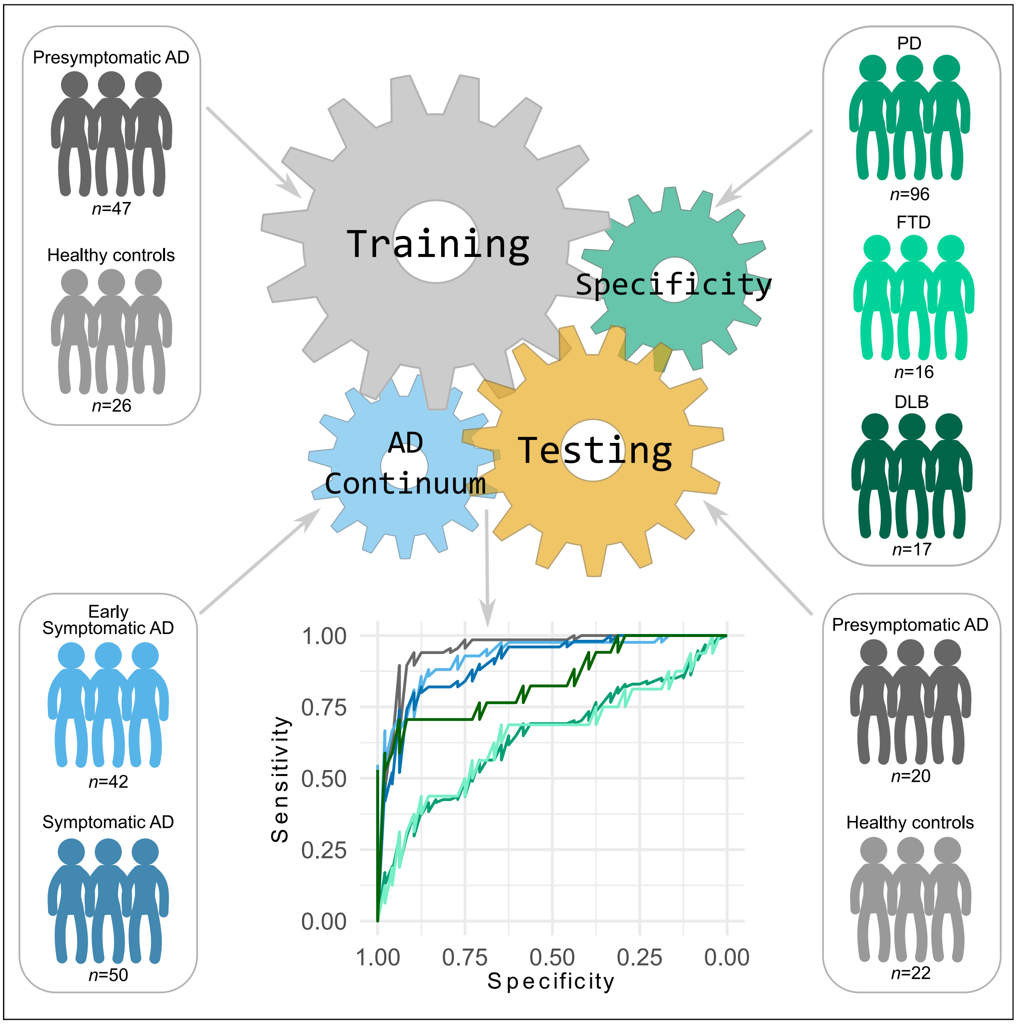
iScience

CellPress
OPEN ACCESS

Article

Cell-free RNA signatures predict Alzheimer's disease



Alejandro Cisterna-García, Aleksandra Beric, Muhammad Ali, ..., Carlos Cruchaga, Juan A. Botia, Laura Ibanez

ibanezl@wustl.edu

Highlights

Plasma cfRNA reflects transcriptional changes related to AD pathology

Plasma cfRNA predictive models accurately predict presymptomatic AD

Plasma cfRNA is informative of brain amyloid positivity

Plasma cfRNA models have limited power to predict non-AD neurodegenerative diseases

Cisterna-García et al., iScience 26, 108534
December 15, 2023 © 2023 The Author(s).
https://doi.org/10.1016/j.isci.2023.108534

iScience

CellPress
OPEN ACCESS

Article

Cell-free RNA signatures predict Alzheimer's disease

Alejandro Cisterna-García,[1,2,3] Aleksandra Beric,[1,3] Muhammad Ali,[1,3] Jose Adrian Pardo,[2] Hsiang-Han Chen,[1,3] Maria Victoria Fernandez,[1,3] Joanne Norton,[1,3,4] Jen Gentsch,[1,3,4] Kristy Bergmann,[1,3] John Budde,[1,3] Joel S. Perlmutter,[5,6,7] John C. Morris,[4,5,8] Carlos Cruchaga,[1,3,4,5,6,9] Juan A. Botia,[2] and Laura Ibanez[1,3,5,10,*]

SUMMARY

There is a need for affordable, scalable, and specific blood-based biomarkers for Alzheimer's disease that can be applied to a population level. We have developed and validated disease-specific cell-free transcriptomic blood-based biomarkers composed by a scalable number of transcripts that capture AD pathobiology even in the presymptomatic stages of the disease. Accuracies are in the range of the current CSF and plasma biomarkers, and specificities are high against other neurodegenerative diseases.

INTRODUCTION

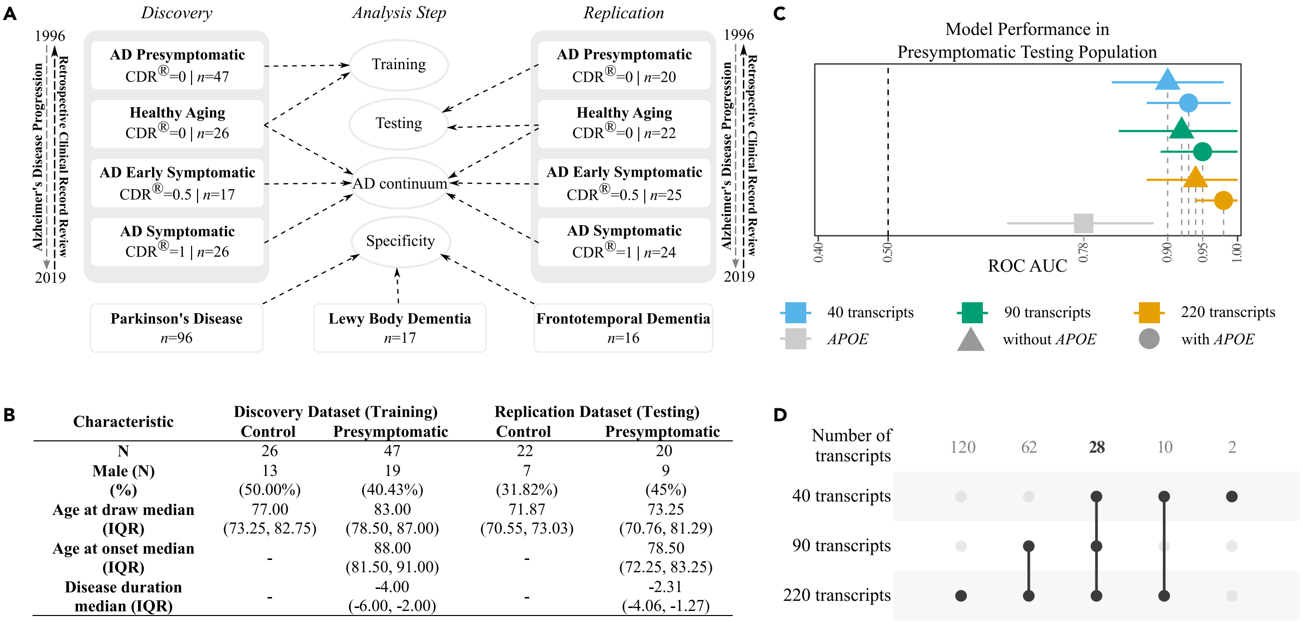
Alzheimer's disease (AD) is a complex neurodegenerative disorder clinically characterized by gradual and progressive memory loss and, pathologically by the presence of senile plaques (amyloid-beta deposits) and neurofibrillary tangles (tau deposits) in the brain.[1] Economically, it has been estimated that AD and other dementias cost approximately $355 billion in 2021, a cost that has been estimated to increase to $1.1 trillion in 2050.[2] The availability of an early and accurate diagnostic tool for AD might save $7.9 trillion in medical and care costs.[3] Currently, many efforts are being directed to find cost-effective and non-invasive biomarkers for AD that can be used to identify individuals at the presymptomatic stage, and patients at early symptomatic stages of the disease (preclinical AD individuals or mild cognitive impairment-MCI.[4,5]

Imaging and cerebrospinal fluid (CSF) biomarkers are commonly used for AD diagnosis.[6–8] The most used and accurate CSF biomarker is the amyloid β42/amyloid β40 (Aβ42/Aβ40) ratio which can correctly diagnose 82.8% of the screened AD patients.[9] Additionally, the Aβ42/Aβ40 measurements in CSF are specific and allow differentiation of AD from dementia with Lewy bodies (DLB), Parkinson's disease (PD), and vascular dementia (VaD).[10] However, the standardization of the measurements to use in the clinical practice has been challenging, mainly due to inter-laboratory differences in sample handling and analytical methods.[11,12] Along with Aβ measurements, CSF levels of phosphorylated tau (p-tau), and total tau (t-tau) in CSF or brain are also used to aid AD diagnosis.[13,14] T-tau is elevated in other neurodegenerative diseases such as DLB, frontotemporal degeneration (FTD), VaD, and Creutzfeldt-Jakob disease (CJD).[15] In contrast, certain CSF p-tau species such as p-tau181 and p-tau231 are more specific to AD and show strong correlations with the tau PET.[16,17] To improve the AD diagnosis, the Amyloid (A) Tau (T) Neurodegeneration (N) framework proposed a biological classification of AD into eight profiles according to positivity/negativity of three biomarkers, Aβ (A), p-tau (T), and t-tau (N).[18] It is accepted that turning positive for Aβ means the beginning of the AD continuum.[19,20] The increase in the number of positive biomarkers for the ATN criteria correlates with more advanced pathology, and it is associated with increased risk of dementia and cognitive decline.[19,20] One of the main challenges for the ATN criteria is the definition of cut-off values for the biomarkers, especially for the triage of presymptomatic AD individual.[21–23]

