


RESEARCH

Open Access

Identification of potential blood biomarkers for early diagnosis of Alzheimer's disease through RNA sequencing analysis



Daichi Shigemizu^{1,2,3*†} , Taiki Mori¹, Shintaro Akiyama¹, Sayuri Higaki¹, Hiroshi Watanabe¹, Takashi Sakurai^{4,5}, Shumpei Niida¹ and Kouichi Ozaki^{1,3†}

Abstract

Background: With demographic shifts toward older populations, the number of people with dementia is steadily increasing. Alzheimer's disease (AD) is the most common cause of dementia, and no curative treatment is available. The current best strategy is to delay disease progression and to practice early intervention to reduce the number of patients that ultimately develop AD. Therefore, promising novel biomarkers for early diagnosis are urgently required.

Methods: To identify blood-based biomarkers for early diagnosis of AD, we performed RNA sequencing (RNA-seq) analysis of 610 blood samples, representing 271 patients with AD, 91 cognitively normal (CN) adults, and 248 subjects with mild cognitive impairment (MCI). We first estimated cell-type proportions among AD, MCI, and CN samples from the bulk RNA-seq data using CIBERSORT and then examined the differentially expressed genes (DEGs) between AD and CN samples. To gain further insight into the biological functions of the DEGs, we performed gene set enrichment analysis (GSEA) and network-based meta-analysis.

Results: In the cell-type distribution analysis, we found a significant association between the proportion of neutrophils and AD prognosis at a false discovery rate (FDR) < 0.05. Furthermore, a similar trend emerged in the results of routine blood tests from a large number of samples ($n = 3,099$: AD, 1,605; MCI, 994; CN, 500). In addition, GSEA and network-based meta-analysis based on DEGs between AD and CN samples revealed functional modules and important hub genes associated with the pathogenesis of AD. The risk prediction model constructed by using the proportion of neutrophils and the most important hub genes (*EEF2* and *RPL7*) achieved a high AUC of 0.878 in a validation cohort; when further applied to a prospective cohort, the model achieved a high accuracy of 0.727.

Conclusions: Our model was demonstrated to be effective in prospective AD risk prediction. These findings indicate the discovery of potential biomarkers for early diagnosis of AD, and their further improvement may lead to future practical clinical use.

Keywords: Alzheimer's disease, RNA sequencing, Biomarkers for early diagnosis

* Correspondence: d.shigemizu@gmail.com

[†]Daichi Shigemizu and Kouichi Ozaki contributed equally to this work.
¹Medical Genome Center, National Center for Geriatrics and Gerontology, 7-430 Morioka-cho, Obu 474-8511, Aichi, Japan

²Department of Medical Science Mathematics, Medical Research Institute, Tokyo Medical and Dental University (TMDU), Tokyo 113-8510, Japan

Full list of author information is available at the end of the article



© The Author(s). 2020 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

Background

With demographic shifts toward older populations, the number of people with dementia is steadily increasing. The total number of people with dementia worldwide has been estimated to be 75 million by 2030 and 135 million by 2050 [1]. Since there is no treatment or prevention for AD, the current best strategy is to delay disease progression and to practice early intervention to reduce the number of patients that ultimately develop AD [2]. Therefore, promising novel biomarkers for early diagnosis are urgently required [3, 4].

Alzheimer's disease (AD) is the most common cause of dementia, accounting for 60 to 80% of dementia cases [5]. Genome-wide association studies (GWAS) have identified several genetic factors that contribute to AD risk [6–8]. However, the cause of the disease still remains to be elucidated. The current AD diagnosis is generally based on assessing patients' cognitive function. These examinations are not performed routinely, because they are time-consuming and the results largely depend on the physician's experience [9, 10]. Alternatively, cerebrospinal fluid (CSF) biomarkers, including amyloid-beta 1–42 ($A\beta_{1-42}$), total tau (T-tau), and phosphorylated tau 181 (P-tau₁₈₁) [11, 12], and positron emission tomography (PET) imaging scans [13–15] are effective for AD diagnosis, but because of the highly invasive nature of CSF collection and high cost of PET, using these biomarkers as part of a general physical examination to facilitate early diagnosis and therapeutic intervention remains challenging.

Compared with CSF biomarkers and PET imaging scans, blood-based biomarkers are attractive as affordable alternatives for the diagnosis of AD. Mattsson et al. recently reported that plasma neurofilament light level (NfL) has the potential to be a noninvasive biomarker to monitor neurodegeneration in AD [16]. Janelidze et al. reported that plasma P-tau₁₈₁ is a noninvasive diagnostic and prognostic biomarker of AD [17]. One of the most powerful tools for detecting those biomarkers, whole RNA sequencing (RNA-seq) of human peripheral blood mononuclear cells (PBMCs) by using a next-generation sequencer, is widely applied and supports comprehensive analysis of the entire transcriptome [18–20]. The most important application of the RNA-seq data analysis is the identification of differentially expressed genes (DEGs) [21–23]. Systems biology analyses using DEGs reveal key functional modules and important hub genes associated with the pathogenesis of diseases (e.g., Gene Ontology [GO] [24, 25], Kyoto Encyclopedia of Genes and Genomes [KEGG] biological pathways [26, 27]). However, to our knowledge, no previous studies have involved comprehensive RNA-seq analysis of a large number of AD samples and applied an mRNA-based risk prediction model to a prospective cohort.

Here, we performed large-scale RNA-seq transcriptome analyses on a large number of AD samples to detect potential blood-based biomarkers for earlier diagnosis of AD. To this end, we used the RNA-seq data to evaluate cell-type composition among samples from subjects with AD, mild cognitive impairment (MCI), and normal cognitive function (CN) and to compare DEGs between AD and CN samples. Subsequent gene set enrichment analyses (GSEA) and network-based meta-analysis using the DEGs revealed new potential biomarkers for AD diagnosis. The risk prediction model using those potential biomarkers achieved a high AUC in a validation cohort and effectively determined AD risk in a prospective cohort. We believe that, once optimized, these new potential biomarkers will be of practical clinical use in the early diagnosis of AD.

Methods

Sample collection

All of the 610 subjects whose blood samples were evaluated for mRNA expression and their associated clinical data were obtained from the National Center for Geriatrics and Gerontology (NCGG) Biobank, which collects human biomaterials and data for geriatrics research. Of them, 271 subjects were AD patients, 91 subjects were elderly CN controls, and 248 patients had mild cognitive impairment (MCI). All of the subjects were 60 years or older (Supplementary Table S1). The AD and MCI subjects were diagnosed with probable or possible AD according to the criteria of the National Institute on Aging Alzheimer's Association workgroups [9, 10]. Patients with probable AD were used as AD subjects in this study. The CN subjects had subjective cognitive abnormalities but normal cognition on a neuropsychological assessment, which included a comprehensive neuropsychological test, Mini-Mental State Examination (MMSE) score > 27. All of the 3,099 subjects (1,605 ADs, 994 MCIs, and 500 CNs) with the proportion of neutrophils measured in routine blood tests were also obtained from the NCGG Biobank. All of these subjects were also ≥ 60 years in age (Supplementary Table S2).

cDNA library preparation and RNA sequencing

Buffy coat samples were isolated from the whole blood according to the standard operating procedure of NCGG Biobank [28]. Buffy coat fractions containing leukocytes were separated by centrifugation (3,500 rpm, 5 min, RT) and were frozen for further use. Total RNAs in buffy coat samples were isolated using the miRNeasy Mini Kit (Qiagen, Hilden, Germany) according to the manufacturer's instructions with slight modification. TRIzol LS reagent (1 mL) (Thermo Fisher Scientific, MA, USA) and 1-bromo-3-chloropropane (100 μ L) (Tokyo Chemical Industry, Tokyo, Japan) were added to each sample. Samples were mixed thoroughly by shaking for more

than 30 s and incubated at room temperature for 3 min. Phase separation was performed by centrifugation at 15,000g at 4 °C for 15 min. The upper aqueous phase was collected and loaded into the miRNeasy mini-column. After washing, total RNAs were extracted by RNase free water (50 µL). Only high-quality samples with an RNA integrity number (RIN) value ≥ 6.0 were used to construct the sequencing library (Supplementary Table S1). Sequencing libraries were prepared by using 500 µg of total RNA for each sample with Illumina TruSeq Stranded Total RNA with Ribo-Zero Globin and IDT for Illumina-TruSeq UD Indexes according to the manufacturer's instructions (Illumina, San Diego, CA). The libraries were subsequently sequenced by using Illumina NovaSeq6000 platform with paired-end reads of 151 bp according to the manufacturer's instructions.

