**Construction of A Dataset for All Expressed Transcripts for Alzheimer's Disease Research**

Zhenyu Huang1,2, Bocheng Shi2,3, Xuechen Mu2,3, Siyu Qiao2, Gangyi Xiao1, Yan Wang1\*, Ying Xu2\*

1College of Computer Science and Technology, Jilin University, Changchun, China, 2Systems Biology Lab for Metabolic Reprogramming, Department of Human Genetics and Cell Biology, School of Medicine, Southern University of Science and Technology, Shenzhen, China, 3School of Mathematics, Jilin University, Changchun, China.

\*Correspondence: wy6868@jlu.edu.cn, [xuy9@sustech.edu.cn](mailto:xuy9@sustech.edu.cn)

**ABSTRACT**

Accurate identification and functional annotation of splicing isoforms and non-coding RNAs (lncRNAs), in addition to full-length protein-encoding transcripts, is essential for detailed understanding of the complexities of gene (mis)regulation and metabolic reprograming in complex diseases like Alzheimer’s disease (AD). In this study, we present a comprehensive set of transcripts, including splicing isoforms and lncRNAs, assembled from publicly available raw RNAseq reads expressed in pre-AD and AD tissues, and their functional annotations. These data and information expand considerably the scope and depth of the relevant data currently available in public databases. Our analyses have identified and corrected numerous false identification and annotations in the existing AD transcripts. We have then demonstrated that our improved transcript identification and function annotation can provide substantially more information related to AD onset and development compared to the currently available data. We anticipate that this new data source, coupled with the associated analysis tools, will provide highly valuable information to AD researchers.

**KEYWORDS**: Alzheimer's disease, lncRNAs, splicing variants, Fenton reaction, metabolic reprogramming

**CONCLUSION**

Transcriptomic data have proven to be the most powerful type of omic data for disease studies in terms of derivation of their drivers and key molecular mechanisms. Our study has demonstrated that the currently available transcriptomic data for AD and pre-AD tissues in the public domain are far from being adequate for detailed studies of the disease. Our new data source has proven to be significantly more informative than the existing one for AD studies.

**1. INTRODUCTION**

Vast majority of the transcriptomic data analyses of human diseases focus on expression data of full-length transcripts of protein-encoding genes. However, such data may not be adequate in full elucidation of the disease drivers and key mechanisms as demonstrated in studies of autism spectrum disorder (ASD), bipolar disorder (BD) and Schizophrenia (SCZ) [1]. Inspired by such studies, we present here a computational framework for accurate identification of all expressed transcripts of protein genes, namely all identifiable full-length and splicing isoforms, and of long non-coding RNA genes or lncRNAs, and demonstrate how the splicing variants and lncRNAs could add substantially to our understanding of the disease drivers and development, for Alzheimer's disease (AD).

AD is the most common form of dementia, characterized by progressive decline of the cognitive capability as a result of massive neuronal cell death, with no effective treatment currently available [2]. While RNA sequencing (RNAseq) data have been generated from postmortem brain samples of both healthy individuals and those with AD [3, 4, 5], published studies predominantly rely on full-length transcripts of protein genes [4, 5, 6]. Similar to the aforementioned studies on other diseases, some recent works on AD and pre-AD MCI (mild cognitive impairment) have shown that inclusion of lncRNAs and splicing variants in AD studies can considerably improve the level of understanding about the disease biology [7, 8, 9].

Our first goal here is to derive all the splicing isoforms and lncRNAs expressed in MCI and AD tissues based on the raw RNAseq reads available in the public domain, along with their expression levels. To accomplish this, we have collected raw RNAseq reads from three public datasets: (i) data generated from the dorsolateral prefrontal cortex region in the ROSMAP cohort [3]; (ii) datasets from regions 22 and 36 of the temporal lobe and from regions 10 and 44 of the frontal lobe in the MSBB database [4]; and (iii) the GSE95587 dataset [5]. De novo assembly of these RNA reads were conducted using StringTie, currently the best-performing assembly software for transcripts publicly available [10]. Then functional annotation of the assembled transcripts was conducted through mapping onto GENCODE, for functional annotations of the newly assembled transcripts in support of MCI/AD studies.

Compared to the publicly available transcriptomic data for MCI and AD tissues, our dataset is 1.89 times larger than the currently available datasets along with 1.845 times more functionally assigned transcripts, enabling functional studies of AD with much expanded scope and higher depth. To the best of our knowledge, this work represents the first effort to build a most comprehensive transcript-level dataset for MCI and AD studies.

1. **RESULTS**

**2.1 Create a transcriptomic dataset for Alzheimer's disease**

We have selected 639 samples from the DLPC region in the ROSMAP database [3], which is an earliest region affected by AD [6]. Then a filtering process is conducted on these samples to remove samples meeting the following criteria: samples diagnosed as another form of dementia by itself or in combination with AD. Then a quality control (QC) step is executed to remove samples with low quality as defined in Methods 4.2.1, giving rise to 528 samples (Figures 1A).

The raw RNA reads of each sample are assembled to full transcripts and mapped onto the reference human genome, using the methods in Methods 4.2.2. Then all expressed lncRNAs and splicing isoforms are identified and extracted using the methods given in Methods 4.2.3 and 4.2.5 (also Figures 1B -C).

For lncRNAs, we have also done the following: 1) identification of novel lncRNAs and classification of them as described in Methods 4.2.3; 2) regulatory target prediction for each lncRNA; and 3) pathways possibly regulated by each lncRNA, using the methods in Methods 4.2.8.

For each splicing variant, we have conducted functional (re-)annotation as follows. Search protein sequence database [11] to find the closest homolog, in human or another organism, that has no other transcript in our assembled dataset being assigned to its function – this guarantees no two splicing variants being assigned to the same function. The following summarizes the transcripts we assembled:

The computer code for data generation is available at <https://github.com/zhenyuh19/A-Model-for-the-Development-of-Alzheimer-s-Disease>.

