

OPEN

Shared Causal Paths underlying Alzheimer's dementia and Type 2 Diabetes

Zixin Hu^{1,3}, Rong Jiao², Panpan Wang¹, Yun Zhu⁴, Jinying Zhao⁴, Phil De Jager⁵, David A. Bennett⁶, Li Jin^{1,3} & Momiao Xiong^{2*}

Although Alzheimer's disease (AD) is a central nervous system disease and type 2 diabetes MELLITUS (T2DM) is a metabolic disorder, an increasing number of genetic epidemiological studies show clear link between AD and T2DM. The current approach to uncovering the shared pathways between AD and T2DM involves association analysis; however such analyses lack power to discover the mechanisms of the diseases. As an alternative, we developed novel causal inference methods for genetic studies of AD and T2DM and pipelines for systematic multi-omic casual analysis to infer multilevel omics causal networks for the discovery of common paths from genetic variants to AD and T2DM. The proposed pipelines were applied to 448 individuals from the ROSMAP Project. We identified 13 shared causal genes, 16 shared causal pathways between AD and T2DM, and 754 gene expression and 101 gene methylation nodes that were connected to both AD and T2DM in multi-omics causal networks.

Although Alzheimer's dementia is a central nervous system disease and type 2 diabetes MELLITUS (T2DM) is a metabolic disorder, an increasing number of epidemiological and genetic epidemiological studies show clear link between Alzheimer's dementia and T2DM. Alzheimer's dementia with great economic, political and social consequences is a progressive, irreversible degenerative disease of the brain and is the most common cause of dementia due to the gradual accumulation of amyloid-beta ($A\beta$) and twisting of tau protein^{1,2}, and other common brain pathologies³. Alzheimer's dementia is also involved in inflammation and oxidative stress and exhibits memory loss and cognitive dysfunction^{4,5}.

Two mechanisms underlying T2DM are insulin resistance and insufficient insulin secretion from pancreatic β -cells⁴. T2DM patients are unable to process insulin signaling correctly. In response to insulin resistance, pancreatic β -cells increase insulin production. However, when pancreatic β -cells gradually lose function; insulin production cannot be increased to maintain normal glucose levels. The brain is a target organ for insulin⁶. Insulin signaling plays an important role in the organization and function of the brain and impaired insulin signaling induces an overactivation of GSK-3 kinase, increases tau phosphorylation, alters tau modification and neurofibrillary degeneration⁷. T2DM also suffer from mild to severe nervous system damage. Persistent blood glucose may impair blood flow to the brain⁸.

Prior work in ROSMP found an association of T2DM with incident Alzheimer's dementia and rate of cognitive decline⁹. However, we did not find an association with Alzheimer's disease (AD) pathology¹⁰. Rather, we found an association with cerebral infarcts. Other evidence from ROSMP continue to point to potential common mechanisms. For example, we found that brain insulin signaling was associated with AD pathology¹¹. We also found interactions between $GSK\beta$ polymorphisms associated with β -amyloid deposition¹².

The current approaches to identifying several shared pathophysiology processes between Alzheimer's dementia and T2DM have several limitations. Firstly, the most previous works have focused on identifying biological pathways underlying AD and T2DM. Few attempts to discover the role of dysregulated SNPs, gene expressions and methylations have been carried out. Secondly, the conventional evidences for linking AD and T2DM purely

¹State Key Laboratory of Genetic Engineering and Innovation Center of Genetics and Development, School of Life Sciences, Fudan University, Shanghai, China. ²Department of Biostatistics and Data Science, School of Public Health, University of Texas Health Science Center at Houston, Houston, Texas, USA. ³Human Phenome Institute, Fudan University, Shanghai, China. ⁴Department of Epidemiology, University of Florida, Florida, USA. ⁵Center for Translational & Computational Neuroimmunology, Department of Neurology, Columbia University Medical Center, New York, 10033, USA. ⁶Rush Alzheimer's Disease Center, Rush University Medical Center, Chicago, IL, 60612, USA. *email: Momiao.Xiong@uth.tmc.edu

depend on the statistical association¹³. There has been increasing recognition that association and causation are different concepts¹⁴. Association attempts to measure dependence between two variables, while causation is to study the distribution of the variable (effect) after taking action on the another variable (cause). The statistical tool for association analysis is the conditional distribution, while the tool for the causal analysis is the intervention calculus. Many association signals may not be causal signals and some causal signals may not show strong association. If causation loci were searched only from association loci, many causation loci might be missed. The widely used gene expression networks are co-expression networks and phenotype networks are correlation networks. The major tools for integrated omics analysis are based on association analysis. The networks in the most multilevel omics analysis are undirected graphs. It is difficult to use undirected graphs for identifying the causal paths from genetic variants to diseases.

We are facing a great challenge to shift the current analytic platforms of genetic analysis from genetic association analysis to multilevel omics causal analysis for unraveling the mechanic link between AD and T2DM. To meet this challenge, we need (1) to develop and implement causation analysis methods for genetic studies of AD and T2DM; (2) to develop a general framework for construction of multilevel causal omics networks to discover common paths from genetic variations to AD and T2DM via methylations, gene expressions and multiple phenotypes. The real data set ROSMAP^{15,16} will be used to valid the multilevel omics causal networks as a useful analytic platform for identifying shared causal paths between AD and T2DM and demonstrates that the proposed methods are capable of identifying the shared pathologic paths between AD and T2DM. A program for construction of multilevel causal networks can be downloaded from https://github.com/wenrurumon/mysrc/tree/master/CNIF_0.3.0.

Results

Simulations. To evaluate the performance of the proposed causal network analysis, we conducted a series of simulation studies to compare the detection power and false discovery rate (FDR) for three methods: (1) weighted gene co-expression network (WGCNA), (2) structural equation model (SEM) and structural equation model coupled with integer programming (SEMIP).

We randomly generated 1,000 directed acyclic graphs (networks) with 20 nodes (15 gene expression or phenotype nodes and 5 genotype nodes) and mean 30 directed edges, 1,000 directed acyclic graphs (networks) with 30 nodes (22 expression/phenotype nodes, 8 genotype nodes), and mean 47 directed edges, and 40 nodes (30 gene expression or phenotype nodes and 10 genotype nodes), and mean 68 directed edges, respectively. Simulation results were summarized in Table 1 where we only listed undirected network results because the WGCNA can only estimate the undirected network. We calculated the power and FDR of three methods for 100, 300, 500 and 1,000 samples. We can observe that in all cases, The SEMIP had the largest power and smallest FDR. When the number of nodes in the networks increased, the power to identify the structure of the networks decreased, while FDR increased. When the number of nodes reached 40, the SEMIP can reach 68.5% power and 7.40% FDR using 1,000 samples.

Shared genetic loci underlying AD and T2DM. The number of AD and T2DM directly connected or indirectly connected genes was summarized in Table 2. The total number of genes connected to both AD and T2DM including directly connected and indirectly connected was 759. The genes that were both directly and indirectly connected to both AD and T2DM were summarized in Table S1. The genes that were indirectly connected to AD and both directly and indirectly connected to T2DM were listed in Table S2. Similarly, the genes that were both directly and indirectly connected to AD and indirectly connected to T2DM were summarized in Table S3.

We also tested causation of 299 pathways in the KEGG pathway database to AD and T2DM (Described in detail in the Methods section). The results were summarized as follows. The number of pathways that were directly connected to both AD and T2DM was 16; the number of pathways that were directly connected to AD and indirectly connected to T2DM was 17; the number of pathways that were directly connected to T2DM and indirectly connected to AD was 18, the number of pathways that were indirectly connected to both AD and T2DM was 114; the number of pathways that were directly connected to AD and not connected to T2DM was 6; the number of pathways that were not connected to AD and directly connected to T2DM was 2.

Then, we investigated shared gene expressions via multilevel causal networks. We summarized the results as follows. The number of expression genes that were directly connected to both AD and T2DM was two genes: GRMD1B, RP1-111D6.3, the number of expression genes that were directly connected to AD, but not directly connected to T2DM was 19 (P-value < 10^{-4} , Table S4) and the number of expression genes that were directly connected to T2DM, but not directly connected to AD was 7 (P-value < 10^{-4} , Table S5). The number of expression genes that were indirectly connected to both AD and T2DM was 725.

Similarly, we can study shared methylation via multilevel causal networks. The number of methylated sites/genes that were directly connected to AD, but not directly connected to T2DM was 17 (Table S6) and the number of methylated sites/genes that were directly connected to T2DM, but not directly connected to AD was 27 (Table S7). The number of methylated sites/genes that were indirectly connected to both AD and T2DM was 117 (Table S8).

The number of phenotypes that were directly connected to both AD and T2DM was six (Age, CHL, HDL ratio, LDL, Semantic memory and working memory).