RNA sequencing data analysis

The quality of the read sequences (fastq files) was assessed by using FastQC (version 0.11.7). The low-quality reads ($< Q20$) and trimmed reads with adaptor sequences (shorter than 50 bp) were discarded by using Cutadapt (version 1.16). The remaining clean sequenced reads were mapped to the human reference genome (GRCh37) by using STAR [29] (2-pass option, version 2.5.2b). By using the featureCounts program [30] from the subread package (version 1.6.6), read counts for each gene were calculated to generate expression levels. Outlier read counts (i.e., the top and bottom 5% of read counts for each gene) were replaced as the maximum and minimum of the remaining effectives, respectively. The read counts from each sample were then combined into a count file, on which differential expression analysis was performed by using edgeR [31] (version 3.18.1). Genes with a threshold CPM (counts per million reads mapped) > 1 in more than one-fourth of all sequenced samples were used for further analysis. The `calcNormFactors` function in edgeR [31] was used to obtain a trimmed mean of M value normalization factors to account for library sizes. Dispersion was calculated by using the `estimateCommonDisp` and `estimateTagwiseDisp` functions in edgeR [31]. The `exactTest` function in edgeR [31] was applied to obtain DEGs between AD and CN samples.

Proportions of immune cell types according to bulk RNA sequencing data

After RNA-seq reads were aligned to the human reference genome by using STAR [29], RSEM [32] (version 1.3.0) was used to quantify transcripts per million (TPM), which were suitable for use with CIBERSORT [33] (version 1.0.1). While CIBERSORT estimated the proportions of 22 immune cell types, we recategorized these 22 cell types into 12 major cell types by summing

the proportions as appropriate. The 12 cell types we evaluated were (1) B cells (naive and memory), (2) plasma cells, (3) CD8⁺ T cells, (4) CD4⁺ T cells (CD4⁺ T cells naive, memory resting, and memory activated; T cells follicular helper; and T cells regulatory), (5) $\gamma\delta$ T cells, (6) NK cells (resting and activated), (7) monocytes, (8) macrophages (M0, M1, and M2), (9) dendritic cells (resting and activated), (10) mast cells (resting and activated), (11) eosinophils, and (12) neutrophils.

In silico biological and functional analysis

Gene Ontology (GO) [24, 25] classification, which is comprised of three major categories—biological process, cellular component, and molecular function—is useful for uncovering the functions of genes of interest. The DAVID [25, 34] (version 6.8) gene functional classification tool (<https://david.ncifcrf.gov>) was used to generate annotations. DAVID was applied to a list of differentially expressed genes with FDR < 0.05 and fold change > 1.2 , and statistically significant GO terms and KEGG biological pathways were identified. Statistically significant GO terms were further expressed as a z -score (the number of upregulated genes minus the number of downregulated genes divided by the square root of the count) and presented in a circular visualization by using the *GOpilot* package (version 1.0.2) in R [35].

Network-based meta-analysis

Network-based analysis was performed by using NetworkAnalyst [36] with the STRING Interactome database [37], which provides comprehensive information regarding interactions between proteins, including prediction and experimental interaction data. The confidence cutoff score was set to 700. The protein–protein interaction (PPI) network was constructed by using zero-order interaction network analysis (direct interaction only) and graphically generated by using Cytoscape v3.7.1 (<http://www.cytoscape.org/>) [38].

Risk prediction model construction

RNA-seq data were split: two-thirds were used for a training data set and one-third for a test data set. Using the training data, we constructed risk prediction models based on clinical information (age, sex, and *APOE* ϵ_4 genotypes), the proportion of neutrophils, and the top-ranked p hub genes using a random forest classifier. The top-ranked p hub genes were then selected stepwise ($p = 1, 2, \dots, 10$). The optimal hyper-parameters in the training data were determined by using 10-fold cross-validation. The adjusted model was then evaluated on the test data, which were completely independent of the training data, by using AUC as the discriminative accuracy of the risk prediction model. The method used in this study was implemented through the *caret* package (version 6.0.76) in R (<https://www.r-project.org/>).

qRT-PCR validation of gene expression

cDNA was synthesized by using a PrimeScriptII 1st Strand cDNA Synthesis Kit (Takara Bio, Shiga, Japan). Quantitative RT-PCR (qRT-PCR) analysis was performed by using customized TaqMan gene expression assays (Applied Biosystems, Waltham, MA) and the Quantstudio7 Flex Real-Time PCR System (Thermo Fisher, Waltham, MA). The following commercially available TaqMan gene expression assays were used: *EEF2* (Hs00157330_m1), *RPL7* (Hs02596927_g1), *LDHB* (Hs00929956_m1), *NR1D2* (Hs00233309_m1), *PDK4* (Hs01037712_m1), *TRIOBP* (Hs00980819_m1), *TAS2R39* (Hs00603443_s1), *BASPI* (Hs00932356_s1), and *ACTB* (Hs01060665_g1). The qRT-PCR conditions were as follows: one cycle of 50 °C for 2 min and 95 °C for 20 s followed by 40 cycles of 95 °C for 1 s, 60 °C for 20 s, and 72 °C for 30 s. Each gene was assayed in duplicate. *ACTB* was pre-selected as a reference gene for normalization of target gene expression levels. Gene expression levels from qRT-PCR were calculated relative to the reference gene *ACTB* using the semi-quantitative method [39]. The gene expressions were obtained for 10 AD and 10 CN randomly selected samples. The log₂ fold change (logFC) was obtained from the average values of the gene expressions.

Results

RNA sequencing data

A total of 610 samples, comprising 271 AD, 248 MCI, and 91 CN samples, were enrolled in this study (Table 1). Using a high-throughput next-generation system to perform RNA sequencing (RNA-seq) analysis, we obtained an average of 44.3, 47.3, and 43.9 million raw read sequences from the AD, MCI, and CN samples, respectively, of which 99.6%, 99.5%, and 99.6% were high-quality (i.e., >Q20) read sequences. After low-quality read sequences were discarded and reads with adaptor sequences were trimmed, 43.8, 47.3, and 43.2 million reads of cleaned data remained for the AD, MCI, and CN samples, respectively, of which 82.7%, 82.1%, and 82.1% uniquely mapped to the human reference genome (GRCh37) (Supplementary Table S3).