**2.2 Information derivation from the assembled transcripts**

To explore how alternative splicing events and isoform usage differs between normal and MCI/AD samples, we first analyzed the genomic distribution of all the assembled transcripts.

**Genomic locations of the assembled transcripts**: We have examined the distribution of sequencing reads across different genomic locations. We note that, in both normal and the disease samples, 47.06%, 28.14%, 18.83% and 5.97% of the RNA reads were mapped to the coding sequence (CDS), introns, the 3' untranslated regions (UTRs), and the upstream regions (10K bases) of transcription start sites (TSS) (Figure 2A).

When looking the two disease groups, MCI and AD, separately, some differences are observed. We note that (i) 18.131% and 18.225% of MCI and AD are mapped to intronic regions; and (ii) 28.596% and 27.8825% of MCI and AD are mapped to 3' UTR regions.

Further examination of read coverage across different disease groups showed notable differences. Specifically, intronic coverage progressively increases from the normal to AD samples, with significant differences: MCI samples vs. normal controls having a p-value 0.082, while the AD samples vs. normal controls having a p-value 0.04 (Figure 2B). For the 3' UTR regions of exons, the AD samples show a significant decrease compared to normal samples (p-value = 0.01), whereas the MCI samples exhibit no significant change (p-value = 0.55) (Figure S1).

**Splicing events and isoforms**: We identified a large number of alternative splicing events, with mutually exclusive exons, skipped exons, and retained introns being predominant in both the MCI and AD samples compared to the normal controls (Figures 2C). Among these splicing events, 3,946 were observed in the MCI samples, and 6,929 in the AD samples. And 2,171 genes harboring alternative splicing events were identified in the MCI samples, and 2,815 such genes were detected in AD samples (Table S1).

**Differential expression analyses**: We performed differential expression analyses using the STAR pipeline, not distinguishing among different splicing isoforms for each gene. In the ROSMAP dataset, we identified 371 differentially expressed genes (DEGs) in the MCI samples and 1,195 DEGs in the AD group (Figures 2D). We note that 35.04% of the DEGs are protein-coding genes in the MCI group and 45.27% in the AD group, while 64.96% and 54.73% of the DEGs are lncRNAs in the MCI and AD group, respectively and the rest are pseudogenes (Table 2).

To ensure the generalizability of these findings, we have conducted similar analyses on other AD datasets. The MSBB dataset exhibited a similar pattern to the above, with 38.89% and 37.84% of DEGs are protein-encoding genes in early-stage and late-stage AD samples (Figure 2E). Similar results are observed in other brain regions within the MSBB dataset and in the GEO dataset GSE95587, as summarized in Figure S2 and detailed in Table 2.

Overall, compared to existing transcriptomic datasets for AD tissues, our assembly significantly expands the coverage: we identified a 4.74-fold more (12,020 vs. 2,537) differentially expressed protein-coding transcripts in MCI tissues, while in the AD samples, the coverage showed a slight decrease (5,150 vs. 7,119), representing a 0.72-fold decrease (Figure 3D).

**Pathway enrichment analyses**: We have then conducted pathway-enrichment analyses over the DEGs, and found that the following are enriched: (1) amyloid production; (2) microtubule transport; (3) lipid oxidation; (4) proteasomal degradation of misfolded proteins (Table S2); (5) immune cell infiltration, including B cells and T cells, in the early-stage AD patients, which are typically absent in the brain (Table S2)；and (6) surprisingly, no amyloid formation, a hallmark of early AD pathology, being detected in the enrichment result. This absence was consistently observed across multiple datasets (Table S2), suggesting that amyloid formation may not be as prominent at the transcriptomic gene level as previously assumed, or that the changes are very subtle and dispersed across different isoform groups.

Our analyses have revealed that our assembled transcripts cover 17.7% more exonic regions, 13.5% more intronic regions, and 39.7% more intergenic regions when mapping our assembled transcripts to the reference genome in GENCODE (Figure 3A), hence all these regions being considered *novel.* The following provides a detailed analysis of our assembled transcripts and functional annotations:

**Statistics on the protein-coding transcripts:** Our analyses revealed that 19.3% (19,410 out of 100,566) of the existing transcripts annotated to be protein-coding were actually non-coding (Figure 3B) and 16.28% (34,806 out of 213,785) of the protein-coding transcripts are incorrectly assigned functionally (Figure 3C), where 100,566 is the number of protein-coding transcripts in the GENCODE dataset, and 213,785 represents the number of protein-coding transcripts identified by StringTie.

Furthermore, we discovered 39.7% more transcripts from the previously unannotated protein-coding genes, consisting of 17.7% novel exons and 13.5% novel introns (Figure 3D). Overall, we assembled 431,781 complete transcripts, corresponding to 213,785 protein-coding genes, out of which 178,979 transcripts have been functionally annotated.

**Statistics on LncRNAs:** We identified 55,098 lncRNAs, including 31,574 novel and 23,524 known ones (Figure 3E). Notably, 1,382 originally annotated as lncRNAs have protein-coding potential (Figure 3F), hence actually protein-coding transcripts, while none of the annotated protein-coding transcripts are annotated as lncRNAs by our study (Figure S3).

**2.3. Applications of our assembled transcripts**

To verify that our assembled transcripts provide considerably more information about the development of AD than the existing transcripts, we focused on several key hallmarks of the disease: amyloid plaque formation [12], Tau fibril aggregation [13], extracellular acidosis [14], intracellular pH increase [15], and iron accumulation [12]. Additionally, we have also examined the functional roles played by lncRNAs in mediating metabolic reprogramming to alleviate extracellular acidosis and intracellular alkalosis. For the simplicity of discussion, we refer to the transcripts expressed in AD and currently available in the public domain as *the existing transcripts*, and our assembled transcripts using expressed RNAseq reads in AD tissues as *our transcripts*.