CSF and imaging biomarkers have proven helpful in the detection of AD; however, they are invasive and expensive. In consequence, the study of blood-based biomarkers has intensified in the last decade. These biomarkers are less invasive and may provide comparable accuracy to CSF and imaging measures.[24,25] For example, plasma t-tau was found to be higher in the advanced stages of AD,[25,26] but as with CSF measurements, it does not seem to be specific to AD.[27] New evidence suggests that phosphorylated species of tau, especially p-tau 217 are specific to AD, with values increasing progressively from healthy individuals to MCI to AD.[28–30] Further, Aβ has also been widely studied in plasma showing that the ratio of plasma Aβ42/Aβ40 highly correlates with brain amyloidosis,[31] especially when measured with high-precision

[1]Department of Psychiatry, Washington University in Saint Louis School of Medicine, Saint Louis, MO, USA
[2]Departamento de Ingeniería de la Información y las Comunicaciones, Universidad de Murcia, Murcia, Spain
[3]NeuroGenomics and Informatics Center, Washington University in Saint Louis School of Medicine, Saint Louis, MO, USA
[4]The Charles F. and Joanne Knight Alzheimer Disease Research Center, Washington University in Saint Louis, Saint Louis, MO, USA
[5]Department of Neurology, Washington University in Saint Louis School of Medicine, Saint Louis, MO, USA
[6]Hope Center for Neurological Disorders, Washington University in Saint Louis School of Medicine, Saint Louis, MO, USA
[7]Department of Radiology, Neuroscience, Physical Therapy, and Occupational Therapy, Washington University in Saint Louis School of Medicine, Saint Louis, MO, USA
[8]Department of Pathology and Immunology, Washington University in Saint Louis School of Medicine, Saint Louis, MO, USA
[9]Department of Genetics, Washington University in Saint Louis School of Medicine, Saint Louis, MO, USA
[10]Lead contact
*Correspondence: ibanezl@wustl.edu

https://doi.org/10.1016/j.isci.2023.108534



Characteristic	Discovery Dataset (Training) Control	Discovery Dataset (Training) Presymptomatic	Replication Dataset (Testing) Control	Replication Dataset (Testing) Presymptomatic
N	26	47	22	20
Male (N)	13	19	7	9
(%)	(50.00%)	(40.43%)	(31.82%)	(45%)
Age at draw median	77.00	83.00	71.87	73.25
(IQR)	(73.25, 82.75)	(78.50, 87.00)	(70.55, 73.03)	(70.76, 81.29)
Age at onset median	-	88.00	-	78.50
(IQR)		(81.50, 91.00)		(72.25, 83.25)
Disease duration	-	-4.00	-	-2.31
median (IQR)		(-6.00, -2.00)		(-4.06, -1.27)

Figure 1. Presymptomatic Alzheimer's disease prediction results

(A) Study design summary showing the sample selection approach (retrospective clinical record review), the groups and subgroups included in the discovery and replication, along with other neurodegenerative diseases.

(B) Summary demographics for the discovery (training) and replication (testing) datasets.

(C) Whisker plot showing the performance of the prediction of presymptomatic AD in the replication/testing dataset for the three predictive models (40, 90, and 220 transcripts) with and without APOE genotype.

(D) APOE genotype predictive power is depicted at the bottom for reference (D) An intersection matrix showing the shared and private transcripts among the three different models for presymptomatic AD.

techniques and combined with *APOE* genotype, the main genetic risk factor for AD.[32] However, cost-effectiveness and scalability of these measurements are not optimal.

Most blood-based biomarker studies measure protein levels; however, nucleic acids can also be used as biomarkers. The cell-free DNA (cfDNA) diagnostic test, which allows the detection of genetic disorders and chromosome abnormalities during pregnancy, revolutionized prenatal screening by avoiding procedure-related miscarriage risks.[33] Plasma also contains ribonucleic acid in its free form (cell-free RNA-cfRNA) that has the potential to capture temporal processes since its source seems to be the result of normal cell death throughout the body.[34] Several species of cfRNA have been intensively investigated as biomarkers for cancer,[35,36] fetal development,[37] and AD.[38–40] While several studies proposed circulating microRNAs as AD biomarkers[39,41–44] only one published study used plasma messenger cfRNA to capture transcriptomic alterations in advanced AD stages.[40] In their comparisons between AD cases (n = 122) and controls (n = 116), they identified 2591 differential expressed (DE) transcripts. Then, they used the transcriptomic information to build classifiers to discriminate AD patients from healthy controls with an area under the curve (AUC) of 0.83.[40] Even though promising, the models include most of the differentially expressed genes (1658), instead of a subset of the most informative genes, which would improve the scalability and facilitate the translation to a clinical setting. No study thus far has evaluated the use of cfRNA as a potential approach to develop clinically useful AD biomarkers for presymptomatic phases of the disease.

Here, different from previous publications,[40,45] we leveraged plasma cfRNA of presymptomatic AD participants to capture the early changes caused by AD pathology and to build unbiased models that were able to balance a good performance with a scalable number of transcripts to facilitate potential clinical applications. We also evaluated the discriminative capabilities of the proposed AD presymptomatic models in the context of AD spectrum, PD, DLB, and Frontotemporal Dementia (FTD) to ensure that the models were selectively capturing changes associated with AD pathobiology (Figure 1A).

RESULTS

Concordance between dysregulated transcripts in plasma cfRNA and brain of AD participants

We analyzed plasma cfRNA from presymptomatic AD participants to capture the early changes caused by AD pathology and build classifiers containing the expression of a scalable number of genes. All presymptomatic AD participants were required to have a sample before onset of symptoms (time of draw), and evidence of Aβ deposition (CSF Aβ<500 ng/L or positive PET scan) and/or evidence of clinical worsening measured by Clinical Dementia Rating (CDR) at the last clinical visit compared to the time of draw (18.0–6.5 years prior) (Figure 1A). We generate two independent datasets, by conducting retrospective sample selection twice (one for discovery and another for replication, separated by four years) from the Knight-ADRC, a deeply phenotyped cohort with longitudinal data and samples available. Due to time difference, RNA extraction and library preparation protocols were different for the discovery compared to the replication datasets (see STAR methods section). In summary, we extracted and sequenced RNA of non-fasted plasma samples from a total of 67 presymptomatic AD participants ($n_{\text{discovery}}$ = 47; $n_{\text{replication}}$ = 20) and 47 controls ($n_{\text{discovery}}$ = 26; $n_{\text{replication}}$ = 22) (Figures 1A and 1B).

After stringent quality control (QC), we performed DE analyses comparing presymptomatic AD participants and controls using DESeq2.[46] We identified 190 DE transcripts while controlling for sex and age at draw (Figure S1; Table S1). We used previously identified DE transcripts found in plasma of advanced symptomatic AD[40] to replicate 37 of our findings, which showed statistical significance for the overlap (p = 0.01 – Table S1). Most importantly, we wanted to know if the cfRNA was potentially capturing changes taking place in the brain. Using an in-house dataset[47] we found that 23 out of 190 transcripts were DE in both the brain and plasma of AD participants (Figure S1; Table S1). The overlap was statistically significant (p = 0.03) with a fold enrichment of 1.6. On top of that, the effect sizes of the 23 genes in the brain and the plasma were highly correlated (cor = 0.83; $p = 7.55 \times 10^{-7}$). Overall, seven out of 190 the plasma DE transcripts were common between the 37 transcripts replicated in plasma[40] and the 23 replicated in brain[47] (e.g., *MBOAT2*, *SLC9A9*, *RHOBTB3*, *RUNX1M*, *POC1B*, *SRBD1*, and *HIPK3*). Furthermore, 73 out of 190 DE transcripts identified here were also found to be differentially expressed in brains from three distinct cohorts[48] (Figure S1; Table S1). To further investigate if the transcripts identified in this study were expressed in the brain, we accessed the GTEx portal and found 176 out of the 190 genes to be expressed in brain cortex tissue (Table S1), adding evidence to the brain as a potential source of the DE cfRNA transcripts.

To assess the potential biological relevance of the 190 DE transcripts, we explored the Kyoto Encyclopedia of Genes and Genomes (KEGG) and found that the identified transcripts were enriched and significantly overlapped with the AD pathway (nine genes, $p = 8.92 \times 10^{-3}$ – Table S1). We also used the ToppFun tool from ToppGene Suite and found that the 190 transcripts are in concordance with transcripts up-regulated in the brains of patients with AD ($p = 1.40 \times 10^{-4}$).[49] We also identified an enrichment in Gene Ontology (GO) terms *cellular component neuronal synapse* ($p = 6.69 \times 10^{-3}$) and *postsynapse* ($p = 1.58 \times 10^{-2}$). Finally, we performed a co-expression analysis using networks from the frontal cortex of AD cases from ROSMAP in the CoExpWeb.[50] We found a statistically significant overlap between the 190 transcripts and two co-expression modules (thistle1 and darkgray). The thistle1 module ($p = 2.00 \times 10^{-4}$), was associated with oligodendrocytes in the cortex whereas the darkgrey module (p = 0.03) was associated with vasculature development and endothelial-external cells. Taken together, plasma cfRNA is capturing metabolic processes happening in the brain and might be reflecting transcriptional changes related to AD pathology of presymptomatic AD participants.