Comparison of cell-type distribution among AD, MCI, and CN samples

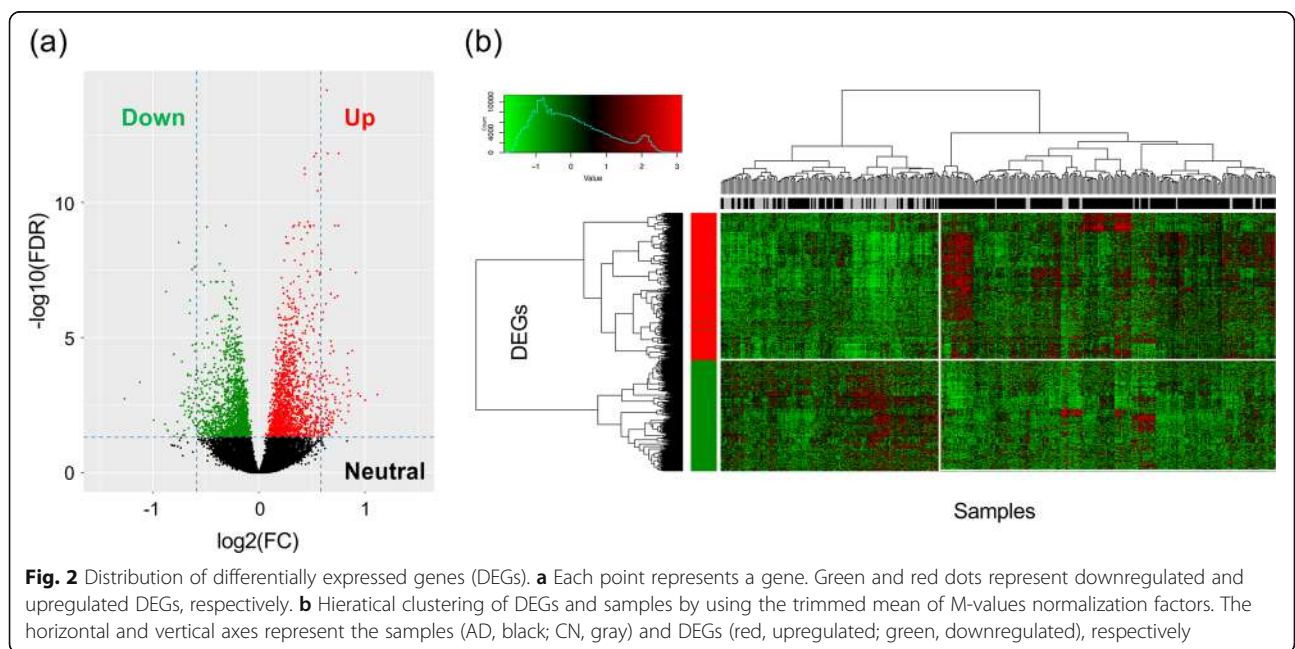
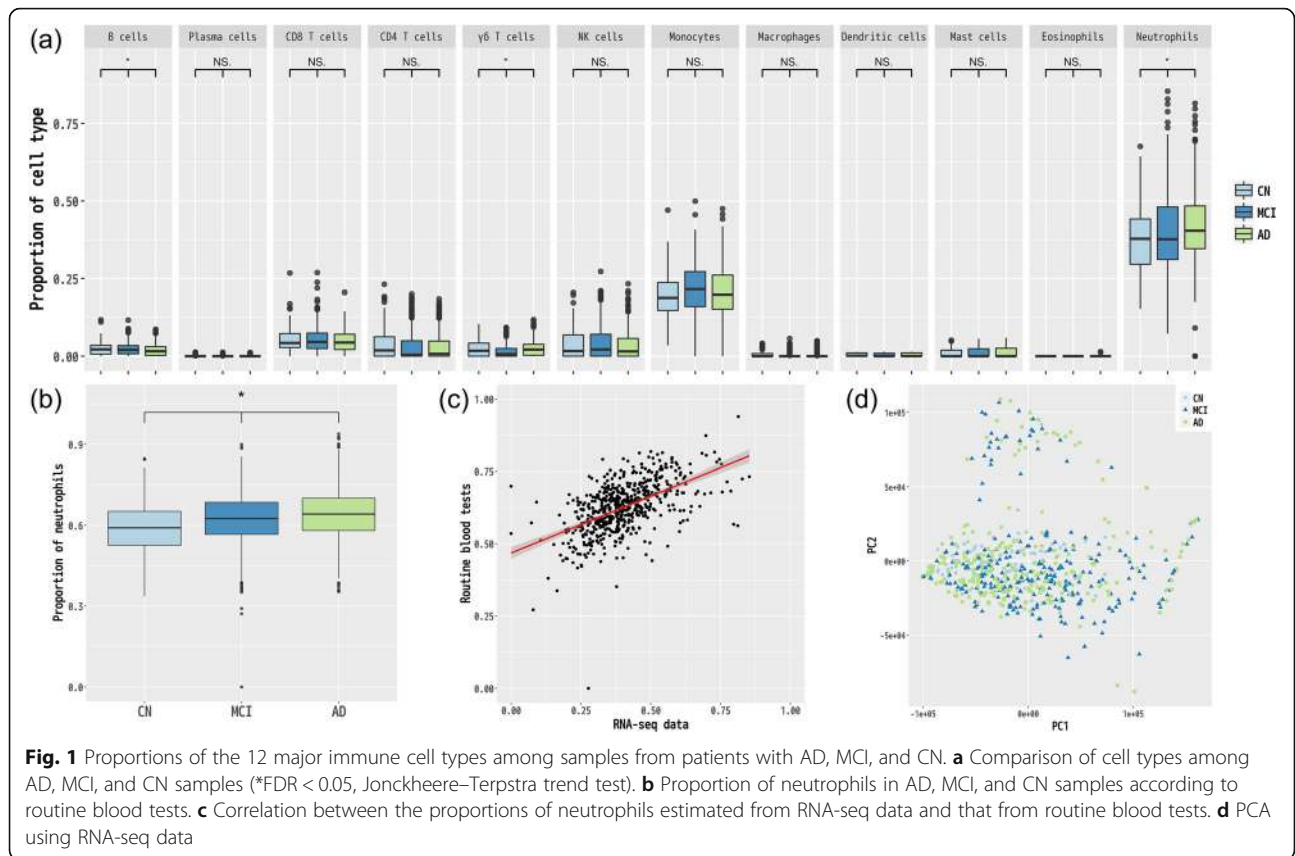
To detect blood-based biomarkers, we first used the bulk RNA-seq data to compare cell-type distribution among AD, MCI, and CN samples. Specifically, CIBERSORT [33] estimated the relative proportions (as transcripts per million [TPM]) of 12 major types of immune cells (i.e., B cells, plasma cells, CD8⁺ T cells, CD4⁺ T cells, $\gamma\delta$ T cells, NK cells, monocytes, macrophages, dendritic cells, mast cells, eosinophils, and neutrophils) in each sample. We used the Jonckheere–Terpstra trend test to identify a statistically significant increase or decrease in cell-type proportion among AD, MCI, and CN samples. Accordingly, the proportion of neutrophils was significantly increased in AD prognosis at an FDR < 0.05 (neutrophils, 0.007; Fig. 1a and Supplementary Table S1). The proportions of B cells and $\gamma\delta$ T cells also showed significant differences in AD prognosis at an FDR < 0.05 (B cells, 0.019; $\gamma\delta$ T cells, 0.007; Fig. 1a and Supplementary Table S1), but these proportions were very low in all samples and too difficult to determine if they were truly associated with the AD prognosis.

To further investigate the association between an increased neutrophil count and AD prognosis, we used a larger number of samples ($n = 3,099$: AD, 1,605; MCI, 994; and CN, 500) to examine the neutrophil population determined through routine blood tests. Interestingly, these data sets obtained by using routine blood tests revealed the same increase in the neutrophil proportion as the RNA-seq data ($P = 0.002$, Jonckheere–Terpstra trend test; Fig. 1b and Supplementary Table S2). Therefore, these results provided strong evidence that an increased neutrophil proportion might be useful as a blood-based biomarker for the diagnosis of AD. The proportion of neutrophils estimated from RNA-seq data was positively correlated with that calculated through routine blood tests (254 ADs, 232 MCIs, and 85 CNs; Pearson $r = 0.56$, $P < 0.01$, Fig. 1c). We also performed a principal component analysis with RNA-seq data of the three groups, but we could not observe the significant difference among the three (Fig. 1d).

Table 1 Summary of characteristics for AD, MCI, and CN samples

| Characteristic | AD | MCI | CN |
|-------------------------|---|---|--|
| Sample number | 271 | 248 | 91 |
| Male:female | 1:2.15 | 1:1.30 | 1:0.82 |
| Age (mean \pm 1 S.D.) | 79.55 \pm 5.83 | 77.37 \pm 6.12 | 71.29 \pm 5.07 |
| MMSE (mean \pm S.D.) | 18.09 \pm 4.49 | 24.54 \pm 2.98 | 29.32 \pm 0.94 |
| APOE genotypes | E2/2 = 2, E3/2 = 14, E3/3 = 148, E4/2 = 3, E4/3 = 88, E4/4 = 16 | E2/2 = 1, E3/2 = 11, E3/3 = 163, E4/2 = 1, E4/3 = 60, E4/4 = 12 | E3/2 = 5, E3/3 = 73, E4/3 = 12, E4/4 = 1 |

MMSE Mini-Mental State Examination (a comprehensive neuropsychological test)



Detection of DEGs

Focusing on the 19,699 genes with a threshold of >1 CPM (counts per million reads mapped) in more than one-fourth of all sequenced samples, we next examined the DEGs in AD and CN samples. A total of 846 statistically significant DEGs (i.e., FDR < 0.05 and fold change > 1.2) with Entrez gene IDs were identified, of which 480 genes were upregulated and 366 were downregulated in the AD samples (Fig. 2a and Supplementary Table S4). In addition, a heatmap of DEGs using the trimmed mean of *M* value normalization factors showed that the expression profiles of the AD and CN samples clustered separately (Fig. 2b).

