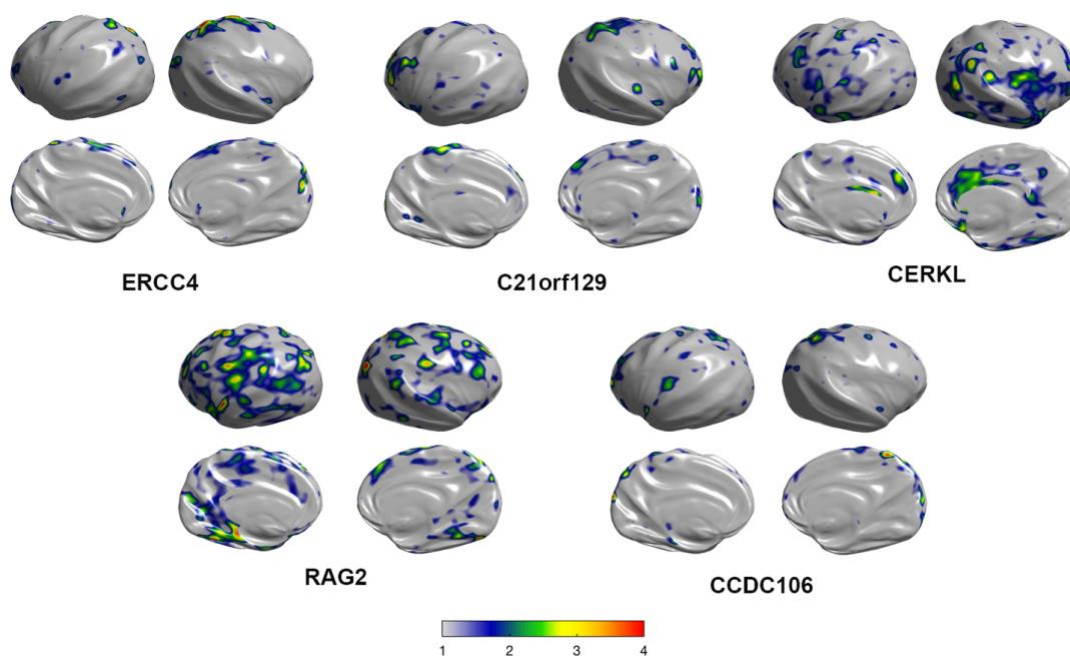
Figure 14: Cluster 37 network representation (blue) with hub nodes in red



Hub genes associated with cortical atrophy:

5 of the 10 hub nodes identified showed significant patterns of association with cortical atrophy (Figure 15). *ERCC4* and *CCDC106* had a positive association with cortical thickness in the parietal lobe. *C21orf129* showed positive association in parietal and frontal regions. *CERKL* showed negative association with regions in the temporal and frontal lobes. *RAG2* showed positive association in temporal, inferior temporal, and occipital regions.

Figure 15: Association of hub genes with cortical thickness (T-maps) ($T > 1.96$, $p < 0.05$)



Discussion:

In the present study, we performed a data driven comprehensive analyses of co-expression networks in mild cognitive impairment, a critical prodromal stage of AD using blood-based transcriptomic data. We constructed a gene co-expression network using differentially expressed genes between diagnostic groups (NC, aMCI, and naMCI), capturing important transcriptomic signatures associated with the disease. We used a novel network backbone construction method, to eliminate redundancy and identify salient features of the network nodes or genes and specifically investigate gene-gene interactions or hub genes that are potentially

important genetic markers in AD pathophysiology. The co-expression network was clustered into 42 clusters or modules, and we further investigated the cluster with highest correlation to disease status (aMCI) containing 46 genes. With the backbone method, we were able to reduce the number of redundant edges in co-expression network significantly and visualize gene-gene associations for the cluster. Further, using degree centrality measures on the cluster backbone, we were able to identify 10 important hub genes or nodes that are potentially important genes driving pathogenic systems within the network. Previous studies have shown that intra-modular hubs, i.e., hubs central to specific modules/clusters in the network, like the ones we identified, are more relevant functionally in biological networks as opposed to inter-modular hubs, which are central to the entire network[155, 156]. We found that 5 of these hub genes also had significant associations with cortical atrophy in our cohort, in commonly affected neurodegenerative regions of the brain[157]. These are potentially promising imaging-genetic signatures which can be used for tracking disease progression and investigated further for therapeutical developments in preventing or controlling neurodegeneration.

An important outcome of the analyses is the identification of these reduced critical mRNA hub measures from a large and inherently noisy transcriptomic dataset and novel gene-gene interactions in backbone metric of the differential co-expression network module. *CCDC106* is a gene which promotes the degradation of p53/TP53 protein and inhibits its activity[158]. P53 is a transcription factor controlling many important cellular pathways including apoptosis. Higher levels of

p53 are associated with AD and other neurodegenerative diseases in both animal models and human samples[159-161]. *C4orf6* and *C21orf129* are affiliated with long non-coding RNA. *C4orf6* has been associated with neuroblastoma and has not been implicated in AD pathogenesis yet[162]. The *RAG2* gene is involved in formation of the RAG complex, which is essential for the activity of B cells and T cells, involved in innate and adaptive immunity. Reduced β -amyloid pathology was observed in *RAG2* knock out mice[163]. *SNTG1*- a member of the syntrophin family, is involved in enhancing neurotrophic activity and has been associated with idiopathic scoliosis but not yet with AD pathogenesis[164]. *NLRP3* codes for a protein that forms the NLRP3 inflammasome, a type of cytosolic multiprotein complex, which plays a crucial role in innate immunity. Experimental evidence has shown that the activation of NLRP3 inflammasome is closely related to neurodegenerative diseases[165, 166]. Due to A β plaques and tau aggregates, microglia and astrocytes mediate chronic neuroinflammatory response, neuronal death and pyroptosis[167] through intracellular NLRP3 inflammasome, thereby driving the progression of AD[168, 169]. Also, pharmacological inhibition of NLRP3 inflammasome has shown to exhibit neuroprotective effects[170]. *OR13A1* is an olfactory receptor which initiates the neuronal response that triggers the perception of smell. Olfactory dysfunction in AD is associated with pathological changes of tau protein in the olfactory bulb and olfactory projection area[171, 172]. A study monitoring predictive value of olfactory recognition and function in AD showed that 47% of MCI patients with olfactory impairment yet only 11% of MCI patients with a normal sense of smell eventually developed AD[173], making olfactory receptors

promising biomarker candidates for risk assessment and early diagnosis. *CERKL* - a gene associated with retinitis pigmentosa that has been shown to protect neuronal cells from apoptosis in oxidative stress conditions[174], is an interesting candidate to prevent neurodegeneration and promote cell survival. *ERCC4* is gene involved in DNA repair through the nucleotide excision repair mechanism[175]. Oxidative DNA damage and age-related decline in DNA repair are associated with progression of AD[176, 177]. *HIST1H2AH* is a gene which codes for a core component of a nucleosome. Nucleosomes are structural units crucial for gene expression and regulation and are altered by epigenetic modifications. Epigenetic regulation is important in learning and memory, mediating critical mechanisms of neuroplasticity[178]. These hub genes and their function shed light on important pathogenic mechanisms during disease progression. Due to them having the greatest number of connections with other genes in the network, they are likely driving or at least indicative of expression/regulation of larger systemic processes dysregulated in AD like DNA damage and/or impaired DNA repair due to oxidative stress and a downstream cascade of interconnected processes such as immune response, neuroinflammation and apoptosis. They warrant further investigation in in vivo/in vitro experiments and as potential biomarker candidates.

