Title

Plasma Proteomics of Genetic Brain Arteriosclerosis and Dementia Syndrome Identifies Signatures of Fibrosis, Angiogenesis, and Metabolic Alterations

Authors

Jonah N. Keller[1,2], Hannah Radabaugh[3,4], Nikolaos Karvelas[2], Stephen Fitzsimons[2], Scott Treiman[3,4,5], Maria F. Palafox[2], Lisa McDonnell[2], Yakeel T. Quiroz[6,7], Francisco J. Lopera[7], Debarag Banerjee[8], Michael M. Wang[9], Joseph F. Arboleda-Velasquez[10, 11], James F. Meschia[12], Adam R. Ferguson[3,13], Fanny M. Elahi[2,14]*

Affiliations

[1] Department of Computational Biology, Cornell University, Ithaca, New York, USA.
[2] Departments of Neurology, Neuroscience, and Pathology, Icahn School of Medicine at Mount Sinai, New York, New York, USA.
[3] Department of Neurological Surgery, University of California, San Francisco, California, USA.
[4] Weill Institute for Neurosciences, University of California, San Francisco, San Francisco, California, USA.
[5] School of Medicine, University of California San Francisco, San Francisco, California, USA.
[6] Departments of Psychiatry and Neurology, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts, USA
[7] Grupo de Neurociencias de Antioquia, University of Antioquia, Medellín, Colombia
[8] Division of Artificial Intelligence, MachinAnimus, Los Altos Hills, California, USA
[9] Departments of Neurology and Physiology, University of Michigan and VA Ann Arbor Healthcare System, Ann Arbor, Michigan, USA.
[10] Schepens Eye Research Institute of Massachusetts Eye and Ear, Boston, Massachusetts, USA.
[11] Department of Ophthalmology, Harvard Medical School, Boston, Massachusetts, USA.
[12] Department of Neurology, Mayo Clinic, Jacksonville, Florida, USA.
[13] San Francisco Veterans Affairs Medical Center, San Francisco, San Francisco, California, USA.
[14] James J. Peters Department of Veterans Affairs Medical Center, Bronx, New York, USA.

* Corresponding Author. Email: fanny.elahi@mssm.edu

Abstract

Cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy (CADASIL) is the most common monogenic form of vascular cognitive impairment and dementia. A genetic arteriolosclerotic disease, the molecular mechanisms driving vascular brain degeneration and decline remain unclear. With the goal of driving discovery of disease-relevant biological perturbations in CADASIL, we used machine learning approaches to extract proteomic disease signatures from large-scale proteomics generated from plasma collected from three distinct cohorts in US and Colombia: CADASIL-Early (N = 53), CADASIL-Late (N = 45), and CADASIL-Colombia (N = 71). We extracted molecular signatures with high predictive value for early and late-stage CADASIL and performed robust cross- and external-validation. We examined the biological and clinical relevance of our findings through pathway enrichment analysis and testing of associations with clinical outcomes. Our study represents a model for unbiased discovery of molecular signatures and disease biomarkers, combining non-invasive

plasma proteomics with clinical data. We report on novel disease-associated molecular signatures for CADASIL, derived from the accessible plasma proteome, with relevance to vascular cognitive impairment and dementia.

Main Text

Introduction

Cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy (CADASIL) is an autosomal dominant form of VCID caused by missense mutations in *NOTCH3*. CADASIL is the leading cause of hereditary stroke and vascular cognitive impairment and dementia. Although the classical Mendelian syndrome is considered rare, with a prevalence of 1.3 - 4.1 per 100,000 adults *(1)*, variants in *NOTCH3* associated with endophenotype of white matter disease are more common, occurring in as many as 1 in 300 individuals *(2)*. Such prevalence suggests that discoveries of mechanisms underlying VCID in CADASIL could be relevant to vascular white matter disorders and dementia syndromes.

Research on the molecular pathogenesis of CADASIL has largely been limited to mice and postmortem human brain tissue. Consequently, our understanding of the early and evolving molecular pathogenesis of CADASIL, most relevant for development of impactful therapeutics, remains limited *(3)*. NOTCH3 is a transmembrane receptor that is highly expressed in mural cells. A key unresolved question involves understanding how mutations in the extracellular domain (ECD) of NOTCH3 lead to the dysfunction of small-caliber blood vessels *(4)* and multicellular vascular phenotypes such as neurovascular decoupling, hypoperfusion, and blood-brain barrier dysfunction *(5)*. In humans, neuroimaging abnormalities provide indications of affected vascular microenvironments *(6–8)*. As a chronic disease, molecular pathologies in CADASIL unfold over decades. Although NOTCH3 is predominantly expressed by brain vascular mural cells, which molecular dysfunctions drive clinical symptomatology and how to counter disease progression across various disease stages is not understood. A comprehensive and unbiased molecular investigation could capture key molecular drivers of brain dysfunction and degeneration in CADASIL, highlighting potential therapeutic opportunities.

In this study, we generated unbiased plasma proteomics and asked whether CADASIL has a specific disease signature in peripheral blood in early and late stages of disease and whether these signatures are associated with clinically relevant outcomes. We used multivariate analytical methods, such as machine learning (ML), to identify molecular signatures of disease. Leveraging publicly available brain single cell transcriptomics, we assigned proteomic signatures to cells within the neurovascular unit. Finally, using clinical outcomes of relevance to VCID, we demonstrated the clinical relevance of our findings, with implications for future novel biomarker development for risk stratification, prognostication, and disease monitoring across spectrum of disease severity from early to late stage of CADASIL.

Here, we investigated the plasma proteome of CADASIL using data generated from an aptamer-based assay that quantifies over 7,000 proteins (SomaSCAN 7k, Somalogic, Boulder, CO) *(9–12)*. Three distinct CADASIL cohorts with different disease stages were included in our analyses. Considering the age-associated nature of CADASIL, in which cognitive impairment and disability typically manifest after the age of 55 years (with notable patient-to-patient variability), we categorized our cohorts accordingly *(13, 14)*. The older cohort from the Mayo Clinic (*N* = 45; M_{Age} > 55 years) was labeled as CADASIL-Late, while the younger cohort recruited at UCSF (*N* = 53; M_{Age} < 55 years) was designated as CADASIL-Early. We employed the third cohort from Colombia, South America as a holdout validation dataset (CADASIL-Colombia; *N* = 71), owing to technical variations in its collection and processing. To identify proteomic signatures intrinsically linked to the disease, we developed a novel machine learning methodology. Our methodological workflow incorporates consensus aggregation of a suite of statistical evaluators coupled with rigorous cross-validation. Our overarching objectives were to: (1) isolate early- and late-stage CADASIL proteomic signatures; (2) validate these signatures both internally and in external CADASIL populations; (3) elucidate the biological implications of these identified proteins; and (4) correlate these protein signatures with relevant clinical and imaging metrics.

The goal of our study was to uncover disease-associated molecular signatures with a hypothesis-agnostic approach to allow for data-driven discovery of molecular perturbations and identification of novel therapeutic targets. To this end, we leveraged state-of-the-art computational methods to analyze plasma proteome data generated in three distinct CADASIL cohorts spanning early preclinical disease to more advanced clinical stages involving strokes. In addition, we used brain tissue to validate the expression of key proteins and associated molecular pathways in brain tissue donated from individuals with CADASIL.

Results

An overview of the study design is presented in **Figure 1**. The demographic characteristics of each cohort are described in **Table 1**. Designation of CADASIL-Early and CADASIL-Late monikers was based on the mean age of CADASIL participants and their symptom severity in respective cohorts.

Novel ML workflow detects CADASIL-associated proteomics signatures in peripheral blood

To elucidate the proteomic signatures of CADASIL, we developed a hypothesis-agnostic, multivariate analytical workflow, emphasizing robust statistical validation and biological and clinical interpretation of findings (**Fig. 2**). Our initial challenge pertained to reducing the dimensionality of over 7,000 proteins to more manageable protein lists. To achieve this, we implemented the leave-one-out (LOO) method, creating partitions (LOO folds) of our dataset. This method ensured minimal bias from individual samples and enhanced the generalizability of our model. In each LOO fold, only proteins with significant differences across groups ($P < 0.05$) progressed to subsequent analyses, resulting in approximately 1,300 proteins per fold. We then used a diverse array of feature selection algorithms, with different mathematical decision boundaries and solvers, carefully chosen to address the unique challenges posed by the high dimensionality and low sample size of our dataset. Next, we adopted a novel highly stringent, heterogeneous ensemble aggregation technique, which combines various algorithmic predictions to achieve a more accurate and reliable outcome. This method was crucial in ensuring that our protein selection was not biased towards any single machine learning algorithm, nor overfitted to any one sample. Proteins were only included in our final proteomic signature if they passed the highly conservative requirement of selection across all LOO folds by at least two different evaluators. When applied to each discovery cohort, these criteria led to the identification of two definitive protein signatures: the CADASIL-Early signature with 16 unique proteins and the CADASIL-Late signature with 20 unique proteins.