Biological and functional analysis

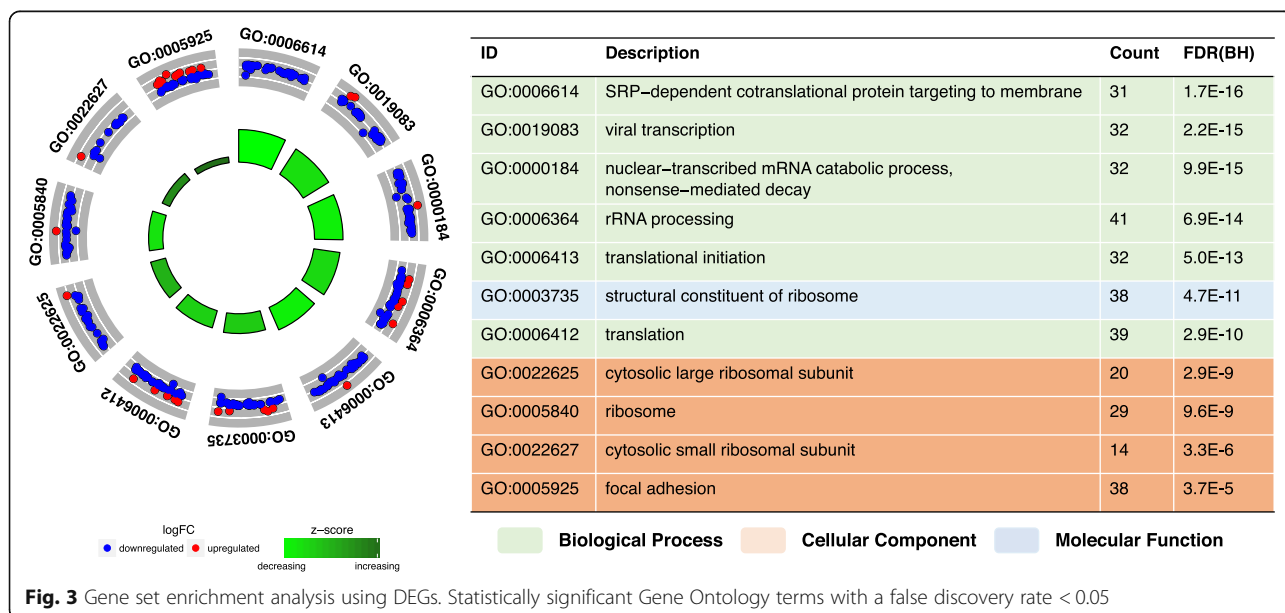
To gain further insight into the biological functions of the DEGs, we performed a gene set enrichment analysis (GSEA) using the DAVID (version 6.8) gene functional classification tool (<https://david.ncifcrf.gov>) [25, 34]. As a result, the DEGs were enriched in 11 GO terms (6 biological processes, 4 cellular components, and 1 molecular function) and one KEGG biological pathway (hsa03010: ribosome), with a significance level set at FDR < 0.05. The enrichment levels of those GO terms are presented in circular visualization (Fig. 3). The GO terms were enriched in many downregulated genes (Fig. 3), and most of them involved ribosomal subunits: 19 RPL genes (*RPL3*, *RPL5*, *RPL6*, *RPL7*, *RPL9*, *RPL10A*, *RPL11*, *RPL18*, *RPL19*, *RPL21*, *RPL22*, *RPL23*, *RPL23A*, *RPL26*, *RPL27*, *RPL29*, *RPL32*, *RPL35*, and *RPL36AL*), 12 RPS genes (*RPS3*, *RPS3A*, *RPS4Y1*, *RPS5*, *RPS6*, *RPS8*, *RPS11*, *RPS12*, *RPS14*, *RPS18*, *RPS24*, and *RPS29*), and 3 MRP genes (*MRPS5*, *MRPL16*, and *MRPL47*).

Network-based meta-analysis

In addition to GSEA, we performed a protein–protein interaction (PPI) network analysis based on the DEGs by using NetworkAnalyst [36] (<http://www.networkanalyst.ca>) with the STRING Interactome database [37]. As a result, we obtained a PPI network comprising 4,164 nodes and 11,886 edges. To prune the network to a more manageable size, we conducted a zero-order interaction network analysis and detected a network containing 161 nodes and 700 edges (Fig. 4). The most highly ranked hub genes were recognized in terms of network topology measures of degree (DC) and betweenness of centrality (BC). The top-ranked 10 hub genes were *EEF2* (eukaryotic elongation factor 2, DC = 38, BC = 883.9, FC = 1.22, FDR = 0.048) and 9 ribosomal proteins: 3 RPL genes (*RPL5*, *RPL7*, and *RPL23A*) and 6 RPS genes (*RPS3*, *RPS3A*, *RPS5*, *RPS6*, *RPS12*, and *RPS24*) (Table 2). Many of the identified genes were common to those obtained through GSEA.

Validation of potential biomarkers of AD in blood

We examined whether many of the top-ranked hub genes could be potential blood biomarkers for AD. For this purpose, two-thirds of all samples were used as a training data set (240 samples: 180 ADs and 60 CNs), and the remaining one-third was used as a test data set (122 samples: 91 ADs and 31 CNs). The top-ranked *p* hub genes were selected stepwise. A risk prediction model was constructed by using clinical information (age, sex, and *APOE* ϵ_4 genotypes), the proportion of neutrophils, and the top-ranked *p* hub genes with a random forest classifier using the training data. The adjusted model was then evaluated on the independent test



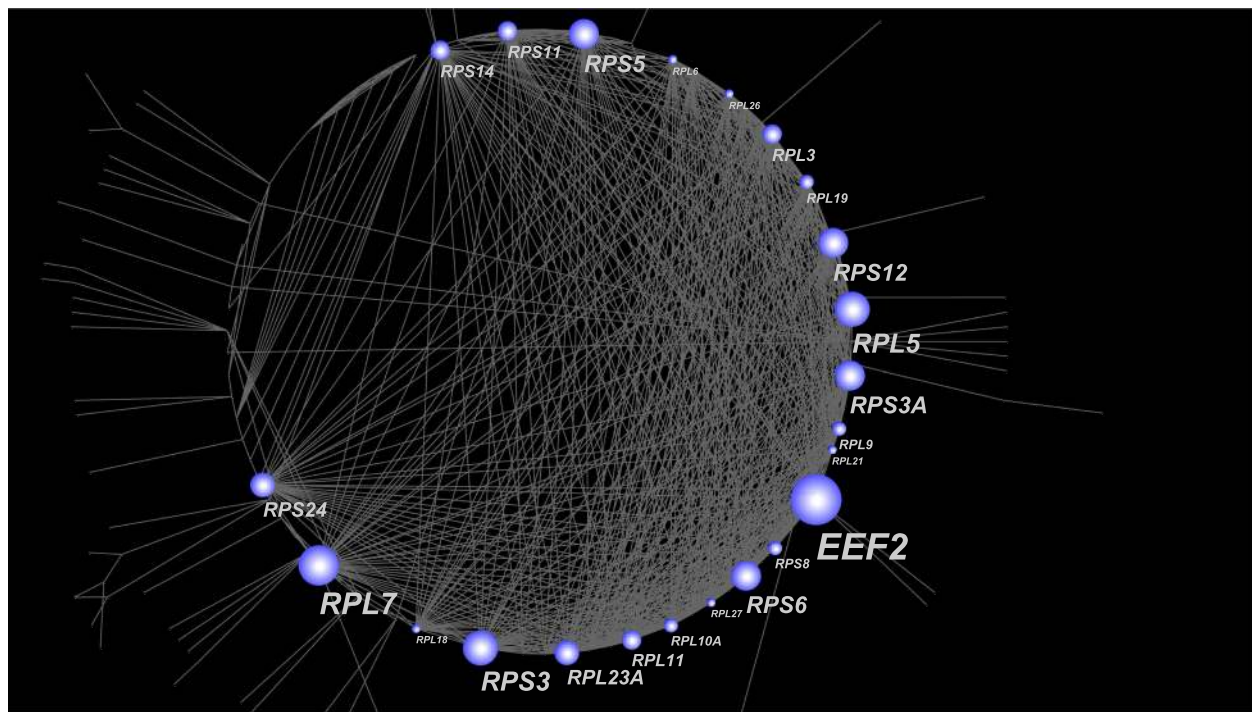


Fig. 4 Network-based meta-analysis using DEGs. A protein–protein interaction network detected in DEGs

data by using the area under the receiver operating characteristic curve (AUC). The best model achieved an AUC of 0.878 (95% CI 0.801–0.955, sensitivity = 0.945, specificity = 0.710, Supplementary Fig. S1a) in the test data when two top-ranked hub genes (*EEF2* and *RPL7*, Fig. 5a) were used. The highest variable importance was age (MeanDecreaseGini = 30.97; *RPL7*, 20.63; Neut, 14.23; *EEF2*, 14.20; *APOE* ϵ_4 genotypes, 5.05; sex, 3.15). The best model had a superior AUC to the model using only clinical information (Fig. 5a). Furthermore, the