There have been other studies that have looked at gene co-expression networks in AD. These studies, focused on gene co-expression networks from brain tissue, have found novel genes and dysregulation of biological processes like immune function in AD relative to normal controls samples[179, 180]. Another study looked at gene co-expression networks from normal, MCI and AD stages

and found novel genes and miRNAs related to AD[149]. Our analyses add to this knowledge base by providing candidate genes identified from peripheral blood gene expression and looking into specifically, the prodromal stage of AD. Additionally, the network backbone method employed in our analyses has helped identify critical measures within the network, reducing noise and redundancy. There are limitations to the study that should be considered. Firstly, there are sensitive A β and p-tau plasma assays which might have more clinical utility than transcriptomic data from blood as biomarkers, but the study of gene co-expression data adds better understanding of critical biological pathways and provides promising targets for therapeutic discovery. Secondly, genes which are co-expressed are not always functionally related[181], but employing measures to reduce noise and incorporating disease specific signals when modeling networks is more advantageous in finding functionally related gene systems.

In summary, we were able to identify novel genes related to the prodromal stage of AD using the backbone metric method applied to co-expression networks. These can be useful in building risk assessment platforms and screening targets for novel therapeutics.

Summary

Overall, the work presented in this thesis focused on analyzing transcriptomic data from peripheral blood from the ImaGene study (which consists of 108 MCI and 52 cognitively normal controls) and the ADNI study, a nationwide multisite study consisting of thousands of participants with genetic, clinical, imaging and fluid biomarker data. As mentioned previously, the multifactorial nature of AD demands integration of multiple layers of clinical, genetic, and imaging data, to build more robust risk assessment platforms and to inform novel therapeutics.

In addition, data-driven approaches may also aid in the precision medicine paradigm in AD as well as help in diagnostic biomarker stratification in clinical trials of AD based on pathogenic profiles and AD progression as defined by combinations of gene expression signatures. With this overall objective, we used neuroimaging endophenotypes, clinical disease status, and genotype data to identify novel biomarkers; the transcriptomic data were the primary outcome variables. Given that blood transcriptomic data is inherently noisy and dynamic, finding robust, disease-specific signals across the entire transcriptome is difficult. To attempt to boost signal detection, we employed novel data reduction methods to detect disease-sensitive genetic signatures.

In chapter 1, we integrated MRI and peripheral blood-based gene expression data using persistent homology followed by kernel-based clustering to characterize gene-expression patterns at the MCI stage to identify critical mRNA measures and gene clusters associated with cortical atrophy, disease status and AD pathogenesis. We identified three clusters of genes significantly associated with

diagnosis of aMCI. The biological processes associated with each cluster were mitochondrial function, NF-kappaB 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, and variant analyses found two genes containing variants, associated with disease. We found a disease-relevant transcriptomic cluster signature sensitive to prodromal AD and cortical atrophy and identified a subset of potential therapeutic targets associated with AD pathogenesis. An important future direction of this work is the validation of persistent homology-kernel clustering pipeline in external larger datasets, to be used as a predictive modelling tool when assessing risk or screening for therapeutics. Additionally, the driver genes and their associated pathways need to be investigated in experimental set ups to evaluate utility as therapeutic targets or biomarker measures.

In chapter 2, we evaluated the efficacy of characterizing and clustering peripheral blood transcriptomic data with respect to brain amyloidosis. 356 ADNI participants with blood gene expression, Florbetapir SUVR, neurodegeneration measures and CSF measures (CN=120, EMCI=130, LMCI=72 and AD=34) were identified. Differentially expressed genes from ~50,000 transcripts ($p < 0.001$) were identified. Pairwise euclidean distance between genes was used to construct a gene-gene distance matrix followed by K-means clustering. Gene enrichment analysis was performed on each cluster to identify biological processes associated with the cluster. “Eigengene”; a data vector representing the whole cluster was

obtained for each cluster using Principal Component Analyses. The top driver genes were identified from each cluster. We used the tool Multi-marker Analysis of GenoMic Annotation (MAGMA) to analyze variants in the driver genes. Voxel-wise multiple regression in SPM12 with age and sex as covariates was used to visualize the pattern of association of driver gene expression with brain amyloidosis. The association of the transcripts identified through our analyses to amyloid SUVR, and disease diagnosis were validated in the external ImaGene dataset. We identified five clusters with distinct biological processes highly relevant in AD. All five clusters were significantly associated with Florbetapir SUVR and CSF A β measures ($p < 0.05$). 28 driver genes were identified (Table 3). Six driver genes (*BAIAP3*, *E2F2*, *PSMF1*, *SMOX*, *UBE20* and *RNF11*) had negative association with brain amyloidosis in the lateral temporal, lateral parietal, and temporo- and parieto-occipital areas, in addition to regions of the frontal lobe. One of the driver genes, *KANK2*, and overall, 29 of the differentially expressed genes had SNPs significantly associated with the AD and MCI phenotypes. The top driver genes showed similar correlation to amyloid SUVR in the ImaGene dataset. A logistic regression model with 28 driver genes, and age and sex as covariates predicted aMCI diagnosis in the external dataset with an AUC of 0.82. Using a data driven approach we identified novel gene targets from peripheral blood which directly correlate to amyloid-related pathogenesis and AD phenotype. An important future direction of this work is investigation of the genes for therapeutic discovery and potential utility in repurposing drugs that target these biological pathways to alleviate symptoms. Their direct association with amyloidosis also makes them

useful candidates in building predictive models for pre symptomatic risk and tracking AD progression.

In chapter 3, we evaluated peripheral blood transcriptomic data as gene co-expression networks constructed from differentially expressed genes between normal controls and MCI subjects to identify co-expressed modules associated with prodromal AD. Data from the ImaGene study was used. This included 160 subjects clinically diagnosed with amnesic MCI (n=70), non-amnesic MCI (n=38) and normal controls (n=52). Quantile-normalized and log-transformed mRNA levels were obtained from peripheral blood. Preliminary analysis was performed by selecting 2062 genes which were significantly differentially expressed between normal controls and MCI (two sample t-test, $p_{\text{fdr}} < 0.05$). Gene co-expression networks were built by creating correlation matrices of the differentially expressed mRNA values. Network backbones were then built using the shortest path computation to remove redundancy and identify important edges. Hierarchical clustering was performed on the co-expression matrix to identify important communities or clusters within the network. We identified the strongest cluster associated with aMCI phenotype ($p_{\text{fdr}} < 0.0001$) by using the first principal component of the cluster as the eigen gene. The genes within the cluster and edges within the backbone were investigated. Upon backbone computation, the number of edges compared to original correlation matrix reduced significantly thereby reducing noise. Hierarchical clustering yielded 42 clusters of genes. One cluster (eigen gene) had the strongest association with aMCI phenotype ($p < 0.0001$) post FDR correction ($r = 0.45$). This cluster consisting of 46

genes was further investigated in terms of biological processes and interactions with other nodes in the cluster backbone. 10 hub genes were found in the cluster based on degree centrality measures, which had highly significant association with aMCI. This approach can also help build models for risk assessment, gene therapy and help identify novel therapeutic targets/ dysregulated pathways. An important future direction of this analysis is looking into gene ontology networks and protein-protein networks that associate with the identified backbone modules, to further learn about the biological systems, which are potentially capturing important information about pathogenic processes and drivers. Additionally, another direction is to investigate expression Quantitative Trait Loci (eQTL) networks associated with gene co-expression network and build bipartite eQTL-gene expression networks. After computation of the network backbone, we can hypothesize finding important communities of SNP-gene pairs associated with AD risk and pathogenesis.