Shared CREBBP, MAPK and PI3K-AKT pathways between AD and T2DM. To assess whether CREBBP is a common genetic factor of AD and T2DM, and how CREBBP mediates the development of AD and T2DM, we searched the all possible paths from gene CREBBP to AD and T2DM in the inferred multilevel causal network. The results were shown in Fig. 1. Figure 1A plotted the path from CREBBP to AD and T2DM via MAPK and PI3K-AKT signaling pathways. The genes in the MAPK and PI3K-AKT signaling pathways, CREBBP,

Methods	Nodes	Sample Sizes	Undirected		Directed	
			Power	FDR	Power	FDR
WGCNA	20	100	51.60%	16.00%		
WGCNA	20	300	53.00%	15.30%		
WGCNA	20	500	66.40%	13.00%		
WGCNA	20	1000	82.60%	13.60%		
SEM	20	100	70.80%	44.40%	50.70%	42.00%
SEM	20	300	77.80%	49.20%	53.90%	23.20%
SEM	20	500	83.50%	46.40%	56.90%	44.90%
SEM	20	1000	98.20%	32.40%	57.30%	26.70%
SEMIP	20	100	64.60%	34.70%	59.50%	15.50%
SEMIP	20	300	73.50%	39.30%	65.40%	17.40%
SEMIP	20	500	77.60%	25.40%	68.30%	12.00%
SEMIP	20	1000	86.60%	22.60%	76.60%	13.20%
WGCNA	30	100	43.30%	21.10%		
WGCNA	30	300	49.80%	15.00%		
WGCNA	30	500	53.60%	21.00%		
WGCNA	30	1000	56.50%	13.20%		
SEM	30	100	64.30%	34.10%	46.70%	26.00%
SEM	30	300	73.60%	41.20%	49.30%	22.50%
SEM	30	500	82.30%	34.80%	52.40%	34.10%
SEM	30	1000	94.50%	36.30%	52.80%	27.60%
SEMIP	30	100	63.30%	15.50%	58.50%	16.40%
SEMIP	30	300	67.40%	27.00%	63.50%	13.50%
SEMIP	30	500	71.50%	18.30%	64.20%	10.80%
SEMIP	30	1000	94.80%	28.60%	71.80%	15.00%
WGCNA	40	100	43.30%	21.40%		
WGCNA	40	300	49.20%	17.00%		
WGCNA	40	500	51.40%	19.70%		
WGCNA	40	1000	54.10%	18.20%		
SEM	40	100	61.70%	37.30%	46.50%	29.50%
SEM	40	300	70.10%	25.70%	49.60%	38.20%
SEM	40	500	79.90%	35.30%	54.50%	17.70%
SEM	40	1000	95.10%	45.90%	62.60%	27.90%
SEMIP	40	100	62.70%	23.20%	58.30%	11.60%
SEMIP	40	300	64.50%	21.10%	62.10%	10.30%
SEMIP	40	500	75.50%	32.30%	66.20%	15.30%
SEMIP	40	1000	82.00%	34.00%	68.50%	7.40%

Table 1. Power and FDR of three methods for construction of causal networks with 20,30 and 40 nodes.

		To T2DM			
		Directly Connected	Indirectly Connected	Both Directly and Indirectly Connected	Not Connected
To AD	Directly Connected	5			13
	Indirectly Connected		682	13	
	Both Directly and Indirectly Connected		20	8	
	Not Connected	17			

Table 2. The number of genes connected to AD and T2DM.

episodic memory, MMSE, AD and T2DM were then used to further infer causal networks using SEMs and IP. The inferred causal network was shown in Fig. 1B. From Fig. 1B we observed a path from CREBBP to AD and T2DM via gene connections: $CREBBP \rightarrow CBL \rightarrow MAP2K4 \rightarrow MAPK8 \rightarrow MAPK1 \rightarrow PI3KCA$. MAPK and PI3K-AKT pathways play critical roles in memory.

Shared TTC3, FoxO, MAPK, and PI3K-AKT Pathways between AD and T2DM. Next we presented an example to illustrate shared causal paths that started a gene directly connected to AD and indirectly connected to T2DM.

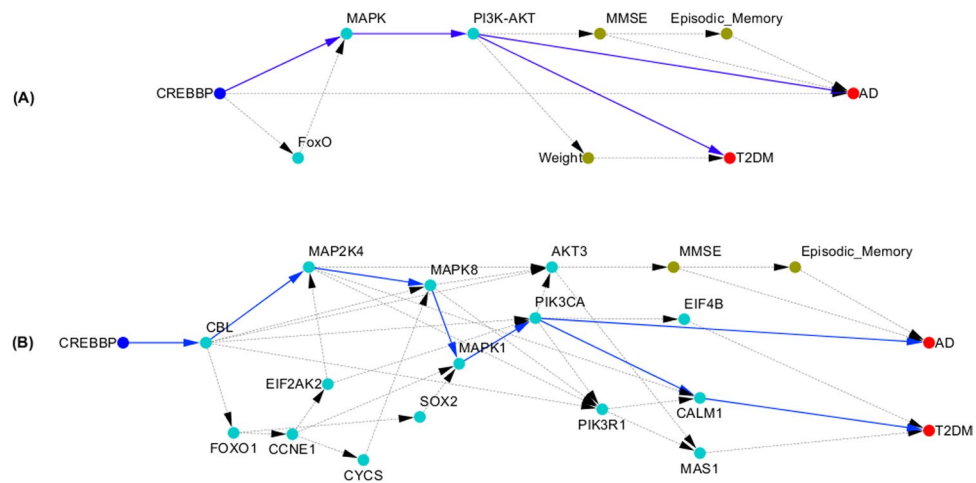


Figure 1. (A) Shared CREBBP, MAPK and PI3K-AKT pathways between AD and T2DM; (B) Shared causal subnetwork structure from CREBBP to AD and T2DM.

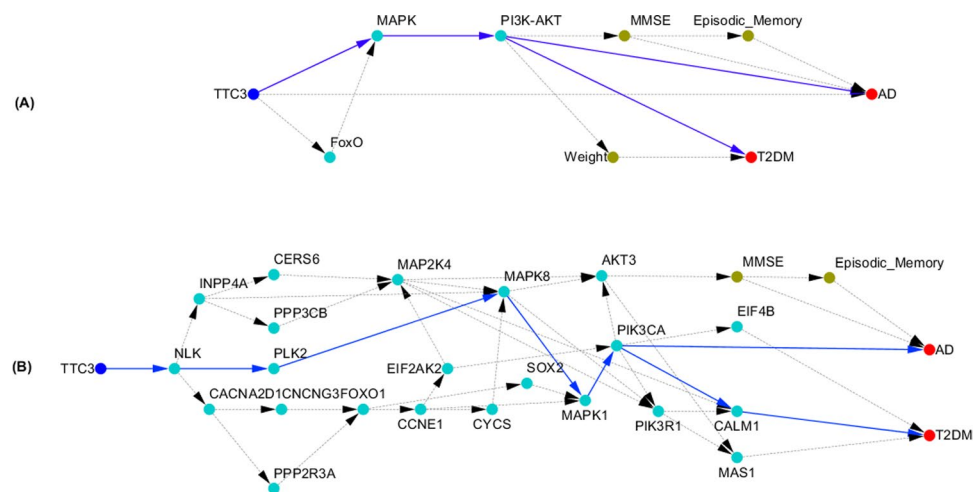


Figure 2. (A) Shared TTC3, FoxO, MAPK, and PI3K-AKT Pathways between AD and T2DM; (B) Shared causal subnetwork structure from TTC3 to AD and T2DM.

Again, we used the DFS algorithm to search the causal paths from multilevel causal networks. The causal paths from TTC3 to AD and T2DM were shown in Fig. 2. The paths from MAPK and PI3K-AKT pathway to AD and T2DM were the same as that in Fig. 1. The genes in the FoxO, MAPK and PI3K-AKT signaling pathways, TTC3, and episodic memory, MMSE, weight, AD and T2DM were then used to further infer causal networks using SEMs and IP. The structure of the inferred network was shown in Fig. 2B. There were a large number of causal paths from TTC3 to either AD or T2DM. The shared common causal paths were $TTC3 \rightarrow NLK \rightarrow CACNA2D1 \rightarrow CNCNG3 \rightarrow FOXO1 \rightarrow CCNE1 \rightarrow CYCS \rightarrow MAPK1 \rightarrow PIK3CA$ and $TTC3 \rightarrow NLK \rightarrow PLK2 \rightarrow MAPK8 \rightarrow MAPK1 \rightarrow PIK3CA$.

Shared morphine addiction and neuroactive ligand receptor interaction pathways. Searching the causal paths from gene *HNF4G* to AD and T2DM via the multilevel causal networks using the DFS algorithm, we found that *HNF4G* was indirectly connected to AD and T2DM. In addition to shared MAPK and PI3K-AKT pathways between AD and T2DM which were discussed in the previous sections, we observed shared two new pathways between AD and T2DM: morphine addiction and neuroactive ligand receptor interaction pathways as shown in Fig. S1A. The structure of the inferred network that consisted of shared morphine addiction and neuroactive ligand receptor interaction pathways between AD and T2DM was shown in Fig. S1B. There were more than 10 shared causal paths. We observed two shared major causal paths: (1) $HNF4G \rightarrow NLK \rightarrow PLK2 \rightarrow MAPK8 \rightarrow MAPK1 \rightarrow PIK3CA \rightarrow AKT1$ and (2) $HNF4G \rightarrow NLK \rightarrow GNGT2 \rightarrow PLCB2 \rightarrow PLCB1 \rightarrow ADRB1$.

