communications biology

A Nature Portfolio journal

Article

https://doi.org/10.1038/s42003-025-07941-z

A personalized metabolic modelling approach through integrated analysis of RNA-Seq-based genomic variants and gene expression levels in Alzheimer's disease

Dilara Uzuner Odongo, Atılay İlgün, Fatma Betül Bozkurt & Tunahan Çakır

Generating condition-specific metabolic models by mapping gene expression data to genome-scale metabolic models (GEMs) is a routine approach to elucidate disease mechanisms from a metabolic perspective. On the other hand, integrating variants that perturb enzyme functionality from the same RNA-seq data may enhance GEM accuracy, offering insights into genome-wide metabolic pathology. Our study pioneers the extraction of both transcriptomic and genomic data from the same RNA-seq data to reconstruct personalized metabolic models. We map genes with significantly higher load of pathogenic variants in Alzheimer's disease (AD) onto a human GEM together with the gene expression data. Comparative analysis of the resulting personalized patient metabolic models with the control models shows enhanced accuracy in detecting AD-associated metabolic pathways compared to the case where only expression data is mapped on the GEM. Besides, several otherwise would-be missed pathways are annotated in AD by considering the effect of genomic variants.

Alzheimer's Disease (AD) is a complex disease without exact treatment. A systems-level perspective is required to discover the molecular mechanisms driving the multifaceted pathogenesis of the disease. Recent research on AD revealed that cellular metabolism is highly affected in the disease. The brain of AD patients shows mitochondrial dysfunction[1], impairment of lipid metabolism[2] and reduced overall energy metabolism[3]. In the light of these findings, therapeutic strategies for AD have also shifted focus to the metabolism. As an example, the Drug Repurposing for Effective Alzheimer's Medicines (DREAM) study is a large scale study for drug repurposing on AD and related dementias that target metabolic abnormalities in AD[4].

Genome-scale metabolic models (GEMs) are powerful instruments to study metabolic events in the cell comprehensively. GEMs describe all known metabolic reactions and their gene associations for an organism, enabling the analysis of alterations in metabolic reactions and pathways at different conditions including disease states. Algorithms are available to integrate transcriptomic data with GEMs to generate reduced, condition-specific metabolic models[5]. iMAT[6] is one of the most widely used integration algorithms for mammalian cells since it does not require any specific measurement constraints and any definition of biological objective in terms of metabolic reactions, which is the case for the majority of such algorithms. Recently, Baloni et al.[7] used iMAT to construct AD models and showed that bile acid and cholesterol metabolism reactions were commonly active in different brain-region specific metabolic models. Also, Moolamalla and Vinod[8] used the iMAT algorithm to generate metabolic models for different neuropsychiatric diseases like schizophrenia, bipolar disorder, and major depressive disorder.

Whole genome sequencing and whole exome sequencing are popular methods to detect single-nucleotide variants (SNVs) on the genome. RNAseq data can be used to identify pathogenic variants based on the sequencing of exonic regions. This makes RNAseq data unique since both gene expression levels and pathogenic variants can be identified from the same sample[9]. Different studies in the literature have already reported the application of variant discovery from RNAseq data in different diseases[10,11]. However, the use of genomic and transcriptomic information content of RNAseq data together to generate condition-specific genome-scale metabolic models have remained unexplored to date.

In this study, we used the iMAT algorithm to generate personalized genome-scale metabolic models for each individual in three different large

Department of Bioengineering, Gebze Technical University, Gebze, Kocaeli, Turkey. e-mail: tcakir@gtu.edu.tr

AD study cohorts; The Religious Orders Study (ROS) and Rush Memory and Aging Project (MAP)[12], Mayo Clinic[13] and Mount Sinai Brain Bank (MSBB)[14], which collectively cover a total of 643 AD individuals. For the first time in the literature, we extracted both transcriptomic and genomic information from the same RNA-seq sample to generate personalized metabolic models. We show that the consideration of genes with significantly higher load of pathogenic variants in AD state considerably improves metabolic models by capturing a number of otherwise missed AD-associated metabolic alterations at pathway level.

Methods

Transcriptome datasets

Three most commonly used RNAseq-derived AD datasets in the literature were used in this study. ROSMAP, Mayo Clinic and MSBB raw FASTQ files were retrieved from the Synapse Platform (https://www.synapse.org) with accession IDs: syn17024112, syn9738945 and syn7416949, respectively. Clinical metadata of the datasets were also retrieved from the Synapse platform. The ROSMAP dataset was derived from the dorsolateral prefrontal cortex, the Mayo Clinic dataset from the temporal cortex and the MSBB dataset from the parahippocampal gyrus. Samples were categorized as AD and control based on the CERAD score. Individuals with CERAD score 4 (No AD) were considered as the control group, and 1 (Definite AD) and 2 (Probable AD) were considered as AD. Since there was no CERAD score information in the Mayo data, sample classification was done based on Braak Stage (0-III as control, IV-VI as AD). ROSMAP, Mayo Clinic and MSBB RNAseq datasets include 404, 82 and 158 AD and 165, 78 and 26 control samples, respectively.

Gene expressions from RNAseq data

The FASTQ files were subjected to FastQC tool (version 0.11.9)[15] for quality control, and low-quality reads, short reads and adapter sequences were trimmed by Trimmomatic (version 0.39)[16]. Trimmed reads were aligned to human reference genome (hg38) by the STAR algorithm (version 2.7.8a)[17]. Then, featureCounts tool (version 2.0.2)[18] was used to obtain raw counts.

Between-sample normalization was applied on the raw counts using Deseq2[19] R package. The log2 values of normalized counts were calculated, and the log-transformed values were adjusted for major covariates (age, gender and post-mortem interval). To this end, the built-in R function *lm* was used to generate linear models and to calculate coefficients of age, gender and post-mortem interval covariates as described before[20]. Subsequently, the effects of the covariates were removed from the normalized and log2-transformed count data.

Principal component analysis (PCA) was performed to detect outlier samples for each dataset. 1 AD (Sample ID: 500_120515) and 1 control (Sample ID: 380_120503) from the ROSMAP dataset and 3 controls from the Mayo Clinic dataset (Sample IDs: 11294, 11396, 11399) were determined to be outliers and were removed from the data.

Variant identification from RNAseq data

By using the trimmed FASTQ files, the STAR tool was re-run to get splice junction information for each sample. Variant calling was then performed after a second indexing and alignment, using GATK tools (version 4.2.0.0) as described in GATK Best Practices for Variant Calling in RNAseq[21]. Briefly, AddOrReplaceReadGroups was used to add read group information to BAM files, and it was followed by MarkDuplicates, which identifies duplicate reads resulting from the PCR step. Then, SplitNCigarReads was used to split mapped reads at the intronic regions, and BSQR (Base Quality Score Recalibration) was used to re-score base quality scores. Afterwards, variants were detected using the HaplotypeCaller tool, and a VCF file was generated for each sample. The variants were filtered and annotated by the VariantFiltration tool of GATK and ANNOVAR[22] respectively. In further analyses, biallelic variants with read depth value equal to or greater than five were used.