In summary, the work presented in this dissertation has contributed to identification of transcriptomic biomarkers associated with baseline diagnosis of MCI and future conversion to AD dementia and improved our understanding of critical disease-related pathways and system-level changes that occur in prodromal AD using blood-based biomarkers. At a systems level, the common biological processes which were identified across the different gene signatures seen in the three chapters were mitochondrial dysfunction, innate immune response, NF- κ B signaling, apoptosis, oxidative stress response, neuroinflammation and dysfunctional proteostasis. These biological processes

and pathways have also been previously reported to be associated with Alzheimer's Disease pathology, but we have successfully contributed to the body of this knowledge and provided interesting gene targets to be further explored for therapeutics. This work has also provided key insight into pathogenesis at the early stages of AD and improved our understanding of systemic processes that drive progression towards dementia. Through an integrated, systematic evaluation of transcriptomic data and neuroimaging endophenotypes, this work contributes towards pre-symptomatic risk assessment, precision medicine and biomarker development in AD while potentially identifying key gene targets and pathways for therapeutic intervention.

References

1. Ausó, E., V. Gómez-Vicente, and G. Esquivá, *Biomarkers for Alzheimer's Disease Early Diagnosis*. J Pers Med, 2020. **10**(3).
2. Petersen, R.C., et al., *Alzheimer's Disease Neuroimaging Initiative (ADNI): clinical characterization*. Neurology, 2010. **74**(3): p. 201-209.
3. Lanoiselée, H.M., et al., *APP, PSEN1, and PSEN2 mutations in early-onset Alzheimer disease: A genetic screening study of familial and sporadic cases*. PLoS Med, 2017. **14**(3): p. e1002270.
4. Lambert, J.C., et al., *Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease*. Nat Genet, 2013. **45**(12): p. 1452-8.
5. Jansen, I.E., et al., *Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk*. Nat Genet, 2019. **51**(3): p. 404-413.
6. Giau, V.V., et al., *Genetic analyses of early-onset Alzheimer's disease using next generation sequencing*. Scientific Reports, 2019. **9**(1): p. 8368.
7. Van Cauwenberghe, C., C. Van Broeckhoven, and K. Sleegers, *The genetic landscape of Alzheimer disease: clinical implications and perspectives*. Genetics in Medicine, 2016. **18**(5): p. 421-430.
8. Scherzer, C.R., 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): p. 955-60.

9. Sullivan, P.F., C. Fan, and C.M. Perou, *Evaluating the comparability of gene expression in blood and brain*. American Journal of Medical Genetics Part B: Neuropsychiatric Genetics, 2006. **141B**(3): p. 261-268.
10. Mapstone, M., et al., *Plasma phospholipids identify antecedent memory impairment in older adults*. Nat Med, 2014. **20**(4): p. 415-8.
11. Hye, A., et al., *Plasma proteins predict conversion to dementia from prodromal disease*. Alzheimers Dement, 2014. **10**(6): p. 799-807.e2.
12. Zetterberg, H., et al., *Plasma tau levels in Alzheimer's disease*. Alzheimer's Research & Therapy, 2013. **5**(2): p. 9.
13. Olsson, B., et al., *CSF and blood biomarkers for the diagnosis of Alzheimer's disease: a systematic review and meta-analysis*. The Lancet Neurology, 2016. **15**(7): p. 673-684.
14. Blennow, K., et al., *Cerebrospinal fluid and plasma biomarkers in Alzheimer disease*. Nat Rev Neurol, 2010. **6**(3): p. 131-44.
15. Zetterberg, H., *Review: Tau in biofluids - relation to pathology, imaging and clinical features*. Neuropathol Appl Neurobiol, 2017. **43**(3): p. 194-199.
16. DeMarshall, C.A., et al., *Detection of Alzheimer's disease at mild cognitive impairment and disease progression using autoantibodies as blood-based biomarkers*. Alzheimers Dement (Amst), 2016. **3**: p. 51-62.
17. Liew, C.C., et al., *The peripheral blood transcriptome dynamically reflects system wide biology: a potential diagnostic tool*. J Lab Clin Med, 2006. **147**(3): p. 126-32.

18. Li, X., et al., *Integrated genomic approaches identify major pathways and upstream regulators in late onset Alzheimer's disease*. Sci Rep, 2015. **5**: p. 12393.
19. Ciryam, P., et al., *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): p. 4753-8.
20. Mills, J.D., et al., *RNA-Seq analysis of the parietal cortex in Alzheimer's disease reveals alternatively spliced isoforms related to lipid metabolism*. Neurosci Lett, 2013. **536**: p. 90-5.
21. Miller, J.A., M.C. Oldham, and D.H. Geschwind, *A Systems Level Analysis of Transcriptional Changes in Alzheimer's Disease and Normal Aging*. The Journal of Neuroscience, 2008. **28**(6): p. 1410-1420.
22. Liang, J.W., et al., *Application of Weighted Gene Co-Expression Network Analysis to Explore the Key Genes in Alzheimer's Disease*. J Alzheimers Dis, 2018. **65**(4): p. 1353-1364.
23. Platig, J., et al., *Bipartite Community Structure of eQTLs*. PLOS Computational Biology, 2016. **12**(9): p. e1005033.
24. Fagny, M., et al., *Exploring regulation in tissues with eQTL networks*. Proc Natl Acad Sci U S A, 2017. **114**(37): p. E7841-e7850.
25. Zhang, B., et al., *Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer's disease*. Cell, 2013. **153**(3): p. 707-20.

26. Pottier, C., et al., *TYROBP genetic variants in early-onset Alzheimer's disease*. *Neurobiol Aging*, 2016. **48**: p. 222.e9-222.e15.
27. Haure-Mirande, J.-V., et al., *Microglial TYROBP/DAP12 in Alzheimer's disease: Transduction of physiological and pathological signals across TREM2*. *Molecular Neurodegeneration*, 2022. **17**(1): p. 55.
28. Li, X., et al., *Integrated genomic approaches identify major pathways and upstream regulators in late onset Alzheimer's disease*. *Scientific Reports*, 2015. **5**(1): p. 12393.
29. De Farias, A.-R.M., et al., *Role of the late-onset Alzheimer's disease risk genes bin1 and ptk2b in the hyperexcitability of hiPSC-derived neurons*. *Alzheimer's & Dementia*, 2021. **17**(S3): p. e053632.
30. Henriksen, K., et al., *The future of blood-based biomarkers for Alzheimer's disease*. *Alzheimers Dement*, 2014. **10**(1): p. 115-31.
31. Lu, D., et al., *Multimodal and Multiscale Deep Neural Networks for the Early Diagnosis of Alzheimer's Disease using structural MR and FDG-PET images*. *Scientific Reports*, 2018. **8**(1): p. 5697.
32. Khan, M.J., et al., *Modern Trends in Hyperspectral Image Analysis: A Review*. *IEEE Access*, 2018. **6**: p. 14118-14129.
33. Huang, W.-Q., et al., *Integrated analysis of microRNA and mRNA expression profiling identifies BAIAP3 as a novel target of dysregulated hsa-miR-1972 in age-related white matter lesions*. *Aging*, 2021. **13**(3): p. 4674-4695.