The following 16 proteins composed the CADASIL-Early proteomic signature: ANP32B, C4A|C4B, ENPP2, FN1, FUT3, GAS7, GPX1, HMBS, HPCAL1, MB, MGP, MYZAP, RRM1, SPINK6, UROS, and VEGFR3 (**Fig. 3A**). The following 20 proteins composed the CADASIL-Late proteomic signature: ABO, ACAA1, B3GAT1, BCAR3, C4A|C4B, CD209, DTNA, ECH1, ENDOU, FABP4, HNMT, KLRF1, KNG1, LTA4H, MASP1, OMG, SEMA3B, SLITRK1, TARDBP, and TMEM132B (**Fig. 3B**).

Machine learning model differentiates individuals affected with CADASIL from Healthy Controls using protein signatures

To further validate the uncovered protein signatures, we visualized associations with disease states using both unsupervised and supervised machine learning approaches. Principal component analysis (PCA), a non-supervised multivariate method, was performed on the selected set of proteins included in the Early and Late signatures (16 and 20 proteins, respectively). Visualization of the data on the principal component axes highlighted the elimination of non-disease-related signals (**Fig. 3C-D**), following the curation steps. We then trained supervised regularized linear discriminant analysis (rLDA) models using the protein sets as the input and disease status as the output. This approach resulted in ideal classification of disease instances (**Fig. 3E-F**). Rigorous cross-validation of the rLDA models was performed, which showcased their capacity for high accuracy and precision in distinguishing CADASIL from control samples, with statistical significance surpassing that of the permuted (i.e. random protein) models ($P < 0.00001$; **Fig. 3G-H**).

Internal and external validation of the CADASIL-Late protein signature for distinguishing CADASIL patients from Controls

To assess the generalizability and reproducibility of our protein signatures, we performed both an internal validation using the opposing cohort (i.e., Late vs Early) in addition to an external validation using an independently collected dataset shared by the Neuroscience Group of Antioquia (Colombia) with data generated using plasma samples collected from a Colombian cohort (CADASIL-Colombia; demographics in **Table 1**). In the internal validation label permutation testing, we found that the CADASIL-Early signature was noisy in discriminating CADASIL from control groups in the CADASIL-Late cohort ($P > 0.05$; **Fig. 4A**). However, the CADASIL-Late signature performed at highly significant levels when distinguishing between CADASIL and control subjects in the CADASIL-Early cohort ($P = 0.004$; **Fig. 4B**).

In our external validation, we trained supervised classifiers to distinguish disease status of the CADASIL-Colombia cohort using the Early and Late protein signatures as input. The CADASIL-Early plasma signature was marginally significant in discriminating in the CADASIL-Colombia dataset ($P = 0.055$; **Fig. 4C**). However, the CADASIL-Late plasma signature was significant in discriminating in the CADASIL-Colombia dataset ($P = 6.6 \times 10^{-3}$; **Fig. 4D**). To further substantiate these findings, we conducted permutation-based testing on the CADASIL-Late results and investigated separation in the high dimensional space by plotting ROC curves and LDA coordinates in biplots (**Fig 3E-H**). Label permutation testing confirmed significant discrimination between CADASIL and disease for both cohorts when provided protein level information for proteins in the CADASIL-Late signature (*P* = 0.049 for Late → Early; *P* = 0.017 for Late → Colombia; **Table S3**). The encouraging results of these tests provided further validation for our machine learning approach, indicating that our disease-associated protein set held reliable predictive capacity across diverse CADASIL populations.

Early and Late CADASIL proteomic signatures have both overlapping and distinct network components suggesting evolution and progression in patho-mechanisms of CADASIL across disease stages

In order to attain a better understanding of the molecular pathways, associated putative mechanisms, and interactions between proteins captured by our protein signatures, we used the web based STRING platform *(15)*. Both the Early and Late signatures served as input, and the resulting networks were overlapped to assess similarities and differences. The resulting network revealed 33 nodes and 55 edges, highlighting intricate relationships among proteins (**Fig. 5A**). Several proteins in the signature served as network hubs, including FN1 with 12 interactions, ITGB1 with 8, and SDC4, KNG1, and VEGFR3 each with 7 interactions. Notably, 7 interactions were found between proteins specifically associated with CADASIL-Early and CADASIL-Late, providing insights into potential transitional pathways and the multifactorial nature of CADASIL.

We then made use of Ingenuity Pathway Analysis (IPA) to predict upstream regulators of each protein signature, respectively. 17 proteins were found to be significant upstream regulators of the Early and Late networks (overlap *P* < 0.05). The overlapped regulator network showed notable interconnectedness between CADASIL-Early and CADASIL-Late signatures (**Fig. 5B**). The predicted activation Z-scores of all upstream regulators are presented in **Table S4**. TGF-β1, a protein with hypothesized involvement in CADASIL *(16–18)*, emerged as a key regulator protein. FN1, a hub protein in the STRING analyses, also emerged as a hub in the IPA network, sharing many upstream regulators with other CADASIL-Early and CADASIL-Late proteins. The upstream regulator analysis provides potential insights into therapeutic targets and yet to be explored adjacent pathways.

Specific investigation of signature proteins and their corresponding genes identified links to perturbations in metabolic, neuronal, cell differentiation, and inflammatory pathways

Noting unique components when contrasting the CADASIL-Early with CADASIL-Late signature networks, we then sought to identify specific perturbations linked to the enrichment of pathways or differential expression of proteins. We investigated these potentially pathological associations using both the EnrichR *(19)* platform and a Protein-centric Reverse GEO Search *(20)*. For the CADASIL-Early signature (**Fig. 5C**), several enriched pathways were noted. These included: heme biosynthesis ($P = 3.3 \times 10^{-5}$), specifically the porphyrin metabolism pathway ($P = 5.3 \times 10^{-4}$), glutathione metabolism ($P = 9.3 \times 10^{-4}$) and phosphodiesterase activity ($P = 4.0 \times 10^{-3}$), regulation of axonogenesis ($P = 1.2 \times 10^{-3}$), and Lewy body enrichment ($P = 4.0 \times 10^{-3}$). These findings suggest notable changes encompassing metabolic, oxidative stress, and neuronal dysfunction related pathologies. Further, the Early signature Reverse GEO Search (**Figure S1A**) highlighted FN1 as the top hit and linked this association to upregulation in a hepatocellular carcinoma model (P = 5.50 x 10^{-39}), cytokine-treated insulinoma cells (P = 1.29 x 10^{-28}) and a proliferating glioblastoma (P = 6.31 x 10^{-26}).

For the CADASIL-Late signature (**Fig. 5D**), we noted significant enrichment for complement pathways (P = 8.0 x 10^{-5}), such as the classical complement cascade (P = 1.4 x 10^{-3}), peroxisome pathways (P = 2.5 x 10^{-3}), including fatty acid metabolism (P = 3.1 x 10^{-3}), manganese binding (P = 1.0 x 10^{-3}), PPAR signaling (P = 2.5 x 10^{-3}), staphylococcal infection (P = 4.0 x 10^{-3}), and acetyl-CoA metabolism (P = 6.0 x 10^{-3}). The broad diversity of enriched pathways observed in this signature reflect key pathologies associated with neuronal dysfunction, chronic inflammation, and metabolic changes. Further, the Late signature Reverse GEO Search (**Figure S1B**) noted significant associations with the upregulation of LTA4H in intrahepatic cholangiocarcinoma (P = 2.04 x 10^{-30}) and in antifibrotic models (P = 8.32 x 10^{-27}) as well as elevated ACAA1 in cholangiocarcinoma (P = 1.70 x 10^{-21}).

Dynamic transcriptomic changes and cell-specific expression patterns uncovered in CADASIL proteomic signature

To further validate clinical relevance, we compared peripheral blood proteomic signatures against brain tissue bulk transcriptomics signatures, using transcriptomics data generated from 5 CADASIL and 7 control brains from the BA4/6 area of the human cortical region. From the Early signature, HPCAL1 and VEGFR3 showed upregulation in CADASIL patients ($\log_2$FC = 1.3 and $\log_2$FC = 0.7, respectively; **Fig. 6A**) while ENPP2, GAS7, and RRM1 were significantly downregulated ($\log_2$FC = -2.6, $\log_2$FC = -1.1, and $\log_2$FC = -0.8, respectively; **Fig. 6A**). From the Late signature, BCAR3 was significantly upregulated in CADASIL ($\log_2$FC = 1.0, **Fig. 6B**) whereas SEMA3B and OMG displayed significant downregulation in CADASIL brain tissue in comparison to control brain tissue ($\log_2$FC = -2.2 and $\log_2$FC = -1.3, respectively; **Fig. 6B**). These results reveal dynamic transcriptomic changes for signature proteins in CADASIL patients compared to controls. It should be noted that all brain tissue represents the end stage.