Gene-level pathogenicity score calculation

The GenePy[23] algorithm was employed to transform variant-level pathogenicity scores into gene-level pathogenicity scores. This algorithm operates by aggregating the impact of all variants within a gene in an additive fashion, thereby generating a cumulative pathogenicity score that considers the combined effects of numerous small to moderate effects exerted by each variant. This method closely follows the non-Mendelian inheritance patterns typically observed in complex diseases. GenePy scores of each gene for each individual were computed in the R environment by using the in-house script implemented based on the GenePy score formula and GenePy algorithm.

In this study, Rare Exome Variant Ensembl Learner (REVEL) scores[24] were used as the deleteriousness value. REVEL predicts variant pathogenicity by combining scores from 13 distinct in silico pathogenicity prediction tools. The REVEL score of a variant takes values between 0 and 1, with higher scores reflecting a higher likelihood of the variant being pathogenic. Structural variants (e.g., frameshift indels, stop loss/gain mutations) that lead to protein truncation/elongation are not assigned REVEL scores. Therefore, the maximum deleteriousness value 1 was assigned to all structural variants due to their profoundly damaging impact on protein function. The Genome Aggregation Database (gnomAD)[25] exome allele frequencies were used as the allele frequency in the GenePy score calculation. Since the gnomAD exome database encompasses 125,748 individuals, variants lacking frequency data in gnomAD were assigned an allele frequency of 3.98×10^{-6}, indicating one allele present among 125,748 individuals. The GenePy scores were computed for each sample in each dataset.

Larger genes can accumulate higher GenePy scores as they tend to harbor a greater number of variants, resulting in inflated GenePy scores. To address this, GenePy scores were divided by gene length and multiplied by the median observed gene length in the data for gene length correction[23]. Finally, the GenePy scores were adjusted to account for gender effect using the *lm* function in R, as mentioned in the previous section. To determine genes with significantly higher pathogenic variants in a personalized manner for each AD individual, the GenePy score of each gene in a given AD sample was combined with the scores of that gene in all the control samples in that dataset, and the values were ranked. If the score is higher than 95% of the scores from the control individuals, the gene was marked as a gene with higher load of pathogenic variants in AD.

Personalized metabolic model reconstruction

To reconstruct personalized genome-scale metabolic models for all controls and AD cases, genes in the covariate-adjusted expression data were mapped individually to the human genome-scale metabolic model (Human-GEM) using the integrative metabolic analysis tool (iMAT)[6]. The iMAT algorithm was run with the parameters described in our previous study[26]. Due to the covariate adjustment of the expression data, there are negative values in the data. iMAT algorithm available through the COBRA Toolbox[27] ignores genes with negative expression values. Hence, the absolute value of the smallest negative expression value in each dataset was added to the whole data. In this way, all expression values were converted to positive values. Since the parameters used by the iMAT algorithm to determine active and inactive reactions are based on percentiles, this process does not affect the results. Human-GEM (version 1.12.0)[28] was retrieved from the GitHub repository (https://github.com/SysBioChalmers/Human-GEM). This comprehensive metabolic network includes 13,070 reactions associated with 3067 genes, 8369 metabolites and 143 pathways. 3055 out of 3067 genes in the Human-GEM are included in the transcriptome datasets. Simulations were performed using MATLAB R2020a with the Gurobi solver.

To incorporate the effects of the variants that disrupt protein function into personalized metabolic models (Fig. 1A), the following steps were followed: (i) It was hypothesized that genes with significantly higher load of pathogenic variants in AD, as identified by the GenePy algorithm, would not be able to form functional proteins due to mutations on them, even if their mRNA expression levels were not low. Based on this assumption, the

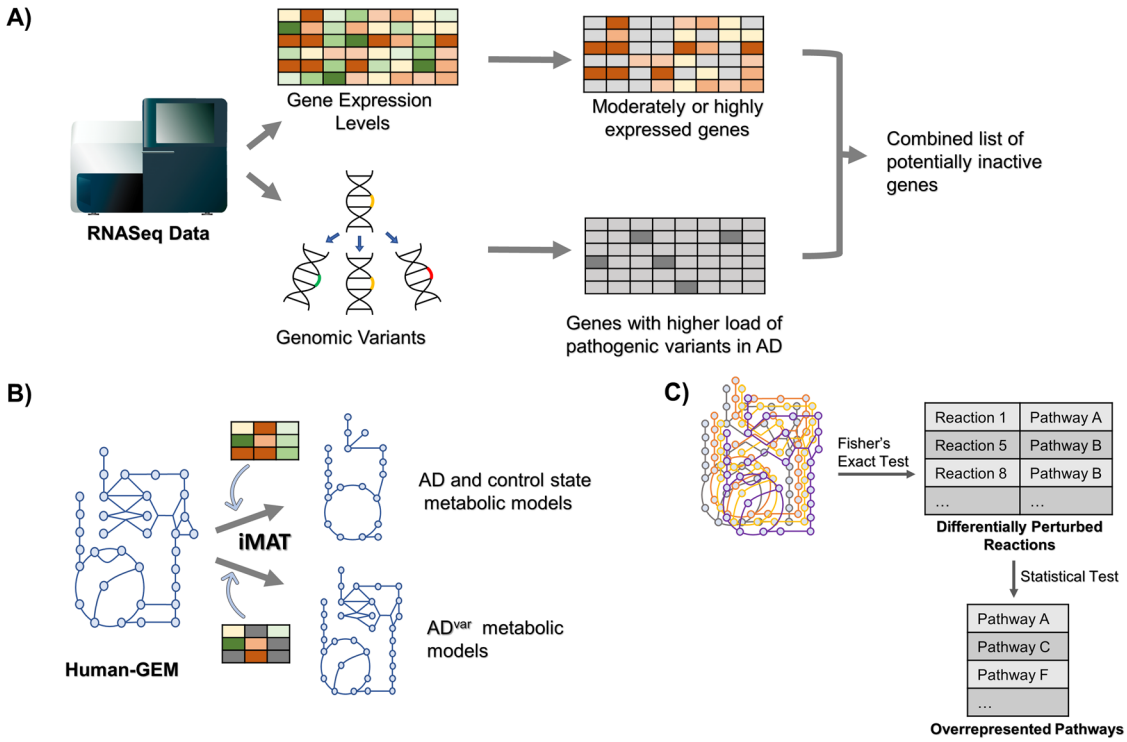

Fig. 1 | Summary of iMAT-based personalized metabolic model generation and incorporation of the impact of pathogenic variants to models. A RNAseq data are obtained in Fastq format and processed with two different pipelines to find gene expression values and genomic variants. The genes with higher load of pathogenic variants and the genes with high or moderate expression levels were selected as inactive genes for each AD sample. **B** Gene expression values are mapped to the genome-scale metabolic model and AD, control and AD^{var} models are created with the iMAT algorithm. To create AD^{var} models, unlike the classical method, the genes carrying higher load of pathogenicity in AD were additionally used. **C** Differentially perturbed reactions between AD and control are identified by Fisher's Exact Test after representing each personalized metabolic model as a binary vector, where "1" means the reaction is active in that model. Then, pathways overrepresented with the differentially perturbed reactions were predicted.

expression levels of pathogenic variant genes were set to zero in the covariate adjusted expression data of that AD sample. This was repeated for all AD samples. (ii) Then, sample-based iMAT models were generated by forcing the removal of the reactions controlled by the combined list of the lowly expressed genes and the genes with pathogenic variants from the metabolic model, subject to mass-balance constraints around the intracellular metabolites. iMAT restores the reactions back if they violate mass-balance constraints or if the removal decreases the consistency of reaction rates with the transcriptome data. The resulting metabolic models were referred as AD^{var} models in this study.