Shared fatty acid biosynthesis and primary bile acid biosynthesis pathways. Our data also provided evidence to show that fatty acid biosynthesis and primary bile acid biosynthesis pathways were shared

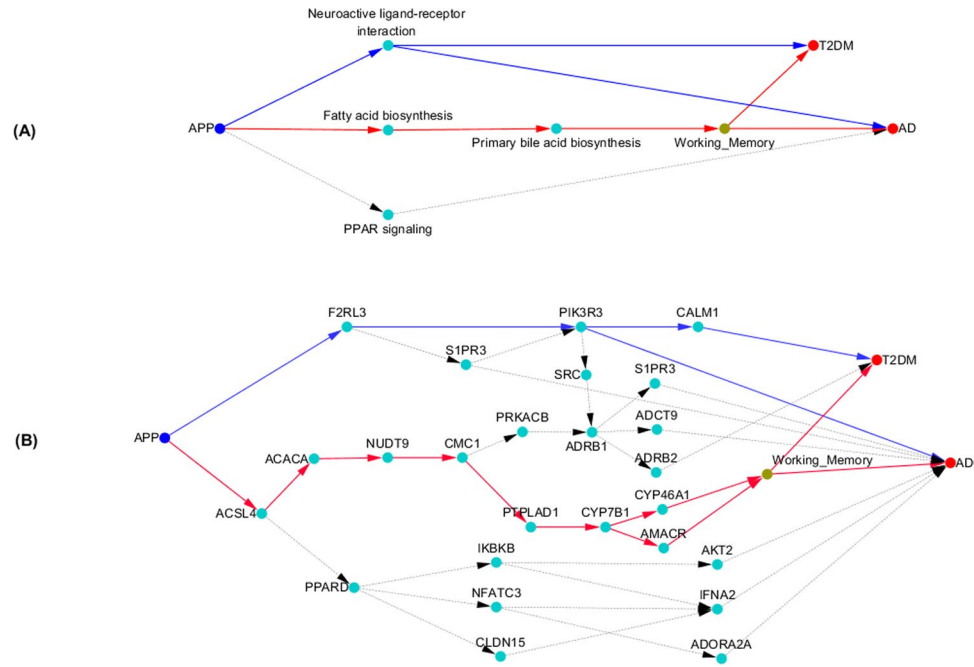


Figure 3. (A) Shared APP, Fatty Acid Biosynthesis and Primary Bile Acid Biosynthesis Pathways between AD and T2DM; (B) Shared causal subnetwork structure from APP to AD and T2DM.

pathways between AD and T2DM. Searching the multilevel causal networks from APP to AD and T2DM using the DFS algorithm, we identified the shared causal paths from APP to both AD and T2DM, shown in Fig. 3A. There were two shared causal paths between AD and T2DM: $APP \rightarrow \text{neuroactive ligand receptor interaction}$ and $APP \rightarrow \text{fatty acid biosynthesis} \rightarrow \text{primary bile acid biosynthesis}$. Neuroactive ligand receptor interaction pathway was discussed in the previous section.

Next we presented the causal network structure of the shared genes between AD and T2DM in the two shared causal paths in Fig. 3B. We observed two major shared paths from APP to AD and T2DM. One path was $APP \rightarrow ACSLA \rightarrow ACACA \rightarrow NUDT9 \rightarrow CMC1 \rightarrow PTPLAD1 \rightarrow CYP781 \rightarrow CYP46A1 \rightarrow \text{working memory}$ (or $CYP781 \rightarrow AMACA \rightarrow \text{working memory}$). Another causal path was $APP \rightarrow F2RL3 \rightarrow PIK3R3 \rightarrow (\text{or } F2RL3 \rightarrow S1PR3 \rightarrow PIK3R3)$.

To further illustrate the validity of the inferred causal paths, we presented Fig. S2 that showed the average levels of expression of the genes in Fig. 3 for AD, T2DM and normal individuals. From Fig. 3, Figs. S2 and S3, we can observe that the genes along the path $APP \rightarrow F2RL3 \rightarrow PIK3R3$ (or $F2RL3 \rightarrow S1PR3 \rightarrow PIK3R3$) of the individuals with AD were over expressed, and the genes along the path $APP \rightarrow ACSLA \rightarrow ACACA \rightarrow NUDT9 \rightarrow CMC1 \rightarrow PTPLAD1 \rightarrow CYP781 \rightarrow CYP46A1 \rightarrow \text{working memory}$ (or $CYP781 \rightarrow AMACA \rightarrow \text{working memory}$) of the individuals with AD were under expressed. Genetic variation in gene APP either regulated over expressed genes or regulated under expressed genes. Both of them caused AD. For the individuals with T2DM, the majority of gene expressions along the causal paths from APP to T2DM which were regulated by genetic variation in gene APP was under expressed.

Shared methylated genes POU3F2, KIF4B and TNSL3, and dopaminergic synapse and AMPK pathways.

In this section, we illustrate how a shared gene regulates three shared gene methylations, which in turn regulate the shared pathways. Our results showed that genetic variation in gene POU3F2 regulated gene expressions in dopaminergic synapse and AMPK pathways via methylations of POU3F2, KIF4B and TMSL3, which in turn influences CHL/HDL Ratio, and finally led to AD and T2DM (Fig. 4A).

Again, we presented the causal network structure of the shared genes between AD and T2DM in the two shared dopaminergic synapse and AMPK pathways in Fig. 4B. There were multiple shared directed paths from POU3F2 to AD and T2DM. A major shared directed path: $m: POU3F2 \rightarrow m: LOC644649 \rightarrow KDM5C \rightarrow PDPK2 \rightarrow XPA \rightarrow MK3R2 \rightarrow ELK1 \rightarrow AD$ (or $CHL \rightarrow T2DM$).

Discussion

This paper addresses several issues for uncovering causal paths shared between AD and T2DM. The first issue is to shift the current paradigm of genetic analysis from association analysis to deep causal inference for uncovering the shared mechanisms between AD and T2DM. The current paradigm for discovering mechanisms of diseases is association analysis. There is increasing recognition that a large proportion of association signals are not causal signals and causal signals may not be association signals. A large number of causal signals cannot be derived from set of association signals. Only searching causal signals from association analysis, a large proportion of causal

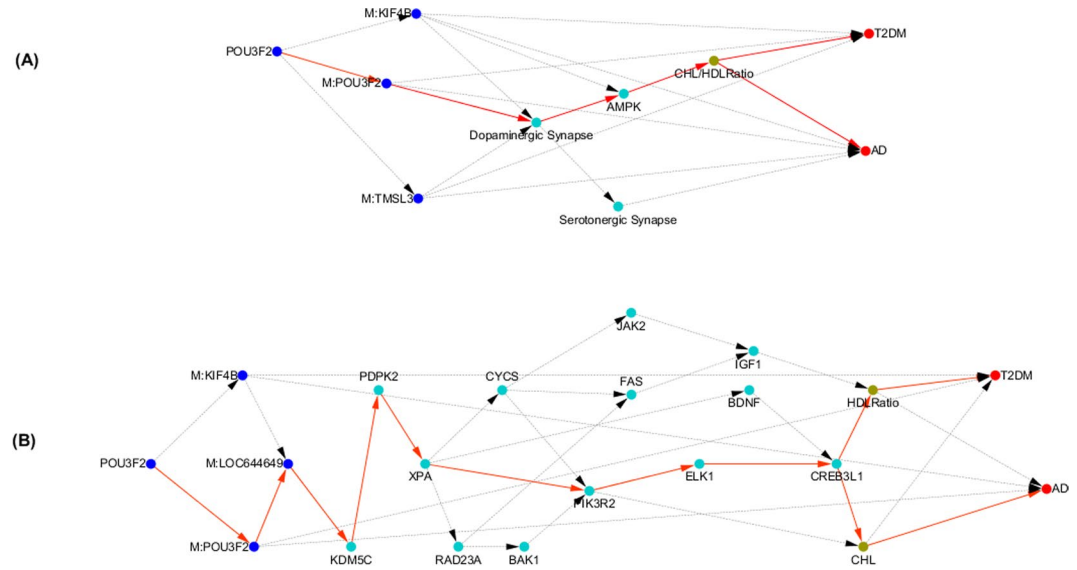


Figure 4. (A) Shared Methylated Genes POU3F2, KIF4B and TMSL3, and Dopaminergic Synapse and AMPK Pathways between AD and T2DM; (B) Shared causal subnetwork structure from POU3F2 to AD and T2DM.

signals will be missing. Therefore, the ANMs were developed as practical causal inference methods to identify the genetic variants that cause disease.

Second issue is to shift the current paradigm of genetic analysis from genetic analysis alone to integrated causal genomic, epigenomic, transcriptional and phenotypic data analysis for unraveling the mechanisms of AD and T2DM. The widespread existing omics networks that are essentially undirected graphs. Using undirected graphs, we cannot to identify direct causal relations among diversified types of variables at multilevel and the causal routes from genetic variants to complex phenotypes via omics. In this paper, we develop novel statistical methods for multilevel causal omics network construction and provide pipelines for uncovering shared causal paths between AD and T2DM via gene expressions, DNA methylations, environments and multiple phenotypes.