**Amyloid plaque formation:** The existing transcripts consist of 774 amyloid-related transcripts vs. 854 such transcripts in our transcripts (Table 3). The existing transcripts expressed in the MCI group did not enrich the pathway of amyloid formation while our transcripts enrich multiple related pathways such as decline in amyloid clearance capability and significantly elevated expression of APP precursor protein (Figure 4A). These differences enable us to accurately classify normal, MCI, and AD samples using amyloid-related transcripts (Figure 4B). Overall, our improved transcript dataset provides more information leading to deeper insights about the molecular mechanisms driving amyloid pathology, particularly in early stages of AD.

One example is that extracellular acidosis is a hallmark of AD and contributes to neuronal death. The accumulation of extracellular copper ions triggers Fenton reactions that generate OH⁻, which helps mitigate extracellular acidosis. Additionally, the hydroxyl radicals (˙OH) produced during these reactions promote the aggregation of Aβ into plaques. These amyloid plaques, in turn, form alkaline structures that further contribute to the progression of AD by stabilizing Aβ monomers and altering the extracellular environment [16].

**Tau fibril aggregation:** Our transcripts cover 347 Tau fibril-related transcripts, compared to 154 in the existing transcripts (Table 3), which give more accurate classification among normal, MCI and AD samples using such transcripts as shown in Figure S4. In addition, the significantly higher number of Tau-fibril related differentially expressed transcripts (DETs) enable more accurate understanding about the roles played by Tau hyperphosphorylation and Tau-fibril aggregation in AD progression as shown in Figure S5.

A related example suggests that Tau hyperphosphorylation releases H+, helping mitigate intracellular pH elevation, while enriched pathways show that Tau fibril formation significantly suppresses synapse formation, leading to impaired neurotransmitter release and contributing to extracellular acidosis [16].

**Intracellular alkalosis and extracellular acidosis:** Intracellular alkalosis and extracellular acidosis are established features of AD [14, 15]. Our transcripts consist of 10 more intracellular alkalosis related transcripts than the existing transcripts (Table 3). These transcripts also exhibit significantly higher expressions in AD samples than the normal and MCI groups (Figure 4C), suggesting that intracellular alkalosis is more pronounced in AD than previously thought [17].

The expression levels of ASIC2 and ASIC3, widely used biomarkers for extracellular acidosis in our transcripts are significantly higher in AD and MCI samples than the normal samples, which are bot detectable in the existing data (Figure 4D). This suggests that our transcripts provide more information enabling for detailed studies of the pH dysregulation in AD.

**Iron accumulation and Fenton reactions:** Previous studies have reported persistent Fenton reactions in AD tissue cells [12] as follows: , where serves as the catalyst, where iron accumulation is observed [18]. Our new transcripts reveal a significant increase in iron accumulation in AD (Figure 4E), covering 371 more iron-related transcripts, specifically 1,265 vs. 904 in matching normal controls. Furthermore, we have confirmed the availability of the other reactants of the Fenton reaction, namely superoxide anion and hydrogen peroxide based on our previous work [17] (Figure S6).

**Roles played by lncRNAs in pH regulation:** Our analysis of lncRNAs revealed their critical roles in keeping the intracellular pH stable, namely in response to intracellular alkalosis and extracellular acidosis. Compared to previous datasets, our assembled transcripts consist of 496 unique lncRNAs that are directly involved in pH regulation pathways, a significant increase from the 119 lncRNAs documented in traditional datasets (Table 3). These transcripts were found to mediate key cellular processes related to pH balance, as outlined below:

1. Intracellular alkalosis: LncRNAs in our dataset promote mitochondrial iron accumulation, which in turn produces hydroxide () and hydroxyl radical ( ) through Fenton reactions (Figure S7), causing intracellular alkalization. Specifically, 2,471 lncRNAs are involved in regulating Fenton reactions. On the other side, lncRNAs are also involved in actions counterbalancing the alkalization via activating acidification process. For example, lncRNAs MSTRG.44295.1 and MSTRG.14265.27 contribute to the activation of acid-loading transporters that lower the intracellular pH. Overall, our transcripts cover 646 lncRNAs vs. 309 lncRNAs in the existing transcripts relevant to intracellular pH (Table 3).
2. Extracellular acidosis: Our transcripts cover 812 vs 229 lncRNAs by the existing transcripts related to regulating bicarbonate (HCO3-) transporters, which help to slow down the process of extracellular acidosis (Table 3).

We also note thatin the early stage of AD, lncRNAs are involved in promoting the formation of amyloid plaques and Tau protein hyperphosphorylation. Our previous work has discovered that the formation of amyloid plaques slows down the process of extracellular acidosis since the folded structure of amyloid plaques tend to keep their alkaline sidechains pointing outwards while having their acidic sidechains folded inside the structure [16,17]. Similarly, we have also discovered that hyperphosphorylation on Tau proteins reduces intracellular alkalinity since each phosphorylation act releases one net proton [16, 17]. Overall, our transcripts over 990 lncRNAs involved in these two processes compared to 286 such lncRNAs in the existing transcripts.

Overall, our transcripts not only increase the number of relevant transcripts and enrich more related pathways but also improve the level of details in AD-related mechanisms that can be studied. All phenotypes and related transcripts involved in this framework are listed in Table S3, and a comparison between the numbers of transcripts related to each of the key mechanisms in our transcripts and the existing ones are listed in Table 3.

**3. DISCUSSION**

In this study, we constructed a transcriptomic dataset for AD and MCI tissues by assembling raw RNA reads in the public domain, which considerably expands the existing transcript dataset [18, 19], enabling more in-depth and comprehensive studies of AD biology. Specifically, our data covers 178,979 protein-coding transcripts and 55,098 lncRNAs. In addition to this new and the state-of-the-art transcriptomic dataset, we have developed a computational pipeline for assembling full-length transcripts from raw RNAseq reads and for functional annotation for each transcript through mapping it uniquely to a known protein sequence in human or other organisms with known functions. Furthermore, a few analysis tools are also included in the pipeline, including pathway enrichment analysis and prediction of the regulatory target genes or pathways of lncRNAs. By applying our transcript data and the computational pipeline, we have discovered:

**1. Correction of false positives and negatives in protein-coding transcripts**: By applying our computational pipeline, we identified and corrected a large number of false positive and false negative protein-coding transcripts in the existing AD transcripts, significantly improving both the size and the quality of the transcriptomic data for AD research.