34. Hu, Y., et al., *Identification of Alzheimer's Disease-Related Genes Based on Data Integration Method*. *Frontiers in Genetics*, 2019. **9**.
35. Jack, C.R., Jr., et al., *NIA-AA Research Framework: Toward a biological definition of Alzheimer's disease*. *Alzheimers Dement*, 2018. **14**(4): p. 535-562.
36. Liu, W., et al., *MRI-based automated volumetric segmentation tool in the detection of early Alzheimer's disease*. *Alzheimer's & Dementia*, 2020. **16**(S5): p. e042340.
37. Uylings, H.B.M. and J.M. de Brabander, *Neuronal Changes in Normal Human Aging and Alzheimer's Disease*. *Brain and Cognition*, 2002. **49**(3): p. 268-276.
38. Jack, C.R., et al., *Rates of hippocampal atrophy correlate with change in clinical status in aging and AD*. *Neurology*, 2000. **55**(4): p. 484-490.
39. Raz, N., et al., *Regional Brain Changes in Aging Healthy Adults: General Trends, Individual Differences and Modifiers*. *Cerebral Cortex*, 2005. **15**(11): p. 1676-1689.
40. Rodrigue, K.M. and N. Raz, *Shrinkage of the Entorhinal Cortex over Five Years Predicts Memory Performance in Healthy Adults*. *The Journal of Neuroscience*, 2004. **24**(4): p. 956-963.
41. Dickerson, B.C., et al., *The Cortical Signature of Alzheimer's Disease: Regionally Specific Cortical Thinning Relates to Symptom Severity in Very*

- Mild to Mild AD Dementia and is Detectable in Asymptomatic Amyloid-Positive Individuals*. Cerebral Cortex, 2008. **19**(3): p. 497-510.
42. Ranganath, C. and M. Ritchey, *Two cortical systems for memory-guided behaviour*. Nature Reviews Neuroscience, 2012. **13**(10): p. 713-726.
43. Maass, A., et al., *Comparison of multiple tau-PET measures as biomarkers in aging and Alzheimer's disease*. NeuroImage, 2017. **157**: p. 448-463.
44. Braak, H. and E. Braak, *Frequency of Stages of Alzheimer-Related Lesions in Different Age Categories*. Neurobiology of Aging, 1997. **18**(4): p. 351-357.
45. Jansen, I.E., et al., *Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk*. Nature Genetics, 2019. **51**(3): p. 404-413.
46. Leandro, G.S., et al., *Changes in Expression Profiles Revealed by Transcriptomic Analysis in Peripheral Blood Mononuclear Cells of Alzheimer's Disease Patients*. J Alzheimers Dis, 2018. **66**(4): p. 1483-1495.
47. Rye, P.D., et al., *A novel blood test for the early detection of Alzheimer's disease*. J Alzheimers Dis, 2011. **23**(1): p. 121-9.
48. Booij, B.B., et al., *A gene expression pattern in blood for the early detection of Alzheimer's disease*. J Alzheimers Dis, 2010. **23**(1): p. 109-19.
49. Reuter, M., et al., *Within-subject template estimation for unbiased longitudinal image analysis*. NeuroImage, 2012. **61**(4): p. 1402-1418.

50. Wilkin Chau, A.R.M., *The Talairach coordinate of a point in the MNI space: how to interpret it*. NeuroImage, 2005. **25**(2): p. 408-416.
51. Chung, M.K., et al., *General multivariate linear modeling of surface shapes using SurfStat*. Neuroimage, 2010. **53**(2): p. 491-505.
52. Edelsbrunner, Letscher, and Zomorodian, *Topological Persistence and Simplification*. Discrete & Computational Geometry, 2002. **28**(4): p. 511-533.
53. Craw, S., *Manhattan Distance*, in *Encyclopedia of Machine Learning and Data Mining*, C. Sammut and G.I. Webb, Editors. 2017, Springer US: Boston, MA. p. 790-791.
54. Gönen, M., & Alpaydin, E. , *Multiple kernel learning algorithms*. Journal of Machine Learning Research, 2011. **12**: p. 2211-2268.
55. Reininghaus J, H.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: p. 4741–4748.
56. Zeileis, A.K.a.A.S.a.K.H.a.A., *kernlab - An S4 Package for Kernel Methods in R*. Journal of Statistical Software, Articles, 2004. **11**(9): p. 1-20.
57. Langfelder, P. and S. Horvath, *Eigengene networks for studying the relationships between co-expression modules*. BMC Syst Biol, 2007. **1**: p. 54.
58. Alexa A, R.J. *topGO: Enrichment Analysis for Gene Ontology. R package version 2.40.0*. 2020.

59. Supek, F., et al., *REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms*. PLOS ONE, 2011. **6**(7): p. e21800.
60. de Leeuw, C.A., et al., *MAGMA: Generalized Gene-Set Analysis of GWAS Data*. PLOS Computational Biology, 2015. **11**(4): p. e1004219.
61. Saykin, A.J., et al., *Alzheimer's Disease Neuroimaging Initiative biomarkers as quantitative phenotypes: Genetics core aims, progress, and plans*. Alzheimers Dement, 2010. **6**(3): p. 265-73.
62. Thurfjell, L., 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): p. 1623-8.
63. Du, A.T., et al., *Different regional patterns of cortical thinning in Alzheimer's disease and frontotemporal dementia*. Brain, 2007. **130**(Pt 4): p. 1159-66.
64. Purcell, S., et al., *PLINK: a tool set for whole-genome association and population-based linkage analyses*. Am J Hum Genet, 2007. **81**(3): p. 559-75.
65. Yates, T.M., et al., *ZMYND11-related syndromic intellectual disability: 16 patients delineating and expanding the phenotypic spectrum*. Human Mutation, 2020. **41**(5): p. 1042-1050.
66. Meyer-Hoffert, U., et al., *Isolation of SPINK6 in human skin: selective inhibitor of kallikrein-related peptidases*. The Journal of biological chemistry, 2010. **285**(42): p. 32174-32181.

67. von Bernhardt, R., et al., *Role of TGF β signaling in the pathogenesis of Alzheimer's disease*. *Frontiers in cellular neuroscience*, 2015. **9**: p. 426-426.
68. Henderson, B.W., et al., *Rho-associated protein kinase 1 (ROCK1) is increased in Alzheimer's disease and ROCK1 depletion reduces amyloid- β levels in brain*. *Journal of neurochemistry*, 2016. **138**(4): p. 525-531.
69. Du, X., et al., *SMAD4 feedback regulates the canonical TGF- β signaling pathway to control granulosa cell apoptosis*. *Cell Death & Disease*, 2018. **9**(2): p. 151.
70. Caccamo, A., 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): p. 14042-56.
71. Tian, J., et al., *Lower Expression of Ndfip1 Is Associated With Alzheimer Disease Pathogenesis Through Decreasing DMT1 Degradation and Increasing Iron Influx*. *Frontiers in aging neuroscience*, 2018. **10**: p. 165-165.
72. Calvo-Rodriguez, M., et al., *Role of Toll Like Receptor 4 in Alzheimer's Disease*. *Frontiers in Immunology*, 2020. **11**(1588).
73. Cho, E. and M. Park, *Palmitoylation in Alzheimer's disease and other neurodegenerative diseases*. *Pharmacological Research*, 2016. **111**: p. 133-151.

74. Boehm, E., et al., *Role of FAST Kinase Domains 3 (FASTKD3) in Post-transcriptional Regulation of Mitochondrial Gene Expression*. The Journal of biological chemistry, 2016. **291**(50): p. 25877-25887.
75. Simarro, M., et al., *Fast kinase domain-containing protein 3 is a mitochondrial protein essential for cellular respiration*. Biochemical and biophysical research communications, 2010. **401**: p. 440-6.
76. Ramanan, V.K., et al., *FASTKD2 is associated with memory and hippocampal structure in older adults*. Molecular psychiatry, 2015. **20**(10): p. 1197-1204.
77. Juraszek, B. and K.A. Nałęcz, *SLC22A5 (OCTN2) Carnitine Transporter-Indispensable for Cell Metabolism, a Jekyll and Hyde of Human Cancer*. Molecules (Basel, Switzerland), 2019. **25**(1): p. 14.
78. Ferreira, G.C. and M.C. McKenna, *I-Carnitine and Acetyl-I-carnitine Roles and Neuroprotection in Developing Brain*. Neurochemical Research, 2017. **42**(6): p. 1661-1675.
79. Serrat, R., et al., *The Armc10/SVH gene: genome context, regulation of mitochondrial dynamics and protection against A β -induced mitochondrial fragmentation*. Cell death & disease, 2014. **5**(4): p. e1163-e1163.
80. Sobajima, T., et al., *The Rab11-binding protein RELCH/KIAA1468 controls intracellular cholesterol distribution*. Journal of Cell Biology, 2018. **217**(5): p. 1777-1796.