The third issue is to develop algorithms that can automatically search the causal routes from genetic variations to the complex phenotypes. The size of multilevel causal omics network is large. The number of nodes of such networks can reach ten thousand. The number of causal paths is huge. Manually searching causal paths from large causal networks is infeasible. To meet the challenge of searching causal paths from large causal networks, we develop computer representation of large causal networks and algorithms for searching the causal paths.

The results of application of the proposed pipelines for identifying causal paths to real data analysis of AD and T2DM provided strong evidence to support the link between AD and T2DM and unraveled causal mechanism to explain this link. We identified the shared causal genes, gene expressions, DNA methylations and pathways between AD and T2DM. Some of them can be supported by literature and some of them are new.

Specifically, we identified the shared CREBBP, MAPK and PI3K-AKT pathways between AD and T2DM. Binding of transcription factors to the cyclic Adenosine Monophosphate (cAMP) response element (CRE) regulates the activity of RNA polymerase. cAMP Response Element binding protein (CREB) is a cellular transcription factor that binds the CRE¹⁷. CREB-binding protein (CREBBP) and CREB together mediate the conversion of short-term memory to long-term memory and alternate the activity of the β -amyloid (A β) peptide, which in turn regulates hippocampal-dependent synaptic plasticity^{18,19}. Cognitive function such as working memory is involved in insulin signaling dysfunction and blood glucose levels. It was reported that working memory is linked with T2DM^{20–22}.

The shared TTC3, FoxO, MAPK, and PI3K-AKT Pathways between AD and T2DM were also identified. The tetratricopeptide repeat domain 3 (TTC3) gene was an AD causing gene (P-value for causation of AD < 0.0001), but not directly connected to T2DM (P-value for causation of T2DM = 0.47). TTC3 is associated with differentiation of neurons²³. It is reported that a rare TTC3 variant is related with AD²⁴. The TTC3-RhoA pathway could be a key determinant of the neuronal development, resulting in detrimental effects on the normal differentiation program²⁵. Rho regulates the activation of MAPK pathway²⁶. The Forkhead box O (FoxO) transcription factors that affect nervous system amyloid (A β) production, are implicated in the regulation of cell apoptosis and survival, and accelerate the progression of degenerative disease. FoxO pathway is involved in the PI3K/Akt and mitogen-activated protein kinase (MAPK) pathways in neuronal apoptosis in the brain.

FoxOs also can offer protection in the nervous system, reduce toxic intracellular protein accumulations and potentially effect A β toxicity^{19,27,28}. Akt-FoxO that suppresses TLR4 signaling in Human Leukocytes is implicated in the development of T2DM²⁹. Increasing evidences indicate that PI3K/AKT pathway are implicated in the development of T2DM^{30,31}.

Our results further supported that morphine addiction and neuroactive ligand receptor interaction pathways were shared between AD and T2DM. Morphine addiction has neurotoxic effects and damages to the brain regions that function for learning, memory and emotions³². High dose of morphine may increase risk to T2DM³³. It is also reported that neuroactive ligand receptor interaction pathway is associated with both AD and T2DM³⁴.

The causal network analysis provided evidence that fatty acid biosynthesis and primary bile acid biosynthesis pathways were shared between AD and T2DM. Brain function such as intelligence, memory, behavior and concentration are all influenced by brain nutrition³⁵. Omega-3 fatty acids affect the fluidity of brain cell membranes, neurotransmitter synthesis and signal transmission and are implicated in AD^{36,37}. Bile acids are involved in cell signaling and immune function. It performs as potent inhibitors of apoptosis and regulates transcriptional and post-transcriptional events that affect mitochondrial function in neurons³⁸. A trend of increased bile acids in AD has been observed³⁹. Fatty acid utilization induces insulin resistance⁴⁰. Bile acids are signal molecules and play an important role in regulating metabolism and inflammation. The abnormal bile acids are correlated with changes in insulin secretion, which lead to T2DM^{41,42}. The amyloid precursor protein (APP) is a transmembrane protein. The aggregated amyloid- β (A β) peptides are generated by sequential proteolytic processing of the APP. Accumulation of A β and the APP play an important role in regulating lipid homeostasis including fatty acids, which finally affect the development of AD⁴³.

Finally, we showed that how the causal analysis identified the shared methylated genes POU3F2, KIF4B and TNSL3, and dopaminergic synapse and AMPK pathways between AD and T2DM. Emerging evidences indicate that methylation alternations to DNA of the brain are linked to Alzheimer's disease^{44,45}. DNA methylation also plays an important role in the pathogenesis of T2DM^{45,46}. In order to better understand the etiology of AD and T2DM, we jointly investigated the genetic variants, DNA methylation and gene expression profiles, multiple phenotypes, AD and T2DM using causal inference pipelines. We found that gene *POU3F2* regulated methylations of POU3F2, KIF4B and TMSL3. Alternations in methylation of three genes directly caused the development of AD and T2DM. Furthermore, methylation levels of three genes regulated gene expressions in dopaminergic synapse and AMPK pathways, which in turn caused AD and T2DM via CHL/HDL Ratio (Fig. 4A). Recent advance revealed that alterations of the dopaminergic system contributes to memory and reward dysfunction and the dopaminergic system may well be involved in the occurrence of AD^{47,48}. Recent studies also unravel that the brain damage in AD is linked to an over-activation of AMPK, which leads to the loss of the ability of neurons to grow axons and the modification of the tau proteins resulting in tangles of tau⁴⁹. AMPK functions as a key energy sensor. AMPK signaling elicits insulin-sensitizing effects and may be implicated in stimulating glucose up taking in skeletal muscles, fatty acid oxidation in adipose (and other) tissues⁵⁰.

We identified an extremely large number of shared causal paths from genetic variants to both AD and T2DM via DNA methylation, gene expressions and phenotypes. This deep knowledge that uncovered the large number of causal mechanisms of AD and T2DM has profound implications in prevention and treatments of AD and T2DM. This explained why the drugs that were based on inhibition or activation of limited number of paths often failed simply because these limited number of paths cannot cover all causal paths to the diseases. Finally, the empirical evidence that the AD and T2DM shared a large number of causal genes, gene expressions, methylations and pathways supported hypothesis that AD can be considered as “type 3 diabetes”.

Methods

All methods were carried out in accordance with relevant guidelines and regulations.

ROSMAP data. The data came from two longitudinal cohort studies of older persons. ROS started in 1994 and enrolled Catholic nuns, priests, and brothers from more than 40 communities across United States, and MAP started in 1997 and enrolled participants with diverse backgrounds and socioeconomic status from continuous care retirement communities throughout northeastern Illinois, as well as from individual homes across the Chicago metropolitan area¹⁹. These two studies are managed by the same team of investigators. Structured, quantitative neuropathological examinations are performed at a single site. Therefore, the data can be combined for analysis. Multi-layered omics datasets are generated from biospecimens donated by ROS and MAP participants, including genotypes, DNA methylation profiles and RNA-seq. The genotype data were generated by Affymetrix or the Illumina OmniQuad express gene chips and were imputed using the 1000 Genomes Project data as reference. DNA methylation profiles were measured using the Illumina Infinium HumanMethylation450 beadset. RNA-seq data were generated using the Illumina HiSeq with 101 bp paired-end reads. Multiple phenotypes including clinical diagnosis, cognitive function, measures of lifestyle, behavior, and activity, chronic medical conditions and risk factors were measured. A total of 432 individuals who simultaneously had genotype, RNA-seq, DNA methylation and some phenotypes were included in analysis. We considered 19 phenotypes and environments, two diseases (AD, T2DM), 299 pathways with RNA-Seq in KEGG pathway database, 20,242 methylation genes with 364,661 CpG sites, and 51,060 genotyped genes with 5,711,541 SNPs (4,283,876 common SNPs, 1,427,665 rare SNPs). All the data were downloaded from <https://www.radc.rush.edu/>.

The ROSMAP studies were approved by the Institutional Review Board of Rush University Medical Center. Written informed consent was obtained from all subjects, followed by an Anatomic Gift Act for organ donation.