Statistics and reproducibility

The generated iMAT models were converted to binary format to apply the statistical tests. This conversion was based on the reactions in Human-GEM. Reactions active in the iMAT models were represented by 1 and inactive reactions were represented by 0 for each sample. Then, using these binary matrices, a contingency table was created for each reaction between AD and control conditions, and Fisher's Exact Test was performed to identify significantly altered reactions in each dataset. *P* value cut-off of 0.05 was used for AD-Control models while the cut-off was 0.01 for AD^{var}-Control models. We used a more stringent *p* value cut-off in comparing AD^{var}-Control metabolic models when identifying affected reactions. Here, we aimed to make the number of affected reactions more comparable to the case where AD models ignoring genomic variants were used. We show that we still captured many more AD related mechanisms in AD^{var}-Control comparisons than AD-Control comparisons although we used a more stringent *p* value cut-off.

The reaction-subsystem assignments in Human-GEM were used to find significantly affected pathways. Fisher Exact Test based *p* values of each reaction were converted to z-scores using the inverse cumulative density function. Each pathway z-score was then calculated by averaging the z-scores of the reactions associated with that pathway and corrected with 1000 random permutations as described elsewhere[29,30]. Corrected pathway z-scores were converted to pathway *p* values using the cumulative density function. Significantly affected pathways were identified using the *p* value < 0.05 cut-off. This approach enabled consideration of all pathway reactions in the calculation compared to a hypergeometric test where only reactions with significant *p* values are considered.

Identification of AD-related genes associated with the affected metabolic reactions

Using the Gene-Protein-Reaction (GPR) rules of Human-GEM, the genes controlling significantly perturbed reactions were extracted. Experimentally confirmed and curated AD-related gene lists were collected from another study[31] and DisGeNET[32]. There are 166 AD-related metabolic genes in the final list. The hypergeometric test was used to test whether the number of AD-related genes was significantly overrepresented in the identified lists of significantly perturbed reactions from AD or AD^{var} models.

ROSMAP proteome data analysis

ROSMAP proteome data was obtained from the Synapse platform with accession ID: syn21448334. There are 400 samples and 8817 proteins in the data. One protein with NA reads in more than 50% of the samples was excluded from the data. Other NA reads were filled with half of the minimum value in the data (limit of detection/2). Quantile normalization was performed using the preprocessCore R package. Then, 1545 proteins in common with the proteins (enzymes) covered by HumanGEM v.1.12 were selected. Similar to RNAseq analysis, AD and control samples were separated according to CERAD score. 251 AD and 106 control samples were identified. Covariate adjustment was performed with respect to age, sex and

PMI. Using the rank-based approach, the abundance of each protein in an AD sample was compared with the abundance of that protein in all control samples. Thus, the abundance ranks of the proteins in each AD sample compared to the control were determined on a sample-based manner.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Results and discussion

iMAT models of AD, AD^{var} and control conditions

Each sample in the ROSMAP, Mayo Clinic and MSBB post-mortem brain RNAseq datasets were mapped to Human-GEM using the iMAT algorithm to reconstruct personalized genome-scale metabolic models for both AD and control samples. This led to 908 models in total spanning the three datasets.

The iMAT algorithm generates metabolic models by inactivating reactions controlled by enzymes encoded by genes with low expression, while keeping reactions controlled by enzymes encoded by genes with high expression in the model, subject to mass-balance constraints around intracellular metabolites[6]. It also ensures maximum consistency between gene expression levels and reaction rates. On the other hand, some genomic variants are pathogenic, i.e. they lead to dysfunctional enzymes, although they do not affect expression at the mRNA level[33,34].

In this study, first, genomic variants in the form of single-nucleotide variants and insertions/deletions were predicted from the RNAseq datasets, and, then, their in silico pathogenicity scores were used to calculate gene-level scores. The correlation between expression levels and pathogenicity scores of metabolic genes was calculated for each sample. The correlation coefficients ranged between -0.03-0.23 for all the samples from three different cohorts (Supplementary Fig. 1). Therefore, the pathogenicity scores of a sample do not depend on gene expression levels and provide new knowledge. To integrate this knowledge into metabolic models, genes with significantly higher pathogenicity in AD samples were marked as inactive in generating iMAT-derived metabolic models, termed AD^{var} models (Fig. 1A-B).

The number of metabolic genes carrying pathogenic variants in AD samples was about 50-150 for the studied datasets (Fig. 2A). This means that the genes additionally included in the list of genes to be removed from the metabolic models to create AD^{var} models is about ~1.5%-5% of the total number of genes in Human-GEM, which is not a high fraction. In other words, our approach of incorporating genes with higher load of pathogenic variants in AD introduces minimal intervention to the original iMAT approach. We also checked the expression levels of these genes to see how many of them are expressed higher than the high-expression cut-off used by iMAT. On average, we found 37 pathogenic variant carrying genes in ROSMAP, 44 in Mayo and 31 in MSBB to have expression levels higher than 75% of all the genes in transcriptome data (Fig. 2B). Thus, the number of