**2. LncRNA and pH imbalance**: The substantially increased pool and the improved quality of the AD transcripts have enabled our in-depth studies of AD drivers and molecular mechanisms as highlighted here and detailed in our previous study [16, 17], particularly how lncRNAs are involved in keeping the pH stable [16, 17]

Our data also provide new types of information regarding the metabolic reprogramming initiated in response to pH imbalance in AD. Specifically, our findings support the notion that AD faces pH imbalance, which triggers a series of metabolic reprogramming events, such as the activation of cytosolic acidifying transporters [16, 20]. Notably, the main mechanism for alleviating intracellular alkalosis involves H+-producingenzymes [21]. Therefore, our data enable the identification of acidifying metabolic pathways involved in keeping pH stable in AD, offering a novel perspective on the metabolic reprogramming associated with the disease progression.

There are several limitations in this study. First, our analysis relies on postmortem brain tissue, which may not fully capture the dynamic changes occurring during the earlier stages of the disease [22]. Additionally, while our dataset provides significant improvements in transcript coverage, there may still be rare or low-abundance transcripts that remain undetected. Finally, functional validation of the newly identified transcripts and their roles in AD progression remains to be explored, and hence further experimental studies will be needed to confirm their biological relevance.

In summary, our assembled transcriptomic dataset represents a significant advance in AD research, enabling new tools and insights that could enhance the understanding of disease mechanisms and improve the study of MCI and AD biology.

**4. MATERIALS AND METHODS**

**4.1 Data collection**

The raw RNAseq data were obtained from the ROSMAP cohort [3], consisting of 639 samples. Based on the "Physician's overall cognitive diagnostic category" score, 529 samples (cogdx values of 1, 2, and 4) were selected, including 146 MCI samples, 193 AD samples, and 189 region-matched control samples. In addition, transcriptomic data from various brain regions in the MSBB cohort [4] and data from GSE95587 for the fusiform gyrus [5] were used in this study for identification of differently expressed transcripts.

Clinical staging in the MSBB cohort was based on the Clinical Dementia Rating (CDR), while that in the GSE95587 data is based on the Braak staging system. Detailed information on the reassembled samples is provided in Table S6.

**4.2. Methods**

**4.2.1 Quality control and preprocessing**

To ensure high data quality, we filtered out low-quality reads from samples in the DLPC brain region of the ROSMAP cohort, resulting in approximately 2.11 TB clean bases. Specifically, for each sample, at least 90% of the reads were required to have a quality score at least Q30 (error rate ≤ 0.1%). SOAPnuke was used to filter out reads with GC distribution fluctuations exceeding 10% [23]. Clean reads were mapped to the rRNA database using RSEQC [24] to remove rRNA-derived reads. Sequencing randomness and gene saturation were also identified and filtered out using RSEQC.

**4.2.2 Genome mapping, transcript assembly, and quantification**

After quality control, the remaining reads were aligned to the reference human genome (GENCODE version 34) using HISAT2 (version 2.7) [25] with default parameters. StringTie was used to assemble transcripts [26], and the "merge" option was applied to unify all output files into a single transcriptome. Gffcompare [27] was used to compare the assembled transcriptome (in GTF format) with reference annotations to obtain the genomic location of each transcript. Transcript expression levels were quantified in transcripts per million (TPM) using Ballgown [28]. The genome FASTA file and the annotation of the assembled transcripts are available in Data S1.

**4.2.3 Identification and classification of lncRNAs**

To identify lncRNA genes, we employed a two-step approach involving protein-coding potential evaluation followed by lncRNA classification.

**Step 1: Protein-coding potential evaluation.** Transcripts were initially filtered based on length and exon count, with only transcripts consisting of at least 200 nucleotides and having two exons or more retained for further analyses [29]. Their protein-coding potential was assessed using five computational methods: CNCI [30], CPAT [31], CPC [32], CPPred [33] and PLEK [34]. Transcripts were classified as lncRNAs only if all five methods all predict them as noncoding.

**Step 2: LncRNA classification.** New transcripts that were not annotated in GENCODE [35] were categorized based on their genomic location as follows:

* "i": transcripts entirely within an intron;
* "p": transcripts located within 2 Kb of a previously known transcript;
* "y": transcripts containing at least one reference intron; and
* "u": transcripts in intergenic regions.

Additionally, all identified lncRNAs were further classified into intergenic, antisense, or intronic lncRNAs based on their genomic positions.

**4.2.4. Differential expression and pathway enrichment analyses**

Differential expression analyses were performed to identify differentially expressed transcripts (DETs) between (MCI ⋃ AD) samples and controls using the "DESeq2" function [36]. Transcripts with a fold change (|FC|) ≥ 1.3 and a false discovery rate (FDR) < 0.05 were considered differentially expressed, while transcripts with normalized read counts below 1 were excluded.

The identified DETs were subject to pathway enrichment analysis against KEGG, REACTOME, and GO Biological Process databases. Only pathways with an enrichment P-value < 0.05 were considered for further study.

**4.2.5 Identification of splicing isoforms**

Assembled transcripts were mapped onto the reference genome (GENCODE v34), using STAR (v. 2.7.10b), which generates BAM files. Then rMATS-Turbo was run on the BAM files using the following parameters: “--chimSegmentMin 2 –outFilterMismatchNmax 3 --twopassMode Basic --alignEndsType EndToEnd --alignSJDBoverhangMin 1 --alignIntronMax 299999”.