General procedures for identifying shared genetic loci underlying AD and T2DM. AD and T2DM result from the interplay of DNA sequence variation and nongenetic factors acting through molecular networks^{51–53}. Their etiology is complex with many intermediate steps between genetic variation and diseases. Neither traditional GWAS, nor classical multi-omics analysis can identify the causal passes of complex diseases because not all these analyses can identify directed routes from genetic loci to diseases through environments, methylations, gene expressions, and phenotypes. To overcome these limitations, we developed a novel general framework for identifying all possible causal passes from genetic loci to diseases. The framework consists of three steps. The first step is to perform genome-wide causation studies (GWCS) where we test causation of each SNP across the genome to the disease. The additive noise model (ANM) with discrete variants will be used to test for causation⁵⁴ (Methods). We focused on the rare variants in the paper. The second step is to use integer programming (IP) and various modern causal models^{55–57} (Methods) for inferring multilevel genome-wide omics causal

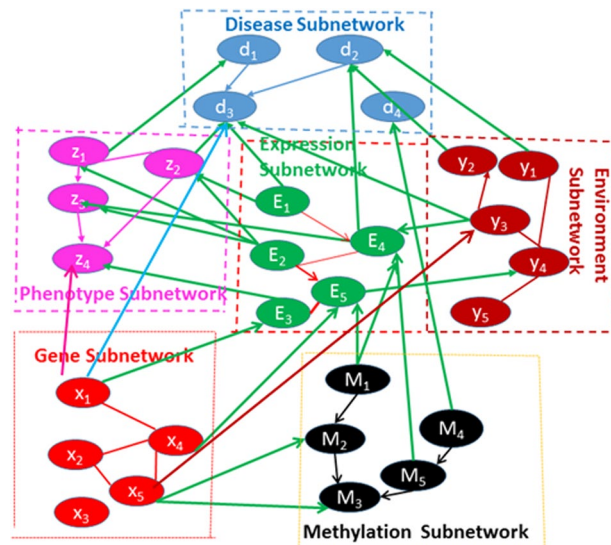


Figure 5. Scheme of multilevel omic networks.

networks that integrate genotype subnetworks, environmental subnetworks, methylation subnetworks, gene regulatory subnetworks, intermediate phenotype subnetworks and multiple disease subnetworks into a single connected multilevel genotype-disease network as shown in Fig. 5. The third step is to augment graph theoretical approaches with approximations for developing efficient search algorithms that discover all possible routes starting from the genetic variant node directed to the disease node, including classical Depth First Search (DFS) and Breadth First Search (BFS) algorithms^{58–61}.

There are two ways to identify shared dysfunctional genes (SNPs) between AD and T2DM. One way is to use ANM with discrete variables and functional data analysis to conduct genome-wide causation analysis^{54,62–64} for unravelling the direct connections between gene nodes and disease nodes to identify the shared dysfunctional genes between AD and T2DM.

Another way is to search the paths from the gene nodes to AD and T2DM in multilevel causal omics networks.

Association and causation are different concepts. Association between two variables is often characterized by dependence between two variables. Causation is a connection of phenomena where one variable acts or intervenes on another variables and leads to its changes. Therefore, the key component of causation is the generation and determination of values of one variable by another. The mechanism of causation is related to the transference of matter, motion and information. Causation as part of universe connection, is well known that nature consists of autonomous and independent causal generating process modules. These modules will not influence each other^{63,65}. In other words, while output of one module may inform or influence input of another module, the events between modules are independent. In the probabilistic language, mechanism is often represented by conditional distribution. Independent mechanism states that “the conditional distribution of each variable given its causes (i.e., its mechanism) does not inform or influence the other conditional distributions”⁶³. In GWCS, we only consider two variables. In this case, independence of cause and mechanism (ICM) indicates that the conditional distribution of the effect given its cause is independent of distribution of cause. Consider the genetic analysis of alleles (A) with a disease allele A a normal allele a and with the disease (D) (disease D and normal d). The joint density function $P(a, d)$ can be decomposed into

$$\begin{aligned} P(A, D) &= P(A)P(D|A) \\ &= P(D)P(A|D). \end{aligned}$$

In the association analysis, we assess whether A is independent of D or not. The relationship between A and D is symmetric. However, in causal analysis, causations $A \rightarrow D$ and $D \rightarrow A$ are different. They are asymmetric. Assessing causation is to consider the effect of intervention. Causation $A \rightarrow D$ indicates that the effect of A is to give rise to disease. However, disease status D will not generate allele A . Suppose that locus A is disease locus and $A \rightarrow D$. If we change the allele a to allele A , then we assume that biological mechanism $P(D|A)$ responsible for giving rise to disease. This would hold true independent of the distribution (frequencies) of allele A . If the locus A is disease locus, we can find that the distributions (frequencies) of allele A in two different populations are different, but the mechanism $P(D|A)$ would apply in two populations. The conditional probability $P(D|A)$ can also be viewed as penetrance of the allele. The marginal distribution $P(A)$ and conditional distribution $P(D|A)$ contain no information about each other. Both continuous and discrete ANMs satisfy the ICM and will be used for GWCS. The proposed method for genome-wide causation analysis and inferring multilevel causal genotype-methylation-expression-phenotype-disease network was applied to the ROSMAP dataset¹⁹ with 432 individuals, 19 phenotypes and environments, two diseases (AD, T2DM), 299 pathways with RNA-Seq in KEGG pathway database, 20,242 methylation genes with 364,661 CpG sites, and 51,060 genotyped genes with 5,711,541

SNPs (4,283,876 common snp, 1,427,665 rare snp) (imputed by 1000 Gnome Data). The inferred genotype-expression-methylation-phenotype-disease network consisted of 2,814 nodes and 22,184 edges where the edges were presented in the network if the path coefficients were significantly from zero with P-values < 0.05.

There were two ways to connect a gene (or SNP) to AD (T2DM). If a gene (or SNP) showed causation to AD (T2DM) by statistical causal test, then the gene (SNP) was directly connected to AD (T2DM) in the causal network. Such gene (SNP) was called AD (T2DM) directly connected gene (SNP). We may observe the connection between a gene (SNP) and AD (T2DM) via multiple edges (paths) in the constructed multilevel causal network. Then, the gene (SNP) that was indirectly connected to AD (T2DM) via paths in the multilevel causal network was called AD (T2DM) indirectly connected gene (SNP).

Genome-wide causation studies. Unlike GWAS where we test the association of each variant across the genome with the disease, genome-wide causation studies (GWCS) test the causation of each variant across the genome to the disease. Additive noise models (ANMs) with discrete variables will be used for GWCS^{54,62–64}. The procedures that use the ANMs for GWCS are summarized as follows^{14,54,63}.

Procedures for causal genetic analysis using ANM.

1. Fit the following nonlinear integer regression to the data.

$$Y = f(X) + N_Y.$$

Calculate the residuals $\hat{N}_Y = Y - \hat{f}(X)$.

2. Fit the following nonlinear integer regression to the data.

$$X = g(Y) + N_X.$$

Calculate the residuals $\hat{N}_X = X - \hat{g}(Y)$.

3. Test for independence.

The contingency table and Fisher's exact test can be used to test independence. Let the statistic for testing the independence between \hat{N}_Y and X as $\Delta_{X \rightarrow Y}$ and the statistic for testing the independence between \hat{N}_X and Y as $\Delta_{Y \rightarrow X}$.

The null hypothesis for testing the causation of the variant is

H_0 : no causation between variables X and Y .

The statistic for testing the causation between two X and Y is defined as

$$T_C = |\Delta_{X \rightarrow Y} - \Delta_{Y \rightarrow X}|.$$

When T_C is large, the causation between genetic variant X and disease status Y exists. When $T_C \approx 0$, this indicates that no causal decision can be made. Since the distribution of the test statistic T_C is difficult to calculate, P-value for testing the causation of the variant X can be calculated by permutations.

To improve the performance of causation analysis of rare variants, we first calculate the functional principle component score (FPCS) of the rare variants within a gene⁶⁴ to summarize information of all rare variants within the gene. Then, the continuous FPCS are discretized. Finally, the ANMs with discrete variables can be used to test causation of discretized FPCS with the disease.

Structural equations for construction of causal networks. Directed graphical models and structural equations can be used as a tool to model the complex causal structures among variables^{64,66}. A graphical model consists of nodes and edges. The nodes represent variables and edges represent the dependence structures among variables. A directed graphic model is defined as the graph in which all the inter-node connections have a direction visually denoted by an arrowhead. Directed acyclic graphics (DAGs) are defined as directed graphics with no cycles. In other words, we can never start at a node X , travel edges in the directions of the arrows and get back to the node X . A DAG with nodes encodes conditional dependence structure of the variables Y_1, \dots, Y_n . We define the parents of a node as the nodes pointing directly to it. The concept of parents provides an easy way to read off conditional independence from DAGs.

Traditional regressions describe one-way or unidirectional relationships among variables in which the variables on the left sides of the equations are dependent variables and the variables on the right sides of the equations are explanatory variables or independent variables. The explanatory variables are used to predict the outcomes of the dependent variables. However, in many cases, there are two ways, or simultaneous relationships between the variables. Variables in some equations are response variables, but will be predictors in other equations. The variables in equations may influence each other. It is difficult to distinguish dependent variables and explanatory variables. The structural equation models (SEMs) are a powerful mathematic tool to describe such data generating mechanism and infer causal relationships among the variables.

The SEMs classify variables into two class variables: endogenous and exogenous variables. The jointly dependent variables that are determined in the model are called endogenous variables. The explanatory variables that are determined outside the model or predetermined are called exogenous variables. In the genotype-phenotype networks, the phenotype variables such as BMI, cognitive function, working memory, are endogenous variables, age, sex, race, environments and genotypes are exogenous variables. In the genotype-expression networks, the gene expressions are endogenous variables and genotypes are exogenous variables. In the methylation-expression networks, gene expressions are endogenous variables and methylations are exogenous variables.

where Q and q are two alleles of the marker at the genomic position t , $P_Q(t)$ and $P_q(t)$ are the frequencies of the alleles Q and q , respectively. Suppose that we are interested in k genomic regions or genes $[a_j, b_j]$ denoted as T_j , $j = 1, \dots, K$. We consider the following functional structural equation models (FSEMs):

$$\begin{aligned} y_1\gamma_{11} + y_2\gamma_{21} + \dots + y_M\gamma_{M1} + \int_{T_1} x_1(t)\beta_{11}(t)dt + \dots + \int_{T_k} x_k(t)\beta_{k1}(t)dt + e_1 &= 0 \\ \vdots & \\ y_1\gamma_{1M} + y_2\gamma_{2M} + \dots + y_M\gamma_{MM} + \int_{T_1} x_1(t)\beta_{1M}(t)dt + \dots + \int_{T_k} x_k(t)\beta_{kM}(t)dt + e_M &= 0 \end{aligned} \quad (6)$$

where $\beta_{ij}(t)$, $j = 1, \dots, k$, $i = 1, \dots, M$ are genetic effect functions.