We then analyzed publicly available single-cell RNA-seq data from brain vascular cells to map neurovascular expression of the protein signatures *(21, 22)*. These data were generated *ex-vivo* from normal cerebral cortex from patients undergoing surgery for refractory epilepsy and cortical dysplasia. We made use of *ex-vivo* cell atlases, rather than postmortem atlases, due to notable changes in gene expression between living and postmortem brains *(23)*.For the Early signature, endothelial expression was prominent for VEGFR3, FN1, MGP, and HPCAL1 (**Fig 5C-D**). ENPP2 was expressed in oligodendrocyte lineage cells. For the CADASIL-Late signature, endothelial expression was noted for SEMA3B, TMEM132B, and ABO (**Fig 5E-F**). OMG and B3GAT1 showed oligodendrocyte expression. KLRF1 was expressed in T cells and BCAR3 in astrocytes. The remainder of the proteins were not significantly cell specific. These cell-specific expression patterns provide insight into which brain vascular cells display dysregulation of signature proteins in CADASIL progression.

CADASIL signature proteins exhibit strong associations with quantitative clinical measures and serve as predictors of disease-related traits

Lastly, we sought to better understand the clinical relevance of blood proteomic signatures obtained. We began this final investigation by testing the association of our protein signatures with characteristic MRI findings from CADASIL-Early patients. The assessed MRI metrics included white matter hyperintensities (WMH), a measure of white matter injury, enlarged perivascular space volume (ePVS), which are a more specific radiographic measure of small vessel disease*(24)*, and brain atrophy measured by the negative log of brain-parenchymal fraction (-log(BPF); with higher -log(BPF) corresponding to higher brain atrophy) (**Fig. 7A-B**). In the Early signature (**Fig. 7A**), ANP32B showed the strongest positive association with WMH (*R* = 0.52, *P* = 0.008) and ePVS (*R* = 0.46, *P* = 0.022). Meanwhile, GAS7 was associated negatively with ePVS (*R* = -0.46, *P* = 0.021) but positively with brain atrophy (*R* = -0.24, *P* > 0.05). For the Late signature (**Fig. 7B**), C4A|C4B was positively associated with measures of ePVS load (*R* = 0.38, *P* > 0.05) and DTNA was positively associated with ePVS (*R* = 0.43, *P* < 0.05). TMEM132B was positively associated with brain atrophy (*R* = -0.35, *P* > 0.05).

With regression analyses, we show associations between proteins in the signatures and measures of cognitive impairment in CADASIL-Early patients. The clinical outcomes include functional decline (Clinical Dementia Rating Score (CDR)) *(25, 26)*, cognitive processing time (Trail Making Test Part B Completion Time (TRAILB)), and global cognitive score (Montreal Cognitive Assessment score (MOCA)). Due to limited statistical power in clinical data, our regression analyses were underpowered to survive Bonferroni correction. Moreover, given the hypothesis-driven nature of these analyses, we did not think it pertinent to look at multiple comparisons corrected p-values. Instead, our investigation aimed to determine if there was consistency in the associations of proteins across MRI findings and cognitive measures. We sought to understand whether the values of proteins were consistently related across various measures, termed as "congruent" associations. For example, if a protein showed a positive association with increased brain atrophy, clinical congruence would suggest it also showed a positive association with increased functional decline and slower processing speed. Conversely, it would exhibit a negative association with global cognition, as higher global cognition indicates better cognitive outcomes. Thus, we defined a protein as clinically congruent for disease progression if it showed a positive association with brain atrophy, functional decline, and processing speed, and a negative association with global cognition. Conversely, proteins with a potentially disease-ameliorating or neuro-protective role in the neurovascular unit were expected to exhibit the opposite pattern: a negative association with brain atrophy, functional decline, and processing speed, and a positive association with global cognition. FN1, a protein also highlighted in the transcriptomic data, demonstrated clinical congruence with disease progression. FN1 was positively associated with brain atrophy (*R* = 0.42, *P* = 0.038; **Fig. 7C**), functional decline (*R* = 0.41, *P* = 0.045; **Fig. 7D**), and slowed processing speed (*R* = 0.33, *P* > 0.05; **Fig. 7F**). As well, FN1 had a strong negative association with global cognition (*R* = -0.24, *P* > 0.05; **Fig. 7E**). ENPP2, on the other hand, was clinically congruent with a possible disease-meliorating role. ENPP2 was negatively associated with brain atrophy (*R* = -0.39, *P* > 0.05; **Fig. 7G**), functional decline (*R* = -0.62, *P* = 0.001; **Fig. 7H**), and slowed processing time (*R* = -0.58, *P* = 0.006; **Fig. 7I**). Furthermore, ENPP2 was positively associated with global cognition (*R* = 0.69, *P* = 0.0005; **Fig. 7J**).

Other proteins showed associations with cognitive measures. For the CADASIL-Early signature (**Fig. 7A**), HPCAL1 was clinically congruent and negatively associated with functional decline (*R* = -0.41, *P* = 0.043) but positively associated with global cognition (*R* = 0.47, *P* = 0.030). HMBS exhibited a strong negative association with global cognition (*R* = -0.47, *P* = 0.029). For the CADASIL-Late signature (**Fig. 7B**), SLITRK1 showed a consistent negative association with cognition, including higher functional decline (*R* = -0.28, *P* > 0.05) and higher processing time (*R* = -0.18, *P* > 0.05). TMEM132B exhibited positive associations with global cognition (*R* = 0.46, *P* = 0.032).

Discussion

The goal of our study was to take an unbiased approach to discovery of blood molecular signatures of CADASIL, a monogenic form of VCID, as a critical first step toward development of biomarkers and formulation of mechanistic hypotheses for development of impactful treatments. To this end we formulated a novel experimental design and analytical approach aimed at extracting proteomic signatures in a methodologically unbiased manner, specifically tailored to monogenic diseases or disease states with clear categorization of disease and control groups. This research marks, to our knowledge, the first instance of unveiling a plasma proteomic signature associated with a monogenic form of VCID. Our findings shed light on molecular pathologies in CADASIL and serve as a stepping stone towards a broader investigation of VCID blood signatures.

We undertook an approach for capturing multivariate associations inherent in “-omics” datasets that traditional differential expression analyses reliant on univariate significance cannot capture *(27)*. Chowdhury et al. (2023) used a technically analogous approach to probe the proteomics of high-grade ovarian cancer, revealing a highly predictive, externally validated biomarker panel *(28)*. Our independently constructed methodology demonstrated a comparably high efficacy. By implementing a heterogenous bagging methodology, we amalgamated multiple algorithms, including recursive feature elimination, linear methods (logistic regression and rLDA), and non-linear methods (random forests and Markov blankets) *(29–33)*. This combination mitigates the risk of overfitting to a singular method’s decision boundary, especially given the disparity between our limited sample size and the exceedingly large feature count inherent in proteomic research *(34)*. Rigorous cross-validation, permutation testing, and validation through independent cohorts further curtailed the potential for overfitting *(35)*. Finally, we used brain tissue transcriptomics and clinical data to ensure that proteins identified by ML in the blood were relevant to the brain and clinical phenotypes of the disease. The advantage of this approach over univariate differential expression analysis encompasses enhanced adaptability in representing nonlinear dynamics and interactions, a reduced propensity for false positives owing to collinearity, a focus on disease categorization beyond mere association, inherent feature selection for pinpointing crucial subsets, and augmented resilience against batch effects through ensemble aggregation *(36, 37)*.

Applying this multi-step machine learning workflow resulted in the identification of robust protein signatures in both cohorts. The CADASIL-Early cohort yielded a concise 16-protein signature that primarily illuminated alterations in metabolic pathways. The CADASIL-Late cohort revealed a 20-protein signature characterized by chronic inflammation, immune alterations, and metabolic dysfunction. The CADASIL-Colombia cohort was utilized for external validation as an independently collected cohort. This validation cohort was characterized by large heterogeneity with regard to the age of participants and the age of the samples. Therefore, some technical noise was expected regarding plasma protein levels. Despite this limitation, we externally validated the discriminatory ability of the Late signal (*AUC* = 0.716, *P* < 0.005; **Table**

S3), whereas the AUC score of the Early signal was marginally significant (AUC = 0.610, P = 0.055; **Table S2**). This finding supports our hypothesis that robust proteomic signatures and molecular signals could be identified using our unbiased and multivariate analytical approaches.

Specific investigation of the proteins comprising each signature using a variety of methods, ranging from pathway enrichment analysis to single-cell RNA-seq, highlighted several additional noteworthy observations that may also serve as guides for the development of new therapeutic strategies. Perturbed pathways suggested by the CADASIL-Early signature emphasized disruptions in heme, porphyrin, and glutathione metabolic processes, aligning with pathways such as oxidative stress resistance via glutathione and biosynthesis of porphyrin-containing compounds. CADASIL compromises redox equilibrium and increases the production of reactive oxygen species *(38–40)*. This is consistent with the enrichment of the glutathione pathway, which is known for its role in neutralizing oxidative stress.