We specified the reads type, length, and strand specificity with “-t paired --readLength 101 --variable-read-length --libType fr-firststrand”. rMATS is used to identify splicing events with p-values < 0.05 and classify them into one of the following five types: SE (skipped exon), A5SS (alternative 5’ splice site), A3SS (alternative 3’ splice site), MXE (mutually exclusive exon), and RI (retained intron). Then one more round of filtering is conducted to remove splicing events with |ΔPSI| > 0.05, where PSI is for Percent Spliced In [37], used to measure the proportion of transcripts that include a particular exon or splicing event compared to the total number of transcripts.

**4.2.6. Gene set enrichment analyses for individual samples using gsva**

To perform gene set enrichment analysis (GSEA) on individual samples, we used the "gsva()" function in R with the "ssgsea" method [38]. This approach ranks genes based on their expression levels and calculates the enrichment score for each gene set, where a positive score indicates that genes in the set are highly expressed, while a negative score otherwise.

**4.2.7. Regression analyses**

To assess the co-expression level between an individual gene g and a gene set M, we employed a principal component-based approach. Principal components (PCs) were calculated using the expression levels of the genes in the set across the specified samples. PCs that can collectively explain at least 60% of the cumulative variance are retained. Then a linear regression of the expression level of g is conducted against all the retained PCs, namely

where { are parameters to be determined through minimizing ||.

**4.2.8 Prediction of target genes and pathways regulated by lncRNAs**

LncRNAs can regulate gene expressions either in their genomic neighborhood (cis-regulation) or at distant locations (trans-regulation) [39, 40]. Cis-regulated target genes of a lncRNA were predicted based on locational proximity within 100kb upstream or downstream of the lncRNA if co-expressed with the lncRNA [41, 42]. Trans-regulated genes were predicted based on sequence complementarity and co-expression between the lncRNA and the target genes, using RIsearch [43]. Prediction criteria also include the free energy of the bound structure between the lncRNA and each DNA sequence it binds (energy < -10) [44].

A lncRNA was predicted to be a regulator of a pathway if (1) an optimal regression model of the lncRNA against the PCs of the pathway yielded a high R² value with an adjusted p-value < 0.05, and (2) the lncRNA is predicted to be a cis or trans regulator of some genes in the pathway, with a Spearman correlation coefficient |r| > 0 and an adjusted p-value < 0.05.

**4.2.8. Multi-class classification analysis using LGBM**

A multi-class classification was constructed using LightGBM (LGBM) [45] to classify a tissue sample as a normal, MCI, or AD sample based on the gene expression data of a pre-defined gene set. The classification performance of the gene set was evaluated using the ROC curve and the AUC value, without feature elimination.

**Reference**

[1] M. J. Gandal *et al.*, “Transcriptome-wide isoform-level dysregulation in ASD, schizophrenia, and bipolar disorder,” *Science*, vol. 362, no. 6420, p. eaat8127, Dec. 2018, doi: 10.1126/science.aat8127.

[2] W. M. van der Flier, M. E. de Vugt, E. M. A. Smets, M. Blom, and C. E. Teunissen, “Towards a future where Alzheimer’s disease pathology is stopped before the onset of dementia,” *Nat. Aging*, vol. 3, no. 5, pp. 494–505, May 2023, doi: 10.1038/s43587-023-00404-2.

[3] D. A. Bennett, A. S. Buchman, P. A. Boyle, L. L. Barnes, R. S. Wilson, and J. A. Schneider, “Religious Orders Study and Rush Memory and Aging Project,” *J. Alzheimers Dis. JAD*, vol. 64, no. Suppl 1, pp. S161–S189, 2018, doi: 10.3233/JAD-179939.

[4] M. Wang *et al.*, “The Mount Sinai cohort of large-scale genomic, transcriptomic and proteomic data in Alzheimer’s disease,” *Sci. Data*, vol. 5, no. 1, p. 180185, Sep. 2018, doi: 10.1038/sdata.2018.185.

[5] B. A. Friedman *et al.*, “Diverse Brain Myeloid Expression Profiles Reveal Distinct Microglial Activation States and Aspects of Alzheimer’s Disease Not Evident in Mouse Models,” *Cell Rep.*, vol. 22, no. 3, pp. 832–847, Jan. 2018, doi: 10.1016/j.celrep.2017.12.066.

[6] A. P. Pérez-González *et al.*, “The ROSMAP project: aging and neurodegenerative diseases through omic sciences,” *Front. Neuroinformatics*, vol. 18, Sep. 2024, doi: 10.3389/fninf.2024.1443865.

[7] D. Marques-Coelho, L. da C. C. Iohan, A. R. Melo de Farias, A. Flaig, J.-C. Lambert, and M. R. Costa, “Differential transcript usage unravels gene expression alterations in Alzheimer’s disease human brains,” *Npj Aging Mech. Dis.*, vol. 7, no. 1, pp. 1–15, Jan. 2021, doi: 10.1038/s41514-020-00052-5.

[8] H. Wu *et al.*, “Comprehensive transcript-level analysis reveals transcriptional reprogramming during the progression of Alzheimer’s disease,” *Front. Aging Neurosci.*, vol. 15, Jun. 2023, doi: 10.3389/fnagi.2023.1191680.

[9] S. K. Leung *et al.*, “Long-read transcript sequencing identifies differential isoform expression in the entorhinal cortex in a transgenic model of tau pathology,” *Nat. Commun.*, vol. 15, no. 1, p. 6458, Aug. 2024, doi: 10.1038/s41467-024-50486-8.

[10] S. Kovaka, A. V. Zimin, G. M. Pertea, R. Razaghi, S. L. Salzberg, and M. Pertea, “Transcriptome assembly from long-read RNA-seq alignments with StringTie2,” *Genome Biol.*, vol. 20, no. 1, p. 278, Dec. 2019, doi: 10.1186/s13059-019-1910-1.

[11] J. Huerta-Cepas *et al.*, “eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses,” *Nucleic Acids Res.*, vol. 47, no. Database issue, pp. D309–D314, Jan. 2019, doi: 10.1093/nar/gky1085.