Table 2 Top-ranked 10 hub genes detected in the network-based meta-analysis using DEGs

| Gene name | DC | BC | FC | FDR |
|---------------|----|--------|------|-----------------------|
| <i>EEF2</i> | 38 | 883.9 | 1.22 | 4.80×10^{-3} |
| <i>RPL7</i> | 36 | 350.3 | 1.51 | 1.75×10^{-5} |
| <i>RPL5</i> | 35 | 4567.0 | 1.23 | 2.07×10^{-5} |
| <i>RPS3</i> | 35 | 1599.7 | 1.38 | 1.89×10^{-4} |
| <i>RPS5</i> | 34 | 741.0 | 1.35 | 0.024 |
| <i>RPS12</i> | 34 | 427.3 | 1.44 | 5.20×10^{-4} |
| <i>RPS3A</i> | 34 | 23.1 | 1.31 | 1.10×10^{-3} |
| <i>RPS6</i> | 34 | 23.1 | 1.23 | 3.31×10^{-5} |
| <i>RPL23A</i> | 33 | 48.1 | 1.27 | 8.50×10^{-3} |
| <i>RPS24</i> | 33 | 22.0 | 1.23 | 8.00×10^{-4} |

The most highly ranked hub genes in terms of network topology measures of degree (DC) and betweenness of centrality (BC)
FC fold change, FDR false discovery rate

expression of two hub genes, *EEF2* and *RPL7*, were associated with a significant decrease and increase in AD prognosis, respectively ($P = 0.015$ in *EEF2*, $P = 0.032$ in *RPL7*, Jonckheere–Terpstra trend test, Fig. 5b and Supplementary Table S5). These results suggested that these two hub genes could serve as potential diagnostic blood biomarkers of AD. In a similar way, risk prediction models were constructed by using clinical features (age, sex, and *APOE* ϵ_4 genotypes), the proportion of neutrophils, and the top-ranked two hub genes with a random forest classifier using the training data. The adjusted models were then evaluated on the independent test data for a MCI and CN set and a MCI and AD set. The best models achieved an AUC of 0.683 (95% CI = 0.559–0.807, sensitivity = 0.744, specificity = 0.633, Supplementary Fig. S1b) and an AUC of 0.645 (95% CI 0.562–0.728, sensitivity = 0.622, specificity = 0.671, Supplementary Fig. S1c) for the MCI and CN set and the MCI and AD set in the test data, respectively.

Validation in a prospective cohort

We measured mRNA expression in 248 MCI samples. Of them, 55 MCI samples were obtained from the prospective data; 17 patients who contributed samples progressed to AD, whereas 38 of the patients corresponding to these samples have not yet been diagnosed with bona fide AD after at least 1 year. Our risk prediction model based on clinical information (age, sex, and *APOE* ϵ_4 genotypes) and three potential biomarkers we obtained

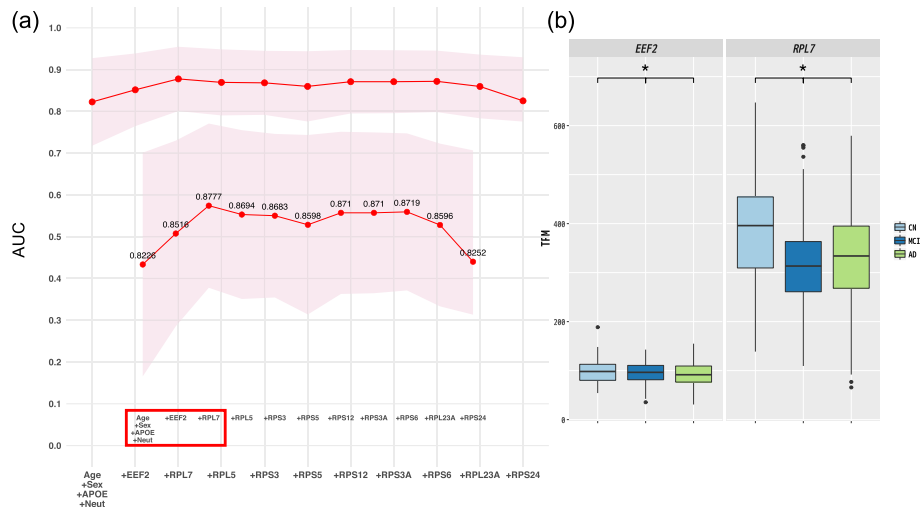


Fig. 5 Potential biomarkers of AD in the blood by using the most important hub genes. **a** Identification of the most important hub genes by using a random forest classifier. Neut, neutrophils. **b** Expression of two hub genes (*EEF2* and *RPL7*) among AD, MCI and CN samples

(i.e., the proportion of neutrophils, *EEF2*, and *RPL7*) was applied to the prospective data. Because our prediction model provides a probability of AD conversion for each MCI sample, we set the MCIs at the probability of > 0.9 for conversion to AD. Survival probabilities were calculated by using the Kaplan–Meier method in the *survival* package (version 2.41.3) for the statistical software R. Our risk prediction model significantly classified the MCI samples into two categories (high and low risk). The Kaplan–Meier curves showed improved outcome for AD conversion-free survival (Fig. 6, log rank trend test = 0.039), which achieved a high accuracy of 0.727 on the prospective cohort (sensitivity = 0.706, specificity = 0.737). Our present model predicted that 33 samples would not convert to AD, of which 5 did convert (negative predictive value (NPV) = 0.848). Although this clear classification of samples might be helpful for future practical use in healthcare, we would have to follow those samples to improve the further predictive value.

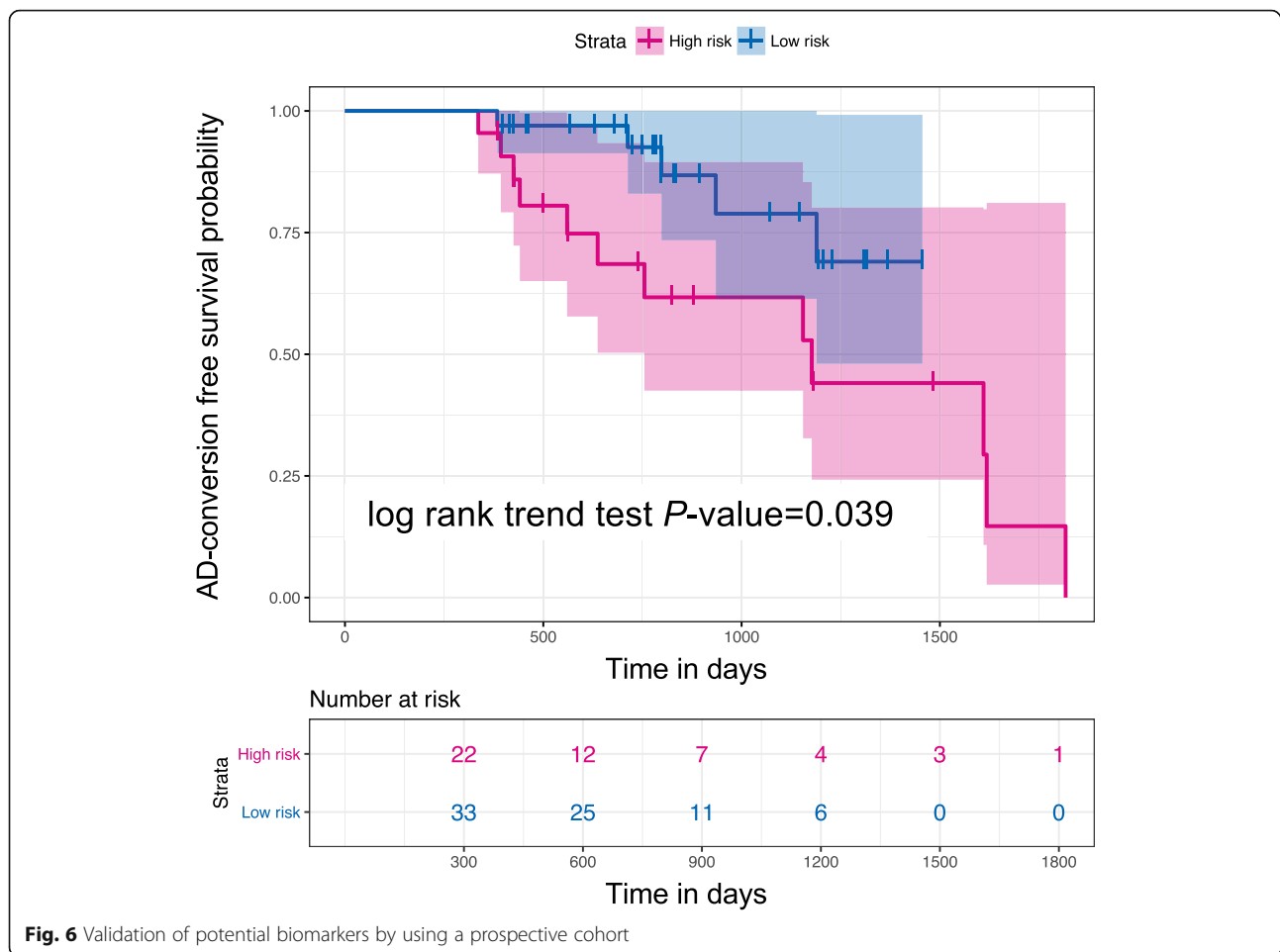
Verification of quantitative RT-PCR assay