Interestingly, TDP-43 was identified in the CADASIL-Late signature (TARDP, a multimeric TDP-43 protein). The cytoplasmic mislocalization of TDP-43 and its aggregation are prominent pathologies in many neurodegenerative diseases (Frontotemporal Lobar Degeneration, Amyotrophic Lateral Sclerosis, and Alzheimer's Disease) *(41, 42)*. It is possible that TDP-43 related molecular dysregulations may serve as an explanatory factor for the observed associations between CADASIL and other neurodegenerative diseases (e.g., ALS, FTD) *(43–45)*. However, its specific role in CADASIL and its links to *NOTCH3* mutations have yet to be investigated. TDP-43 was found to be downregulated in plasma of CADASIL patients. It is not clear whether it is mis-localized. Similarly, TDP-43 levels were found to be decreased in the plasma of FTD patients *(46)*. The presence of TDP-43 in the CADASIL-Late signature suggests a potentially progressive pathology primed for therapeutic targeting, and thus warrants further investigation.

Other hallmark pathologies known to be associated with other neurodegenerative conditions were also highlighted, particularly in pathway analysis of the CADASIL-Early signature. Pathways associated with Lewy bodies, the defining characteristic of Lewy body dementia *(47)* and often observed in Parkinson's disease *(48)*, were found to be significantly enriched (P = 4.0 x 10^{-3}). This result is intriguing given prior findings that link *NOTCH3* mutations to worsened clinical outcomes in Parkinson's disease *(49)* and numerous reports observing parkinsonism in late CADASIL *(50)*. The finding of Lewy body pathways in CADASIL implicates a possible pathological link between CADASIL and Lewy-body-related neurodegenerative diseases.

TGF-β1 emerged as a primary predicted upstream regulator of both signatures. TGF-β1 signaling has consistently been found to play a functional role in the cerebrovascular system, including vascular senescence *(51–54)*, cerebral angiogenesis and maintenance of brain vessel homeostasis *(16)*. Furthermore, TGF-β1 and its receptors are known to play a key role in fibrosis across several diseases *(55, 56)*. Specific to the pathways affected in our signatures, earlier research has linked TGF-β1 upregulation to heightened oxidative stress and decreased glutathione levels in endothelial cells *(17)*. Additionally, latent TGF-β1 has been found to bind mutated NOTCH3 extracellular aggregates, suggesting its potential involvement in fibrogenesis *(18)*. The confluence of fibrosis-associated TGF-β1 upstream regulation with the proteomic signature of CADASIL provides biological validation, considering that fibrosis is a characteristic feature of

CADASIL histopathology *(57)*. Furthermore, Reverse GEO analyses reported connections with many cases of oncological disease, coupled with recent studies emerging from the oncology field to target the TGF-β pathway *(58)*, suggest potential analogous alterations in processes such as angiogenesis and present an intriguing path forward for further investigation.

In addition to their relevance in neurodegenerative and fibrotic pathways, the Early and Late signatures in our study also revealed associations with past proteomic analyses related to aging and cerebrovascular dysfunction. Oh et al. employed a machine learning-based approach to create a Feature Importance for Biological Aging score for human blood plasma proteins measured on the same SomaLogic platform as our study *(59)*. Their investigation into over 4,000 proteins pinpointed a distinct five protein signature of aging arteries, notably including MGP, a protein also identified in our Early signature. Significantly, MGP showed a substantial interaction with TAGLN in their analysis, a protein under the regulatory influence of TGF-β1 *(60)* and the most significant protein in the Oh et al. organismal aging model. The interaction of MGP and TAGLN, both prominently expressed in fibroblasts and endothelial cells, underscores their potential role in the advanced vascular aging process characteristic of CADASIL, and possibly in the general aging population. Similarly, Walker et al. found that plasma TAGLN is highly significant for an increased risk of dementia, further underscoring the relevance of MGP/TAGLN proteins in aging-related pathologies *(61)*. Intriguingly, the adipose aging signature of Oh et al., which demonstrated the third highest hazard ratio for organ-chronological age-gap, included FABP4, a protein of the Late signature. The co-presence of these signature proteins in our age-matched cohorts is particularly noteworthy. It may characterize CADASIL as an expedited model for studying cerebrovascular aging, extending beyond its primary neurodegenerative context. This aspect of our findings not only reinforces their robustness but also enhances the generalizability and applicability of our results in broader aging research.

We further validated our peripheral blood signature proteins against brain tissue transcriptomics and assigned proteins to specific cells using publicly available single nucleus RNA-seq data. Using the bulk transcriptomic data generated, we found dynamic transcriptomic changes. Additionally, using *ex-vivo* single-cell RNA-seq data from brain vascular cells *(21, 22)*, we observed cell-specific expression patterns consistent with pathological alterations in the cerebrovasculature. Endothelial cells and fibroblasts appeared to be prominently affected in the Early signature, while the involvement of astrocytes, microglia, and oligodendrocytes was noted in the Late signature, consistent with the advancement of disease from brain borders into parenchyma as disease progresses. This suggests that therapeutic targets may differ based on the disease stage. Both signatures showed marked expression of numerous proteins (FN1, MGP, GAS7, DTNA, and TMEM132B) in fibroblasts, perhaps further reflecting fibrosis, collagen protein alterations, and extracellular matrix changes identified by pathway analyses. Numerous of these molecular processes and pathways have been implicated in CADASIL, including collagen/ECM involvement *(4, 62–64)* and fibrosis *(17, 40)*.

Lastly, to examine the clinical relevance of our molecular findings we tested associations between identified proteomic signatures and disease-associated quantitative clinical phenotypes. For this we limited our analysis to CADASIL-Early patient data. We found strong associations with clinical congruence across several metrics. Several proteins from the CADASIL-Early signature were found to be associated with neuroimaging findings. Fibronectin, a protein previously known to be enriched in CADASIL vessels and shown to increase levels in blood vessels of CADASIL patients *(18)*, was upregulated and demonstrated clinical congruence with disease progression in the CADASIL-Early cohort. While FN1 plays a critical role in vascular remodeling after hypoxia- or hypertension-induced vessel injury *(65, 66)*, its aggregation in tissues may be detrimental by promotion of thrombogenesis *(67, 68)*, neuroinflammation *(69)*, and arrest of myelination *(70)*, all thought to be components of chronic cSVD and stroke-related VCID. Upregulated CADASIL FN1 levels were positively associated with increased cerebral atrophy and cognitive deterioration in our Early cohort, emerging as a critical, possibly neurodegenerative factor in CADASIL, which warrants exploration in future studies. Conversely, ENPP2 (autotaxin) displayed clinical congruence with a disease-meliorating character, as it was inversely associated with brain atrophy and cognitive decline measures. ENPP2 was significantly downregulated in both plasma and brain transcriptomics of CADASIL patients. ENPP2 is an abundantly expressed member of the ectonucleotide pyrophosphatase/phosphodiesterase family with lysophospholipase activity, catalyzing lysophosphatidic acid (LPA) formation *(71)*. Despite mounting evidence for an excitotoxic potential in acute stroke *(72)*, other studies have noted potentially beneficial effects in oligodendrocyte maturation *(73)*, protection of endothelial cells from hypoxia *(74)* and suppression of CD8+ T cell infiltration in tumors *(75)*. These results suggest that ENP22 might confer protection in chronic, but not acute injury, and thus its depletion could reflect long-term changes in CADASIL rather than the effects of strokes.

The elucidation of distinct Early and Late signatures that discriminate CADASIL from controls and are associated with clinical outcome metrics represents an advancement in the development of molecular approaches to advance precision diagnostics across disease stages. We compensated for small sample sizes by using novel analytical methods in addition to cross-validation of results in independent cohorts and brain tissue. The methods for integration of advanced computational techniques and independent validation cohorts translate cross-sectional discoveries into robust and generalizable targets. These findings lay the groundwork for the elucidation of CADASIL pathogenesis and the subsequent identification of disease-stage-specific biomarkers and disease-stage agnostic or specific therapeutic targets.

To our knowledge, this is the first study to implement a multi-step machine learning approach to blood proteomics data to uncover molecular signatures as an unbiased starting point for understanding evolving molecular dysregulation across disease stages in CADASIL. The application of multidimensional analytical techniques enabled us to capture interactions between proteins and patterns shared among multiple analytes in a single analysis, rather than limiting our investigation to specific hypothesis testing in instances of univariate significance. This approach can provide a more comprehensive view of the dataset than most proteomic studies currently in the literature by directly interrogating interactions in a high-dimensional space. The broader adoption of this approach or similar multidimensional pipelines may provide the opportunity to gain a more complete understanding of "-omics" datasets and guide future research towards novel therapeutic targets.

Materials & Methods

Study Design - Overview of cohorts

Our study employed plasma proteomics data and clinical information from diverse cohorts, described below. Information regarding demographics (age, sex) and neurological function were provided. Study protocols were approved by their respective Institutional Review Boards. Research was performed in accordance with the Code of Ethics of the World Medical Association. Written informed consent was obtained from all patients before data collection.