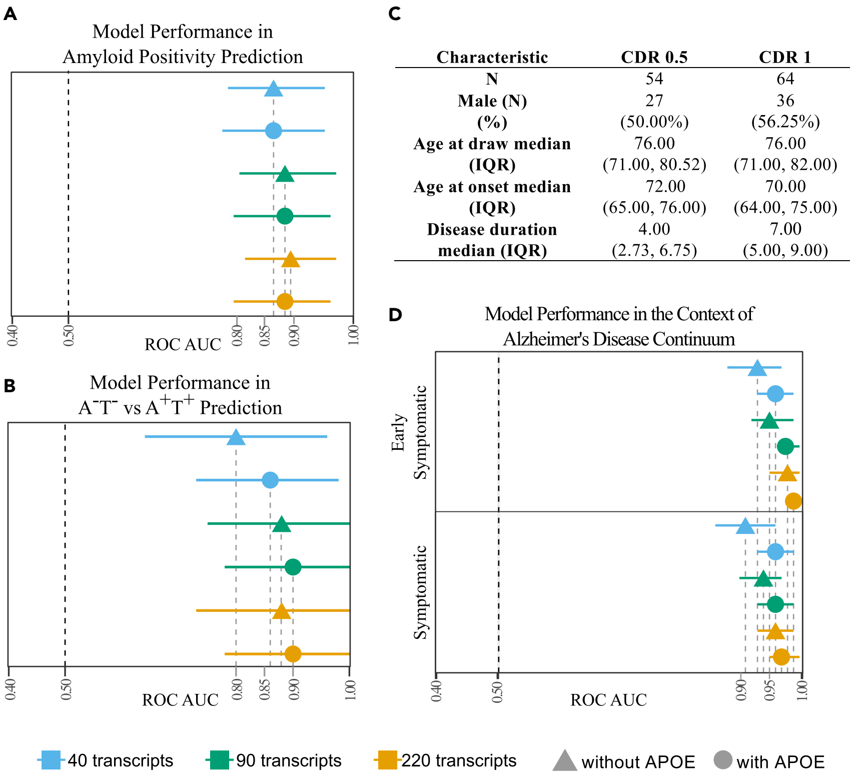
cfRNA recapitulates a transcript signature corresponding to the presymptomatic stages of Alzheimer's disease

To leverage all the RNA data available, we have developed a new approach that allows the use of two independent RNA-seq experiments as training (discovery) and testing (replication) for machine learning model development pipeline (Figure S2). Briefly, we reduced the dimensionality of the two datasets by retaining transcripts showing the same direction of effect in the case-control comparison. Then, we calculated the distribution overlap of each transcript within the two datasets using the Kullback-Leibler divergence (KLD) and used their absolute values to rank the transcripts and generate eight subsets with diverse number of genes. To comprehensively investigate the best mathematical approach to model presymptomatic AD participants, we evaluated L1 regularization, L2 regularization, and the non-linear (random forest) approach. Owing to its superior performance in terms of the mean ROC-AUC and testing balanced accuracy out of 10 repeated 5-fold cross-validation experiments (Figure S3), we decided to proceed with L2 regularized linear models. Thus, within each subset we used KLD thresholds (from 0.06 to 0.36 by increments of 0.02) and L2 regularization linear models (ridge regression) to predict presymptomatic AD in the training dataset (discovery). Then we assessed the performance on the testing dataset.

We generated a total of 272 models with different number of transcripts and then selected the best three based on the cross-validation experiment (Figure S4). The best models contained 40, 90 and 220 transcripts with an area under the ROC curve (ROC) in the testing dataset of 0.90, 0.92, and 0.94 respectively (Figure 1C; Table S2). We observed that the transcript overlap across models was significant ($p < 2.16 \times 10^{-16}$ - Figure 1D), suggesting that the new standardization, described in detail in the STAR methods section, and feature selection strategy implemented here tends to select good predictors in a consistent manner. In fact, the 28 common transcripts across the three models had an AUC of 0.92. After extracting the beta values (i.e., the importance of the predictors) from each gene in each of the predictive models (Figure S5), we observed that the transcript corresponding to the gene *SYNPO* (the top hit from the DE analyses) was the most relevant feature for the models with 90 and 220 genes. In previously published plasma biomarkers, the inclusion of *APOE* genotype in the model improved the performance. In our case, the addition of *APOE* genotype did not change the predictive power of the models, implying that we are already capturing the risk associated with *APOE* genotype (Figure 1C; Table S2).

Predictive models are capturing pathways related to AD early in the disease pathobiology

To understand the link between the transcripts included in the predictive models and their potential involvement in the pathobiology of AD we performed gene enrichment analyses for each of the models separately. Given the limited number of transcripts included in each of the three models, in order to add robustness to the enrichment analyses, we expanded each transcript set to include transcripts that show significant correlation ($p < 0.05$ and $r > 0.95$) with the transcripts of each predictive model (see STAR methods section). Thus, the sets increased to 844, 1054, and 2436 for the predictive models including 40, 90, and 220 transcripts respectively, with several transcripts present in all of the sets. We identified 1201, 1111, and 494 overrepresented GO terms (Table S3). Relevant terms known to be associated with AD such as



Characteristic	CDR 0.5	CDR 1
N	54	64
Male (N)	27	36
(%)	(50.00%)	(56.25%)
Age at draw median	76.00	76.00
(IQR)	(71.00, 80.52)	(71.00, 82.00)
Age at onset median	72.00	70.00
(IQR)	(65.00, 76.00)	(64.00, 75.00)
Disease duration	4.00	7.00
median (IQR)	(2.73, 6.75)	(5.00, 9.00)

Figure 2. Sensitivity analyses for the predictive models in the AD continuum and in the context of the ATN framework

(A) Whisker plot showing the performance of the prediction of A+T+ vs. A-T- for the three predictive models (40, 90, and 220 transcripts) with and without APOE genotype.

(B) Whisker plot showing the performance of the prediction of CSF amyloid beta positivity for the three predictive models (40, 90, and 220 transcripts) with and without APOE genotype.

(C) Summary demographics for the early symptomatic and the symptomatic individuals.

(D) Whisker plot showing the performance of the prediction of early symptomatic AD and symptomatic AD for the three predictive models (40, 90, and 220 transcripts) with and without APOE genotype.

immune-related pathways and processes (GO term IDs: 0002218, 0002753, 0002757, and 0002764), or lysosome (GO term IDs: 0005765 and 0005766) were significant in all three analyses. We identified significant enrichment in terms related to the regulation of neuronal apoptosis and death (GO term IDs: 0043523, 0051402, 0070997, 1901214, 1901215, and 1901216) in all three analyses, supporting the capture of early neuropathological processes taking place in the brain by the predictive models. Similarly, KEGG enrichment analyses identified 78, 68, and 40 significantly overrepresented terms (Table S4) for each of the sets generated for each predictive model. Among others, neurodegenerative diseases including AD and PD were significantly enriched suggesting that we are in fact capturing processes related to the known biology of neurodegenerative disease early in the course of the disease.

Predictive models trained with presymptomatic AD participants can accurately predict amyloid positivity

Current biomarkers evaluate the levels of Aβ42 in CSF or plasma to predict brain amyloidosis. We investigated if the estimated risk of AD calculated using the three models generated here (i.e., a number in the [0,1] interval) correlated with CSF Aβ42 levels. For those controls (n = 43) and presymptomatic AD participants (n = 28) with CSF measurements available at the time of blood draw, we tested if the AD risk calculated using the three models correlated with CSF Aβ42, tau, and p-tau (Figure S6). We found significant associations with CSF Aβ42 levels, especially for the model containing 220 transcripts ($r^2 = -0.54$; p = 1.27 × 10-6), but not with other CSF biomarkers or AD risk factors (Figure S6). Associations with CSF Aβ42 had a negative direction, as expected. Finally, we classified these samples following the ATN criteria. Out of 72 samples, 49 were A^{-}, and 23 A^{+}, whereas 23 were T^{-} and 49 T^{+}. Using the three transcriptomic models, we predicted A positivity status with AUCs of 0.89, 0.88 and 0.86 for the models with 220, 90 and 40 transcripts respectively (Figure 2A; Table S5). When including *APOE* to the models, we observed no changes to the AUCs, adding evidence to the fact that the transcriptomic model is capturing changes related to AD and to its main pathology. We also tested the predictive performance for A^{+}T^{+} compared to A^{-}T^{-}, even though the sample size was limited (n = 13 participants in each group). The model with 40 transcripts was the one with the poorer performance with an AUC of 0.80, whereas the models with 90 and 220 transcripts had an AUC of 0.88 (Figure 2B; Table S5). In this case, including *APOE* improved the AUC in all cases, 0.86 for the model with 40 transcripts and 0.90 for the models including 90 and 220 transcripts.

A

Characteristic	Parkinson's Disease	Lewy Body Dementia	Frontotemporal Dementia
N	96	17	16
Male (N) (%)	61 (63.54%)	11 (64.71%)	11 (68.75%)
Age at draw median (IQR)	72.00 (67.00, 77.00)	79 (74.00, 83.00)	62.50 (59.50, 67.25)
Age at onset median (IQR)	64.00 (60.00, 70.00)	71.00 (69.00, 73.00)	57.50 (54.50, 62.25)
Disease duration median (IQR)	7.00 (4.00, 10.00)	6.00 (3.00, 10.00)	4.50 (3.00, 7.25)

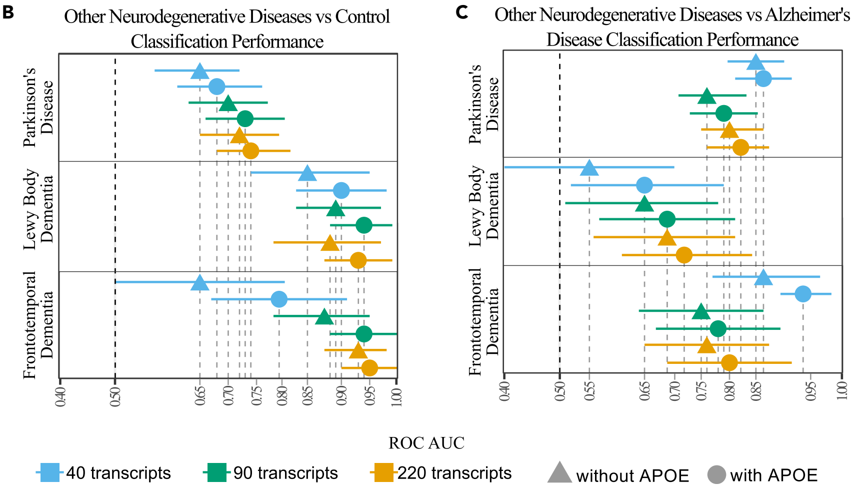


Figure 3. Sensitivity analyses for the predictive models in other neurodegenerative diseases

(A) Summary demographics for the individuals from other neurodegenerative diseases.