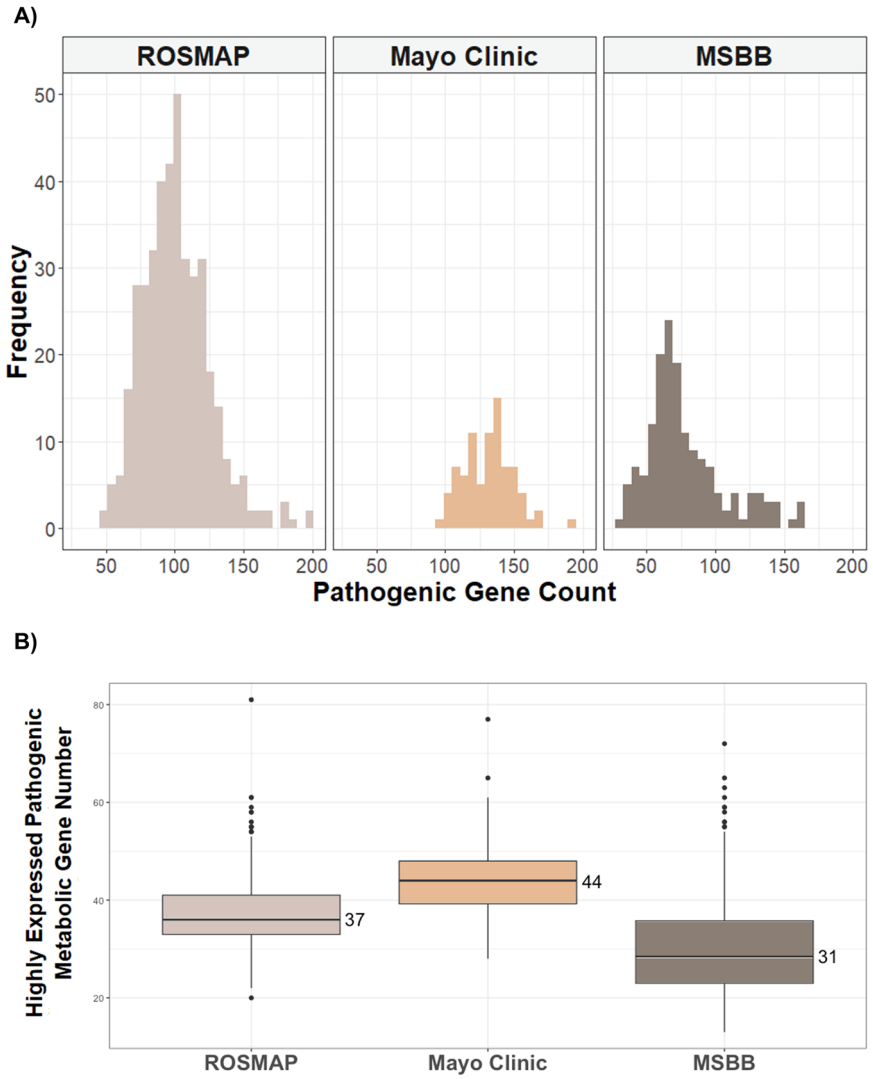

Fig. 2 | Number of metabolic genes carrying significantly higher load of pathogenic variants in AD for each dataset. A Histogram of the number of metabolic genes with higher load of pathogenic variants in each AD sample for each dataset. Frequency axis shows the number of samples having a specific number of pathogenic genes as indicated in the x-axis. **B** Boxplot of the number of metabolic genes carrying pathogenic variants and having expression levels higher than 75% of the genes (the iMAT high-expression cut-off) in the whole transcriptome.

metabolic genes manipulated in the transcriptome data is considerably less than the number of genes in the model.

Our major hypothesis in this study is that genes carrying higher pathogenic load in an AD patient compared to controls will not encode a functional protein although the gene is highly expressed. To validate this hypothesis, we focused on the genes that were assumed to be inactive in the iMAT analysis because of their high pathogenic load in AD although they were highly expressed (expression higher than 75% of the averaged expressions of the genes across all samples). For the ROSMAP dataset, there were 1126 such genes that were set to inactive state in at least one sample. Of note, setting a gene with high pathogenic variant load in AD to inactive state does not necessarily mean that the corresponding reaction will be removed from the model since a high number of reactions in the Human-GEM are indeed controlled by multiple enzymes/genes. Only 24 of these genes were set to inactive state in more than 10% of the ROSMAP AD samples (Supplementary Table 1). Besides, these 24 genes were not highly expressed in 85% of the AD samples. These genes were examined in more detail as explained below.

Firstly, we checked the expression values of these 24 genes in the Mayo Clinic and MSBB datasets, and we determined for each dataset the number of samples in which their expressions were lower than the average expression. Among these 24 genes, six of them (PGGHG, PNPLA2, LPL, PLCD3, SLC52A3, and MPST) were below the average expression in at least 40% of the MSBB samples. This shows that these genes are not always highly expressed in AD, and the behavior we see in the ROSMAP dataset can be cohort-specific or brain-region specific. Supportively, Jia et al. [35] reported that mRNA levels of TCA cycle genes ACO2, SDHB and PDK2, which are among these 24 genes, were decreased in different brain regions in AD. Another study applied a transcriptome-wide association study (TWAS) on SLC25 family genes, and found that decreased expression of the SLC25A22 gene is associated with AD[36]. Bis et al. [37] created a list of AD-associated genes from GWAS studies reporting also the type of mutation causing the association. When these 24 genes were compared with their list, it was found that missense mutation in PNPLA2 and POLR2E genes have been associated with AD. In addition, the ROSMAP proteome data[38] (synapse accession ID: syn21448334) was analyzed. 14 of these 24 genes have measured protein abundance in the proteome data. 5 (MEPCE, RNF13, POLR2E, UBE2E2, and SMPD1) out of these 14 genes have abundances below the average protein abundance in all AD samples in the ROSMAP proteome data. This implies that these genes are translated into non-functional or weakly functional proteins. In summary, for 15 of the 24 genes, there is evidence that their expression may be low in AD or that they may form non-functional proteins. Therefore, this supports inactivation of these genes by the approach proposed in this study.

Subsequently, the iMAT models of each dataset were compared to investigate whether the numbers of active reactions between the control, AD^{var} and AD models are different. As shown in Fig. 3A, all models consisted of ~7000–8000 reactions while Human-GEM includes 13070 reactions. Accordingly, it was observed that the distribution of the number of active reactions was more similar in the AD and control models compared to the AD^{var} models. In the AD^{var} models, although there was no dramatic difference in terms of the number of reactions in the created personalized models, the distribution became more spread out. For each sample, we also looked at the difference between the number of reactions between the AD and AD^{var} models of that sample (Fig. 3B) and found that there was a difference of about 50 reactions in ROSMAP and MSBB and about 200 reactions for Mayo Clinic. Differing from the other datasets, the AD^{var} metabolic models of the Mayo Clinic dataset had fewer reactions than the