Early CADASIL cohort (*N* = 53)

CADASIL patients (*N* = 25) were consecutively recruited, so as to avoid bias, at the Memory and Aging Center, UCSF, between February 25, 2019, to August 2, 2021. Patients were evaluated by neuropsychological testing, subjected to a blood draw (for plasma collection), and, except for one, underwent MRI neuroimaging. Neurocognitive testing included measures of functional decline (Global Clinical Dementia Rating - CDRTot, and Sum of Boxes - CDRBox), global cognition (Montreal Cognitive Assessment - MOCA), and processing speed (Modified Trail Making Test completion time - MTTime, Trail Making Test B completion time - TRAILB). MRI-derived measurements included white matter hyperintensity (WMH) volume, enlarged perivascular space (ePVS) volume (measured by LOAD), and Brain Parenchymal Fraction (BPF), which were quantified according to a previously described image processing pipeline *(76)*. CADASIL status was confirmed based on *NOTCH3* sequencing. The inclusion criteria for all control subjects (*N* = 28) were intact daily functioning per an informant (Clinical Dementia Rating = 0), neuropsychological performances within normative standards, and absence of significant clinical neurological disease assessed by history and physical exam. Control subjects underwent blood collection but no neuropsychological evaluation or medical imaging.

Mayo Clinic CADASIL cohort (*N* = 45)

CADASIL patients (*N* = 20) were recruited at the Department of Neurology, Mayo Clinic, Jacksonville between July 29, 2014, to February 2, 2021. Patients underwent clinical evaluation and blood draw. CADASIL status was confirmed based on *NOTCH3* mutations discovered via genetic sequencing. Control subjects (*N* = 25) were selected based on absence of significant clinical neurological disease assessed by history and physical exam.

Colombia CADASIL cohort (*N* = 66)

Patients with CADASIL from Colombia (N = 25) and controls with no *NOTCH3* mutations (*N* = 41) who were family members of the CADASIL patients, were recruited from a cohort in Colombia during two periods, from August 3, 2000, to July 14, 2005 and from January 12, 2015 to December 11, 2016. A subset of participants (10 CADASIL, 20 control participants) was longitudinally evaluated during both periods and constituted the longitudinal cohort for this study. Sequencing the *NOTCH3* gene confirmed CADASIL status.

Plasma Collection and Proteomic Analysis

For proteomics characterization, plasma samples from all cohorts were analyzed through the SOMAscan 7k assay (SomaLogic, Inc., Boulder, CO). The SomaScan assay offers the advantage of unbiased protein expression analysis of a wide range of proteins, covering all biological functional domains. As described previously, the SOMAscan assay kit employs highly selective single-stranded modified Slow Off-rate Modified DNA Aptamers (SOMAmer) for protein identification and quantification. A custom DNA microarray (Agilent) was used for quantification, which is reported as relative fluorescence units (RFU). Raw data then underwent quality control, calibration, and normalization. Prior to data analysis, we performed sample pre-processing. All non-human SOMAmers (307 proteins) were removed from the dataset leaving 7,289 proteins of the 7,596 proteins measured by SomaLogic. The data was then transformed by a natural log.

Machine Learning Pipeline

Recursive Feature Elimination, Leave-One-Out-Cross-Validation: Definition and Analysis

In our study, we aimed to uncover CADASIL disease-associated changes in proteomic networks using comprehensive and unbiased machine learning (ML) techniques. Our approach's central basis was that protein sets which are crucial in distinguishing disease states may be key biological drivers of the disease *(32)*. We developed a novel ML methodology that employs auxiliary Markov blanket feature selection *(77, 78)* combined with multiple recursive feature selection algorithms to mitigate bias towards any specific algorithm *(79)* and reduce overfitting, which is the fundamental challenge considering the inherent low sample size and high dimensionality of our, and many others, proteomics datasets. The first step of our method was the creation of Leave-One-Out (LOO) partitions of our data *(35)*. For a dataset with N samples, we generated N partitions (LOO folds), each excluding one unique sample while including the rest. This approach ensured that the influence of any single sample is minimized in the model, addressing overfitting, and improving the model's generalizability.

Before feature selection, every protein in each LOO fold was subjected to a univariate t-test at an alpha level of 0.05. This step is vital to reduce the risk of including proteins that show apparent but spurious associations with the disease due to random variation, a common issue in datasets where $n \ll p$. For each fold, ~1,300 proteins survived the filtering step and proceeded to feature selection.

For feature selection, we utilized a diverse array of algorithms, some employing Recursive Feature Elimination (RFE) and others independent of it. RFE is a technique that systematically removes the least significant features to identify the most relevant ones. In the context of our high-dimensional data, RFE helps in reducing the feature space, making the model more robust and less prone to overfitting. The RFE algorithm was coupled with repeated 10-fold cross-validation during each elimination step to minimize variance in selection. The suite of algorithms employed included RFE with Logistic Regression (LR) with L1 and L2 regularization penalties, respectively *(30, 31)*, RFE with regularized Linear Discriminant Analysis (rLDA) *(80)*, RFE with Random Forests (RF) *(29)*, Boruta - Random Forests *(81)*, and Maximum-Relevance-Minimum-Redundancy (MRMR) with an F-Statistic evaluator *(82)*. Markov blanket feature selection was employed separately on the original datasets, due to computational expense and subsequently incorporated during the later aggregation steps *(77, 78)*.

Each algorithm was utilized to address specific challenges in the $n \ll p$ problem, where variables significantly outnumber observations. RFE-LR with L1 regularization was employed for its sparsity-inducing property, efficiently eliminating less significant features in well-defined classes

(31). RFE-LR with L2 regularization was used to manage multicollinearity, shrinking coefficients without excluding any features, thus preserving the contributions of all variables even in the presence of high inter-correlations *(83)*. Regularized-LDA (rLDA) was selected for its enhanced accuracy in settings with normally distributed within-class proteins and small sample sizes, addressing the instability issues of logistic regression in scenarios with significant class separation *(80)*. Additionally, the Boruta-RF method was integrated to effectively identify crucial features in high-dimensional datasets by comparing real features against randomly generated shadow features, optimizing feature selection under the n <<< p constraint. The Markov Blanket approach was effective for focusing on complex, non-linear variables most relevant to the target, thus efficiently reducing dimensionality *(77)*. Finally, the MRMR method was used for balancing feature relevance and minimizing redundancy, crucial for predictive accuracy in datasets with numerous features. This comprehensive approach demonstrates a nuanced handling of feature selection in complex, high-dimensional datasets.

Each aforementioned algorithm was applied to all *N* LOO folds, producing *N* sets of proteins for each algorithm. The final step in our methodology was a consensus aggregated feature selection approach to combine these results. First, we identified a ‘bag of features’ for each algorithm, selecting only those proteins that appeared in every set produced by that algorithm across all LOO partitions. Next, we cross-referenced these bags of features across different algorithms, further minimizing bias towards any single ML approach. A protein was included in our final proteomic signature only if it appeared in the ‘bag of features’ sets of at least two different models, often appearing in several. This method ensures a consensus among different algorithms, further reducing the risk of model overfitting and bias towards any single ML approach and providing a more reliable indicator of disease-associated proteomic changes.

Subsequently, we performed Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) on this protein set for visualization of disease discrimination. PCA was executed to display the remaining variance in the data, emphasizing that it is based on the disease state *(84)*. In contrast, LDA was utilized to visually demonstrate that the protein set harbors information crucial for delineating and clustering the disease. Lastly, we performed permutation analysis to evaluate the significance of our findings. We randomly selected a protein signature of the same length from the dataset and evaluated its performance. Similarly, we permuted the class labels and evaluated the resulting classifier performance to ensure the signature was disease specific.

Cross-Validation Between Cohorts and External Validation with Independent Cohort

Following the determination of our protein signatures, we proceeded to evaluate their capacity to classify diseases. This exploration was underpinned by the hypothesis that proteins integral to successful disease classification models could potentially serve as disease-associated markers. To this end, we scrutinized the efficacy of an array of machine learning algorithms, inclusive of Linear Support Vector Machines (SVM), Random Forests (RF), Regularized Linear Discriminant Analysis (rLDA), Logistic Regression (LR), Ridge Classifier, Perceptron and Decision Trees. To ensure robustness and applicability of our findings, we sought to validate our proteomic signature specific to CADASIL. This validation was undertaken by refitting the Early disease signature to the Late disease cohort and vice versa. As well, an external CADASIL dataset (CADASIL-Colombia) was tested. This cross-validation and external validation served to substantiate the generalizability of our CADASIL proteomic signatures across different patient cohorts.

Programming Languages and Packages Used

Data processing was handled in the R environment version 4.2.1 (R Foundation for Statistical Computing, 2022). The more complex machine learning data analyses, including Boruta-Random Forests and MRMR feature selection, were performed using Jupyter Notebooks, leveraging pertinent libraries such as SciKit-Learn for machine learning algorithms and MRMR_Selection for MRMR analyses. SciKit-Learn was also the tool of choice for performing all correlation calculations. For data organization and basic numerical operations, we employed the Pandas and Numpy libraries respectively. Finally, visualizations pertaining to Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) were generated using Matplotlib.