81. Casanova, R., et al., *Blood metabolite markers of preclinical Alzheimer's disease in two longitudinally followed cohorts of older individuals*. *Alzheimer's & dementia : the journal of the Alzheimer's Association*, 2016. **12**(7): p. 815-822.
82. Lee, T. and H. Lee, *Prediction of Alzheimer's disease using blood gene expression data*. *Scientific Reports*, 2020. **10**(1): p. 3485.
83. Hadar, A. and D. Gurwitz, *Peripheral transcriptomic biomarkers for early detection of sporadic Alzheimer disease?* *Dialogues in clinical neuroscience*, 2018. **20**(4): p. 293-300.
84. Su, L., et al., *Meta-Analysis of Gene Expression and Identification of Biological Regulatory Mechanisms in Alzheimer's Disease*. *Frontiers in Neuroscience*, 2019. **13**(633).
85. Schindler, S.E., et al., *High-precision plasma β -amyloid 42/40 predicts current and future brain amyloidosis*. *Neurology*, 2019. **93**(17): p. e1647.
86. Yang, C.C., et al., *Assay of Plasma Phosphorylated Tau Protein (Threonine 181) and Total Tau Protein in Early-Stage Alzheimer's Disease*. *J Alzheimers Dis*, 2018. **61**(4): p. 1323-1332.
87. Grimm, M.O.W., et al., *Plasmalogen synthesis is regulated via alkyl-dihydroxyacetonephosphate-synthase by amyloid precursor protein processing and is affected in Alzheimer's disease*. *Journal of Neurochemistry*, 2011. **116**(5): p. 916-925.

88. Ogaki, K., et al., *Analysis of COQ2 gene in multiple system atrophy*. Molecular neurodegeneration, 2014. **9**: p. 44-44.
89. Arefin, A.S., et al., *Unveiling clusters of RNA transcript pairs associated with markers of Alzheimer's disease progression*. PloS one, 2012. **7**(9): p. e45535-e45535.
90. Huttunen, H.J., et al., *HtrA2 regulates beta-amyloid precursor protein (APP) metabolism through endoplasmic reticulum-associated degradation*. J Biol Chem, 2007. **282**(38): p. 28285-95.
91. Mok, S.-A., et al., *Mapping interactions with the chaperone network reveals factors that protect against tau aggregation*. Nature structural & molecular biology, 2018. **25**(5): p. 384-393.
92. Astarita, G., et al., *Deficient liver biosynthesis of docosahexaenoic acid correlates with cognitive impairment in Alzheimer's disease*. PloS one, 2010. **5**(9): p. e12538-e12538.
93. Katsel, P., C. Li, and V. Haroutunian, *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): p. 845-56.
94. Jayapalan, S., D. Subramanian, and J. Natarajan, *Computational identification and analysis of neurodegenerative disease associated protein kinases in hominid genomes*. Genes & diseases, 2016. **3**(3): p. 228-237.

95. Kim, N.-Y., et al., *Sorting nexin-4 regulates β -amyloid production by modulating β -site-activating cleavage enzyme-1*. *Alzheimer's research & therapy*, 2017. **9**(1): p. 4-4.
96. Fujiwara, H., et al., *Inhibition of microtubule assembly competent tubulin synthesis leads to accumulation of phosphorylated tau in neuronal cell bodies*. *Biochemical and Biophysical Research Communications*, 2020. **521**(3): p. 779-785.
97. Pietrzak, M., et al., *Gene expression profiling of brain samples from patients with Lewy body dementia*. *Biochemical and biophysical research communications*, 2016. **479**(4): p. 875-880.
98. Milde, S. and M.P. Coleman, *Identification of palmitoyltransferase and thioesterase enzymes that control the subcellular localization of axon survival factor nicotinamide mononucleotide adenylyltransferase 2 (NMNAT2)*. *The Journal of biological chemistry*, 2014. **289**(47): p. 32858-32870.
99. *2020 Alzheimer's disease facts and figures*. *Alzheimer's & Dementia*, 2020. **16**(3): p. 391-460.
100. Binder, L.I., et al., *Tau, tangles, and Alzheimer's disease*. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, 2005. **1739**(2): p. 216-223.
101. Ricciarelli, R. and E. Fedele, *The Amyloid Cascade Hypothesis in Alzheimer's Disease: It's Time to Change Our Mind*. *Current neuropharmacology*, 2017. **15**(6): p. 926-935.

102. Karran, E., M. Mercken, and B.D. Strooper, *The amyloid cascade hypothesis for Alzheimer's disease: an appraisal for the development of therapeutics*. Nature Reviews Drug Discovery, 2011. **10**(9): p. 698-712.
103. Forloni, G., *Alzheimer's disease: from basic science to precision medicine approach*. BMJ Neurology Open, 2020. **2**(2): p. e000079.
104. Saykin, A.J., et al., *Genetic studies of quantitative MCI and AD phenotypes in ADNI: Progress, opportunities, and plans*. Alzheimer's & dementia : the journal of the Alzheimer's Association, 2015. **11**(7): p. 792-814.
105. Landau, S.M., et al., *Measurement of longitudinal β -amyloid change with ¹⁸F-florbetapir PET and standardized uptake value ratios*. Journal of nuclear medicine : official publication, Society of Nuclear Medicine, 2015. **56**(4): p. 567-574.
106. Clark, C.M., et al., *Use of florbetapir-PET for imaging beta-amyloid pathology*. Jama, 2011. **305**(3): p. 275-83.
107. Jagust, W.J., et al., *The Alzheimer's Disease Neuroimaging Initiative 2 PET Core: 2015*. Alzheimer's & dementia : the journal of the Alzheimer's Association, 2015. **11**(7): p. 757-771.
108. Dale, A.M., B. Fischl, and M.I. Sereno, *Cortical Surface-Based Analysis: I. Segmentation and Surface Reconstruction*. NeuroImage, 1999. **9**(2): p. 179-194.

109. Desikan, R.S., et al., *An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest*. Neuroimage, 2006. **31**(3): p. 968-80.
110. Perez Ortiz, J.M. and R.H. Swerdlow, *Mitochondrial dysfunction in Alzheimer's disease: Role in pathogenesis and novel therapeutic opportunities*. Br J Pharmacol, 2019. **176**(18): p. 3489-3507.
111. Castillo, D.S., et al., *E2F1 and E2F2 induction in response to DNA damage preserves genomic stability in neuronal cells*. Cell cycle (Georgetown, Tex.), 2015. **14**(8): p. 1300-1314.
112. Sekar, S., et al., *Alzheimer's disease is associated with altered expression of genes involved in immune response and mitochondrial processes in astrocytes*. Neurobiology of aging, 2015. **36**(2): p. 583-591.
113. Wang, Y. and Z. Wang, *An Integrated Network Analysis of mRNA and Gene Expression Profiles in Parkinson's Disease*. Medical science monitor : international medical journal of experimental and clinical research, 2020. **26**: p. e920846-e920846.
114. Yao, W., et al., *The Familial dementia gene ITM2b/BRI2 facilitates glutamate transmission via both presynaptic and postsynaptic mechanisms*. Scientific Reports, 2019. **9**(1): p. 4862.
115. Milton, R.H., et al., *CLIC1 function is required for beta-amyloid-induced generation of reactive oxygen species by microglia*. J Neurosci, 2008. **28**(45): p. 11488-99.