(B) Whisker plot showing the performance of the prediction of other neurodegenerative diseases compared to controls for the three predictive models (40, 90, and 220 transcripts) with and without APOE genotype.

(C) Whisker plot showing the performance of the prediction of other neurodegenerative diseases compared to AD for the three predictive models (40, 90, and 220 transcripts) with and without APOE genotype.

Predictive models trained with presymptomatic AD participants can also predict AD in the symptomatic phases of the disease

Besides the ability to differentiate between presymptomatic AD participants and controls; we also tested the performance of the models within the AD continuum (Figure 1A). We evaluated the accuracy of the three models in early symptomatic (CDR = 0.5, n = 42) and symptomatic (CDR = 1, n = 50) AD compared to controls (n = 48) (Figure 2C). For early symptomatic AD participants, the AUC of the models composed by 40, 90, and 220 transcripts was 0.93, 0.95, and 0.98 respectively while for symptomatic AD the AUC were 0.91, 0.94, and 0.96 (Figure 2B; Table S6). Differently from our results with the presymptomatic group, the addition of *APOE* genotype did improve the accuracy of the three models in the AD continuum (Figure 2D; Table S6). However, the improvement was not accentuated, suggesting that we are still capturing some of the *APOE* genotype in the plasma transcriptome. Overall, the model with 40 transcripts showed less predictive power than models including additional transcripts. Although this could be a technical artifact of how linear models work (higher number of predictors tend to increase predictive power), our results may alternatively suggest that as AD progresses, the molecular signatures change, which has an impact on the accuracy of the predictive models. However, the addition of transcripts to the signature yield better AUC, suggesting that for some transcripts, the changes accentuate with disease progression, increasing the predictive ability of the models.

Predictive models trained with presymptomatic AD participants have limited ability to predict other neurodegenerative diseases

Lastly, we wanted to assess if the models were specific to AD. We evaluated the performance of our models in samples from Parkinson's disease (PD-*n* = 96), Lewy body dementia (DLB-*n* = 17) and Frontotemporal dementia (FTD-*n* = 16) (Figure 3A). We tested the specificity using two approaches. Firstly, we asked if the models could correctly classify PD/DLB/FTD when compared to controls (Figure 3B; Table S6), and secondly, if they could classify them when compared to AD (Figure 3C; Table S7). The models had low predictive power to differentiate PD from controls (AUC<0.72), while the performance for FTD and DLB varied and depended on the number of transcripts (0.64<AUC<0.93), suggesting that the models are specific to AD, but we might be capturing the same biological process in diseases with high overlap like DLB and

AD. In this case, the addition of *APOE* genotype did not decrease the predictive power, and this did not increase the specificity (Figure 3B; Table S6). Similarly, when differentiating between AD and other neurodegenerative diseases, the models had high predictive power to differentiate PD (0.77<AUC<0.85) and FTD (0.75<AUC<0.86), but not as much for DLB (0.55<AUC<0.69). In contrast with the previous sections, the addition of *APOE* genotype improved the differentiation of AD from other neurodegenerative diseases with AUC>0.70 in all cases (Figures 3B and 3C; Table S7).

The model with 220 transcripts could differentiate PD from AD participants with an AUC of 0.81, whereas the differentiation from DLB or FTD (AUC<0.76) was less accurate. For the models including a smaller number of transcripts (90 and 40), they could differentiate AD from PD (AUC>0.81) and FTD (AUC>0.75), but not from DLB (AUC<0.65), suggesting that there are several transcripts that are commonly dysregulated in the two diseases and thus not useful for the differentiation. In fact, when we evaluated the expression patterns in each neurodegenerative disease compared to controls of all the transcripts included in the three models, we observed that the same transcripts were dysregulated in all three diseases, but in different directions when compared to controls. DLB was the one with the most striking differences in dysregulation for the transcripts selected in the predictive models (Figure S7), suggesting that DLB has several genes commonly dysregulated with AD, but in different directions and proportions. In consequence, it is possible to hypothesize that DLB has, not only more clinical features shared with AD, but also more molecular pathways than those shared with PD or FTD, making it the most difficult to differentiate by the transcription models.

DISCUSSION

This is the first study using plasma cfRNA to create machine learning-based predictive models able to identify AD at the presymptomatic stages in two independent datasets. We identified transcripts in plasma that seem to recapitulate the changes taking place in the brain of AD participants, suggesting that changes taking place in the brain are leaking into the blood stream, most likely due to blood-brain barrier (BBB) breakdown.[51,52] We have also built predictive models that correctly classify presymptomatic AD and control participants with a reasonable number of transcripts, and that showed high accuracy and specificity for AD. Given the reduced number of transcripts that the models include, they are potentially applicable to the clinical setting if further testing supports their beneficial use. Studies with larger sample sizes are needed to improve the performance of the predictive models, however, here, we are demonstrating for the first time that cfRNA not only can be used as an early predictive tool but also that it captures early pathological changes.

We have investigated the early changes in plasma using cfRNA quantification, and identified a significant overlap with those previously published.[40] On top of that, we have also proven that early changes in cfRNA plasma might be originating in the brain since several transcripts are also DE in brains of individuals with AD.[47,53] Additionally, we identified that the cfRNA dysregulated transcripts are part of co-expression modules already identified in the cortex of AD cases and enriched in GO terms associated with the brain such as *synapse* and *postsynapse*. We also found an association with oligodendrocytes and cytoskeleton for the identified transcripts. Cytoskeleton organization seems to play a key role in oligodendrocyte proliferation. For example, *TRAK2*, a DE transcript in plasma of presymptomatic AD participants, is associated with oligodendrocytes by participating in the regulation of the organization of the actin cytoskeleton and with mitochondrial transport, all processes that may be contributing to AD.[54–56] Several studies suggest that there is an early breakdown of the BBB due to the initiation of the disease,[51,52,57] fact that supports the central nervous system origin of our findings, and those from others.[40]

To our knowledge, the study by Toden et al.[40] was the only one that evaluated plasma cfRNA as AD biomarker. However, it has important differences with the present study that make our study design better suited to build predictive models. First, it did not include presymptomatic AD participants, thus their model with 1658 transcripts is only applicable to clinical phases. Second, they did not perform specificity analysis with other neurodegenerative diseases, in consequence, the specificity of the model is unknown. Third, they used the 1658 DE transcripts to build the model, which, given the elevated number of transcripts, makes it difficult to translate it to a clinical setting. On top of that, by using all the DE transcripts the model has the potential to be redundant, overfitted and thus, non-generalizable. We have taken a more conservative approach, by developing several predictive models, considerably simpler in terms of the number of transcripts, to understand how cfRNA behaves in the context of AD. We have also included different neurodegenerative diseases to calculate the specificity and acknowledge the known overlap across diseases. Finally, we have studied the correlation between CSF AD biomarkers and ATN criteria to compare with the tools used in real clinical settings.

To date, CSF biomarkers have proven to be the most effective approach to classifying AD. Combinations of CSF biomarkers classify clinical AD and controls with accuracies ranging from 0.6 to 0.95 depending on age.[58,59] The most used and accurate CSF biomarker is the Aβ42/Aβ40 ratio which can correctly diagnose 82.8% of the screened AD patients.[9] In addition, combinations of CSF biomarkers also showed good accuracy in detecting incipient AD in participants with mild cognitive impairment (MCI).[11] Current plasma biomarkers show accuracies similar to that of CSF. The combination of plasma Aβ42/Aβ40, age, and *APOE* ε4 status is highly correlated with amyloid PET positivity and thus it could be used to screen for individuals prior to lumbar puncture, PET scan, or further testing.[32] Plasma p-tau has been shown to associate significantly with CSF Aβ42/Aβ40 and Aβ-PET, with p-tau217 and/or p-tau231 with p-tau181 showing equal performance in differentiating AD pathology from healthy controls.[60] Studies also found that plasma neurofilament light chain (NfL) increased in AD compared to Subjective Cognitive Decline (SCD) participants, making it a promising biomarker candidate.[61] Further, platelet proteins have been suggested as prospective peripheral AD biomarkers.[62] The three cfRNA models yield similar accuracy to that of plasma Aβ42/Aβ40, without including other variables. In fact, we observe that the addition of *APOE* does not improve the models. The main advantage of cfRNA from the current protein measurements is its potential to be translated to a real-time PCR, which is a more cost-effective technique that can be implemented in all clinical settings, even in those that are remote. Additionally, given the independence from Aβ, cfRNA could be used for therapy monitoring when we evaluated Aβ protein-targeted drugs.