Functional principal components (FPCs) are efficient summary statistics. The FPCs simultaneously employ genetic information of the individual variants and correlation information (LD) among all variants. For each genomic region or gene, we use functional principal component analysis to calculate principal component function. Let N be the number of sampled individuals. We expand $x_{nj}(t)$, $n = 1, \dots, N$, $j = 1, \dots, k$ in each genomic region in terms of orthogonal principal component functions:

$$x_{nj}(t) = \sum_{l=1}^{L_j} \eta_{njl} \phi_{jl}(t), \quad j = 1, \dots, k,$$

where $\phi_{jl}(t)$, $j = 1, \dots, k$, $l = 1, \dots, L_j$ are the l -th principal component function in the j -th genomic region or gene and η_{njl} are the functional principal component scores of the n -th individual. Using the functional principal component expansion of $x_{nj}(t)$, we can transform the FSEMs (6) into the traditional multivariate SEMs (1).

Integer programming for causal network learning. Given the dataset, learning causal networks is the task of finding network structures that best fits the data^{57,64}. We used “score and search” methods to learn causal networks via maximizing the score metrics that characterize the causal networks. The “score and search” algorithms consist of two parts: (1) formulate objective function (global score for the whole network) using the score function for each node and (2) search algorithm.

We collected all nodes with directed edges in the causal network into a DAG, denoted as $G = (V, E)$. The score (objective function) for the DAG G was defined as

$$\text{Score}(G) = \sum_{j \in V} \text{Score}_j(G)$$

where $\text{Score}_j(G)$ was a score for the node j in the network. The $\text{Score}_j(G)$ was calculated as $f(\Delta_j)$ via solving the optimization problem (5). Therefore, the total score can be decomposed into a sum of score for all nodes in the DAG. In addition, the $\text{Score}_j(G)$ is entirely determined by the parent set of the node j in G . A DAG can be encoded by the set $W = \{W_1, \dots, W_p\}$ of parent variables for all nodes V in the graph G . We use $C(j, W_j)$ to denote a score function for the pair of node j and its parent set W_j . Therefore, the total score for the DAG G was given by

$$C(D) = \sum_{i \in V} C(v, W_v).$$

The learning task is to find a DAG that optimizes the global score $C(D)$ over all possible DAGs D or parent sets⁵⁷:

$$\min_D \sum_{i \in V, W_v \in D} C(v, W_v).$$

Integer linear programming (ILP) was used as a search algorithm⁵⁷. A DAG learning was formulated as the ILP as follows. We define a variable $x(W_v \rightarrow v)$ to indicate the presence or absence of the parent set W_v in the DAG. In other words, $x(W_v \rightarrow v) = 1$ if and only if it is the parent set for the node v . The parent set W_v can be an empty set. The objective function for the ILP formulation of a DAG learning can be defined as

$$\sum_{v=1}^p \sum_{j_v=1}^{I_v} C(v, W_{j_v}) x(W_{j_v} \rightarrow v). \quad (7)$$

The goal was to find a candidate parent set W_v for each node v by optimizing the objective function in (7). It is clear that every DAG can be encoded by a zero-one indicator variable. However, any set of zero-one numbers may not encode a DAG. A set of linear constraints must be posted to make the set of indicator variables to represent a DAG. Without constraints all indicator variables for the parent sets will be equal to either zero or one. These solutions will not form a DAG. The constraints need to be imposed to ensure that the solutions encode a DAG. This constraint that is referred to as convexity constraint, can be expressed as

$$\sum_{i_j=1}^{I_j} x(W_{i_j} \rightarrow j) = 1, \quad j = 1, \dots, p \quad (8)$$

The convexity constraints (8) can define a directed graph. However, the generated directed graph may have cycles. To eliminate a cycle, we need to impose the following constraint to ensure that any subset C of the nodes V in a DAG must contain at least one node that has no parent in the subset C

$$\forall C \subseteq \sum_{j \in CW: W \cap C = \emptyset} x(W \rightarrow j) \geq 1, \tag{9}$$

which is referred to as cluster-based constraints. Our goal is to find a candidate parent set W_j for each node j by optimizing objective function (7) subject to the constraints (8) and (9).

The branch and bound method is a popular algorithm ensured to find an optimal solution to the 0–1 ILP problem⁵⁷. Let the LP solution represent “solution of the current linear relaxation”. The basic idea of the branch and bound method is to successively divide the ILP problem into smaller problems that are easy to solve and reduce the search space. Briefly, the branch and bound algorithm is summarized as follows. Step 1: Let \hat{x} be the LP solution. Step 2: if there are, valid constraints not satisfied by \hat{x} add them and go to Step 1; otherwise if the solution \hat{x} is an integer then stop, the current problem is solved; otherwise branch on a variable with a non-integer part in \hat{x} to generate two new sub-IP problems. We then again use branch and bound algorithms to solve two sub-ILP problems⁵⁷.

Multilevel causal networks. Multilevel causal omics networks integrated genotype subnetworks, methylation subnetworks, gene expression subnetworks, the intermediate phenotype subnetworks and multiple disease subnetworks into a single connected multilevel genotype-disease networks to reveal the deep causal chain of mechanisms underlying the diseases⁶⁴. ILP was extended from a single causal network estimation to joint multiple causal network estimations to integrate genomic, epigenomic and phenotype data.

For the convenience of discussion, consider M gene expression variables Y_1, \dots, Y_M , Q methylation variables Z_1, \dots, Z_Q , and K genotype variables X_1, \dots, X_K . Let $pa_D(D)$ be the parent set of the node d including gene expression, methylation and genotype variables. Consider three types of SEMs. First, we consider a general SEM model for the gene expression:

$$Y_d = \sum_{i \in pa_D(D)} f_{di}(Y_i) + \sum_{q \in pa_D(D)} f_{dq}(Z_q) + \sum_{j \in pa_D(D)} f_{dj}(X_j) + \varepsilon_d, \quad d = 1, \dots, M \tag{10}$$

And

$$Z_q = \sum_{l \in pa_Q(q)} f_{ql}(Z_l) + \sum_{m \in pa_Q(q)} f_{qm}(X_m) + \varepsilon_q, \quad q = 1, \dots, Q \tag{11}$$

where f_d and f_q are linear functions from $R^{|pa_D|} \rightarrow R$ and $R^{|pa_Q|} \rightarrow R$, respectively, and the errors ε_d and ε_q are independent, following distributions P_{ε_d} and P_{ε_q} , respectively. Equation (10) define a causal network that connects gene expressions, methylations and genotypes. Equation (11) define a causal network that connects methylations and genotypes.

Integer programming as a general framework for joint estimation of multiple causal networks. We collected multiple types of data: genotype, gene expression, methylation, and phenotype and disease data. We wanted to estimate multiple causal networks with different types of data⁶⁴.

The scores of the nodes Y_d and Z_q were, respectively, given by

$$C(Y_d, W_{di}) = Y_d^T (I - D_Y^i (D_Y^i)^T D_Y^i)^{-1} (D_Y^i)^T Y_d \tag{12}$$

and

$$C(Z_q, W_{qi}) = Z_q^T (I - D_Z^l (D_Z^l)^T D_Z^l)^{-1} (D_Z^l)^T Z_q \tag{13}$$

where matrices D_Y^i and D_Z^l corresponded to the parent sets W_{di} and W_{qi} .

Let V_E be the set of nodes in the gene expression network and V_M be the set of nodes in the methylation network. Let C_E be a subset of nodes in V_E and C_M be a subset of nodes in V_M . A joint expression and methylation causal network can be formulated as the following ILP:

$$\begin{aligned} \min & \sum_{d=1}^M \sum_{i \in pa_D(d)} C(d, W_{di}) x(W_{di} \rightarrow d) + \sum_{q=1}^Q \sum_{l \in pa_Q(q)} C(q, W_{qi}) x(W_{qi} \rightarrow q) \\ \text{s.t.} & \sum_{i \in pa_D(d)} x(W_{di} \rightarrow d) = 1, \quad d = 1, \dots, M, \\ & \sum_{i \in pa_Q(q)} x(W_{qi} \rightarrow q) = 1, \quad q = 1, \dots, Q, \\ & \forall C_E \subseteq V_E: \sum_{d \in C_E} \sum_{W_{di}: W_{di} \cap C_E = \emptyset} x(W_{di} \rightarrow d) \geq 1, \\ & \forall C_M \subseteq V_M: \sum_{q \in C_M} \sum_{W_{qi}: W_{qi} \cap C_M = \emptyset} x(W_{qi} \rightarrow q) \geq 1. \end{aligned} \tag{14}$$

Using branch and bound and other methods for solving the ILP, we can solve the ILP problem (14) to obtain the best joint causal genotype-methylation-expression and genotype-methylation network fitting the data.