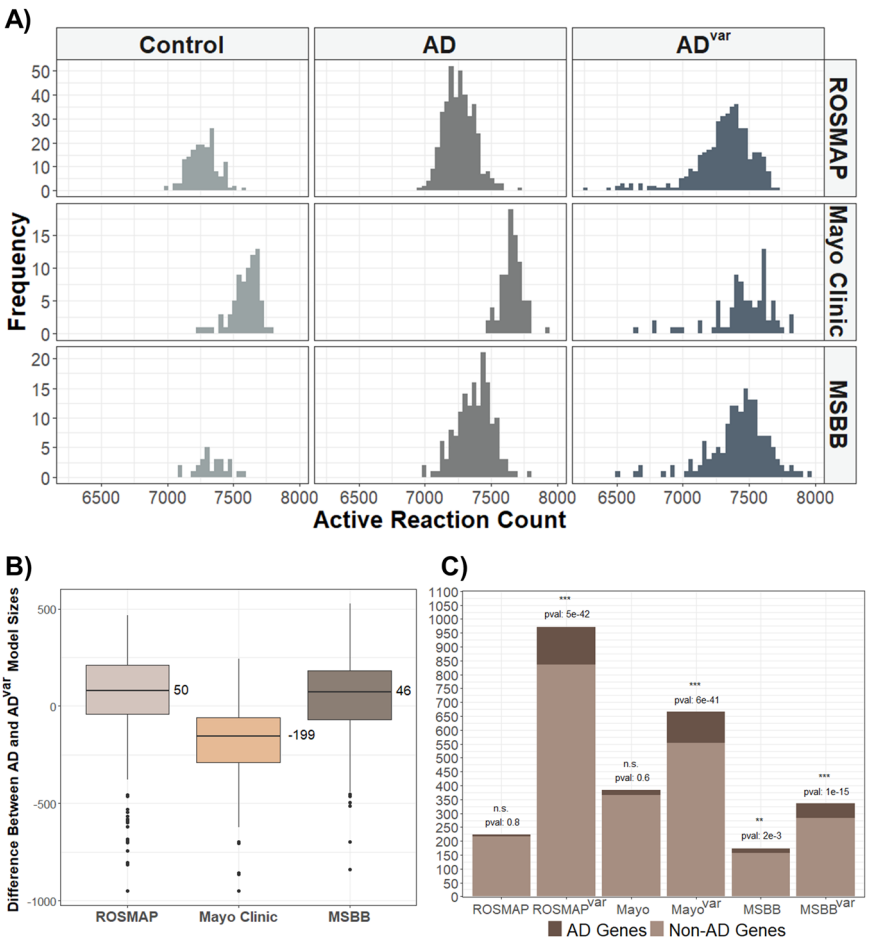

Fig. 3 | Condition specific model generation By iMAT. A Histogram of the number of active reactions in each personalized model for each dataset. Frequency shows the number of instances with the count. **B** Boxplot of the difference in reaction numbers between AD and AD^{var} iMAT models for each sample. **C** Bar plot showing the total number of genes associated with the lists of significantly altered reactions, and the proportion of AD-related genes in these lists for each dataset and for each of AD and AD^{var} models. pval shows the hypergeometric *p* value obtained from the over-representation analysis of AD-related genes in each list. n.s: not significant, ** *p* val < 0.01, *** *p* val < 0.001.

AD models. To better understand this difference, Jaccard similarity indices were calculated between the binary matrices representing AD and AD^{var} models for each dataset, and the similarity of the binary models was around 85% for all the datasets. These results indicate that although there are differences between the models at the individual level, in general the AD and AD^{var} models are similar in terms of the majority of active reactions.

After the significantly altered reactions were identified using Fisher's Exact test, gene compositions controlling these reactions were analyzed to determine the number of genes associated with AD for each reaction list (Fig. 3C). Accordingly, while the significant reaction lists were significantly enriched with the AD-related genes in all AD^{var} models, a significant result was obtained only in MSBB among the models constructed with only gene-expression data, and with a much higher p value compared to the AD^{var} counterpart. This shows that the integration of pathogenic-variant data into the metabolic models contributes to a better representation of AD status at the gene level.

Differentially perturbed pathways

After the personalized models were generated, Fisher's Exact Test was applied to identify differentially affected reactions between AD-Control comparison and AD^{var}-Control comparison (Fig. 1C). Although the model sizes and pathogenic gene counts were not much different in the AD and AD^{var} models, the number of differentially perturbed reactions between AD-Control and AD^{var}-Control were considerably different (Supplementary Fig. 2). For all the cohorts, adding the effect of pathogenic variants to the models increased the number of differentially altered reactions (Table 1). Using reaction-pathway associations of Human-GEM, we also identified the number of unique pathways associated with the significantly affected reactions for each dataset (Table 1, Supplementary Fig. 3). The statistical significance of the intersection of AD and AD^{var} pathways was checked for each dataset using the pathways associated with the significantly affected reactions. The intersections were found to be highly significant based on Super Exact Test[39], with p values being 7.76×10^{-5}, 1.30×10^{-8} and 1.27×10^{-8} for ROSMAP, Mayo Clinic and MSBB datasets, respectively (Supplementary Table 2). Likewise, Jaccard similarity indices between AD and AD^{var} pathways were 0.46, 0.57 and 0.45 for ROSMAP, Mayo Clinic and MSBB, respectively. A modified version, Overlap similarity index, showed very high similarity between AD and AD^{var} pathways (~0.90 on average), implying that AD^{var} pathways were almost inclusive of AD pathways (Supplementary Table 3).

We applied a statistical analysis to identify pathways significantly overrepresented among the affected reactions (Fig. 4, Table 1). In general, only few pathways were identified to be significantly perturbed based on the AD models, regardless of the dataset used. We identified a number of pathways whose perturbation was only captured with the AD^{var} models. Below, we provide a detailed analysis of the pathways identified to be significantly perturbed only in AD^{var} models.

The sphingolipid metabolism was significantly perturbed only in the AD^{var} models in all the three cohorts. Sphingolipids are a member of lipid family, and they are crucial elements of membrane biology and have roles in

Table 1 | The number of significantly affected reactions and associated pathways

Datasets and Compared Conditions	# of significantly affected reactions	# of associated pathways	# of overrepresented pathways
ROSMAP AD-Control	290	55	6
ROSMAP AD^{var}-Control	3089	108	28
Mayo Clinic AD-Control	674	64	7
Mayo Clinic AD^{var}-Control	1343	98	17
MSBB AD-Control	220	40	4
MSBB AD^{var}-Control	793	72	17

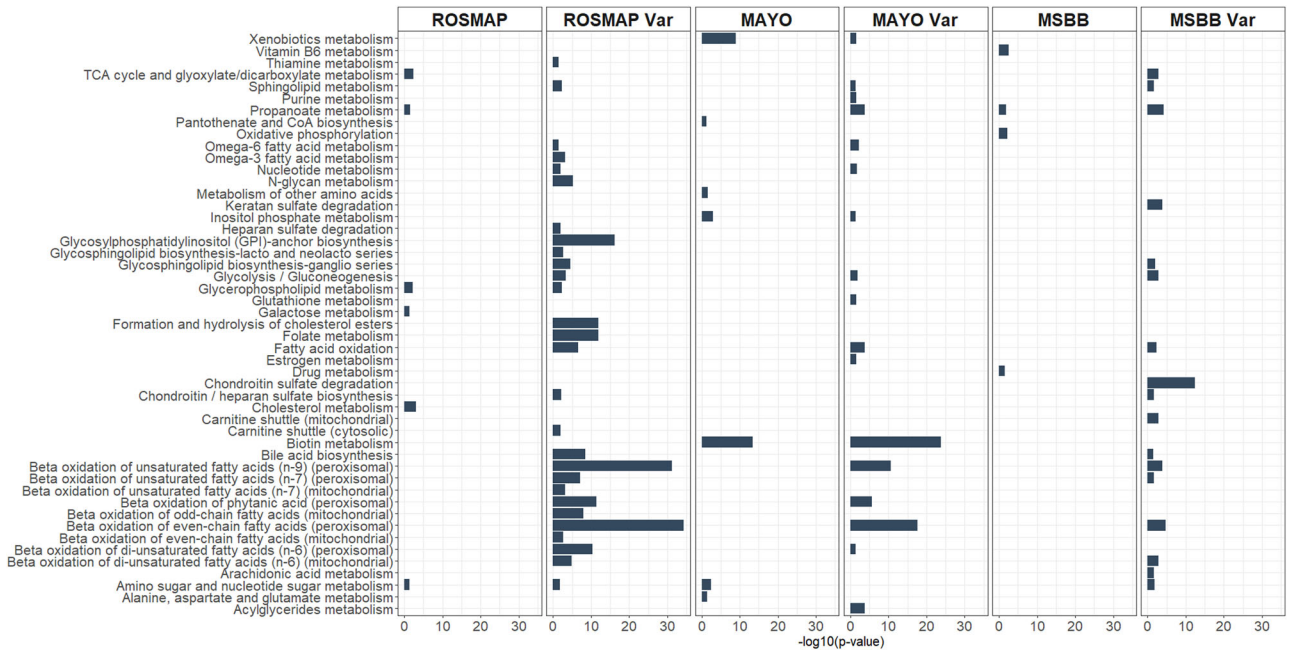

Fig. 4 | Differentially perturbed pathways between AD-Control models and AD^{var}-Control models. Pathway enrichment analysis was applied on the differentially altered reactions. x-axis shows -log10(p value) of the pathways. Only significant pathways are shown in the figure for each dataset.