Predictive models built using machine learning approaches in neurodegenerative diseases tend to contain a large number of features,[40,45] or contain only the identified DE genes, transcripts, or proteins.[63,64] In here, we focused on models with a relatively low number of transcripts without compromising accuracy to maximize their potential to be translated to the clinic. Previous models using cfRNA reported an AUC of 0.83 for clinical AD; by substantially reducing the number of transcripts, we have increased the AUC up to 0.94 for presymptomatic AD and using only 220 transcripts. For clinical phases, the model with 220 genes showed a stable accuracy, outperforming the previously published. We also demonstrated that the cfRNA predictive model is specific to AD, since it is not able to predict PD, or FTD. Even though the model is able to predict DLB with a reasonable accuracy, we expect some degree of overlap in the prediction of these diseases due to the existing clinical and pathological overlap, along with the clinical misdiagnosis of DLB for AD, which can have an impact on the construction of the predictive models using supervised techniques.[49,65–67]

Limitations of the study

This study has several limitations. The sample size we reached for presymptomatic AD is rather limited and sourced from one repository. However, to the best of our knowledge, this is the largest sample of presymptomatic sporadic AD with clinical retrospective data there is. Due to the sample selection strategy, the samples have been stored in the freezer for long periods of time, which might affect our findings. Considering this, we have removed any transcript that showed selective degradation to minimize this effect. The use of RNA-seq techniques is very sensitive to bias, especially to using machine learning afterward. While the two datasets are methodologically independent according to the machine learning field, they originate from the same site, which adds to the potential bias effect of the present study. Nonetheless, for modeling purposes, the generation of two independent datasets and the use of mathematical approaches have mitigated the presence of potential methodological bias. Additionally, we have proposed a new approach to integrate RNA-seq datasets applicable in large studies utilizing multiple datasets and different data types. Finally, larger sample sizes for all AD stages and other neurodegenerative diseases are needed to confirm our DE findings and generalize and improve the accuracy and specificity of the model. Nevertheless, we believe that this study serves as proof that cfRNA has the potential to detect changes related to AD pathobiology, even before the onset of symptoms.

Despite the aforementioned limitations, in this study we were able to model and replicate in an independent dataset a predictor that can identify presymptomatic AD. On top of that, the predictor has been designed independently of Aβ42, which makes it an excellent candidate to monitor potential disease modifying therapies. The use of plasma cfRNA as biomarker is very advantageous due to its cost-effectiveness compared to current CSF and plasma measures and the fact that cfRNA models have the potential to be transformed into real-time PCR panels, therefore cfRNA could be smoothly implemented in the clinic without additional equipment or training, also in remote settings, something impossible with the current tools. Overall, we believe that further longitudinal studies with larger sample sizes are needed to confirm the use of cfRNA as a biomarker, but the current results show unprecedented potential.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2023.108534.

ACKNOWLEDGMENTS

We thank all the participants and their families along with the institutions and all the staff who provided plasma tissue, without whom this study would not have been possible. This work was supported by access to equipment made possible by the Hope Center for Neurological Disorders, the Neurogenomics and Informatics Center (NGI: https://neurogenomics.wustl.edu/) and the Departments of Neurology and

Psychiatry at Washington University School of Medicine. This work was supported by grants from the National Institutes of Health K99/R00-AG062723 (L.I.), R01AG044546 (C.C.), P01AG003991, RF1AG053303 (C.C.), R01AG058501 (C.C.), RO1 NS075321 (J.S.P.), U01AG058922 (C.C.), P30 AG066444 (J.C.M.), P01 AG03991 (J.C.M.), and P01 AG026276 (J.C.M.), the Michael J. Fox Foundation (L.I./C.C.), the Alzheimer's Drug Discovery Foundation (L.I.), BrightFocus Foundation (L.I.), and the Department of Defense (L.I.). A.C. was supported by the Science and Technology Agency, Séneca Foundation, Comunidad Autónoma Región de Murcia, Spain through the Grant [20762/FPI/18].

AUTHOR CONTRIBUTIONS

L.I., A.C., A.B., and J.B. conceived and wrote this article. L.I. and C.C. conceptualized and designed the research plan. L.I., A.C., M.F., J.B., and C.C. designed the analysis plan. A.C., A.B., M.A., J.P., and H.C. processed all the data and performed the analyses. J.N., J.G., and K.B. obtained and processed the samples. A.C., A.B., M.A., J.A., H.C., M.F., J.N., J.G., K.B., J.B., J.P., J.M., C.C., J.B., and L.I. discussed the project, revised the manuscript, and provided critical feedback.

DECLARATION OF INTERESTS

The funders of the study had no role in the collection, analysis, or interpretation of data; in the writing of the report; or in the decision to submit the paper for publication. CC is a member of the advisory board of Vivid Genetics, Halia Therapeutics, and ADx Healthcare and has received research support from Biogen, EISAI, Alector and Parabon. The rest of the authors report no conflict of interest. Patent Pending.

INCLUSION AND DIVERSITY

We support inclusive, diverse, and equitable conduct of research.

Received: August 22, 2023
Revised: November 9, 2023
Accepted: November 20, 2023
Published: November 23, 2023

REFERENCES

1. LaFerla, F.M., and Oddo, S. (2005). Alzheimer's disease: Abeta, tau and synaptic dysfunction. Trends Mol. Med. *11*, 170–176.
2. 2021 Alzheimer's Disease Facts and Figures - 2021 - Alzheimer's & Dementia - Wiley Online Library. https://alz-journals.onlinelibrary.wiley.com/doi/10.1002/alz.12328.
3. (2018). 2018 Alzheimer's disease facts and figures. Alzheimers Dement. *14*, 367–429.
4. Dubois, B., Hampel, H., Feldman, H.H., Scheltens, P., Aisen, P., Andrieu, S., Bakardjian, H., Benali, H., Bertram, L., Blennow, K., et al. (2016). Preclinical Alzheimer's disease: Definition, natural history, and diagnostic criteria. Alzheimers Dement. *12*, 292–323.
5. Paraskevaidi, M., Allsop, D., Karim, S., Martin, F.L., and Crean, S. (2020). Diagnostic Biomarkers for Alzheimer's Disease Using Non-Invasive Specimens. J. Clin. Med. *9*, 1673.
6. Simonsen, A.H., McGuire, J., Hansson, O., Zetterberg, H., Podust, V.N., Davies, H.A., Waldemar, G., Minthon, L., and Blennow, K. (2007). Novel Panel of Cerebrospinal Fluid Biomarkers for the Prediction of Progression to Alzheimer Dementia in Patients With Mild Cognitive Impairment. Arch. Neurol. *64*, 366–370.
7. Osorio, R.S., Pirraglia, E., Gumb, T., Mantua, J., Ayappa, I., Williams, S., Mosconi, L., Glodzik, L., and de Leon, M.J. (2014). Imaging and CSF Biomarkers in the Search for Alzheimer's Disease Mechanisms. Neurodegener. Dis. *13*, 163–165.
8. Young, P.N.E., Estarellas, M., Coomans, E., Srikrishna, M., Beaumont, H., Maass, A., Venkataraman, A.V., Lissaman, R., Jiménez, D., Betts, M.J., et al. (2020). Imaging biomarkers in neurodegeneration: current and future practices. Alzheimer's Res. Ther. *12*, 49.
9. Biscetti, L., Salvadori, N., Farotti, L., Cataldi, S., Eusebi, P., Paciotti, S., and Parnetti, L. (2019). The added value of Aβ42/Aβ40 in the CSF signature for routine diagnostics of Alzheimer's disease. Clin. Chim. Acta *494*, 71–73.
10. Janelidze, S., Zetterberg, H., Mattsson, N., Palmqvist, S., Vanderstichele, H., Lindberg, O., van Westen, D., Stomrud, E., Minthon, L., Blennow, K., et al. (2016). CSF Aβ42/Aβ40 and Aβ42/Aβ38 ratios: better diagnostic markers of Alzheimer disease. Ann. Clin. Transl. Neurol. *3*, 154–165.
11. Hansson, O., Lehmann, S., Otto, M., Zetterberg, H., and Lewczuk, P. (2019). Advantages and disadvantages of the use of the CSF Amyloid β (Aβ) 42/40 ratio in the diagnosis of Alzheimer's Disease. Alzheimer's Res. Ther. *11*, 34.
12. Kuhlmann, J., Andreasson, U., Pannee, J., Bjerke, M., Portelius, E., Leinenbach, A., Bittner, T., Korecka, M., Jenkins, R.G., Vanderstichele, H., et al. (2017). CSF Aβ1–42 – an excellent but complicated Alzheimer's biomarker – a route to standardisation. Clin. Chim. Acta *467*, 27–33.
13. Brier, M.R., Gordon, B., Friedrichsen, K., McCarthy, J., Stern, A., Christensen, J., Owen, C., Aldea, P., Su, Y., Hassenstab, J., et al. (2016). Tau and Aβ imaging, CSF measures, and cognition in Alzheimer's disease. Sci. Transl. Med. *8*, 338ra66.
14. Buerger, K., Ewers, M., Pirttilä, T., Zinkowski, R., Alafuzoff, I., Teipel, S.J., DeBernardis, J., Kerkman, D., McCulloch, C., Soininen, H., and Hampel, H. (2006). CSF phosphorylated tau protein correlates with neocortical neurofibrillary pathology in Alzheimer's disease. Brain *129*, 3035–3041.
15. van Harten, A.C., Kester, M.I., Visser, P.-J., Blankenstein, M.A., Pijnenburg, Y.A.L., van der Flier, W.M., and Scheltens, P. (2011). Tau and p-tau as CSF biomarkers in dementia: a meta-analysis. Clin. Chem. Lab. Med. *49*, 353–366.
16. Janelidze, S., Stomrud, E., Smith, R., Palmqvist, S., Mattsson, N., Airey, D.C., Proctor, N.K., Chai, X., Shcherbinin, S., Sims, J.R., et al. (2020). Cerebrospinal fluid p-tau217 performs better than p-tau181 as a biomarker of Alzheimer's disease. Nat. Commun. *11*, 1683.
17. Kandimalla, R.J.L., Prabhakar, S., Wani, W.Y., Kaushal, A., Gupta, N., Sharma, D.R., Grover, V.K., Bhardwaj, N., Jain, K., and Gill, K.D. (2013). CSF p-Tau levels in the prediction of Alzheimer's disease. Biol. Open *2*, 1119–1124.
18. Jack, C.R., Bennett, D.A., Blennow, K., Carrillo, M.C., Dunn, B., Haeberlein, S.B., Holtzman, D.M., Jagust, W., Jessen, F., Karlawish, J., et al. (2018). NIA-AA Research Framework: Toward a biological definition of Alzheimer's disease. Alzheimers Dement. *14*, 535–562.
19. Ebenau, J.L., Timmers, T., Wesselman, L.M.P., Verberk, I.M.W., Verfaillie, S.C.J., Slot, R.E.R., van Harten, A.C., Teunissen, C.E., Barkhof, F., van den Bosch, K.A., et al. (2020). ATN classification and clinical progression in subjective cognitive decline: The SCIENCe project. Neurology *95*, e46–e58.
20. Delmotte, K., Schaeverbeke, J., Poesen, K., and Vandenberghe, R. (2021). Prognostic value of amyloid/tau/neurodegeneration (ATN) classification based on diagnostic

cerebrospinal fluid samples for Alzheimer's disease. Alzheimer's Res. Ther. *13*, 84.