116. Jakhmola-Mani, R., A. Islam, and D.P. Katare, *Liver-Brain Axis in Sporadic Alzheimer's Disease: Role of Ten Signature Genes in a Mouse model*. CNS Neurol Disord Drug Targets, 2020.
117. Lee, M.H., et al., *TGF- β induces TIAF1 self-aggregation via type II receptor-independent signaling that leads to generation of amyloid β plaques in Alzheimer's disease*. Cell death & disease, 2010. **1**(12): p. e110-e110.
118. Bixel, M.G., et al., *CD99 and CD99L2 act at the same site as, but independently of, PECAM-1 during leukocyte diapedesis*. Blood, 2010. **116**(7): p. 1172-84.
119. Begum, G., et al., *Inhibition of WNK3 Kinase Signaling Reduces Brain Damage and Accelerates Neurological Recovery After Stroke*. Stroke, 2015. **46**(7): p. 1956-1965.
120. Altuna, M., et al., *DNA methylation signature of human hippocampus in Alzheimer's disease is linked to neurogenesis*. Clinical Epigenetics, 2019. **11**(1): p. 91.
121. Dokter, M. and O. von Bohlen und Halbach, *Neurogenesis within the adult hippocampus under physiological conditions and in depression*. Neural Regen Res, 2012. **7**(7): p. 552-9.
122. Ming, G.L. and H. Song, *Adult neurogenesis in the mammalian brain: significant answers and significant questions*. Neuron, 2011. **70**(4): p. 687-702.

123. Drapeau, E., et al., *Spatial memory performances of aged rats in the water maze predict levels of hippocampal neurogenesis*. Proc Natl Acad Sci U S A, 2003. **100**(24): p. 14385-90.
124. Spalding, K.L., et al., *Dynamics of hippocampal neurogenesis in adult humans*. Cell, 2013. **153**(6): p. 1219-1227.
125. Mathews, K.J., et al., *Evidence for reduced neurogenesis in the aging human hippocampus despite stable stem cell markers*. Aging Cell, 2017. **16**(5): p. 1195-1199.
126. Su, L., et al., *PRDM16 orchestrates angiogenesis via neural differentiation in the developing brain*. Cell Death & Differentiation, 2020. **27**(8): p. 2313-2329.
127. Ansoleaga, B., et al., *Dysregulation of brain olfactory and taste receptors in AD, PSP and CJD, and AD-related model*. Neuroscience, 2013. **248**: p. 369-82.
128. Oliveira, J., et al., *Protein Phosphorylation is a Key Mechanism in Alzheimer's Disease*. J Alzheimers Dis, 2017. **58**(4): p. 953-978.
129. Henriques, A.G., et al., *Altered protein phosphorylation as a resource for potential AD biomarkers*. Scientific Reports, 2016. **6**(1): p. 30319.
130. Gonatopoulos-Pournatzis, T., et al., *Autism-Misregulated eIF4G Microexons Control Synaptic Translation and Higher Order Cognitive Functions*. Molecular Cell, 2020. **77**(6): p. 1176-1192.e16.

131. Quimbaya, M., et al., *Deregulation of the replisome factor MCMBP prompts oncogenesis in colorectal carcinomas through chromosomal instability*. Neoplasia (New York, N.Y.), 2014. **16**(9): p. 694-709.
132. Zhu, X., et al., *The role of mitogen-activated protein kinase pathways in Alzheimer's disease*. Neurosignals, 2002. **11**(5): p. 270-81.
133. Gee, M.S., et al., *A selective p38 α / β MAPK inhibitor alleviates neuropathology and cognitive impairment, and modulates microglia function in 5XFAD mouse*. Alzheimer's Research & Therapy, 2020. **12**(1): p. 45.
134. Kim, E.K. and E.-J. Choi, *Pathological roles of MAPK signaling pathways in human diseases*. Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease, 2010. **1802**(4): p. 396-405.
135. Daniels, R.J., et al., *Sequence, structure and pathology of the fully annotated terminal 2 Mb of the short arm of human chromosome 16*. Hum Mol Genet, 2001. **10**(4): p. 339-52.
136. Fan, J., et al., *Targeting Smox Is Neuroprotective and Ameliorates Brain Inflammation in Cerebral Ischemia/Reperfusion Rats*. Toxicological Sciences, 2018. **168**(2): p. 381-393.
137. Zhang, X., et al., *UBE2O negatively regulates TRAF6-mediated NF- κ B activation by inhibiting TRAF6 polyubiquitination*. Cell Res, 2013. **23**(3): p. 366-77.
138. Pranski, E.L., et al., *Neuronal RING finger protein 11 (RNF11) regulates canonical NF- κ B signaling*. J Neuroinflammation, 2012. **9**: p. 67.

139. Jones, S.V. and I. Kounatidis, *Nuclear Factor-Kappa B and Alzheimer Disease, Unifying Genetic and Environmental Risk Factors from Cell to Humans*. Front Immunol, 2017. **8**: p. 1805.
140. Drache, B., et al., *Bcl-xl-specific antibody labels activated microglia associated with Alzheimer's disease and other pathological states*. J Neurosci Res, 1997. **47**(1): p. 98-108.
141. Fukaya, T., et al., *Conditional ablation of CD205+ conventional dendritic cells impacts the regulation of T-cell immunity and homeostasis in vivo*. Proc Natl Acad Sci U S A, 2012. **109**(28): p. 11288-93.
142. Bonifaz, L., et al., *Efficient targeting of protein antigen to the dendritic cell receptor DEC-205 in the steady state leads to antigen presentation on major histocompatibility complex class I products and peripheral CD8+ T cell tolerance*. J Exp Med, 2002. **196**(12): p. 1627-38.
143. Bettayeb, K., et al., *Relevance of the COPI complex for Alzheimer's disease progression in vivo*. Proc Natl Acad Sci U S A, 2016. **113**(19): p. 5418-23.
144. Fontalba, A., et al., *Deficiency of CARD8 is associated with increased Alzheimer's disease risk in women*. Dement Geriatr Cogn Disord, 2008. **26**(3): p. 247-50.
145. McCaughey, J., et al., *A general role for TANGO1, encoded by MIA3, in secretory pathway organization and function*. J Cell Sci, 2021. **134**(17).

146. Lue, L.-F., A. Guerra, and D.G. Walker, *Amyloid Beta and Tau as Alzheimer's Disease Blood Biomarkers: Promise From New Technologies*. *Neurology and therapy*, 2017. **6**(Suppl 1): p. 25-36.
147. *2021 Alzheimer's disease facts and figures*. *Alzheimers Dement*, 2021. **17**(3): p. 327-406.
148. Arvanitakis, Z., R.C. Shah, and D.A. Bennett, *Diagnosis and Management of Dementia: Review*. *Jama*, 2019. **322**(16): p. 1589-1599.
149. Soleimani Zakeri, N.S., S. Pashazadeh, and H. MotieGhader, *Gene biomarker discovery at different stages of Alzheimer using gene co-expression network approach*. *Scientific Reports*, 2020. **10**(1): p. 12210.
150. Langfelder, P. and S. Horvath, *WGCNA: an R package for weighted correlation network analysis*. *BMC Bioinformatics*, 2008. **9**(1): p. 559.
151. Gates, A.J., et al., *The effective graph reveals redundancy, canalization, and control pathways in biochemical regulation and signaling*. *Proceedings of the National Academy of Sciences*, 2021. **118**(12): p. e2022598118.
152. Simas, T., R.B. Correia, and L.M. Rocha, *The distance backbone of complex networks*. *Journal of Complex Networks*, 2021. **9**(6).
153. Kaufman, L. and P.J. Rousseeuw, *Finding groups in data: an introduction to cluster analysis*. 2009: John Wiley & Sons.
154. Langfelder, P., B. Zhang, and S. Horvath, *Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R*. *Bioinformatics*, 2007. **24**(5): p. 719-720.