Pathway and Gene Ontology Analysis

Biological function analysis and pathway analyses via the Gene Ontology, KEGG, Panther, and the Reactome databases were performed through the use of an Appyter-based version of the EnrichR web tool *(19)*. Similar EnrichR terms from separate databases were combined under one term and depicted as overlapping plots. We used Ingenuity® Pathway Analysis (QIAGEN, Redwood City) software for functional pathway and upstream regulatory analysis (URA) of the proteins–of-interest identified in this study *(85)*. For the above we set the significance level threshold for the Benjamin-Hochberg adjusted p-value at 0.05. We further explored proteins of interest in the STRING database version 12.0 for protein-protein functional and physical interaction analysis, the results of which were displayed as a functional network. Interactions were considered with a medium confidence score of 0.4 or higher. We used the Reverse GEO Search Appyter to investigate the disease relevance of the protein signature *(20)*. We searched the results for each protein in the signature and plotted the top 25 terms based on the negative log of the p-value for upregulated and downregulated signatures respectively.

Human Brain Tissue and Bulk Transcriptome Sequencing

We performed RNA sequencing (RNA-seq) analysis on postmortem frontal cortex samples from CADASIL patients (*N* = 5, provided by the Wang Laboratory, University of Michigan) and controls (*N* = 7, provided by the Mount Sinai Neuropathology Brain Bank). In short, deep-frozen samples from the BA4/6 area of the human cortex were micro-dissected. RNA extraction, preparation of cDNA libraries and transcriptome sequencing using the Illumina NovaSeq (paired-end 150-nucleotide read length) was conducted by Novogene Co., LTD (Beijing, China). All samples were assessed to have an RNA integrity number (RIN) above 3.5, so as to ensure a sufficient relative abundance of full-length transcripts. Reads of low quality or containing adapter sequences were filtered. Raw fastq files were analyzed through an in-house transcriptomics pipeline, RAPiD-nf, implemented in the NextFlow framework. Briefly, remaining adapter sequences were filtered by Trimmomatic v0.36 *(86)*, STAR v2.7.1 *(87)* was used to align to the hg38 build of the human reference genome (GRCh38), and featureCounts performed BAM-level quantification *(88)*. Results were subjected to quality control using FastQC and Picard. For Differential Expression Analysis we employed the DESeq2 package in R *(89)*. Before analysis, sequences with sum of sequence counts < 10 across all participants were removed. We calculated the average log2 fold change for each protein between CADASIL and control groups and then plotted the results in a bar plot to visualize transcriptional alterations of the signature proteins.

Investigation of Brain Vasculature Expression of CADASIL Plasma Signature

To elucidate cell type–specific dysregulation of the CADASIL signature proteins, we analyzed published single-cell RNA-seq data from Winkler et al. and Garcia et al. Heatmaps were generated to visualize the signed log2 fold changes of signature proteins across different brain vascular cell types.

Statistical Analysis

Mean demographics, imaging, and cognitive test measures were compared between cohorts with Student's t-test for continuous variables and Chi-square tests for categorical variables. Average signed log2 fold change and Student's t-test was calculated for old versus new time point CADASIL and control samples in longitudinal Colombia cohort. Linear regression models, with proteins derived from the ML pipeline as predictors, were used to test for associations with imaging and clinical variables. As well, we developed multiple linear regression models with backward feature selection of the protein set to generate feature subsets to predict clinical variables. All statistical analyses were performed using Numpy, Scipy, and Pandas libraries in Python. All statistical tests were unpaired. A 2-sided p-value ≤ 0.05 was considered statistically significant, and a p-value < 0.10 but greater than 0.05 marginally significant.

References

1. Y. Yamamoto, Y.-C. Liao, Y.-C. Lee, M. Ihara, J. C. Choi, Update on the Epidemiology, Pathogenesis, and Biomarkers of Cerebral Autosomal Dominant Arteriopathy With Subcortical Infarcts and Leukoencephalopathy. *Journal of Clinical Neurology* **19**, 12–27 (2023).

2. J. W. Rutten, H. G. Dauwerse, G. Gravesteijn, M. J. van Belzen, J. van der Grond, J. M. Polke, M. Bernal-Quiros, S. A. J. Lesnik Oberstein, Archetypal NOTCH3 mutations frequent in public exome: implications for CADASIL. *Ann Clin Transl Neurol* **3**, 844–853 (2016).

3. M. Locatelli, A. Padovani, A. Pezzini, Pathophysiological Mechanisms and Potential Therapeutic Targets in Cerebral Autosomal Dominant Arteriopathy With Subcortical Infarcts and Leukoencephalopathy (CADASIL). *Front Pharmacol* **11**, 321 (2020).

4. M. Monet-Leprêtre, I. Haddad, C. Baron-Menguy, M. Fouillot-Panchal, M. Riani, V. Domenga-Denier, C. Dussaule, E. Cognat, J. Vinh, A. Joutel, Abnormal recruitment of extracellular matrix proteins by excess Notch3 ECD: a new pathomechanism in CADASIL. *Brain* **136**, 1830–1845 (2013).

5. J. Kelleher, A. Dickinson, S. Cain, Y. Hu, N. Bates, A. Harvey, J. Ren, W. Zhang, F. C. Moreton, K. W. Muir, C. Ward, R. M. Touyz, P. Sharma, Q. Xu, S. J. Kimber, T. Wang, Patient-Specific iPSC Model of a Genetic Vascular Dementia Syndrome Reveals Failure of Mural Cells to Stabilize Capillary Structures. *Stem Cell Reports* **13**, 817–831 (2019).

6. R. van den Boom, S. A. J. Lesnik Oberstein, M. D. Ferrari, J. Haan, M. A. van Buchem,

Cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy: MR imaging findings at different ages--3rd-6th decades. *Radiology* **229**, 683–690 (2003).

7. M. K. Liem, S. A. J. Lesnik Oberstein, J. Haan, I. L. van der Neut, R. van den Boom, M. D. Ferrari, M. A. van Buchem, J. van der Grond, Cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy: progression of MR abnormalities in prospective 7-year follow-up study. *Radiology* **249**, 964–971 (2008).

8. E. Jouvent, A. Viswanathan, J.-F. Mangin, M. O'Sullivan, J.-P. Guichard, A. Gschwendtner, R. Cumurciuc, F. Buffon, N. Peters, C. Pachaï, M.-G. Bousser, M. Dichgans, H. Chabriat, Brain atrophy is related to lacunar lesions and tissue microstructural changes in CADASIL. *Stroke* **38**, 1786–1790 (2007).

9. L. Gold, D. Ayers, J. Bertino, C. Bock, A. Bock, E. N. Brody, J. Carter, A. B. Dalby, B. E. Eaton, T. Fitzwater, D. Flather, A. Forbes, T. Foreman, C. Fowler, B. Gawande, M. Goss, M. Gunn, S. Gupta, D. Halladay, J. Heil, J. Heilig, B. Hicke, G. Husar, N. Janjic, T. Jarvis, S. Jennings, E. Katilius, T. R. Keeney, N. Kim, T. H. Koch, S. Kraemer, L. Kroiss, N. Le, D. Levine, W. Lindsey, B. Lollo, W. Mayfield, M. Mehan, R. Mehler, S. K. Nelson, M. Nelson, D. Nieuwlandt, M. Nikrad, U. Ochsner, R. M. Ostroff, M. Otis, T. Parker, S. Pietrasiewicz, D. I. Resnicow, J. Rohloff, G. Sanders, S. Sattin, D. Schneider, B. Singer, M. Stanton, A. Sterkel, A. Stewart, S. Stratford, J. D. Vaught, M. Vrkljan, J. J. Walker, M. Watrobka, S. Waugh, A. Weiss, S. K. Wilcox, A. Wolfson, S. K. Wolk, C. Zhang, D. Zichi, Aptamer-based multiplexed proteomic technology for biomarker discovery. *PLoS One* **5**, e15004 (2010).

10. J. C. Rohloff, A. D. Gelinas, T. C. Jarvis, U. A. Ochsner, D. J. Schneider, L. Gold, N. Janjic, Nucleic Acid Ligands With Protein-like Side Chains: Modified Aptamers and Their Use as Diagnostic and Therapeutic Agents. *Mol Ther Nucleic Acids* **3**, e201 (2014).

11. C. H. Kim, S. S. Tworoger, M. J. Stampfer, S. T. Dillon, X. Gu, S. J. Sawyer, A. T. Chan, T. A. Libermann, A. H. Eliassen, Stability and reproducibility of proteomic profiles measured with an aptamer-based platform. *Sci Rep* **8**, 8382 (2018).

12. J. Candia, F. Cheung, Y. Kotliarov, G. Fantoni, B. Sellers, T. Griesman, J. Huang, S. Stuccio, A. Zingone, B. M. Ryan, J. S. Tsang, A. Biancotto, Assessment of Variability in the SOMAscan Assay. *Sci Rep* **7**, 14248 (2017).

13. S. Brice, A. Jabouley, S. Reyes, C. Machado, C. Rogan, N. Dias-Gastellier, H. Chabriat, S. T. du Montcel, Modeling the Cognitive Trajectory in CADASIL. *J Alzheimers Dis* **77**, 291–300 (2020).