Summary statistics for representation of groups of gene expressions. Generalized low rank models were used to segment (cluster) the data. Principal component analysis (PCA) was used to reduce data dimensions. The PCs were used to summarize the gene expression data in pathways and clusters⁶⁹.

Simulations of causal networks. We simulated causal networks with genes (genotype) and gene expressions as the nodes of the networks. We randomly selected 8 genes (30 node model) and 10 genes (50 node model) from the ROSMAP dataset. The genotype information of multiple SNPs within a gene was summarized by FPCA scores which were taken as the values of the gene node. We used R package PCALG^{70,71} to randomly generate DAG with 30 nodes (edges ranging from 70 to 90), and with 50 nodes (edges ranging from 80 to 110). The values of the gene expression nodes were generated by the following model^{66,70-72}

$$y_i = \sum_{j \in pa(y_i)} \gamma_{ji} y_j + \sum_{k \in pa(x_i)} \beta_{ki} x_k + e_i, \quad i = 1, \dots, M,$$

where $pa(y_i)$ is the set of parents of the node y_i , the coefficients γ_{ji} and β_{ki} followed a uniform distribution $u(1, 2)$, e_i followed a normal distribution $N(0, 1)$.

A total of 100, 300, 500 and 1,000 DAGs were generated. The number of replication was 1,000. Let N_i be the total number of edges among simulated DAGs, N_0 the total number of edges that were not presented in the simulated DAGs, N_{True} the total number of edges detected by the algorithm and N_{False} the false edges directed among N_0 . Then the false discovery rate (FDR) was defined as $\frac{N_{False}}{N_0}$ and power of detection defined as $\frac{N_{True}}{N_i}$.

Ethical approval and informed consent. The ROSMAP studies were approved by the Institutional Review Board of Rush University Medical Center. Written informed consent was obtained from all subjects, followed by an Anatomic Gift Act for organ donation.

Data availability

All data are publicly available and can be downloaded from RADc Research Resource Sharing Hub (<https://www.radc.rush.edu/>).

Received: 19 September 2019; Accepted: 3 February 2020;

Published online: 05 March 2020

References

- Zhuang, Q.-S., Zheng, H., Gu, X.-D., Shen, L. & Ji, H.-F. Detecting the genetic link between Alzheimer's disease and obesity using bioinformatics analysis of GWAS data. *Oncotarget* **8** (2017).
- Song, M. K., Bischoff, D. S., Song, A. M., Uyemura, K. & Yamaguchi, D. T. Metabolic relationship between diabetes and Alzheimers Disease affected by Cyclo(His-Pro) plus zinc treatment. *BBA Clinical* **7**, 41–54 (2017).
- Lashley, T. *et al.* Molecular biomarkers of Alzheimers disease: progress and prospects. *Disease Models & Mechanisms* **11**, dmm031781 (2018).
- Fischer, R. & Maier, O. Interrelation of Oxidative Stress and Inflammation in Neurodegenerative Disease: Role of TNF. *Oxidative Medicine and Cellular Longevity* **2015**, 1–18 (2015).
- Li, X., Leng, S. & Song, D. Link between type 2 diabetes and Alzheimer's disease: from epidemiology to mechanism and treatment. *Clinical Interventions in Aging* **549**. <https://doi.org/10.2147/cia.s74042> (2015).
- Baglietto-Vargas, D., Shi, J., Yaeger, D. M., Ager, R. & Laferla, F. M. Diabetes and Alzheimer's disease crosstalk. *Neuroscience & Biobehavioral Reviews* **64**, 272–287 (2016).
- Pugazhenthil, S., Qin, L. & Reddy, P. H. Common neurodegenerative pathways in obesity, diabetes, and Alzheimers disease. *Biochimica et Biophysica Acta (BBA). Molecular Basis of Disease* **1863**, 1037–1045 (2017).
- Akter, K. *et al.* Diabetes mellitus and Alzheimers disease: shared pathology and treatment? *British Journal of Clinical Pharmacology* **71**, 365–376 (2011).
- Arvanitakis, Z., Wilson, R. S., Bienias, J. L., Evans, D. A. & Bennett, D. A. Diabetes Mellitus and Risk of Alzheimer Disease and Decline in Cognitive Function. *Archives of Neurology* **61**, 661 (2004).
- Arvanitakis, Z. *et al.* Diabetes is related to cerebral infarction but not to AD pathology in older persons. *Neurology* **67**, 1960–1965 (2006).
- Talbot, K. *et al.* Demonstrated brain insulin resistance in Alzheimer's disease patients is associated with IGF-1 resistance, IRS-1 dysregulation, and cognitive decline. *Journal of Clinical Investigation* **122**, 1316–1338 (2012).
- Hohman, T. J. *et al.* GSK3 β Interactions with Amyloid Genes: An Autopsy Verification and Extension. *Neurotoxicity Research* **28**, 232–238 (2015).
- Karki, R., Kodamullil, A. T. & Hofmann-Apitius, M. Comorbidity Analysis between Alzheimer's Disease and Type 2 Diabetes Mellitus (T2DM) Based on Shared Pathways and the Role of T2DM Drugs. *Journal of Alzheimers Disease* **60**, 721–731 (2017).
- Jiao, R. *et al.* Bivariate Causal Discovery and Its Applications to Gene Expression and Imaging Data Analysis. *Frontiers in Genetics* **9** (2018).
- Bennett, D. A., Schneider, J. A., Arvanitakis, Z. & Wilson, R. S. Overview and Findings from the Religious Orders Study. *Current Alzheimer Research* **9**, 628–645 (2012).
- Bennett, D. A. *et al.* Overview and Findings from the Rush Memory and Aging Project. *Current Alzheimer Research* **9**, 646–663 (2012).
- Kandel, E. R. The molecular biology of memory: cAMP, PKA, CRE, CREB-1, CREB-2, and CPEB. *Molecular Brain* **5**, 14 (2012).
- Saura, C. A. & Valero, J. The role of CREB signaling in Alzheimer's disease and other cognitive disorders. *Reviews in the Neurosciences* **22** (2011).
- White, C. C. *et al.* Identification of genes associated with dissociation of cognitive performance and neuropathological burden: Multistep analysis of genetic, epigenetic, and transcriptional data. *PLOS Medicine* **14** (2017).
- Cerasuolo, J. & Izzo, A. Persistent impairment in working memory following severe hyperglycemia in newly diagnosed type 2 diabetes. *Endocrinology, Diabetes & Metabolism Case Reports* **2017** (2017).