[12] T. Rival *et al.*, “Fenton chemistry and oxidative stress mediate the toxicity of the beta-amyloid peptide in a Drosophila model of Alzheimer’s disease,” *Eur. J. Neurosci.*, vol. 29, no. 7, pp. 1335–1347, Apr. 2009, doi: 10.1111/j.1460-9568.2009.06701.x.

[13] G. S. Bloom, “Amyloid-β and tau: the trigger and bullet in Alzheimer disease pathogenesis,” *JAMA Neurol.*, vol. 71, no. 4, pp. 505–508, Apr. 2014, doi: 10.1001/jamaneurol.2013.5847.

[14] Z.-G. Xiong, G. Pignataro, M. Li, S. Chang, and R. P. Simon, “Acid-Sensing Ion Channels (ASICs) as Pharmacological Targets for Neurodegenerative Diseases,” *Curr. Opin. Pharmacol.*, vol. 8, no. 1, pp. 25–32, Feb. 2008, doi: 10.1016/j.coph.2007.09.001.

[15] A. Rijpma, M. van der Graaf, O. Meulenbroek, M. G. M. Olde Rikkert, and A. Heerschap, “Altered brain high-energy phosphate metabolism in mild Alzheimer’s disease: A 3-dimensional 31P MR spectroscopic imaging study,” *NeuroImage Clin.*, vol. 18, pp. 254–261, Feb. 2018, doi: 10.1016/j.nicl.2018.01.031.

[16] Y. Xu, Z. Huang, and X. Mu, “Chapter 14 - Understanding Alzheimer’s disease at the chemistry level,” in *Introduction to Condensed Matter Chemistry*, R. Xu, J. Yu, and W. Yan, Eds., Elsevier, 2024, pp. 379–405. doi: 10.1016/B978-0-443-16140-7.00008-0.

[17] Z. Huang, Q. Chen, X. Mu, Z. An, and Y. Xu, “Elucidating the Functional Roles of Long Non-Coding RNAs in Alzheimer’s Disease,” *Int. J. Mol. Sci.*, vol. 25, no. 17, Art. no. 17, Jan. 2024, doi: 10.3390/ijms25179211.

[18] “Introduction to Condensed Matter Chemistry - 1st Edition | Elsevier Shop.” Accessed: Apr. 16, 2024. [Online]. Available: https://shop.elsevier.com/books/introduction-to-condensed-matter-chemistry/yu/978-0-443-16140-7

[19] D. Marques-Coelho, L. da C. C. Iohan, A. R. Melo de Farias, A. Flaig, J.-C. Lambert, and M. R. Costa, “Differential transcript usage unravels gene expression alterations in Alzheimer’s disease human brains,” *Npj Aging Mech. Dis.*, vol. 7, no. 1, pp. 1–15, Jan. 2021, doi: 10.1038/s41514-020-00052-5.

[20] C. Fan *et al.*, “Systematic analysis to identify transcriptome-wide dysregulation of Alzheimer’s disease in genes and isoforms,” *Hum. Genet.*, vol. 140, no. 4, pp. 609–623, Apr. 2021, doi: 10.1007/s00439-020-02230-7.

[21] J. R. Casey, S. Grinstein, and J. Orlowski, “Sensors and regulators of intracellular pH,” *Nat. Rev. Mol. Cell Biol.*, vol. 11, no. 1, Art. no. 1, Jan. 2010, doi: 10.1038/nrm2820.

[22] D. Trabzuni *et al.*, “Quality control parameters on a large dataset of regionally dissected human control brains for whole genome expression studies,” *J. Neurochem.*, vol. 119, no. 2, pp. 275–282, Oct. 2011, doi: 10.1111/j.1471-4159.2011.07432.x.

[23] Y. Chen *et al.*, “SOAPnuke: a MapReduce acceleration-supported software for integrated quality control and preprocessing of high-throughput sequencing data,” *GigaScience*, vol. 7, no. 1, Dec. 2017, doi: 10.1093/gigascience/gix120.

[24] L. Wang, S. Wang, and W. Li, “RSeQC: quality control of RNA-seq experiments,” *Bioinformatics*, vol. 28, no. 16, pp. 2184–2185, Aug. 2012, doi: 10.1093/bioinformatics/bts356.

[25] Y. Zhang, C. Park, C. Bennett, M. Thornton, and D. Kim, “Rapid and accurate alignment of nucleotide conversion sequencing reads with HISAT-3N,” *Genome Res.*, vol. 31, no. 7, pp. 1290–1295, Jan. 2021, doi: 10.1101/gr.275193.120.

[26] A. Shumate, B. Wong, G. Pertea, and M. Pertea, “Improved transcriptome assembly using a hybrid of long and short reads with StringTie,” *PLOS Comput. Biol.*, vol. 18, no. 6, p. e1009730, Jun. 2022, doi: 10.1371/journal.pcbi.1009730.

[27] G. Pertea and M. Pertea, “GFF Utilities: GffRead and GffCompare,” *F1000Research*, vol. 9, Apr. 2020, doi: 10.12688/f1000research.23297.1.

[28] A. C. Frazee, G. Pertea, A. E. Jaffe, B. Langmead, S. L. Salzberg, and J. T. Leek, “Ballgown bridges the gap between transcriptome assembly and expression analysis,” *Nat. Biotechnol.*, vol. 33, no. 3, pp. 243–246, Mar. 2015, doi: 10.1038/nbt.3172.

[29] Y. Liu *et al.*, “Long non-coding RNA and circular RNA and coding RNA profiling of plasma exosomes of osteosarcoma by RNA seq,” *Sci. Data*, vol. 10, no. 1, Art. no. 1, Jun. 2023, doi: 10.1038/s41597-023-02295-9.