21. Shaw, L.M., Waligorska, T., Fields, L., Korecka, M., Figurski, M., Trojanowski, J.Q., Eichenlaub, U., Wahl, S., Quan, M., Pontecorvo, M.J., et al. (2018). Derivation of cutoffs for the Elecsys® amyloid β (1–42) assay in Alzheimer's disease. Alzheimers Dement. *10*, 698–705.
22. Vogelgsang, J., Wedekind, D., Bouter, C., Klafki, H.-W., and Wiltfang, J. (2018). Reproducibility of Alzheimer's Disease Cerebrospinal Fluid-Biomarker Measurements under Clinical Routine Conditions. J. Alzheimers Dis. *62*, 203–212.
23. Ingala, S., De Boer, C., Masselink, L.A., Vergari, I., Lorenzini, L., Blennow, K., Chételat, G., Di Perri, C., Ewers, M., van der Flier, W.M., et al. (2021). Application of the ATN classification scheme in a population without dementia: Findings from the EPAD cohort. Alzheimers Dement. *17*, 1189–1204.
24. Fiandaca, M.S., Kapogiannis, D., Mapstone, M., Boxer, A., Eitan, E., Schwartz, J.B., Abner, E.L., Petersen, R.C., Federoff, H.J., Miller, B.L., and Goetzl, E.J. (2015). Identification of pre-clinical Alzheimer's disease by a profile of pathogenic proteins in neurally-derived blood exosomes: a case-control study. Alzheimers Dement. *11*, 600–607.e1.
25. Olsson, B., Lautner, R., Andreasson, U., Öhrfelt, A., Portelius, E., Bjerke, M., Hölttä, M., Rosén, C., Olsson, C., Strobel, G., et al. (2016). CSF and blood biomarkers for the diagnosis of Alzheimer's disease: a systematic review and meta-analysis. Lancet Neurol. *15*, 673–684.
26. Dage, J.L., Wennberg, A.M.V., Airey, D.C., Hagen, C.E., Knopman, D.S., Machulda, M.M., Roberts, R.O., Jack, C.R., Petersen, R.C., and Mielke, M.M. (2016). Levels of tau protein in plasma are associated with neurodegeneration and cognitive function in a population-based elderly cohort. Alzheimers Dement. *12*, 1226–1234.
27. Pase, M.P., Beiser, A.S., Himali, J.J., Satizabal, C.L., Aparicio, H.J., DeCarli, C., Chêne, G., Dufouil, C., and Seshadri, S. (2019). Assessment of Plasma Total Tau Level as a Predictive Biomarker for Dementia and Related Endophenotypes. JAMA Neurol. *76*, 598–606.
28. Karikari, T.K., Pascoal, T.A., Ashton, N.J., Janelidze, S., Benedet, A.L., Rodriguez, J.L., Chamoun, M., Savard, M., Kang, M.S., Therriault, J., et al. (2020). Blood phosphorylated tau 181 as a biomarker for Alzheimer's disease: a diagnostic performance and prediction modelling study using data from four prospective cohorts. Lancet Neurol. *19*, 422–433.
29. Thijssen, E.H., La Joie, R., Wolf, A., Strom, A., Wang, P., Iaccarino, L., Bourakova, V., Cobigo, Y., Heuer, H., Spina, S., et al. (2020). Diagnostic value of plasma phosphorylated tau181 in Alzheimer's disease and frontotemporal lobar degeneration. Nat. Med. *26*, 387–397.
30. Janelidze, S., Bali, D., Ashton, N.J., Barthélemy, N.R., Vanbrabant, J., Stoops, E., Vanmechelen, E., He, Y., Dolado, A.O., Triana-Baltzer, G., et al. (2023). Head-to-head comparison of 10 plasma phospho-tau assays in prodromal Alzheimer's disease. Brain *146*, 1592–1601.
31. Risacher, S.L., Fandos, N., Romero, J., Sherriff, I., Pesini, P., Saykin, A.J., and Apostolova, L.G. (2019). Plasma amyloid beta levels are associated with cerebral amyloid and tau deposition. Alzheimers Dement. *11*, 510–519.
32. Schindler, S.E., Bollinger, J.G., Ovod, V., Mawuenyega, K.G., Li, Y., Gordon, B.A., Holtzman, D.M., Morris, J.C., Benzinger, T.L.S., Xiong, C., et al. (2019). High-precision plasma β-amyloid 42/40 predicts current and future brain amyloidosis. Neurology *93*, e1647–e1659.
33. Everett, T.R., and Chitty, L.S. (2015). Cell-free fetal DNA: the new tool in fetal medicine. Ultrasound Obstet. Gynecol. *45*, 499–507.
34. Tzimagiorgis, G., Michailidou, E.Z., Kritis, A., Markopoulos, A.K., and Kouidou, S. (2011). Recovering circulating extracellular or cell-free RNA from bodily fluids. Cancer Epidemiol. *35*, 580–589.
35. Snyder, M.W., Kircher, M., Hill, A.J., Daza, R.M., and Shendure, J. (2016). Cell-free DNA comprises an in vivo nucleosome footprint that informs its tissues-of-origin. Cell *164*, 57–68.
36. Rapisuwon, S., Vietsch, E.E., and Wellstein, A. (2016). Circulating biomarkers to monitor cancer progression and treatment. Comput. Struct. Biotechnol. J. *14*, 211–222.
37. Koh, W., Pan, W., Gawad, C., Fan, H.C., Kerchner, G.A., Wyss-Coray, T., Blumenfeld, Y.J., El-Sayed, Y.Y., and Quake, S.R. (2014). Noninvasive in vivo monitoring of tissue-specific global gene expression in humans. Proc. Natl. Acad. Sci. USA *111*, 7361–7366.
38. Kumar, S., and Reddy, P.H. (2016). Are circulating microRNAs peripheral biomarkers for Alzheimer's disease? Biochim. Biophys. Acta *1862*, 1617–1627.
39. Sheinerman, K.S., Toledo, J.B., Tsivinsky, V.G., Irwin, D., Grossman, M., Weintraub, D., Hurtig, H.I., Chen-Plotkin, A., Wolk, D.A., McCluskey, L.F., et al. (2017). Circulating brain-enriched microRNAs as novel biomarkers for detection and differentiation of neurodegenerative diseases. Alzheimer's Res. Ther. *9*, 89.
40. Toden, S., Zhuang, J., Acosta, A.D., Karns, A.P., Salathia, N.S., Brewer, J.B., Wilcock, D.M., Aballi, J., Nerenberg, M., Quake, S.R., and Ibarra, A. (2020). Noninvasive characterization of Alzheimer's disease by circulating, cell-free messenger RNA next-generation sequencing. Sci. Adv. *6*, eabb1654.
41. Satoh, J.I., Kino, Y., and Niida, S. (2015). MicroRNA-Seq Data Analysis Pipeline to Identify Blood Biomarkers for Alzheimer's Disease from Public Data. Biomark. Insights *10*, 21–31.
42. Schipper, H.M., Maes, O.C., Chertkow, H.M., and Wang, E. (2007). MicroRNA Expression in Alzheimer Blood Mononuclear Cells. Gene Regul. Syst. Biol. *1*, 263–274.
43. Galimberti, D., Villa, C., Fenoglio, C., Serpente, M., Ghezzi, L., Cioffi, S.M.G., Arighi, A., Fumagalli, G., and Scarpini, E. (2014). Circulating miRNAs as potential biomarkers in Alzheimer's disease. J. Alzheimers Dis. *42*, 1261–1267.
44. Tan, L., Yu, J.-T., Liu, Q.-Y., Tan, M.-S., Zhang, W., Hu, N., Wang, Y.-L., Sun, L., Jiang, T., and Tan, L. (2014). Circulating miR-125b as a biomarker of Alzheimer's disease. J. Neurol. Sci. *336*, 52–56.
45. Makarious, M.B., Leonard, H.L., Vitale, D., Iwaki, H., Sargent, L., Dadu, A., Violich, I., Hutchins, E., Saffo, D., Bandres-Ciga, S., et al. (2022). Multi-modality machine learning predicting Parkinson's disease. NPJ Parkinsons Dis. *8*, 35.
46. Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550.
47. Chen, H.-H., Eteleeb, A., Wang, C., Fernandez, M.V., Budde, J.P., Bergmann, K., Norton, J., Wang, F., Ebl, C., Morris, J.C., et al. (2022). Circular RNA detection identifies circPSEN1 alterations in brain specific to autosomal dominant Alzheimer's disease. Acta Neuropathol. Commun. *10*, 29.
48. Marques-Coelho, D., Iohan, L.d.C.C., Melo de Farias, A.R., Flaig, A., Brainbank Neuro–CEB Neuropathology Network, Lambert, J.C., and Costa, M.R. (2021). Differential transcript usage unravels gene expression alterations in Alzheimer's disease human brains. NPJ Aging Mech. Dis. *7*, 2–15.
49. Wakasugi, N., and Hanakawa, T. (2021). It Is Time to Study Overlapping Molecular and Circuit Pathophysiologies in Alzheimer's and Lewy Body Disease Spectra. Front. Syst. Neurosci. *15*, 777706.
50. García-Ruiz, S., Gil-Martínez, A.L., Cisterna, A., Jurado-Ruiz, F., Reynolds, R.H., Botía, J.A., NABEC North America Brain Expression Consortium, Cookson, M.R., Hardy, J., and Botía, J.A. (2021). CoExp: A Web Tool for the Exploitation of Co-expression Networks. Front. Genet. *12*, 630187.
51. Montagne, A., Zhao, Z., and Zlokovic, B.V. (2017). Alzheimer's disease: A matter of blood-brain barrier dysfunction? J. Exp. Med. *214*, 3151–3169.
52. Sweeney, M.D., Sagare, A.P., and Zlokovic, B.V. (2018). Blood-brain barrier breakdown in Alzheimer disease and other neurodegenerative disorders. Nat. Rev. Neurol. *14*, 133–150.
53. Blalock, E.M., Geddes, J.W., Chen, K.C., Porter, N.M., Markesbery, W.R., and Landfield, P.W. (2004). Incipient Alzheimer's disease: Microarray correlation analyses reveal major transcriptional and tumor suppressor responses. Proc. Natl. Acad. Sci. USA *101*, 2173–2178.
54. Fenton, A.R., Jongens, T.A., and Holzbaur, E.L.F. (2021). Mitochondrial adaptor TRAK2 activates and functionally links opposing kinesin and dynein motors. Nat. Commun. *12*, 4578.
55. Quintanilla, R.A., Tapia-Monsalves, C., Vergara, E.H., Pérez, M.J., and Aranguiz, A. (2020). Truncated Tau Induces Mitochondrial Transport Failure Through the Impairment of TRAK2 Protein and Bioenergetics Decline in Neuronal Cells. Front. Cell. Neurosci. *14*, 175.
56. Correia, S.C., Perry, G., and Moreira, P.I. (2016). Mitochondrial traffic jams in Alzheimer's disease - pinpointing the roadblocks. Biochim. Biophys. Acta *1862*, 1909–1917.
57. Preische, O., Schultz, S.A., Apel, A., Kuhle, J., Kaeser, S.A., Barro, C., Gräber, S., Kuder-Buletta, E., LaFougere, C., Laske, C., et al. (2019). Serum neurofilament dynamics predicts neurodegeneration and clinical progression in presymptomatic Alzheimer's disease. Nat. Med. *25*, 277–283.
58. Mattsson, N., Rosén, E., Hansson, O., Andreasen, N., Parnetti, L., Jonsson, M., Herukka, S.-K., van der Flier, W.M., Blankenstein, M.A., Ewers, M., et al. (2012). Age and diagnostic performance of Alzheimer disease CSF biomarkers. Neurology *78*, 468–476.
59. Molinuevo, J.L., Gispert, J.D., Dubois, B., Heneka, M.T., Lleo, A., Engelborghs, S., Pujol, J., de Souza, L.C., Alcolea, D., Jessen, F., et al.