14. S. Brice, S. Reyes, A. Jabouley, C. Machado, C. Rogan, N. Gastellier, N. Alili, S. Guey, E. Jouvent, D. Hervé, S. Tezenas du Montcel, H. Chabriat, Trajectory Pattern of Cognitive Decline in Cerebral Autosomal Dominant Arteriopathy With Subcortical Infarcts and Leukoencephalopathy. *Neurology* **99**, e1019–e1031 (2022).

15. D. Szklarczyk, A. L. Gable, K. C. Nastou, D. Lyon, R. Kirsch, S. Pyysalo, N. T. Doncheva, M. Legeay, T. Fang, P. Bork, L. J. Jensen, C. von Mering, The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res* **49**, D605–D612 (2021).

16. G. Ferrari, B. D. Cook, V. Terushkin, G. Pintucci, P. Mignatti, Transforming growth factor-beta 1 (TGF-beta1) induces angiogenesis through vascular endothelial growth factor (VEGF)-mediated apoptosis. *J Cell Physiol* **219**, 449–458 (2009).

17. I. Montorfano, A. Becerra, R. Cerro, C. Echeverría, E. Sáez, M. G. Morales, R. Fernández, C. Cabello-Verrugio, F. Simon, Oxidative stress mediates the conversion of endothelial cells into myofibroblasts via a TGF-β1 and TGF-β2-dependent pathway. *Lab Invest* **94**, 1068–1082 (2014).

18. J. Kast, P. Hanecker, N. Beaufort, A. Giese, A. Joutel, M. Dichgans, C. Opherk, C. Haffner, Sequestration of latent TGF-β binding protein 1 into CADASIL-related Notch3-ECD deposits. *Acta Neuropathologica Communications* **2**, 96 (2014).

19. E. Y. Chen, C. M. Tan, Y. Kou, Q. Duan, Z. Wang, G. V. Meirelles, N. R. Clark, A. Ma'ayan, Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* **14**, 128 (2013).

20. D. J. B. Clarke, M. Jeon, D. J. Stein, N. Moiseyev, E. Kropiwnicki, C. Dai, Z. Xie, M. L. Wojciechowicz, S. Litz, J. Hom, J. E. Evangelista, L. Goldman, S. Zhang, C. Yoon, T. Ahamed, S. Bhuiyan, M. Cheng, J. Karam, K. M. Jagodnik, I. Shu, A. Lachmann, S. Ayling, S. L. Jenkins, A. Ma'ayan, Appyters: Turning Jupyter Notebooks into data-driven web apps. *Patterns (N Y)* **2**, 100213 (2021).

21. F. J. Garcia, N. Sun, H. Lee, B. Godlewski, H. Mathys, K. Galani, B. Zhou, X. Jiang, A. P. Ng, J. Mantero, L.-H. Tsai, D. A. Bennett, M. Sahin, M. Kellis, M. Heiman, Single-cell dissection of the human brain vasculature. *Nature* **603**, 893–899 (2022).

22. E. A. Winkler, C. N. Kim, J. M. Ross, J. H. Garcia, E. Gil, I. Oh, L. Q. Chen, D. Wu, J. S. Catapano, K. Raygor, K. Narsinh, H. Kim, S. Weinsheimer, D. L. Cooke, B. P. Walcott, M. T. Lawton, N. Gupta, B. V. Zlokovic, E. F. Chang, A. A. Abla, D. A. Lim, T. J. Nowakowski, A single-cell atlas of the normal and malformed human brain vasculature. *Science* **375**, eabi7377 (2022).

23. L. E. Liharska, Y. J. Park, K. Ziafat, L. Wilkins, H. Silk, L. M. Linares, R. C. Thompson, E. Vornholt, B. Sullivan, V. Cohen, P. Kota, C. Feng, E. Cheng, J. S. Johnson, M.-K. Rieder, J. Huang, J. Scarpa, J. Polanco, E. Moya, A. Hashemi, J. Bendl, G. E. Hoffman, P. Roussos, M. A. Levin, G. N. Nadkarni, R. Sebra, J. Crary, P. Sklar, E. E. Schadt, N. D. Beckmann, B. H. Kopell, A. W. Charney, A study of gene expression in the living human brain, 2023.04.21.23288916 (2023).

24. J. Gutierrez, M. S. V. Elkind, C. Dong, M. Di Tullio, T. Rundek, R. L. Sacco, C. B. Wright, Brain Perivascular Spaces as Biomarkers of Vascular Risk: Results from the Northern Manhattan Study. *AJNR Am J Neuroradiol* **38**, 862–867 (2017).

25. J. C. Morris, C. Ernesto, K. Schafer, M. Coats, S. Leon, M. Sano, L. J. Thal, P. Woodbury, Clinical dementia rating training and reliability in multicenter studies: the Alzheimer's Disease Cooperative Study experience. *Neurology* **48**, 1508–1510 (1997).

26. S. E. O'Bryant, S. C. Waring, C. M. Cullum, J. Hall, L. Lacritz, P. J. Massman, P. J. Lupo, J.

S. Reisch, R. Doody, Texas Alzheimer's Research Consortium, Staging Dementia Using Clinical Dementia Rating Scale Sum of Boxes Scores: A Texas Alzheimer's Research Consortium Study. *Archives of Neurology* **65**, 1091–1095 (2008).

27. N. Perakakis, A. Yazdani, G. E. Karniadakis, C. Mantzoros, Omics, big data and machine learning as tools to propel understanding of biological mechanisms and to discover novel diagnostics and therapeutics. *Metabolism* **87**, A1–A9 (2018).

28. S. Chowdhury, J. J. Kennedy, R. G. Ivey, O. D. Murillo, N. Hosseini, X. Song, F. Petralia, A. Calinawan, S. R. Savage, A. B. Berry, B. Reva, U. Ozbek, A. Krek, W. Ma, F. da Veiga Leprevost, J. Ji, S. Yoo, C. Lin, U. J. Voytovich, Y. Huang, S.-H. Lee, L. Bergan, T. D. Lorentzen, M. Mesri, H. Rodriguez, A. N. Hoofnagle, Z. T. Herbert, A. I. Nesvizhskii, B. Zhang, J. R. Whiteaker, D. Fenyo, W. McKerrow, J. Wang, S. C. Schürer, V. Stathias, X. S. Chen, M. H. Barcellos-Hoff, T. K. Starr, B. J. Winterhoff, A. C. Nelson, S. C. Mok, S. H. Kaufmann, C. Drescher, M. Cieslik, P. Wang, M. J. Birrer, A. G. Paulovich, Proteogenomic analysis of chemo-refractory high-grade serous ovarian cancer. *Cell* **186**, 3476-3498.e35 (2023).

29. L. Breiman, Random Forests. *Machine Learning* **45**, 5–32 (2001).

30. J. Berkson, Application of the Logistic Function to Bio-Assay. *Journal of the American Statistical Association* **39**, 357–365 (1944).

31. I. Guyon, A. Elisseeff, An introduction to variable and feature selection*The Journal of Machine Learning Research* (2003) (available at https://dl.acm.org/doi/10.5555/944919.944968).

32. T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, E. S. Lander, Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**, 531–537 (1999).

33. D. Koller, M. Sahami, in (1996), pp. 284–292.

34. C. Strobl, J. Malley, G. Tutz, An Introduction to Recursive Partitioning: Rationale, Application and Characteristics of Classification and Regression Trees, Bagging and Random Forests. *Psychol Methods* **14**, 323–348 (2009).

35. R. Kohavi, in *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence (II)*, (1995).

36. S. Ng, S. Masarone, D. Watson, M. R. Barnes, The benefits and pitfalls of machine learning for biomarker discovery. *Cell Tissue Res* (2023), doi:10.1007/s00441-023-03816-z.

37. J. Niu, J. Yang, Y. Guo, K. Qian, Q. Wang, Joint deep learning for batch effect removal and classification toward MALDI MS based metabolomics. *BMC Bioinformatics* **23**, 270 (2022).

38. J. Campolo, R. De Maria, C. Mariotti, C. Tomasello, M. Parolini, M. Frontali, D. Inzitari, R. Valenti, A. Federico, F. Taroni, O. Parodi, Is the Oxidant/Antioxidant Status Altered in CADASIL Patients? *PLoS One* **8**, e67077 (2013).

39. K. B. Neves, A. P. Harvey, F. Moreton, A. C. Montezano, F. J. Rios, R. Alves-Lopes, A.

Nguyen Dinh Cat, P. Rocchicciolli, C. Delles, A. Joutel, K. Muir, R. M. Touyz, ER stress and Rho kinase activation underlie the vasculopathy of CADASIL. *JCI Insight* **4**, 131344 (2019).

40. P. Formichi, E. Radi, C. Battisti, G. Di Maio, E. Tarquini, A. Leonini, A. Di Stefano, M. T. Dotti, A. Federico, Apoptosis in CADASIL: an in vitro study of lymphocytes and fibroblasts from a cohort of Italian patients. *J Cell Physiol* **219**, 494–502 (2009).