155. Langfelder, P., P.S. Mischel, and S. Horvath, *When Is Hub Gene Selection Better than Standard Meta-Analysis?* PLOS ONE, 2013. **8**(4): p. e61505.
156. Chen, Y., et al., *Variations in DNA elucidate molecular networks that cause disease.* Nature, 2008. **452**(7186): p. 429-35.
157. Pini, L., et al., *Brain atrophy in Alzheimer's Disease and aging.* Ageing Res Rev, 2016. **30**: p. 25-48.
158. Zhou, J., et al., *Identification and characterization of the novel protein CCDC106 that interacts with p53 and promotes its degradation.* FEBS Lett, 2010. **584**(6): p. 1085-90.
159. Kitamura, Y., et al., *Changes of p53 in the brains of patients with Alzheimer's disease.* Biochem Biophys Res Commun, 1997. **232**(2): p. 418-21.
160. Ohyaigi, Y., et al., *Intracellular Abeta42 activates p53 promoter: a pathway to neurodegeneration in Alzheimer's disease.* Faseb j, 2005. **19**(2): p. 255-7.
161. Cenini, G., et al., *Elevated levels of pro-apoptotic p53 and its oxidative modification by the lipid peroxidation product, HNE, in brain from subjects with amnesic mild cognitive impairment and Alzheimer's disease.* J Cell Mol Med, 2008. **12**(3): p. 987-94.
162. Fagerberg, L., et al., *Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics.* Mol Cell Proteomics, 2014. **13**(2): p. 397-406.

163. Späni, C., et al., *Reduced β -amyloid pathology in an APP transgenic mouse model of Alzheimer's disease lacking functional B and T cells*. *Acta neuropathologica communications*, 2015. **3**: p. 71-71.
164. Bashiardes, S., et al., *SNTG1, the gene encoding γ 1-syntrophin: a candidate gene for idiopathic scoliosis*. *Human Genetics*, 2004. **115**(1): p. 81-89.
165. Duan, Y., N. Kelley, and Y. He, *Role of the NLRP3 inflammasome in neurodegenerative diseases and therapeutic implications*. *Neural Regen Res*, 2020. **15**(7): p. 1249-1250.
166. Feng, Y.S., et al., *The involvement of NLRP3 inflammasome in the treatment of neurodegenerative diseases*. *Biomed Pharmacother*, 2021. **138**: p. 111428.
167. Bergsbaken, T., S.L. Fink, and B.T. Cookson, *Pyroptosis: host cell death and inflammation*. *Nature Reviews Microbiology*, 2009. **7**(2): p. 99-109.
168. Han, C., et al., *New mechanism of nerve injury in Alzheimer's disease: β -amyloid-induced neuronal pyroptosis*. *Journal of Cellular and Molecular Medicine*, 2020. **24**(14): p. 8078-8090.
169. Van Zeller, M., et al., *NLRP3 Inflammasome: A Starring Role in Amyloid- β - and Tau-Driven Pathological Events in Alzheimer's Disease*. *Journal of Alzheimer's Disease*, 2021. **83**: p. 939-961.

170. Yan, S., et al., *CSB6B prevents β -amyloid-associated neuroinflammation and cognitive impairments via inhibiting NF- κ B and NLRP3 in microglia cells*. *Int Immunopharmacol*, 2020. **81**: p. 106263.
171. Attems, J. and K.A. Jellinger, *Olfactory tau pathology in Alzheimer disease and mild cognitive impairment*. *Clin Neuropathol*, 2006. **25**(6): p. 265-71.
172. Sun, G.H., et al., *Olfactory identification testing as a predictor of the development of Alzheimer's dementia: a systematic review*. *Laryngoscope*, 2012. **122**(7): p. 1455-62.
173. Conti, M.Z., et al., *Odor identification deficit predicts clinical conversion from mild cognitive impairment to dementia due to Alzheimer's disease*. *Arch Clin Neuropsychol*, 2013. **28**(5): p. 391-9.
174. Tuson, M., et al., *Overexpression of CERKL, a gene responsible for retinitis pigmentosa in humans, protects cells from apoptosis induced by oxidative stress*. *Mol Vis*, 2009. **15**: p. 168-80.
175. Manandhar, M., K.S. Boulware, and R.D. Wood, *The ERCC1 and ERCC4 (XPF) genes and gene products*. *Gene*, 2015. **569**(2): p. 153-61.
176. Canugovi, C., et al., *The role of DNA repair in brain related disease pathology*. *DNA Repair (Amst)*, 2013. **12**(8): p. 578-87.
177. Madabhushi, R., L. Pan, and L.H. Tsai, *DNA damage and its links to neurodegeneration*. *Neuron*, 2014. **83**(2): p. 266-282.

178. Kosik, K.S., et al., *Mechanisms of age-related cognitive change and targets for intervention: epigenetics*. J Gerontol A Biol Sci Med Sci, 2012. **67**(7): p. 741-6.
179. Malamon, J.S. and A. Kriete, *Erosion of Gene Co-expression Networks Reveal Deregulation of Immune System Processes in Late-Onset Alzheimer's Disease*. Frontiers in Neuroscience, 2020. **14**.
180. Lancour, D., et al., *Analysis of brain region-specific co-expression networks reveals clustering of established and novel genes associated with Alzheimer disease*. Alzheimer's Research & Therapy, 2020. **12**(1): p. 103.
181. Fionda, V., *Networks in Biology*, in *Encyclopedia of Bioinformatics and Computational Biology*, S. Ranganathan, et al., Editors. 2019, Academic Press: Oxford. p. 915

CURRICULUM VITAE
APOORVA BHARTHUR SANJAY

EDUCATION

- **PhD in Medical Neuroscience:** Aug 2017 - Dec 2022
Indiana University | Indianapolis, IN, USA
- **PhD (Minor) Biostatistics:** Aug 2018- Dec 2020
Indiana University | Indianapolis, IN, USA
- **M.S in Medical and Molecular Genetics:** Aug 2013- Aug 2015
Indiana University | Indianapolis, IN, USA
- **Bachelor of Technology in Biotechnology and Genetic Engineering**
Aug 2009- May 2013
SRM University | Tamil Nadu, India

WORK EXPERIENCE

- **UNIVERSITY OF PENNSYLVANIA** | Philadelphia, PA, USA
Research Specialist- Department of Biology 2015 - 2017
- **INDIANA UNIVERSITY SCHOOL OF MEDICINE** | Indianapolis, IN, USA
Graduate Research Assistant – Stark Neuroscience Institute 2013–2015
- **NATIONAL INSTITUTE OF MENTAL HEALTH AND NEUROSCIENCE (NIMHANS)** Bangalore, India
Research Intern Jan 2013-June 2013

SKILLS

- Excellent organizational and interpersonal skills
- Strong communication skills as demonstrated through numerous oral and poster presentations
- Cross functional leadership: Mentored and trained students, worked with collaborators on projects, and writing successfully funded grants and manuscripts.
- Programming Languages (MATLAB, R, SAS, and Python)
- Operating Systems (UNIX, Linux, and Windows)
- Knowledge of statistical and machine learning frameworks, such as deep neural networks, random forests, support vector machines.
- Experience working with real world healthcare data and clinical research data (biomarkers, GWAS, omics data).
- Experience with High Performance Computing (HPC) systems
- Experience applying analytics, modeling, and statistics to large biological datasets, such as dbSNP
- Neuroimaging programming (Freesurfer, FSL, AFNI, SPM)
- Microsoft Office Suite (Word, Excel, and PowerPoint)
- Strong Knowledge in Neuroscience, Biochemistry, Alzheimer's Disease, Biostatistics

- Knowledge of sequence analysis techniques especially variant detection and annotation
- Biochemical methods (DNA/RNA extraction, protein purification, PCR, Western Blot, ELISA, Cloning)

AWARDS

- Travel Scholarship Award for Human Amyloid Imaging (HAI), Miami, FL, 2019
- Travel Scholarship Award for Alzheimer's Association International Conference (AAIC), Los Angeles, CA, 2019
- Student Travel Award, ACNN (Advanced Computational Neuroscience Network), Ann Arbor, MI, 2019
- Poster selected for **Student Science Spotlight** program, Alzheimer's Association International Conference (virtual conference), 2020
- Poster selected for poster award competition, Alzheimer's Association International Conference, 2021, Denver, CO, USA