[30] S. Han, Y. Liang, Y. Li, and W. Du, “Long Noncoding RNA Identification: Comparing Machine Learning Based Tools for Long Noncoding Transcripts Discrimination,” *BioMed Res. Int.*, vol. 2016, 2016, doi: 10.1155/2016/8496165.

[31] L. Wang, H. J. Park, S. Dasari, S. Wang, J.-P. Kocher, and W. Li, “CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model,” *Nucleic Acids Res.*, vol. 41, no. 6, p. e74, Apr. 2013, doi: 10.1093/nar/gkt006.

[32] Y.-J. Kang *et al.*, “CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features,” *Nucleic Acids Res.*, vol. 45, no. W1, pp. W12–W16, Jul. 2017, doi: 10.1093/nar/gkx428.

[33] X. Tong and S. Liu, “CPPred: coding potential prediction based on the global description of RNA sequence,” *Nucleic Acids Res.*, vol. 47, no. 8, pp. e43–e43, May 2019, doi: 10.1093/nar/gkz087.

[34] A. Li, J. Zhang, and Z. Zhou, “PLEK: a tool for predicting long non-coding RNAs and messenger RNAs based on an improved k-mer scheme,” *BMC Bioinformatics*, vol. 15, no. 1, p. 311, Sep. 2014, doi: 10.1186/1471-2105-15-311.

[35] J. Harrow *et al.*, “GENCODE: The reference human genome annotation for The ENCODE Project,” *Genome Res.*, vol. 22, no. 9, pp. 1760–1774, Sep. 2012, doi: 10.1101/gr.135350.111.

[36] M. I. Love, W. Huber, and S. Anders, “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2,” *Genome Biol.*, vol. 15, no. 12, p. 550, Dec. 2014, doi: 10.1186/s13059-014-0550-8.

[37] J. Zhou, S. Ma, D. Wang, J. Zeng, and T. Jiang, “FreePSI: an alignment-free approach to estimating exon-inclusion ratios without a reference transcriptome,” *Nucleic Acids Res.*, vol. 46, no. 2, p. e11, Jan. 2018, doi: 10.1093/nar/gkx1059.

[38] A. Subramanian *et al.*, “Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles,” *Proc. Natl. Acad. Sci.*, vol. 102, no. 43, pp. 15545–15550, Oct. 2005, doi: 10.1073/pnas.0506580102.

[39] C. P. Ponting, P. L. Oliver, and W. Reik, “Evolution and functions of long noncoding RNAs,” *Cell*, vol. 136, no. 4, pp. 629–641, Feb. 2009, doi: 10.1016/j.cell.2009.02.006.

[40] A. E. Kornienko, P. M. Guenzl, D. P. Barlow, and F. M. Pauler, “Gene regulation by the act of long non-coding RNA transcription,” *BMC Biol.*, vol. 11, p. 59, May 2013, doi: 10.1186/1741-7007-11-59.

[41] S. Liu *et al.*, “Whole-Transcriptome RNA Sequencing Uncovers the Global Expression Changes and RNA Regulatory Networks in Duck Embryonic Myogenesis,” *Int. J. Mol. Sci.*, vol. 24, no. 22, p. 16387, Nov. 2023, doi: 10.3390/ijms242216387.

[42] X. Li *et al.*, “Global identification and analysis revealed differentially expressed lncRNAs associated with meiosis and low fertility in autotetraploid rice,” *BMC Plant Biol.*, vol. 20, p. 82, Feb. 2020, doi: 10.1186/s12870-020-2290-0.

[43] A. Wenzel, E. Akbaşli, and J. Gorodkin, “RIsearch: fast RNA–RNA interaction search using a simplified nearest-neighbor energy model,” *Bioinformatics*, vol. 28, no. 21, pp. 2738–2746, Nov. 2012, doi: 10.1093/bioinformatics/bts519.

[44] H. Sun *et al.*, “Genome-wide transcriptional profiling and functional analysis of long noncoding RNAs and mRNAs in chicken macrophages associated with the infection of avian pathogenic E. coli,” *BMC Vet. Res.*, vol. 20, 2024, doi: 10.1186/s12917-024-03890-7.

[45] G. Ke *et al.*, “LightGBM: A Highly Efficient Gradient Boosting Decision Tree,” in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper\_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf

**Table1：**Summary of assembled transcript.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Cohort** | **Region** | **Stage** | **#Samples** | **#Protein-encoding transcripts** | **#lncRNAs** |
| ROSMAP | Dorsolateral prefrontal cortex | Normal | 189 | 206,573 | 49,965 |
| ROSMAP | Dorsolateral prefrontal cortex | MCI | 146 | 204,800 | 49,259 |
| ROSMAP | Dorsolateral prefrontal cortex | AD | 193 | 206,175 | 49,988 |