To validate the RNA-seq results, we used quantitative RT-PCR (qRT-PCR) analysis to evaluate the two most significant hub genes (*EEF2* and *RPL7*) as potential biomarkers of AD for early diagnosis, three upregulated DEGs (*TRIOBP*, *TAS2R39*, and *BASPI*), and three down-regulated DEGs (*LDHB*, *NR1D2*, and *PDK4*). Figure 7 summarizes the RNA-seq and qRT-PCR results. Although the 8 DEGs were not expressed at precisely the same levels in both RNA-seq and qRT-PCR analyses, the regulated trends of the 8 DEGs were entirely consistent (Fig. 7). These results demonstrated our RNA-seq data accurately estimates gene expression.

Discussion

Peripheral blood biomarkers for early diagnosis have been examined in many diseases including AD [40–42]. In addition, various blood biomarkers associated with neurocognitive impairments have been reported, for example, glucose [43–45] and atherogenic index of plasma (AIP) [46]. However, no reliable and sensitive blood biomarkers are routinely used in clinical practice yet. One powerful and widely used approach to detect blood-based biomarkers, next-generation RNA-seq in human PBMCs, allows a comprehensive analysis of the entire transcriptome, but many of the previous studies were conducted in a small number of samples, particularly for AD.

Here, we performed comprehensive RNA-seq analysis using a large number of samples, to detect potential blood-based biomarkers associated with early diagnosis of AD. First, we used the bulk RNA-seq data to evaluate the difference in cell-type composition among AD, MCI, and CN samples. Of the 12 major immune cell types (B cells, plasma cells, CD8⁺ T cells, CD4⁺ T cells, $\gamma\delta$ T cells, NK cells, monocytes, macrophages, dendritic cells, mast cells, eosinophils, and neutrophils), we found a statistically significant difference in the proportion of neutrophils; that is, an increase in the proportion of neutrophils was significantly associated with AD prognosis. In addition, the association of this increase with prognosis was further confirmed using a large number of additional samples obtained from routine blood tests. Although a recent report suggested that the neutrophil phenotype could be associated with the rate of cognitive decline and therefore might be a prognostic blood biomarker in patients with AD [47], the study involved only a few samples ($n = 42$). In contrast, our current results were obtained from not only different data sets (RNA-seq and routine blood tests) but also a far



larger sample population ($n = 3,099$), providing stronger evidence that the proportion of neutrophils has the potential to be a blood biomarker of AD.

We also examined the DEGs between AD and CN samples. Of the 846 total statistically significant DEGs identified, 480 genes were upregulated and 366 were downregulated in AD. To gain further insight into the biological functions of the identified DEGs, we performed GSEA and PPI network analysis. Multiple statistically significant GO terms, one KEGG biological pathway, and several important hub genes were identified. A risk prediction model using the top two hub genes (*EEF2* and *RPL7*) and the proportion of neutrophils increased the model's AUC, compared with that of a model using clinical information only. Therefore, our model provides an effective and precise prediction of AD risk.

One of the potential biomarkers, *EEF2*, is a member of the GTP-binding translation elongation factor family and an essential factor for protein synthesis and cell survival. In recent studies, *EEF2* kinase reduction alleviated AD-associated defects in AD model mice [48]. In addition, *RPL7* is reported to be a tau-dependent T cell

intracellular antigen 1 (*TIA1*)-interacting protein [49, 50]. *TIA1* co-localizes with neuropathology in brain tissue of subjects with AD, frontotemporal lobar dementia, and amyotrophic lateral sclerosis, as well as in animal models of these diseases [51–53], all of which are associated with pathological tau misfolding and aggregation. These results suggest that these two hub genes could play a key role in the pathogenesis of AD.

We applied our risk prediction model—constructed by using these three potential biomarkers (proportion of neutrophils, *EEF2*, and *RPL7*) and three clinical features (age, sex, and *APOE* ϵ_4 genotypes)—to prospective cohort data. Although the highest variable importance was age among the six features, the three potential biomarkers interestingly showed a higher variable importance than the other clinical features (age and *APOE* ϵ_4 genotypes). In general, when a risk prediction model is constructed by using AD and CN samples, it is difficult to apply to MCI samples. However, because our prediction model provides a probability of AD conversion for each sample, we were able to make it applicable to MCI samples simply by adjusting the cutoff probability for conversion. Our risk prediction model significantly

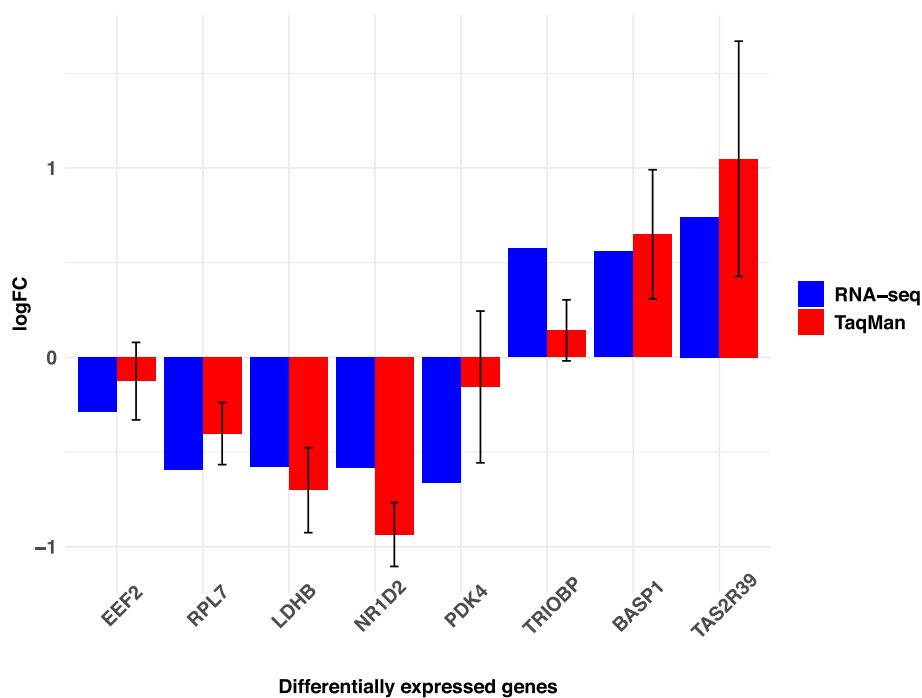


Fig. 7 qRT-PCR verification of DEGs detected by RNA-seq. Fold change values obtained from RNA-seq data: *TRIOBP*, *TAS2R39*, and *BASP1* were upregulated, and *EEF2*, *RPL7*, *LDHB*, *NR1D2*, and *PDK4* were downregulated. The expression determined by qRT-PCR was similar to that obtained by RNA-seq. Error bars in qRT-PCR indicate the standard error

classified MCI samples into two categories (high and low risks) and yielded a high NPV of 0.848. For clinical use, this prospective prediction model must have high NPV because it likely will be used at the first screening for AD conversion. This risk prediction model requires further refinement before its practical use in healthcare. One improvement would be to consider genetic variations, such as single-nucleotide variants, short insertions and deletions, and copy number variations, because GWAS have revealed many types of genetic variation that contribute to AD risk [6–8]. In addition, the combination of genetic variation and gene expression—expression quantitative trait loci (eQTLs) [54–56], which are genetic variants that affect gene expression levels—should be considered for the improvement of AD risk prediction models. Integration of that genetic variation, along with eQTL effects, likely will further improve the prospective AD risk prediction model.