the regulation of cell function[40]. Sphingolipid metabolism was previously associated with dementia and the AD risk factor gene APOE genotype[41]. Also, sphingomyelin metabolism was recently proposed as a therapeutic target for AD based on multi-omics profiling of the AD brain[42]. Therefore, identifying sphingolipid metabolism only in AD^{var} models shows the increased accuracy in capturing AD-related processes in the metabolic models.

Glycolysis/glucogenesis is another pathway that was significantly affected only in AD^{var} models in the three cohorts. Glucose metabolism is important in energy production, which is vital for all cells, and it was reported that glucose metabolism and glycolysis defects contribute to AD pathogenesis[43–45]. In the brain, astrocyte cells prefer aerobic glycolysis to produce lactate. Lactate produced from glycolysis or glycogenolysis in astrocytes is transferred to neurons to be used in oxidative phosphorylation. However, it was observed that the lactate production capacity of astrocytes decreased in AD[46]. Consequently, energy deficiency in neurons and glia play an essential role in AD.

Fatty acid oxidation is another pathway that was significantly affected only in AD^{var} models in the three cohorts. Brain is the organ where fatty acids are the most abundant in our body. It was shown that the levels of all subclasses of fatty acids are different in the cerebrospinal fluid of AD patients compared to controls[47]. Qi et al. reported that APOE4 genotype caused a shift in both glucose and fatty acid oxidation in astrocytes[48]. Overall, these results indicate that integrating pathogenic variants into the models allows capturing of some important AD-related pathways in different cohorts.

In models generated with the samples obtained from different brain regions, there were also differences among significantly perturbed pathways that were captured only in the variant-aware models. Bile acid, chondroitin/heparan sulfate and glycosphingolipid biosynthesis were differentially perturbed only in the variant models (MSBB and ROSMAP cohorts). Among these three pathways, bile acid metabolism is highly integrated with cholesterol metabolism and has been associated with AD. Recently, Baloni et al. used metabolic modeling approach and reported that bile acid metabolism was altered in AD brains[7]. Similarly, Varma et al. showed that altered brain bile acid metabolism was associated with AD[49]. Heparan sulfate is another important metabolic pathway affecting amyloid-beta clearance, and high levels of heparan sulfate proteoglycans were found in post-mortem AD brains[50,51]. Glycosphingolipids are a heterogeneous group of membrane lipids. They comprise a high amount of the brain lipid composition. Tang et al. [52] analysed the expression levels of glycosylation related genes in the AD brains and found alterations in the levels of the genes that play a role in glycosphingolipid biosynthesis in human.

Similarly, omega-6 fatty acid, fatty acid beta-oxidation and nucleotide metabolisms were differentially perturbed only in the variant models of Mayo Clinic and ROSMAP. Lipid metabolism including fatty acids is one of the most studied metabolic pathways in AD[53,54]. Lipid metabolism is known to be impaired in the early stages of AD[47]. On the other hand, since nucleotide metabolism is closely related to cell proliferation pathways, it is interesting that it was also captured in our results although it is a pathway generally studied in cancer[55]. As a result, different datasets have the capacity to capture different pathways that have been associated with AD or that may be candidates for AD. These results are also important for understanding how different brain regions are affected in AD.

Additionally, propanoate, inositol phosphate, glycerophospholipid and biotin metabolic pathways were identified to be perturbed in both variant and non-variant models of the same dataset. Propanoate metabolite is known to be neuroprotective and associated with brain-gut axis[56]. A study showed that changes in the levels of inositol phosphate metabolic enzymes were correlated with amyloid beta peptide formation and tau hyperphosphorylation in AD[57]. A recent study using Drosophila model of tauopathy suggested biotin metabolism as a druggable target for neuroprotection[58]. Glycerophospholipids are complex species of fatty acids, and they are the most abundant class of lipids in the human brain[59]. These results also indicated that adding the variant effect to the models does not cause the loss of previously annotated metabolic processes.

As an additional perspective, we reconstructed personalized iMAT models of AD patients by using a more stringent cut-off of gene pathogenicity. In this analysis, a gene was considered pathogenic if its pathogenicity score was higher than 99% of the controls. Significantly affected reactions and significantly enriched pathways were identified accordingly. Similar to our findings with the 95% cut-off, variant-integrated models again revealed more pathways than expression-only models. Also, the pathways known to be associated with AD such as glycolysis, fatty acid oxidation and bile acid biosynthesis were identified only in the ADvar models (Supplementary Fig. 4).

Conclusion

Genome-scale metabolic models in complex diseases such as Alzheimer's disease, in which many different metabolisms in the cell are affected, provide a great advantage for systemic investigation of the etiology of the disease. For this purpose, the generation of disease models by integrating gene expression data into GEMs has become widespread in the literature in the last 20 years, and it is a frequently used method today. In these models, information from mRNA level is incorporated into the model as the abundance of enzyme-coding proteins, and optimization problems based on mass-balance around intracellular metabolites are solved by aiming to keep the reactions catalyzed by enzymes with high abundance in the model and excluding the reactions catalyzed by enzymes with low abundance from the model. Although these models are very convenient, possible loss of function in the mRNA-to-protein transition is neglected. In this study, for the first time in the literature, we constructed personalized metabolic models by determining both gene expression levels and pathogenic variants from the same AD and control RNAseq samples from different brain regions. When we compared these models with the models constructed with only gene-expression data, we showed that pathways such as fatty acid oxidation, bile acid metabolism, sphingolipid metabolism and glycolysis, which are known to play crucial roles in AD metabolism, were found to be significantly altered only in the variant integrated models. In addition, gene-level analysis showed that the list of significantly perturbed reactions was significantly enriched with AD genes for all the three datasets in the AD^{var} models, and with a much higher significance level compared to the models constructed with only gene expression data.

The three data sets analyzed in this study were obtained from three different brain regions. The pathways associated with the significantly altered reactions of these datasets were compared, and it was observed that the pathways of variant-integrated models were more similar to each other across the three datasets (Jaccard similarity index >0.6) than the pathways of only expression-integrated models (Supplementary Fig. 5). This is expected since genetic effects should be more consistent across different brain regions than gene expression. Although the results of this study have brought a new perspective to the use of metabolic modeling and transcriptome data in AD studies, it has some limitations. Within the framework of this study, only the effect of genomic variants that are predicted to affect protein expression was considered. However, some variants in intronic regions called eQTLs are known to regulate the expression of target genes. Transcriptome-wide association studies (TWAS) aim to reveal such variant-expression-trait relationships[60]. In future studies, metabolic models can be generated by also adding the information of metabolic genes whose expressions are regulated by variants based on TWAS studies.