**Table 2.** Composition statistics of differential gene types across various AD datasets.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Cohort** | **Region** | **Group** | **Gene types** | **Number** | **Percentage** |
| ROSMAP | Dorsolateral Prefrontal Cortex | MCI | lncRNA | 154 | 0.415 |
| ROSMAP | Dorsolateral Prefrontal Cortex | MCI | protein-coding | 130 | 0.350 |
| ROSMAP | Dorsolateral Prefrontal Cortex | MCI | other non-coding RNA | 87 | 0.235 |
| ROSMAP | Dorsolateral Prefrontal Cortex | AD | lncRNA | 516 | 0.432 |
| ROSMAP | Dorsolateral Prefrontal Cortex | AD | protein-coding | 541 | 0.453 |
| ROSMAP | Dorsolateral Prefrontal Cortex | AD | other non-coding RNA | 138 | 0.115 |
| MSBB | Superior Temporal Gyrus | CDR I-II | lncRNA | 261 | 0.110 |
| MSBB | Superior Temporal Gyrus | CDR I-II | protein-coding | 1067 | 0.448 |
| MSBB | Superior Temporal Gyrus | CDR I-II | other non-coding RNA | 1055 | 0.443 |
| MSBB | Parahippocampal Gyrus | CDR I-II | lncRNA | 491 | 0.123 |
| MSBB | Parahippocampal Gyrus | CDR I-II | protein-coding | 1671 | 0.417 |
| MSBB | Parahippocampal Gyrus | CDR I-II | other non-coding RNA | 1844 | 0.460 |
| MSBB | Inferior Frontal Gyrus | CDR I-II | lncRNA | 506 | 0.116 |
| MSBB | Inferior Frontal Gyrus | CDR I-II | protein-coding | 1741 | 0.399 |
| MSBB | Inferior Frontal Gyrus | CDR I-II | other non-coding RNA | 2119 | 0.485 |
| MSBB | Superior Temporal Gyrus | CDR III-IV | lncRNA | 1206 | 0.146 |
| MSBB | Superior Temporal Gyrus | CDR III-IV | protein-coding | 3395 | 0.411 |
| MSBB | Superior Temporal Gyrus | CDR III-IV | other non-coding RNA | 3666 | 0.443 |
| MSBB | Parahippocampal Gyrus | CDR III-IV | lncRNA | 1356 | 0.117 |
| MSBB | Parahippocampal Gyrus | CDR III-IV | protein-coding | 5927 | 0.513 |
| MSBB | Parahippocampal Gyrus | CDR III-IV | other non-coding RNA | 4273 | 0.370 |
| MSBB | Inferior Frontal Gyrus | CDR III-IV | lncRNA | 976 | 0.121 |
| MSBB | Inferior Frontal Gyrus | CDR III-IV | protein-coding | 2649 | 0.329 |
| MSBB | Inferior Frontal Gyrus | CDR III-IV | other non-coding RNA | 4436 | 0.550 |
| GSE95587 | Fusiform Gyrus | Braak III-IV | lncRNA | 316 | 0.257 |
| GSE95587 | Fusiform Gyrus | Braak III-IV | protein-coding | 722 | 0.587 |
| GSE95587 | Fusiform Gyrus | Braak III-IV | other non-coding RNA | 191 | 0.155 |
| GSE95587 | Fusiform Gyrus | Braak V-VI | lncRNA | 323 | 0.211 |
| GSE95587 | Fusiform Gyrus | Braak V-VI | protein-coding | 754 | 0.492 |
| GSE95587 | Fusiform Gyrus | Braak V-VI | other non-coding RNA | 456 | 0.297 |

**Note：**CDR (Clinical Dementia Rating): a standard used to assess the severity of dementia of patients. Braak staging is a system for measuring the progression level of AD based on neuropathological changes in the brain.

**Table 3.** The number of transcripts associated with AD hallmarks in our vs. the existing transcripts.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Phenotype** | **Our protein-coding transcripts** | **The existing protein-coding transcripts** | **Our** **lncRNA transcripts** | **The existing lncRNA transcripts** |
| Intracellular alkalosis | 125 | 115 | 496 | 119 |
| Extracellular acidosis | 39 | 28 | 605 | 310 |
| Mitochondrial iron accumulation | 1265 | 904 | 819 | 340 |
| Amyloid formation | 854 | 774 | 833 | 286 |
| Tau fiber formation | 347 | 154 | 1077 | 306 |
| Superoxide generation | 314 | 376 | 825 | 333 |
| Hydrogen peroxide generation | 213 | 186 | 827 | 324 |
| Acid loading transporter | 182 | 117 | 646 | 309 |
| Bicarbonate transporter | 146 | 122 | 812 | 229 |

日程表

中度可信度描述已自动生成

**Figure 1**. Bioinformatics analysis workflow in our study. **A**. The orange box represents data quality control criteria, including quality control of reads, filtering of rRNA reads, transcript assembly, and transcript type identification. **B**. The green box represents the criteria for LncRNA identification and analysis, considering the relationship between a transfrag and the closest reference, transcript length, exon number, read coverage, five non-coding potential predictions, and target gene prediction. **C.** The blue box represents the criteria for protein-coding RNA identification and analysis, including four protein-coding potential predictions, read coverage, functional reassignment, differential transcript analysis, and functional enrichment.

图示

描述已自动生成

**Figure 2.** Statistical characteristics of the AD dataset. **A.** Distribution of reads across different genomic regions. **B.** Intronic alignment rates across Normal, MCI, and AD groups. **C.** Number of genes affected by five significantly different alternative splicing events in the MCI and AD groups compared to the normal group. **D.** Differential gene statistics between MCI, AD, and Normal groups with significantly different non-coding protein genes specifically labeled. **F.** Composition of differential gene types across different stages in the ROSMAP and MSBB datasets.

图表, 图示, 示意图

描述已自动生成

**Figure 3.** False positive and false negative rates of isoforms in the current transcriptomic datasets. **A.** Improvement in identifying novel exonic, intronic, and intergenic regions using the optimized genome annotation file compared to the GENCODE genome. **B.** False positives in protein-coding transcripts, where the overlap refers to transcript names found in both GENCODE and StringTie annotations. **C.** False negatives in protein-coding transcripts, with the overlap indicating transcript names shared between the GENCODE database and StringTie annotation, but linked to different protein-coding transcripts. **D.** Proportions of differentially expressed protein-coding transcripts (DETs) in GENCODE and StringTie. The x-axis represents the source of the protein-coding transcripts, and the y-axis shows the proportion of each transcript type. **E.** Statistics on the number of known lncRNAs and newly assembled lncRNAs. **F.** False positives in lncRNA transcripts, where the overlap refers to transcript names present in both GENCODE and StringTie annotations.

图表, 箱线图

描述已自动生成

**Figure 4.** Functional validation of the new framework. **A.** Enrichment results of amyloid formation-associated transcripts under the new framework. **B.** AUC evaluation of amyloid formation-related transcripts using GENCODE and StringTie. **C.** Evaluation of intracellular alkalosis under the new framework. **D.** Assessment of extracellular acidosis under the new framework. **E.** Evaluation of mitochondrial iron accumulation under the new framework.