(2013). The AD-CSF-index discriminates Alzheimer's disease patients from healthy controls: a validation study. J. Alzheimers Dis. *36*, 67–77.

60. Gonzalez-Ortiz, F., Kac, P.R., Brum, W.S., Zetterberg, H., Blennow, K., and Karikari, T.K. (2023). Plasma phospho-tau in Alzheimer's disease: towards diagnostic and therapeutic trial applications. Mol. Neurodegener. *18*, 18.

61. Giacomucci, G., Mazzeo, S., Bagnoli, S., Ingannato, A., Leccese, D., Berti, V., Padiglioni, S., Galdo, G., Ferrari, C., Sorbi, S., et al. (2022). Plasma neurofilament light chain as a biomarker of Alzheimer's disease in Subjective Cognitive Decline and Mild Cognitive Impairment. J. Neurol. *269*, 4270–4280.

62. Akingbade, O.E.S., Gibson, C., Kalaria, R.N., and Mukaetova-Ladinska, E.B. (2018). Platelets: Peripheral Biomarkers of Dementia? J. Alzheimers Dis. *63*, 1235–1259.

63. Zhang, Y., Ghose, U., Buckley, N.J., Engelborghs, S., Sleegers, K., Frisoni, G.B., Wallin, A., Lleó, A., Popp, J., Martinez-Lage, P., et al. (2022). Predicting AT(N) pathologies in Alzheimer's disease from blood-based proteomic data using neural networks. Front. Aging Neurosci. *14*, 1040001.

64. Wang, L., Western, D., Timsina, J., Repaci, C., Song, W.-M., Norton, J., Kohlfeld, P., Budde, J., Climer, S., Butt, O.H., et al. (2023). Plasma proteomics of SARS-CoV-2 infection and severity reveals impact on Alzheimer's and coronary disease pathways. iScience *26*, 106408.

65. Galasko, D., Hansen, L.A., Katzman, R., Wiederholt, W., Masliah, E., Terry, R., Hill, L.R., Lessin, P., and Thal, L.J. (1994). Clinical-neuropathological correlations in Alzheimer's disease and related dementias. Arch. Neurol. *51*, 888–895.

66. Fujishiro, H., Iseki, E., Higashi, S., Kasanuki, K., Murayama, N., Togo, T., Katsuse, O., Uchikado, H., Aoki, N., Kosaka, K., et al. (2010). Distribution of cerebral amyloid deposition and its relevance to clinical phenotype in Lewy body dementia. Neurosci. Lett. *486*, 19–23.

67. Padovani, A., Premi, E., Pilotto, A., Gazzina, S., Cosseddu, M., Archetti, S., Cancelli, V., Paghera, B., and Borroni, B. (2013). Overlap between frontotemporal dementia and Alzheimer's disease: cerebrospinal fluid pattern and neuroimaging study. J. Alzheimers Dis. *36*, 49–55.

68. Morris, J.C. (1993). The Clinical Dementia Rating (CDR): current version and scoring rules. Neurology *43*, 2412–2414.

69. Hughes, A.J., Daniel, S.E., Kilford, L., and Lees, A.J. (1992). Accuracy of clinical diagnosis of idiopathic Parkinson's disease: a clinico-pathological study of 100 cases. J. Neurol. Neurosurg. Psychiatry *55*, 181–184.

70. Fagan, A.M., Mintun, M.A., Mach, R.H., Lee, S.-Y., Dence, C.S., Shah, A.R., LaRossa, G.N., Spinner, M.L., Klunk, W.E., Mathis, C.A., et al. (2006). Inverse relation between in vivo amyloid imaging load and cerebrospinal fluid Abeta42 in humans. Ann. Neurol. *59*, 512–519.

71. Ewels, P., Magnusson, M., Lundin, S., and Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics *32*, 3047–3048.

72. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21.

73. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

74. Patro, R., Duggal, G., Love, M.I., Irizarry, R.A., and Kingsford, C. (2017). Salmon: fast and bias-aware quantification of transcript expression using dual-phase inference. Nat. Methods *14*, 417–419.

75. Chen, J., Bardes, E.E., Aronow, B.J., and Jegga, A.G. (2009). ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. Nucleic Acids Res. *37*, W305–W311.

76. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2017). KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. *45*, D353–D361.

77. Bennett, D.A., Buchman, A.S., Boyle, P.A., Barnes, L.L., Wilson, R.S., and Schneider, J.A. (2018). Religious Orders Study and Rush Memory and Aging Project. J. Alzheimers Dis. *64*, S161–S189.

78. Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. J. Stat. Softw. *33*, 1–22.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological samples		
Human plasma samples	Knight ADRC and Movement Disorder Clinic, Washington University in St. Louis	N/A
Chemicals, peptides, and recombinant proteins		
DNAseI	NEB	Cat# M0303
Critical commercial assays		
Maxwell® RSC miRNA Plasma or Serum Kit	Ambion	N/A
NEBNext rRNA Depletion Kit	NEB	Cat# E6310
QIAmp Circulating Nucleic Acid kit	QIAGEN	Cat# 55114
NEBNext Ultra II Directional RNA Library Prep Kit for Illumina	NEB	Cat# E7760
Deposited data		
Raw data	This paper	NIAGADS dss: NG00142
Demographic participant information	Knight ADRC and Movement Disorder Clinic, Washington University in St. Louis	NIAGADS dss: NG00142
Software and algorithms		
Custom R codes	This paper	https://github.com/Ibanez-Lab/PlasmaCellFreeRNA-AlzhiemerDisease
FastQC v0.11.7	Babraham Bioinformatics	https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
STAR v2.7.1a	Dobin et al.[67]	https://github.com/alexdobin/STAR
PICARD v2.26	Broad Institute	https://broadinstitute.github.io/picard/
SAMtools	Li et al.[68]	https://samtools.sourceforge.net/
Salmon v0.11.3	Patro et al.[69]	https://salmon.readthedocs.io/en/latest/index.html
MultiQC v1.9	Ewels et al.[70]	https://multiqc.info/
DESeq2 v1.22.2	Love et al.[46]	https://bioconductor.org/packages/release/bioc/html/DESeq2.html
glmnet v2.0.16	Friedman et al.[71]	https://cran.r-project.org/web/packages/glmnet/index.html

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Laura Ibanez, PhD (ibanezl@wustl.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- RNAseq data, as well as the accompanying demographic data and normalized transcript counts, have been deposited to the NIAGADS Data Sharing Service (NIAGADS dss) and are publicly available as of the date of publication. Accession number is listed in the key resources table.