Conclusions

The current study identified potential biomarkers for early diagnosis of AD from RNA sequencing data. The risk prediction model constructed by using the biomarkers achieved a high AUC for a validation cohort; when further applied to a prospective cohort, the model achieved high accuracy. Our model was demonstrated to be effective in prospective AD risk prediction. These

findings indicate the discovery of potential biomarkers for early diagnosis of AD, and their further improvement may lead to future practical clinical use.

Supplementary information

Supplementary information accompanies this paper at <https://doi.org/10.1186/s13195-020-00654-x>.

Additional file 1: Supplementary Table S1. Clinical information of samples used in RNA-seq data.

Additional file 2: Supplementary Table S2. Clinical information of samples used in routine blood tests.

Additional file 3: Supplementary Table S3. RNA-seq data.

Additional file 4: Supplementary Table S4. A list of all detailed DEGs.

Additional file 5: Supplementary Table S5. TPM of *RPL7* and *EEF2*.

Additional file 6: Supplementary Figure S1. Risk prediction models constructed using clinical information and two hub genes expression. The ROC curves of our risk prediction models in a test set. (a) AUC = 0.878 in AD and CN (b) AUC = 0.683 in MCI and CN (c) AUC = 0.645 in MCI and AD.

Authors' contributions

D.S. developed the method and performed the analyses. T.M. performed the experiments on mRNA expression. S.A. and S.H. provided technical assistance. H.W., T.S., and S.N. contributed to data acquisition and analyses. D.S. wrote the manuscript. D.S. and K.O. organized this work. All authors contributed to and approved the final manuscript.

Funding

This study was supported by The Japan Foundation for Aging and Health and Takeda Science Foundation (to D.S.), Research Funding for Longevity

Sciences (29-45) from the National Center for Geriatrics and Gerontology (to K.O.), and a grant for Research on Dementia from the Japanese Ministry of Health, Labor, and Welfare (to K.O.).

Availability of data and materials

All datasets used or analyzed in the current study are available from the corresponding author on reasonable request.

Ethics approval and consent to participate

This study protocol was approved by the ethics committee of the National Center for Geriatrics and Gerontology (NCGG) of Japan. The design and performance of the current study involving human subjects were clearly described in a research protocol. All participants were volunteers and completed informed consent in writing before registering to the NCGG Biobank.

Consent for publication

Not applicable

Competing interests

The authors declare that they have no competing interests.

Author details

¹Medical Genome Center, National Center for Geriatrics and Gerontology, 7-430 Morioka-cho, Obu 474-8511, Aichi, Japan. ²Department of Medical Science Mathematics, Medical Research Institute, Tokyo Medical and Dental University (TMDU), Tokyo 113-8510, Japan. ³RIKEN Center for Integrative Medical Sciences, Yokohama 230-0045, Kanagawa, Japan. ⁴The Center for Comprehensive Care and Research on Memory Disorders, National Center for Geriatrics and Gerontology, Obu 474-8511, Aichi, Japan. ⁵Department of Cognitive and Behavioral Science, Nagoya University Graduate School of Medicine, Nagoya 466-8550, Aichi, Japan.

Received: 16 February 2020 Accepted: 8 July 2020

Published online: 16 July 2020

References

- Robinson L, Tang E, Taylor JP. Dementia: timely diagnosis and early intervention. *BMJ*. 2015;350:h3029.
- Kim DH, Yeo SH, Park JM, Choi JY, Lee TH, Park SY, Ock MS, Eo J, Kim HS, Cha HJ. Genetic markers for diagnosis and pathogenesis of Alzheimer's disease. *Gene*. 2014;545:185–93.
- Zetterberg H. Applying fluid biomarkers to Alzheimer's disease. *Am J Physiol Cell Physiol*. 2017;313:C3–C10.
- Zverova M. Alzheimer's disease and blood-based biomarkers - potential contexts of use. *Neuropsychiatr Dis Treat*. 2018;14:1877–82.
- Ashraf GM, Chibber S, Mohammad, Zaidi SK, Tabrez S, Ahmad A, Shakil S, Mushtaq G, Baeesa SS, Kamal MA. Recent updates on the association between Alzheimer's disease and vascular dementia. *Med Chem*. 2016;12:226–37.
- Jansen IE, Savage JE, Watanabe K, Bryois J, Williams DM, Steinberg S, Sealock J, Karlsson IK, Hagg S, Athanasiu L, et al. Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. *Nat Genet*. 2019;51:404–13.
- Grau S, de Rojas I, Hernandez I, Quintela I, Montreal L, Alegret M, Hernandez-Olasagarre B, Madrid L, Gonzalez-Perez A, Maronas O, et al. Genome-wide association analysis of dementia and its clinical endophenotypes reveal novel loci associated with Alzheimer's disease and three causal networks: The GR@ACE project. *Alzheimers Dement*. 2019;15:1333–47.
- Kunkle BW, Grenier-Boley B, Sims R, Bis JC, Damotte V, Naj AC, Boland A, Vronskaya M, van der Lee SJ, Amlie-Wolf A, et al. Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing. *Nat Genet*. 2019;51:414–30.
- Albert MS, DeKosky ST, Dickson D, Dubois B, Feldman HH, Fox NC, Gamst A, Holtzman DM, Jagust WJ, Petersen RC, et al. The diagnosis of mild cognitive impairment due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement*. 2011;7:270–9.
- McKhann GM, Knopman DS, Chertkow H, Hyman BT, Jack CR Jr, Kawas CH, Klunk WE, Koroshetz WJ, Manly JJ, Mayeux R, et al. The diagnosis of dementia due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement*. 2011;7:263–9.
- De Meyer G, Shapiro F, Vanderstichele H, Vanmechelen E, Engelborghs S, De Deyn PP, Coart E, Hansson O, Minthon L, Zetterberg H, et al. Diagnosis-independent Alzheimer disease biomarker signature in cognitively normal elderly people. *Arch Neurol*. 2010;67:949–56.
- Fagan AM, Shaw LM, Xiong C, Vanderstichele H, Mintun MA, Trojanowski JQ, Coart E, Morris JC, Holtzman DM. Comparison of analytical platforms for cerebrospinal fluid measures of β-amyloid 1-42, total tau, and p-tau181 for identifying Alzheimer disease amyloid plaque pathology. *Arch Neurol*. 2011; 68:1137–44.
- Miller G. Alzheimer's biomarker initiative hits its stride. *Science*. 2009;326:386–9.
- Mistur R, Mosconi L, Santi SD, Guzman M, Li Y, Tsui W, de Leon MJ. Current challenges for the early detection of Alzheimer's disease: brain imaging and CSF studies. *J Clin Neurol*. 2009;5:153–66.
- Schmand B, Eikelenboom P, van Gool WA, Alzheimer's Disease Neuroimaging I. Value of neuropsychological tests, neuroimaging, and biomarkers for diagnosing Alzheimer's disease in younger and older age cohorts. *J Am Geriatr Soc*. 2011;59:1705–10.
- Mattsson N, Cullen NC, Andreasson U, Zetterberg H, Blennow K. Association between longitudinal plasma neurofilament light and neurodegeneration in patients with Alzheimer disease. *JAMA Neurol*. 2019;76:791–9.
- Janelidze S, Mattsson N, Palmqvist S, Smith R, Beach TG, Serrano GE, Chai X, Proctor NK, Eichenlaub U, Zetterberg H, et al. Plasma P-tau181 in Alzheimer's disease: relationship to other biomarkers, differential diagnosis, neuropathology and longitudinal progression to Alzheimer's dementia. *Nat Med*. 2020;26:379–86.
- Liu T, Yu N, Ding F, Wang S, Li S, Zhang X, Sun X, Chen Y, Liu P. Verifying the markers of ovarian cancer using RNA-seq data. *Mol Med Rep*. 2015;12:1125–30.
- Wang H, Li Y, Ryder JW, Hole JT, Ebert PJ, Airey DC, Qian HR, Logsdon B, Fisher A, Ahmed Z, et al. Genome-wide RNAseq study of the molecular mechanisms underlying microglia activation in response to pathological tau perturbation in the rTg4510 tau transgenic animal model. *Mol Neurodegener*. 2018;13:65.
- Bennett JP Jr, Keeney PM, Brohawn DG. RNA sequencing reveals small and variable contributions of infectious agents to transcriptomes of postmortem nervous tissues from amyotrophic lateral sclerosis, Alzheimer's disease and Parkinson's disease subjects, and increased expression of genes from disease-activated microglia. *Front Neurosci*. 2019;13:235.
- Anders S, Huber W. Differential expression analysis for sequence count data. *Genome Biol*. 2010;11:R106.
- Li J, Tibshirani R. Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-Seq data. *Stat Methods Med Res*. 2013;22:519–36.
- Zhu Q, Sun Y, Zhou Q, He Q, Qian H. Identification of key genes and pathways by bioinformatics analysis with TCGA RNA sequencing data in hepatocellular carcinoma. *Mol Clin Oncol*. 2018;9:597–606.
- Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al. Gene ontology: tool for the unification of biology. *The Gene Ontology Consortium*. *Nat Genet*. 2000;25:25–9.
- Huang da W, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res*. 2009; 37:1–13.
- Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res*. 2017;45:D353–61.
- Kanehisa M, Sato Y, Furumichi M, Morishima K, Tanabe M. New approach for understanding genome variations in KEGG. *Nucleic Acids Res*. 2019;47:D590–5.
- Shigemizu D, Akiyama S, Asanomi Y, Borojevich KA, Sharma A, Tsunoda T, Matsukuma K, Ichikawa M, Sudo H, Takizawa S, et al. Risk prediction models for dementia constructed by supervised principal component analysis using miRNA expression data. *Commun Biol*. 2019;2:77.
- Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013;29:15–21.
- Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*. 2014;30:923–30.
- Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. 2010;26:139–40.

32. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*. 2011;12:323.
33. Chen B, Khodadoust MS, Liu CL, Newman AM, Alizadeh AA. Profiling tumor infiltrating immune cells with CIBERSORT. *Methods Mol Biol*. 2018;1711:243–59.
34. Huang da W, Sherman BT, Lempicki RA: Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* 2009, 4:44–57.
35. Walter W, Sanchez-Cabo F, Ricote M. GOrilla: an R package for visually combining expression data with functional analysis. *Bioinformatics*. 2015;31:2912–4.
36. Santiago JA, Potashkin JA. Network-based metaanalysis identifies HNF4A and PTBP1 as longitudinally dynamic biomarkers for Parkinson's disease. *Proc Natl Acad Sci U S A*. 2015;112:2257–62.
37. Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A, Minguéz P, Doerks T, Stark M, Muller J, Bork P, et al. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res*. 2011;39:D561–8.
38. Su G, Morris JH, Demchak B, Bader GD: Biological network exploration with Cytoscape 3. *Curr Protoc Bioinformatics* 2014, 47:8 13 11–24.
39. Marone M, Mozzetti S, De Ritis D, Pierelli L, Scambia G. Semiquantitative RT-PCR analysis to assess the expression levels of multiple transcripts from the same sample. *Biol Proced Online*. 2001;3:19–25.
40. Long J, Pan G, Ifeachor E, Belshaw R, Li X. Discovery of novel biomarkers for Alzheimer's disease from blood. *Dis Markers*. 2016;2016:4250480.
41. Rai N, Kumar R, Desai GR, Venugopalan G, Shekhar S, Chatterjee P, Tripathi M, Upadhyay AD, Dwivedi S, Dey AB, Dey S. Relative alterations in blood-based levels of sestrin in Alzheimer's disease and mild cognitive impairment patients. *J Alzheimers Dis*. 2016;54:1147–55.
42. San Segundo-Acosta P, Montero-Calle A, Fuentes M, Rabano A, Villalba M, Barderas R. Identification of Alzheimer's disease autoantibodies and their target biomarkers by phage microarrays. *J Proteome Res*. 2019;18:2940–53.
43. Grober E, Hall CB, Hahn SR, Lipton RB. Memory impairment and executive dysfunction are associated with inadequately controlled diabetes in older adults. *J Prim Care Community Health*. 2011;2:229–33.
44. Crane PK, Walker R, Hubbard RA, Li G, Nathan DM, Zheng H, Haneuse S, Craft S, Montine TJ, Kahn SE, et al. Glucose levels and risk of dementia. *N Engl J Med*. 2013;369:540–8.
45. Pappas C, Andel R, Infurna FJ, Seetharaman S. Glycated haemoglobin (HbA1c), diabetes and trajectories of change in episodic memory performance. *J Epidemiol Community Health*. 2017;71:115–20.
46. Aniwattanapong D, Tangwongchai S, Supasitthumrong T, Hemrunroj S, Tunvirachaisakul C, Tawankanjanachot I, Chuchuen P, Snaboon T, Carvalho AF, Maes M. Validation of the Thai version of the short Boston Naming Test (T-BNT) in patients with Alzheimer's dementia and mild cognitive impairment: clinical and biomarker correlates. *Aging Ment Health*. 2019;23:840–50.
47. Dong Y, Lagarde J, Xicota L, Corne H, Chantran Y, Chaigneau T, Crestani B, Bottlaender M, Potier MC, Aucouturier P, et al. Neutrophil hyperactivation correlates with Alzheimer's disease progression. *Ann Neurol*. 2018;83:387–405.
48. Beckelman BC, Yang W, Kasica NP, Zimmermann HR, Zhou X, Keene CD, Ryazanov AG, Ma T. Genetic reduction of eEF2 kinase alleviates pathophysiology in Alzheimer's disease model mice. *J Clin Invest*. 2019;129:820–33.
49. Minjarez B, Valero Rustarazo ML, Sanchez del Pino MM, Gonzalez-Robles A, Sosa-Melgarejo JA, Luna-Munoz J, Mena R, Luna-Arias JP. Identification of polypeptides in neurofibrillary tangles and total homogenates of brains with Alzheimer's disease by tandem mass spectrometry. *J Alzheimers Dis*. 2013;34:239–62.
50. Vanderweyde T, Apicco DJ, Youmans-Kidder K, Ash PEA, Cook C, Lummertz da Rocha E, Jansen-West K, Frame AA, Citro A, Leszyk JD, et al. Interaction of tau with the RNA-binding protein TIA1 regulates tau pathophysiology and toxicity. *Cell Rep*. 2016;15:1455–66.
51. Liu-Yesucevitz L, Bilgutay A, Zhang YJ, Vanderweyde T, Citro A, Mehta T, Zaarur N, McKee A, Bowser R, Sherman M, et al. Tar DNA binding protein-43 (TDP-43) associates with stress granules: analysis of cultured cells and pathological brain tissue. *PLoS One*. 2010;5:e13250.
52. Thomas MG, Loschi M, Desbats MA, Boccaccio GL. RNA granules: the good, the bad and the ugly. *Cell Signal*. 2011;23:324–34.
53. Vanderweyde T, Yu H, Varnum M, Liu-Yesucevitz L, Citro A, Ikezu T, Duff K, Wolozin B. Contrasting pathology of the stress granule proteins TIA-1 and G3BP in tauopathies. *J Neurosci*. 2012;32:8270–83.
54. Schadt EE, Monks SA, Drake TA, Lusk AJ, Che N, Colinayo V, Ruff TG, Milligan SB, Lamb JR, Cavet G, et al. Genetics of gene expression surveyed in maize, mouse and man. *Nature*. 2003;422:297–302.
55. Stranger BE, Forrest MS, Clark AG, Minichiello MJ, Deutsch S, Lyle R, Hunt S, Kahl B, Antonarakis SE, Tavare S, et al. Genome-wide associations of gene expression variation in humans. *PLoS Genet*. 2005;1:e78.
56. Emilsson V, Thorleifsson G, Zhang B, Leonardson AS, Zink F, Zhu J, Carlson S, Helgason A, Walters GB, Gunnarsdottir S, et al. Genetics of gene expression and its effect on disease. *Nature*. 2008;452:423–8.

Publisher's Note

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

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more [biomedcentral.com/submissions](https://www.biomedcentral.com/submissions)