The reactions in GEMs can have complex gene-protein-reaction (GPR) rules. The presence of a significantly higher pathological variant load in a gene in AD may impair gene function, affecting the reaction catalyzed by the enzymes coded by a single gene or enzyme complexes; however, reactions catalyzed by alternative enzymes may remain unaffected if the variant does not impact the alternative genes. On the other hand, for some genes with high pathological variant load in AD, it is possible that one of the known transcripts of these genes indeed does not carry identified pathogenic variants, rendering the gene functional. This possibility was ignored in this

study. This limitation can be eliminated by transcript-based variant identification and data mapping on GEMs instead of gene-based strategy. In addition, post-translational modifications which could have an important role in enzyme activity were ignored in this study. For more comprehensive results, integrating the effect of post-translational modifications in the cell into the genome-scale metabolic models is another important future study.

Data availability

The results published here are in whole or in part based on data obtained from the AD Knowledge Portal (https://adknowledgeportal.synapse.org/) (accession code: syn17024112 (ROSMAP), syn9738945 (Mayo Clinic) and syn7416949 (MSBB)). Data is available for general research use according to the following requirements for data access and data attribution (https://adknowledgeportal.org/DataAccess/Instructions).

Code availability

The scripts are available at https://github.com/SysBioGTU/GenomicVariantsMetabolicModels.

Received: 15 May 2024; Accepted: 17 March 2025;
Published online: 27 March 2025

References

1. Bhatia, S. et al. Mitochondrial dysfunction in Alzheimer's disease: Opportunities for drug development. *Curr. Neuropharmacol.* **20**, 675–692 (2022).
2. Varma, V. R. et al. Abnormal brain cholesterol homeostasis in Alzheimer's disease—a targeted metabolomic and transcriptomic study. *npj Aging Mech. Dis.* **7**, 11 (2021).
3. Yu, L., Jin, J., Xu, Y. & Zhu, X. Aberrant energy metabolism in Alzheimer's disease. *J. Transl. Int Med.* **10**, 197–206 (2022).
4. Desai, R. J. et al. Targeting abnormal metabolism in Alzheimer's disease: The Drug Repurposing for Effective Alzheimer's Medicines (DREAM) study. *Alzheimer's. Dement.: Transl. Res. Clin. Interv.* **6**, e12095 (2020).
5. Gopalakrishnan, S. et al. Guidelines for extracting biologically relevant context-specific metabolic models using gene expression data. *Metab. Eng.* **75**, 181–191 (2023).
6. Zur, H., Ruppin, E. & Shlomi, T. iMAT: an integrative metabolic analysis tool. *Bioinformatics* **26**, 3140–3142 (2010).
7. Baloni, P. et al. Metabolic network analysis reveals altered bile acid synthesis and metabolism in Alzheimer's disease. *Cell Rep. Med* **1**, 100138 (2020).
8. Moolamalla, S. T. R. & Vinod, P. K. Genome-scale metabolic modelling predicts biomarkers and therapeutic targets for neuropsychiatric disorders. *Comput. Biol. Med.* **125**, 103994 (2020).
9. Zhao, S. Alternative splicing, RNA-seq and drug discovery. *Drug Discov. Today* **24**, 1258–1267 (2019).
10. Berger, K., Arafat, D., Chandrakasan, S., Snapper, S. B. & Gibson, G. Targeted RNAseq improves clinical diagnosis of very early-onset pediatric immune dysregulation. *J. Personalized Med.* **12**, 919 (2022).
11. Tushir, S. et al. Proteo-genomic analysis of SARS-CoV-2: A clinical landscape of single-nucleotide polymorphisms, COVID-19 proteome, and host responses. *J. Proteome Res.* **20**, 1591–1601 (2021).
12. De Jager, P. L. et al. A multi-omic atlas of the human frontal cortex for aging and Alzheimer's disease research. *Sci. Data* **5**, 180142–180142 (2018).
13. Allen, M. et al. Human whole genome genotype and transcriptome data for Alzheimer's and other neurodegenerative diseases. *Sci. Data* **3**, 160089 (2016).
14. Wang, M. et al. Integrative network analysis of nineteen brain regions identifies molecular signatures and networks underlying selective regional vulnerability to Alzheimer's disease. *Genome Med* **8**, 104 (2016).
15. Andrews, S. (Babraham Bioinformatics, Babraham Institute, Cambridge, United Kingdom, 2010).
16. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
17. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
18. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
19. Love, M., Anders, S. & Huber, W. Differential analysis of count data–the DESeq2 package. *Genome Biol.* **15**, 10–1186 (2014).
20. Lüleci, H. B. et al. A benchmark of RNA-seq data normalization methods for transcriptome mapping on human genome-scale metabolic networks. *npj Syst. Biol. Appl.* **10**, 124 (2024).
21. Brouard, J.-S. & Bissonnette, N. *Variant Calling: Methods and Protocols* (eds C. Ng & S. Piscuoglio) 205-233 (Springer US, 2022).
22. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164–e164 (2010).
23. Mossotto, E. et al. GenePy - a score for estimating gene pathogenicity in individuals using next-generation sequencing data. *BMC Bioinforma.* **20**, 254–254 (2019).
24. Ioannidis, N. M. et al. REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. *Am. J. Hum. Genet.* **99**, 877–885 (2016).
25. Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
26. Lüleci, et al. *Alzheimer's Disease: Methods and Protocols* (ed J. Chun) 173–189 (Springer US, 2023).
27. Heirendt, L. et al. Creation and analysis of biochemical constraint-based models using the COBRA Toolbox v.3.0. *Nat. Protoc.* **14**, 639–702 (2019).
28. Wang, H. et al. (Zenodo, 2022).
29. Çakır, T. Reporter pathway analysis from transcriptome data: Metabolite-centric versus Reaction-centric approach. *Sci. Rep.* **5**, 14563 (2015).
30. Patil, K. R. & Nielsen, J. Uncovering transcriptional regulation of metabolism by using metabolic network topology. *Proc. Natl. Acad. Sci. USA* **102**, 2685 (2005).
31. Ceylan, B., Düz, E. & Çakir, T. Personalized protein-protein interaction networks towards unraveling the molecular mechanisms of Alzheimer's disease. *Mol. Neurobiol.* **61**, 2120–2135 (2024).
32. Piñero, J., Saüch, J., Sanz, F. & Furlong, L. I. The DisGeNET cytoscape app: Exploring and visualizing disease genomics data. *Comput. Struct. Biotechnol. J.* **19**, 2960–2967 (2021).
33. Chen, J. et al. Functional analysis of genetic variation in catechol-O-methyltransferase (COMT): Effects on mRNA, protein, and enzyme activity in postmortem human brain. *Am. J. Hum. Genet.* **75**, 807–821 (2004).
34. Lockett, K. L. et al. The ADPRT V762A genetic variant contributes to prostate cancer susceptibility and deficient enzyme function. *Cancer Res.* **64**, 6344–6348 (2004).
35. Jia, D., Wang, F. & Yu, H. Systemic alterations of tricarboxylic acid cycle enzymes in Alzheimer's disease. *Front. Neurosci.* **17**, 1206688 (2023).
36. Tian, J. et al. Hippocampal transcriptome-wide association study and pathway analysis of mitochondrial solute carriers in Alzheimer's disease. *Transl. Psychiatry* **14**, 250 (2024).
37. Bis, J. C. et al. Whole exome sequencing study identifies novel rare and common Alzheimer's-Associated variants involved in immune response and transcriptional regulation. *Mol. Psychiatry* **25**, 1859–1875 (2020).
38. Johnson, E. C. B. et al. Large-scale proteomic analysis of Alzheimer's disease brain and cerebrospinal fluid reveals early changes in energy metabolism associated with microglia and astrocyte activation. *Nat. Med*. **26**, 769–780 (2020).