- All original code has been deposited at GitHub and is publicly available as of the date of publication. The link to the code page is available in the key resources table.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Study design

RNA was extracted from unfasted plasma samples from AD participants, controls, and other neurodegenerative diseases from two independent cohorts, both from the Knight Alzheimer Disease Research Center (Knight-ADRC) and the Movement Disorders Clinic (MDC) at Washington University in Saint Louis. After library preparation, sequencing, and stringent quality control, we compared the presymptomatic AD participants to the controls from the discovery cohort to identify differentially expressed transcripts. We compared our results to those previously published,[40] and to those identified to be differentially expressed in brain[47] to understand the potential origin of the altered transcripts. Then we leveraged machine learning tools to build predictive models that differentiate between presymptomatic AD participants and controls with a scalable number of genes that were replicated in an independent cohort. Finally, we calculated the predictive value of the models in symptomatic AD (CDR = 0.5 and CDR = 1) to test if the models were useful in clinical stages of the diseases, and in other neurodegenerative diseases such as Parkinson's Disease (PD), Lewy Body Dementia (DLB), and Frontotemporal Dementia (FTD) to test their specificity to AD.

Study participants

Plasma samples were obtained from the Knight-ADRC and the MDC at Washington University in Saint Louis repositories. These are deeply phenotyped cohorts, both clinically and molecularly with longitudinal data and samples available. We included 48 samples from healthy nondemented control participants, 67 samples from presymptomatic AD participants (Clinical Dementia Rating[68] (CDR = 0 at draw and current clinical diagnostic of AD), 42 samples from early symptomatic AD participants (CDR = 0.5 at draw and current diagnostic of AD), and 50 samples from symptomatic AD (CDR = 1 at draw, diagnostic of AD at draw, and current diagnostic of AD). All AD participants were required to have evidence of Aβ deposition (CSF Aβ<500 ng/L), positive PET scan and/or evidence of clinical worsening measured by CDR from the time at draw to the last clinical visit. For 71 participants, timely matched CSF biomarker measurements and time at draw are available. We also included participants from other neurodegenerative diseases: 17 DLB participants, 16 FTD participants, and 96 PD participants. AD, DLB, and FTD participants were diagnosed in accordance with clinical criteria that are embodied in the Uniform DataSet (UDS), the standard clinical dataset that is collected in all participants who are enrolled in all of the 37-federally funded ADRCs. PD participants were clinically diagnosed according to the UK Brain Bank criteria.[69] This research was conducted in accordance with the recommended protocols. Written informed consent was obtained from all participants or their family members. The Washington University in Saint Louis Institutional Review Board approved the study (IRB ID 201701124 and 202004010). Detailed demographic data is available from the NIAGADS dss, accession number listed in the key resources table.

METHOD DETAILS

RNA extraction and sequencing

Non-fasted plasma samples are collected as part of the research protocol every two years for all participants. After whole blood is obtained, it is centrifuged within 20 min at 1500 rpm for 10 min to obtain plasma and stored at −80°C until assayed.[70] Selected plasma samples from study participants that meet the inclusion criteria were thawed on ice and centrifuged at 2000 rpm for 5 min prior to RNA extraction to avoid cell RNA contamination. Samples were processed in two batches. For the training batch (n = 245), total plasma cfRNA was extracted from 0.5 mL of plasma using the Maxwell RSC miRNA from plasma or serum kit (Ambion) and ribodepleted (NEBNext rRNA Depletion Kit). For the testing batch (N = 91), total cfRNA was extracted from 1 mL of plasma using the QIAmp Circulating Nucleic Acid kit (QIAGEN) followed by a DNAseI digestion (New England Biolabs). In both cases, libraries were generated using the NEBNext Ultra II Directional RNA Library Prep Kit for Illumina (New England Biolabs) using 1 ng of RNA as input. Libraries were cleaned for possible adapter dimers. We targeted 40 million 100 base pair single-end reads for each sample using an Illumina NovaSeq 6000 for the training batch, and 15 million 100 base pair reads single-end reads Illumina HiSeq 2500 for the testing.

QUANTIFICATION AND STATISTICAL ANALYSIS

Data processing and quality control

We used FastQC (v0.11.7) to evaluate the sequencing quality of each sample. Then, we used STAR (v2.7.1a)[72] to obtain the BAM files and align them to the human reference genome GRCh38. After that, we used PICARD (v2.26) and SAMtools[73] to assess the quality of the sequences and the alignment. Finally, we used Salmon (v0.11.3)[74] to quantify the expression of transcripts. MultiQC (v1.9)[71] was used to gather quality control measures. We applied stringent quality control (QC). Briefly, after removing all transcripts with less than ten reads in more than 90% of the individuals, we calculated transcriptome Principal Component Analysis (PCA) and screen for correlation with technical and methodological variables to detect potential biases. We observed a strong correlation with total reads and coding bases; thus, we removed samples with less than 10% of coding bases and less than 1000000 total reads that were part of the same sequencing round. We also removed outlier samples based on transcriptome PCA and Cook's distances. Plasma samples have been stored for long periods of time prior to usage (up to 20 years), consequently, to address degradation, we used DESeq2 (v1.22.2)[46] to find transcripts associated with storage time in control participants. All trasncripts nominally ($p < 0.05$) associated were removed from the analyses (n = 2,580). Finally, we used DESeq2 to adjust for library complexity and normalize the counts using log transformation for the remaining transcripts (n = 19,830) and obtained the final population we used for the present analyses.

Differential expression and pathway analyses

Differential expression (DE) analyses were performed using DESeq2.[46] All analyses were adjusted by gender and age at draw. We used the Benjamini-Hochberg correction (FDR) to correct for multiple testing. FDR p values below 0.05 were considered significant. To replicate our findings, we used the DE transcripts identified in cfRNA by Toden et al.[40] Additionally, to evaluate if those transcripts were also DE in the brains of AD participants, we used an in-house RNAseq dataset from brains of participants of the Knight-ADRC.[47] To functionally characterize the DE transcripts we carried out gene-ontology enrichment analysis using the ToppGene Suite,[75] disease pathway overlap analysis using KEGG,[76] and gene co-expression network analysis using CoExp Web[50] with the Religious Orders Study and Rush Memory and Aging Project (ROSMAP) data[77] as a background matrix. For these analyses, we also used FDR to correct for multiple testing. Corrected p values below 0.05 were considered as significant.

Predictive models construction and evaluation

We designed a specific machine learning pipeline to produce a suitable classifier to identify presymptomatic AD cases based on gene expression and using two independent datasets. Briefly, we scaled the two datasets using z-scores and generated a linear model comparing the 47 presymptomatic AD cases and the 26 controls in the training dataset. We did the same for the testing dataset (20 presymptomatic AD cases and 22 controls). We kept only those transcripts that had the same direction of effect regardless of the p value. With the remaining transcripts, we calculated the Kullback-Leibler divergence (KLD) between the training and the testing dataset for each transcript, using the entropy R package (v1.3.1). We used the absolute value of the effect size from the linear model and the KLD value to rank the transcripts. Then we generate subsets of 40, 65, 90, 120, 150, 180, 220 and 250 transcripts. For each subset, we generated a model using KLD thresholds between 0.06 and 0.36 by increments of 0.02 and the R package glmnet (v2.0.16).[78] We evaluated different machine learning approaches: L1 regularization, L2 regularization, and a non-linear (random forest) approach. L2 regularization linear model was selected based on the best mean ROC-AUC and testing balanced accuracy out of 10 repeated 5-fold cross-validation experiments. We trained a total of 272 L2 regularization linear models. We selected the best based on the cross-validation error estimated produced by the algorithm on the training dataset.

To understand the biology associated with the predictive models, we performed pathway analyses following the approach described above. To add robustness to per-model pathway analyses, each transcript set was expanded to include transcripts significantly corelated ($p < 0.05$ and $r > 0.95$) to transcripts in each of the predictive models, respectively.

Assessment of AD risk factors

Brain amyloidosis is the biomarker of reference for AD. To assess if the predictive models were correlated with brain amyloidosis, along with other known AD risk factors, we used the Spearman correlation between the estimated risk provided by the classifier and the CSF levels of Aβ42, tau, p-tau. We only included those individuals with CSF measurements available within seven years before or after the draw date (n = 72). Additional to the correlation, we also used the CSF values to classify the participants using the ATN criteria, and then tested the performance of the models using the ATN criteria as outcome. We tested the performance of the cfRNA transcriptome models to differentiate between A positivity and AT positivity. No data was available for the N criteria for these samples.

Sensitivity to AD stages and specificity evaluation

To assess the AD continuum, we calculated the performance of the predictive models in the combined early symptomatic (CDR = 0.5) and symptomatic AD (CDR = 1) participants. We scaled the gene counts to the range of the training population by computing the *Z* score using the mean and standard deviation from the training population. Then, we calculated the risk score for each individual using the L2 regularization formula. Scores higher than 0.50 were considered cases. To calculate the ROC curve, we compared the predicted status to the true status for each group.

Due to the clinical and pathological overlap across neurodegenerative diseases, one of the challenges in the development of biomarkers for neurodegeneration is specificity to disease. To evaluate the performance of the predictive models in the context of other neurodegenerative diseases, we calculated the predictive risk value in 96 PD individuals, 16 DLB and 17 FTD and computed the ROC curves as described above. Additionally, we also calculated the ROC curve using AD samples instead of controls as the comparison group. Finally, we evaluated if the model showed any improvement when adding *APOE* genotype to the cfRNA predictor. *APOE* is the most important genetic risk factor. Thus, to understand if the effect of *APOE* was captured by the predictor, we included the *APOE* genotype in the model coded by two variables representing the number of ε2 alleles and ε4 alleles.