21. Huang, R.-R. *et al.* Spatial working memory impairment in primary onset middle-age type 2 diabetes mellitus: An ethology and BOLD-fMRI study. *Journal of Magnetic Resonance Imaging* **43**, 75–87 (2015).
22. Montoya, J. C. *et al.* Global differential expression of genes located in the Down Syndrome Critical Region in normal human brain. *Colombia Medica* 154–161. <https://doi.org/10.25100/cm.v45i4.1640> (2014).
23. Kohli, M. A. *et al.* Segregation of a rare TTC3 variant in an extended family with late-onset Alzheimer disease. *Neurology Genetics* **2** (2016).
24. Berto, G. *et al.* The Down syndrome critical region protein TTC3 inhibits neuronal differentiation via RhoA and Citron kinase. *Journal of Cell Science* **120**, 1859–1867 (2007).
25. Galaria, I. I., Nicholl, S. M., Roztocil, E. & Davies, M. Differential Regulation Of Erk1/2 And P38Mapk By Components Of The Rho Signaling Pathway During Sphingosine-1-Phosphate (Sip) - Induced Smooth Muscle Cell (Smc) Migration. *Cardiovascular Pathology* **13**, 47 (2004).
26. Maiese, K. Forkhead Transcription Factors: Formulating a FOXO Target for Cognitive Loss. *Current Neurovascular Research* **14**, 415–420 (2018).
27. Kim, J. H., Choi, J. S. & Lee, B. H. PI3K/Akt and MAPK pathways evoke activation of FoxO transcription factor to undergo neuronal apoptosis in brain of the silkworm *Bombyx mori* (Lepidoptera: Bombycidae). *Cell Mol Biol (Noisy-le-grand)*. Suppl.58:OL1780–1785 (2012).
28. Zhang, Z. *et al.* Insulin-Dependent Regulation of mTORC2-Akt-FoxO Suppresses TLR4 Signaling in Human Leukocytes: Relevance to Type 2 Diabetes. *Diabetes* **65**, 2224–2234 (2016).
29. Yin, X. *et al.* Association of PI3K/AKT/mTOR pathway genetic variants with type 2 diabetes mellitus in Chinese. *Diabetes Research and Clinical Practice* **128**, 127–135 (2017).
30. Kitagishi, Y. *et al.* Certain Diet and Lifestyle May Contribute to Islet β -cells Protection in Type-2 Diabetes via the Modulation of Cellular PI3K/AKT Pathway. *The Open Biochemistry Journal* **1**, 74–82 (2014).
31. Soleimani, S. M. A., Ekhtiari, H. & Cadet, J. L. Drug-induced neurotoxicity in addiction medicine. *Progress in Brain Research Neuroscience for Addiction Medicine: From Prevention to Rehabilitation - Constructs and Drugs* 19–41. <https://doi.org/10.1016/bs.pbr.2015.07.004> (2016).
32. Yang, S.-P. *et al.* Risk of type 2 diabetes mellitus in female breast cancer patients treated with morphine: A retrospective population-based time-dependent cohort study. *Diabetes Research and Clinical Practice* **110**, 285–290 (2015).
33. Trucco, M. Genetic and Environmental Pathways in Type 1 Diabetes Complications. <https://doi.org/10.21236/ada544029> (2009).
34. Jager, C. A. D. & Kovatcheva, A. Summary and discussion: Methodologies to assess long-term effects of nutrition on brain function. *Nutrition Reviews* **68** (2010).
35. Hooper, C., De Souto Barreto, P., Pahor, M., Weiner, M. & Vellas, B. The Relationship of Omega 3 Polyunsaturated Fatty Acids in Red Blood Cell Membranes with Cognitive Function and Brain Structure: A Review Focussed on Alzheimer's Disease. *J. Prev. Alzheimers Dis.* **5**(1), 78–84 (2018).
36. Grimm, M. O. W., Michaelson, D. M. & Hartmann, T. Omega-3 fatty acids, lipids, and apoE lipidation in Alzheimer's disease: a rationale for multi-nutrient dementia prevention. *Journal of Lipid Research* **58**, 2083–2101 (2017).
37. Ramalho, R. M., Viana, R. J., Low, W. C., Steer, C. J. & Rodrigues, C. M. Bile acids and apoptosis modulation: an emerging role in experimental Alzheimers disease. *Trends in Molecular Medicine* **14**, 54–62 (2008).
38. Pan, X. *et al.* Metabolomic Profiling of Bile Acids in Clinical and Experimental Samples of Alzheimer's Disease. *Metabolites* **7**, 28 (2017).
39. Bouchouirab, F.-Z., Fortin, M., Noll, C., Dubé, J. & Carpentier, A. C. Plasma Palmitoyl-Carnitine (AC16:0) Is a Marker of Increased Postprandial Nonesterified Incomplete Fatty Acid Oxidation Rate in Adults With Type 2 Diabetes. *Canadian Journal of Diabetes* **42** (2018).
40. Wang, S. *et al.* Plasma bile acid changes in type 2 diabetes correlated with insulin secretion in two-step hyperglycemic clamp. *Journal of Diabetes*. <https://doi.org/10.1111/1753-0407.12771> (2018).
41. Chávez-Talavera, O., Tailleux, A., Lefebvre, P. & Staels, B. Bile Acid Control of Metabolism and Inflammation in Obesity, Type 2 Diabetes, Dyslipidemia, and Nonalcoholic Fatty Liver Disease. *Gastroenterology* **152** (2017).
42. Grimm, M. O. W., Mett, J., Grimm, H. S. & Hartmann, T. APP Function and Lipids: A Bidirectional Link. *Frontiers in Molecular Neuroscience* **10** (2017).
43. Huang, Y.-T., Iwamoto, K., Kurosaki, T., Nasu, M. & Ueda, S. The neuronal POU transcription factor Brn-2 interacts with Jab1, a gene involved in the onset of neurodegenerative diseases. *Neuroscience Letters* **382**, 175–178 (2005).
44. Nagata, K., Mano, T., Murayama, S., Saido, T. C. & Iwata, A. DNA methylation level of the neprilysin promoter in Alzheimers disease brains. *Neuroscience Letters* **670**, 8–13 (2018).
45. Shen, J. & Zhu, B. Integrated analysis of the gene expression profile and DNA methylation profile of obese patients with type 2 diabetes. *Molecular Medicine Reports*. <https://doi.org/10.3892/mmr.2018.8804> (2018).
46. Elliott, H. R. *et al.* Role of DNA Methylation in Type 2 Diabetes Etiology: Using Genotype as a Causal Anchor. *Diabetes* **66**, 1713–1722 (2017).
47. Martorana, A. & Koch, G. Is dopamine involved in Alzheimers disease?. *Frontiers in Aging Neuroscience* **6** (2014).
48. Nobili, A. *et al.* Dopamine neuronal loss contributes to memory and reward dysfunction in a model of Alzheimer's disease. *Nature Communications* **8** (2017).
49. Domise, M. & Vingtdoux, V. AMPK in Neurodegenerative Diseases. *Experientia Supplementum AMP-activated Protein Kinase* 153–177. https://doi.org/10.1007/978-3-319-43589-3_7 (2016).
50. Saha, A., Coughlan, K., Valentine, R. & Ruderman, N. AMPK activation: a therapeutic target for type 2 diabetes? *Diabetes, Metabolic Syndrome and Obesity: Targets and Therapy* **241**. <https://doi.org/10.2147/dms.o.s43731> (2014).
51. Huan, T. *et al.* Integrative network analysis reveals molecular mechanisms of blood pressure regulation. *Molecular Systems Biology* **11**, 799 (2015).
52. Jiang, P. *et al.* A Systems Approach Identifies Networks and Genes Linking Sleep and Stress: Implications for Neuropsychiatric Disorders. *Cell Reports* **11**, 835–848 (2015).
53. Schwartz, S. M., Schwartz, H. T., Horvath, S., Schadt, E. & Lee, S.-I. A Systematic Approach to Multifactorial Cardiovascular Disease. *Arteriosclerosis, Thrombosis, and Vascular Biology* **32**, 2821–2835 (2012).
54. Peters, J., Janzing, D. & Scholkopf, B. Causal Inference on Discrete Data Using Additive Noise Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **33**, 2436–2450 (2011).
55. Peters, J., Janzing, D. & Schölkopf B. *Elements of causal inference: foundations and learning algorithms*. (The MIT Press., 2017).
56. Pearl, J. *Causality*. (Cambridge University Press, 2009).
57. Cussens, J. Integer Programming for Bayesian Network Structure Learning. *Quality Technology & Quantitative Management* **11**, 99–110 (2014).
58. Devasia, J. V. & Chandran, P. Inferring disease causing genes and their pathways: A mathematical perspective. arXiv:1611.02538. (2016).
59. Quek, L.-E. & Nielsen, L. K. A depth-first search algorithm to compute elementary flux modes by linear programming. *BMC Systems Biology* **8** (2014).
60. Jindalertudomdee, J., Hayashida, M. & Akutsu, T. Enumeration Method for Structural Isomers Containing User-Defined Structures Based on Breadth-First Search Approach. *Journal of Computational Biology* **23**, 625–640 (2016).

61. Tang, X., Wang, J., Li, M., He, Y. & Pan, Y. A Novel Algorithm for Detecting Protein Complexes with the Breadth First Search. *BioMed Research International* **2014**, 1–8 (2014).
62. Janzing, D. & Steudel, B. Justifying Additive Noise Model-Based Causal Discovery via Algorithmic Information Theory. *Open Systems & Information Dynamics* **17**, 189–212 (2010).
63. Nonlinear causal discovery with additive noise models. *Advances in Neural Information Processing Systems*. (The MIT Press., 2017).
64. Xiong, M. Big data in omics and imaging: integrated analysis and causal inference. (CRC Press, 2018).
65. Parascandolo, G., Kilbertus, N., Rojas-Carulla, M. & Schölkopf, B. Learning Independent Causal Mechanisms. In Proceedings of the 35th International Conference on Machine Learning (ICML), **80**, pages: 4033–4041, Proceedings of Machine Learning Research, (Editors: Dy, Jennifer and Krause, Andreas), PMLR (2018).
66. Wang, P., Rahman, M., Jin, L. & Xiong, M. A new statistical framework for genetic pleiotropic analysis of high dimensional phenotype data. *BMC Genomics* **17** (2016).
67. Boyd, S. Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers. *Foundations and Trends[®] in Machine Learning* **3**, 1–122 (2010).
68. Parikh, N. Proximal Algorithms. *Foundations and Trends[®] in Optimization* **1**, 127–239 (2014).
69. Udell, M., Horn, C., Zadeh, R. & Boyd, S. Generalized Low Rank Models. *Foundations and Trends[®] in Machine Learning* **9**, 1–118 (2016).
70. Kalisch, M., Mächler, M., Colombo, D., Maathuis, M. H. & Bühlmann, P. Causal Inference Using Graphical Models with the R Packagepcalg. *Journal of Statistical Software* **47** (2012).
71. Hauser, A & Bühlmann, P. Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. arXiv:1104.2808 (2012).
72. Yokoyama, A. S., Rutledge, J. C. & Medici, V. DNA methylation alterations in Alzheimer's disease. *Environmental Epigenetics* **3** (2017).

Acknowledgements

Authors would like to thank Nan Lin, Tao Xu and Kelin Xu for their generous suggestion and support. Authors also thank two anonymous reviewers for the thorough reading of the manuscript and thoughtful suggestions that improved the manuscript.

Author contributions

Z.H., R.J., P.W., Y.Z. developed software and conducted data analysis, M.X. designed project and wrote manuscript, L.J., J.Z. and J.W. designed the project, D.B. provided data and wrote manuscript, P.J. provided data.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41598-020-60682-3>.

Correspondence and requests for materials should be addressed to M.X.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2020