39. Wang, M., Zhao, Y. & Zhang, B. Efficient Test and Visualization of Multi-Set Intersections. *Sci. Rep.* **5**, 16923 (2015).
40. Hannun, Y. A. & Obeid, L. M. Sphingolipids and their metabolism in physiology and disease. *Nat. Rev. Mol. Cell Biol.* **19**, 175–191 (2018).
41. Alaamery, M. et al. Role of sphingolipid metabolism in neurodegeneration. *J. Neurochem.* **158**, 25–35 (2021).
42. Baloni, P. et al. Multi-Omic analyses characterize the ceramide/sphingomyelin pathway as a therapeutic target in Alzheimer's disease. *Commun. Biol.* **5**, 1074 (2022).
43. Bergau, N., Maul, S., Rujescu, D., Simm, A. & Navarrete Santos, A. Reduction of glycolysis intermediate concentrations in the cerebrospinal fluid of Alzheimer's disease patients. *Front. Neurosci.* **13**, 871–871 (2019).
44. Pan, R.-Y. et al. Positive feedback regulation of microglial glucose metabolism by histone H4 lysine 12 lactylation in Alzheimer's disease. *Cell Metab.* **34**, 634–648.e636 (2022).
45. Theurey, P. et al. Systems biology identifies preserved integrity but impaired metabolism of mitochondria due to a glycolytic defect in Alzheimer's disease neurons. *Aging Cell* **18**, e12924–e12924 (2019).
46. Wang, Q. et al. Glucose metabolism, neural cell senescence and Alzheimer's disease. *Int. J. Mol. Sci.* **23**, 4351 (2022).
47. Yin, F. Lipid metabolism and Alzheimer's disease: clinical evidence, mechanistic link and therapeutic promise. *FEBS J.* **290**, 1420–1453 (2023).
48. Qi, G. et al. ApoE4 impairs neuron-astrocyte coupling of fatty acid metabolism. *Cell Rep.* **34**, 108572 (2021).
49. Varma, V. R. et al. Bile acid synthesis, modulation, and dementia: A metabolomic, transcriptomic, and pharmacoepidemiologic study. *PLoS Med* **18**, e1003615 (2021).
50. Liu, C.-C. et al. Neuronal heparan sulfates promote amyloid pathology by modulating brain amyloid-β clearance and aggregation in Alzheimer's disease. *Sci. Transl. Med.* **8**, 332ra344–332ra344 (2016).
51. Ozsan McMillan, I., Li, J.-P. & Wang, L. Heparan sulfate proteoglycan in Alzheimer's disease: aberrant expression and functions in molecular pathways related to amyloid-β metabolism. *Am. J. Physiol. -Cell Physiol.* **324**, C893–C909 (2023).
52. Tang, X. et al. Transcriptomic and glycomic analyses highlight pathway-specific glycosylation alterations unique to Alzheimer's disease. *Sci. Rep.* **13**, (2023). 7816.
53. Chew, H., Solomon, V. A. & Fonteh, A. N. Involvement of lipids in Alzheimer's disease pathology and potential therapies. *Front Physiol.* **11**, 598 (2020).
54. Kao, Y.-C., Ho, P.-C., Tu, Y.-K., Jou, I. M. & Tsai, K.-J. Lipids and Alzheimer's disease. *Int. J. Mol. Sci.* **21**, 1505 (2020).
55. Wu, H.-l. et al. Targeting nucleotide metabolism: A promising approach to enhance cancer immunotherapy. *J. Hematol. Oncol.* **15**, 45 (2022).
56. Bayraktar, A. et al. Revealing the molecular mechanisms of Alzheimer's disease based on network analysis. *Int J. Mol. Sci.* **22**, 11556 (2021).
57. Stygelbout, V. et al. Inositol trisphosphate 3-kinase B is increased in human Alzheimer brain and exacerbates mouse Alzheimer pathology. *Brain* **137**, 537–552 (2014).
58. Lohr, K. M., Frost, B., Scherzer, C. & Feany, M. B. Biotin rescues mitochondrial dysfunction and neurotoxicity in a tauopathy model. *Proc. Natl. Acad. Sci.* **117**, 33608–33618 (2020).
59. Garcia Corrales, A. V., Haidar, M., Bogie, J. F. J. & Hendriks, J. J. A. Fatty acid synthesis in glial cells of the CNS. *Int. J. Mol. Sci.* **22**, 8159 (2021).
60. Li, B. & Ritchie, M. D. From GWAS to gene: Transcriptome-wide association studies and other methods to functionally understand GWAS discoveries. *Front. Genet.* **12**, 713230 (2021).

Acknowledgements

This study was financially supported by a grant from The Scientific and Technological Research Council of Türkiye (TUBITAK) (Project Code: 120S824). Also, Dilara Uzuner Odongo is funded by TUBITAK 2211-A PhD Student Scholarship and Council of Higher Education (CoHE) 100/2000 PhD Scholarship Programs. Study data in ROSMAP cohort were provided by the Rush Alzheimer's Disease Center, Rush University Medical Center, Chicago, IL, USA. The Mayo RNAseq study data was led by Dr. Nilüfer Ertekin-Taner, Mayo Clinic, Jacksonville, FL as part of the multi-PI U01 AG046139 (MPIs Golde, Ertekin-Taner, Younkin, Price). Samples were provided from the following sources: The Mayo Clinic Brain Bank. MSBB data were generated from postmortem brain tissue collected through the Mount Sinai VA Medical Center Brain Bank and were provided by Dr. Eric Schadt from Mount Sinai School of Medicine. The data available in the AD Knowledge Portal would not be possible without the participation of research volunteers and the contribution of data by collaborating researchers.

Author contributions

A.I. processed raw RNAseq data and generated count and variant calling files and wrote the relevant method sections. F.B.B. performed detection of pathogenic variants with GenePy and wrote the relevant method sections. D.U.O. generated personalized genome-scale metabolic models, conducted statistical tests, prepared the figures and codes, and wrote the manuscript. T.Ç. supervised the study, and reviewed and edited the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42003-025-07941-z.

Correspondence and requests for materials should be addressed to Tunahan. Çakır.

Peer review information *Communications Biology* thanks Mohammad Arif, Lucy Bicks and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editors: Hélène Choquet and Aylin Bircan, Benjamin Bessieres.