ARTICLES

- **Apoorva Bharthur Sanjay**, Alice Patania, Xiaoran Yan, Diana Svaldi, Tugce Duran, Niraj Shah, Eric Chen and Liana Apostolova. *Characterization of genetic expression patterns in MCI using a multiomics approach and neuroimaging endophenotypes, Alzheimer's and Dementia*, 2022.
<https://doi.org/10.1002/alz.12587>

- Eva Birgitte Aamodt, Till Schellhorn, Edwin Stage, **Apoorva Bharthur Sanjay**, Paige Elise Logan, Diana Otero Svaldi, Liana G Apostolova, Ingvild Saltvedt and Mona Kristiansen Beyer. *Predicting the Emergence of Major Neurocognitive Disorder Within Three Months After a Stroke (Frontiers Aging Neuroscience)*, 2021.
<https://doi.org/10.3389/fnagi.2021.705889>
- Tugce Duran, Ellen Woo, Diana Otero, Shannon L. Risacher, Eddie Stage, **Apoorva Bharthur Sanjay**, Kwangsik Nho, John D. West, Meredith Phillips, Naira Goukasian, Kristy Hwang and Liana G. Apostolova. *Associations between Cortical Thickness and Metamemory in Alzheimer's Disease (Brain Imaging and Behavior)* 2021.
<https://doi.org/10.1007/s11682-021-00627-0>
- Wie, J., **Apoorva Bharthur Sanjay**, Wolfgang, M. *et al.* Intellectual disability associated UNC80 mutations reveal inter-subunit interaction and dendritic function of the NALCN channel complex. *Nature Communications* 11, 3351 (2020). <https://doi.org/10.1038/s41467-020-17105-8>
- Diana Svaldi, Joaquín Goni, **Apoorva Bharthur Sanjay**, Enrico Amico, Shannon Risacher, John West & M Dzemidzic, Andrew Saykin and Liana Apostolova. (2018). Towards Subject and Diagnostic Identifiability in the Alzheimer's Disease Spectrum Based on Functional Connectomes: *Second International Workshop, GRAIL 2018 and First International Workshop, Beyond MIC 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings*. 10.1007/978-3-030-00689-1_8.

- Nho, K., West, J.D., Li, H., Henschel, R., **Apoorva Bharthur Sanjay**, Tavares, M.C., and Saykin, A.J. (2014). Comparison of multi-sample variant calling methods for whole genome sequencing. *IEEE International Conference on Systems Biology, 2014*, 59–62.

ORAL PRESENTATIONS

- **Apoorva Bharthur Sanjay** (2021). Blood transcriptomic biomarkers in Alzheimer's Disease. Indiana Alzheimer's Disease Research Center (IADRC) Fall Research Symposium 2021.
- **Apoorva Bharthur Sanjay**, Alice Patania, PhD, Xiaoran Yan, PhD, Diana Svaldi, PhD. and Liana G. Apostolova (2020). Characterization of genetic expression patterns in MCI using a multiomics approach and neuroimaging endophenotypes. Alzheimer's Association International Conference (AAIC) 2020
- **Apoorva Bharthur Sanjay**, Kristy S. Hwang, Diana Svaldi, Rowina Hussainali, Tugce Duran, Naira Goukasian, Liana G. Apostolova (2019). Predicting brain amyloidosis using peripheral blood- based gene expression and early-stage neurodegeneration biomarkers. Human Amyloid Imaging (HAI) 2019
- **Apoorva Bharthur Sanjay**, Kristy S. Hwang, Diana Svaldi, Rowina Hussainali, Tugce Duran, Naira Goukasian, Liana G. Apostolova (2019). Predicting brain amyloidosis using peripheral blood- based gene

expression and early-stage neurodegeneration biomarkers. AAIC, Los Angeles, 2019.

POSTER PRESENTATIONS

- **Apoorva Bharthur Sanjay, MS**, Rion Brattig Correia, Luis M Rocha, PhD, and Liana G. Apostolova. *Transcriptomic profiling in Mild Cognitive Impairment using peripheral blood gene co-expression networks*. Alzheimer's Association International Conference, 2021
- **Apoorva Bharthur Sanjay**, Yunyi Li, Sujuan Gao, PhD and Liana G. Apostolova, MD, MS. *Effect of genetic and vascular AD risk factors on rate of cognitive decline in EOAD vs LOAD*. Alzheimer's Association International Conference, 2021.
- Makoto Ishii, Costantino Iadecola, **Apoorva Bharthur Sanjay**, Liana Apostolova. *Sexually dimorphic association of circulating leptin levels with early amyloid-beta pathology assessed by amyloid PET*. Alzheimer's Association International Conference, 2021
- Liana Apostolova, Katie Lane, Paige Logan, Mohit Manchella, **Apoorva Bharthur Sanjay**, Sujuan Gao. *APOE4 is associated with earlier symptom onset in LOAD but later symptom onset in EOAD*. Alzheimer's Association International Conference, 2021
- Paige E. Logan, Kathleen A. Lane, Mohit K. Manchella, **Apoorva Bharthur Sanjay**, Sujuan Gao, Liana G. Apostolova. *Early-onset APOE ε4 carriers show greater decline in memory, language, and executive*

- function than late-onset carriers. Alzheimer's Association International Conference, 2021*
- Aamodt EB, Schellhorn T, Apostolova LG, Svaldi DO, Logan P, **Sanjay AB**, Saltvedt I, Beyer MK. *Prediction of Early Post-stroke Major Neurocognitive Disorder Using Support Vector Machines*. 2021 International Stroke Conference, Denver, Colorado, February 2021.
 - **Apoorva Bharthur Sanjay**, Diana Otero Svaldi, and Liana G. Apostolova (2020). *Transcriptomic profiling of brain amyloidosis using peripheral blood-based gene expression*. Alzheimer's Association International Conference, 2020.
 - Jenna Rae Groh, Diana Svaldi, Eddie Stage, **Apoorva Bharthur Sanjay**, Paige E Logan, Shannon L Risacher, Andrew J Saykin, Liana G Apostolova. *Data-driven Characterization of Tau Accumulation across the Alzheimer's disease spectrum*. Alzheimer's Association International Conference 2020
 - **Apoorva Bharthur**, Kristy S. Hwang, Diana Svaldi, Rowina Hussainali, Tugce Duran, Naira Goukasian, Liana G. Apostolova (2019). *Predicting brain amyloidosis using peripheral blood-based gene expression and early-stage neurodegeneration biomarkers*. Indy SFN, IUPUI, 2019.
 - **Apoorva Bharthur**, Kristy S. Hwang, Diana Svaldi, Rowina Hussainali, Tugce Duran, Naira Goukasian, Liana G. Apostolova (2019). *Predicting brain amyloidosis using peripheral blood-based gene expression and*

early-stage neurodegeneration biomarkers. Alzheimer's Imaging Consortium, Los Angeles, 2019.

- Xiaoran Yan, Alice Patania, Diana O. Svaldi, **Apoorva Bharthur Sanjay**, Tugce Duran, Kristy Hwang, Valentin Penchev, John D. West, Patricia Mabry, Liana Apostolova (2019). *Evaluating the Utility of Homological Structural MRI Features and A Kernel Based Learning Framework for Prediction of MCI*. AAIC, LA, 2019
- Abby Braun, **Apoorva Bharthur Sanjay**, Diana Otero Svaldi, Liana Apostolova. *CD33 and TREM2 peripheral gene expression in relation to cortical thickness*. Proceedings of IMPRS 1, 20

FELLOWSHIPS

- 2012 Summer Research Fellowship, IISER, India
- 2019-2021 CTSI Eli Lilly-Stark Pre-Doctoral Research Fellowship in Neurodegeneration

GRANTS FUNDED

2021 R21 AG072101-01: Integrative Predictive Modeling of Alzheimer's Disease
(PI: Alice Patania)