41. P. Steinacker, P. Barschke, M. Otto, Biomarkers for diseases with TDP-43 pathology. *Mol Cell Neurosci* **97**, 43–59 (2019).

42. M. Jo, S. Lee, Y.-M. Jeon, S. Kim, Y. Kwon, H.-J. Kim, The role of TDP-43 propagation in neurodegenerative diseases: integrating insights from clinical and experimental studies. *Exp Mol Med* **52**, 1652–1662 (2020).

43. J. Praline, N. Limousin, P. Vourc'h, M. Pallix, S. Debiais, A. Guennoc, C. R. Andres, P. Corcia, CADASIL and ALS: a link? *Amyotroph Lateral Scler* **11**, 399–401 (2010).

44. D. W. Sirkis, E. G. Geier, L. W. Bonham, C. M. Karch, J. S. Yokoyama, Recent Advances in the Genetics of Frontotemporal Dementia. *Curr Genet Med Rep* **7**, 41–52 (2019).

45. H.-J. Kim, H. Y. Kim, W. K. Paek, A. Park, M. Young Park, C. S. Ki, H.-M. Park, S. H. Kim, Amyotrophic lateral sclerosis and frontotemporal lobar degeneration in association with CADASIL. *Neurologist* **18**, 92–95 (2012).

46. K. Katisko, N. Huber, T. Kokkola, P. Hartikainen, J. Krüger, A.-L. Heikkinen, V. Paananen, V. Leinonen, V. E. Korhonen, S. Helisalmi, S.-K. Herukka, V. Cantoni, Y. Gadola, S. Archetti, A. M. Remes, A. Haapasalo, B. Borroni, E. Solje, Serum total TDP-43 levels are decreased in frontotemporal dementia patients with C9orf72 repeat expansion or concomitant motoneuron disease phenotype. *Alzheimer's Research & Therapy* **14**, 151 (2022).

47. K. Menšíková, R. Matěj, C. Colosimo, R. Rosales, L. Tučková, J. Ehrmann, D. Hraboš, K. Kolaříková, R. Vodička, R. Vrtěl, M. Procházka, M. Nevrlý, M. Kaiserová, S. Kurčová, P. Otruba, P. Kaňovský, Lewy body disease or diseases with Lewy bodies? *npj Parkinsons Dis.* **8**, 1–11 (2022).

48. R. L. Nussbaum, C. E. Ellis, Alzheimer's Disease and Parkinson's Disease. *New England Journal of Medicine* **348**, 1356–1364 (2003).

49. J. Ramirez, A. A. Dilliott, M. A. Binns, D. P. Breen, E. C. Evans, D. Beaton, P. M. McLaughlin, D. Kwan, M. F. Holmes, M. Ozzoude, C. J. M. Scott, S. C. Strother, S. Symons, R. H. Swartz, D. Grimes, M. Jog, M. Masellis, S. E. Black, A. Joutel, C. Marras, E. Rogaeva, R. A. Hegele, A. E. Lang, Ontario Neurodegenerative Disease Research Initiative Investigators, Parkinson's Disease, NOTCH3 Genetic Variants, and White Matter Hyperintensities. *Mov Disord* **35**, 2090–2095 (2020).

50. M. Ragno, A. Berbellini, G. Cacchiò, A. Manca, F. Di Marzio, L. Pianese, A. De Rosa, S. Silvestri, M. Scarcella, G. De Michele, Parkinsonism is a late, not rare, feature of CADASIL: a study on Italian patients carrying the R1006C mutation. *Stroke* **44**, 1147–1149 (2013).

51. S. Velasco, P. Alvarez-Muñoz, M. Pericacho, P. ten Dijke, C. Bernabéu, J. M. López-Novoa,

A. Rodríguez-Barbero, L- and S-endoglin differentially modulate TGFβ1 signaling mediated by ALK1 and ALK5 in L6E9 myoblasts. *Journal of Cell Science* **121**, 913–919 (2008).

52. F. J. Blanco, M. T. Grande, C. Langa, B. Oujo, S. Velasco, A. Rodriguez-Barbero, E. Perez-Gomez, M. Quintanilla, J. M. López-Novoa, C. Bernabeu, S-Endoglin Expression Is Induced in Senescent Endothelial Cells and Contributes to Vascular Pathology. *Circulation Research* **103**, 1383–1392 (2008).

53. K. Tominaga, H. I. Suzuki, TGF-β Signaling in Cellular Senescence and Aging-Related Pathology. *Int J Mol Sci* **20**, 5002 (2019).

54. S. Matsuda, A. Revandkar, T. D. Dubash, A. Ravi, B. S. Wittner, M. Lin, R. Morris, R. Burr, H. Guo, K. Seeger, A. Szabolcs, D. Che, L. Nieman, G. A. Getz, D. T. Ting, M. S. Lawrence, J. Gainor, D. A. Haber, S. Maheswaran, TGF-β in the microenvironment induces a physiologically occurring immune-suppressive senescent state. *Cell Reports* **42**, 112129 (2023).

55. X. Meng, D. J. Nikolic-Paterson, H. Y. Lan, TGF-β: the master regulator of fibrosis. *Nat Rev Nephrol* **12**, 325–338 (2016).

56. D. Peng, M. Fu, M. Wang, Y. Wei, X. Wei, Targeting TGF-β signal transduction for fibrosis and cancer therapy. *Molecular Cancer* **21**, 104 (2022).

57. Q. Miao, T. Paloneva, S. Tuominen, M. Pöyhönen, S. Tuisku, M. Viitanen, H. Kalimo, Fibrosis and stenosis of the long penetrating cerebral arteries: the cause of the white matter pathology in cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy. *Brain Pathol* **14**, 358–364 (2004).

58. B.-G. Kim, E. Malek, S. H. Choi, J. J. Ignatz-Hoover, J. J. Driscoll, Novel therapies emerging in oncology to target the TGF-β pathway. *Journal of Hematology & Oncology* **14**, 55 (2021).

59. H. S.-H. Oh, J. Rutledge, D. Nachun, R. Pálovics, O. Abiose, P. Moran-Losada, D. Channappa, D. Y. Urey, K. Kim, Y. J. Sung, L. Wang, J. Timsina, D. Western, M. Liu, P. Kohlfeld, J. Budde, E. N. Wilson, Y. Guen, T. M. Maurer, M. Haney, A. C. Yang, Z. He, M. D. Greicius, K. I. Andreasson, S. Sathyan, E. F. Weiss, S. Milman, N. Barzilai, C. Cruchaga, A. D. Wagner, E. Mormino, B. Lehallier, V. W. Henderson, F. M. Longo, S. B. Montgomery, T. Wyss-Coray, Organ aging signatures in the plasma proteome track health and disease. *Nature* **624**, 164–172 (2023).

60. M. Elsafadi, M. Manikandan, R. A. Dawud, N. M. Alajez, R. Hamam, M. Alfayez, M. Kassem, A. Aldahmash, A. Mahmood, Transgelin is a TGFβ-inducible gene that regulates osteoblastic and adipogenic differentiation of human skeletal stem cells through actin cytoskeleston organization. *Cell Death Dis* **7**, e2321–e2321 (2016).

61. K. A. Walker, J. Chen, J. Zhang, M. Fornage, Y. Yang, L. Zhou, M. E. Grams, A. Tin, N. Daya, R. C. Hoogeveen, A. Wu, K. J. Sullivan, P. Ganz, S. L. Zeger, E. F. Gudmundsson, V. Emilsson, L. J. Launer, L. L. Jennings, V. Gudnason, N. Chatterjee, R. F. Gottesman, T. H. Mosley, E. Boerwinkle, C. M. Ballantyne, J. Coresh, Large-scale plasma proteomic analysis identifies proteins and pathways associated with dementia risk. *Nat Aging* **1**, 473–489 (2021).

62. S. J. Lee, X. Zhang, M. M. Wang, Vascular accumulation of the small leucine rich proteoglycan Decorin in CADASIL. *Neuroreport* **25**, 1059–1063 (2014).

63. X. Zhang, S. J. Lee, M. F. Young, M. M. Wang, The small leucine-rich proteoglycan BGN accumulates in CADASIL and binds to NOTCH3. *Transl Stroke Res* **6**, 148–155 (2015).

64. H. Dong, M. Blaivas, M. M. Wang, Bidirectional encroachment of collagen into the tunica media in cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy. *Brain Res* **1456**, 64–71 (2012).

65. Z. Chen, C. Givens, J. S. Reader, E. Tzima, Haemodynamics Regulate Fibronectin Assembly via PECAM. *Sci Rep* **7**, 41223 (2017).

66. J. A. Buczek-Thomas, C. B. Rich, M. A. Nugent, Hypoxia Induced Heparan Sulfate Primes the Extracellular Matrix for Endothelial Cell Recruitment by Facilitating VEGF-Fibronectin Interactions. *Int J Mol Sci* **20**, 5065 (2019).

67. H. Ni, P. S. T. Yuen, J. M. Papalia, J. E. Trevithick, T. Sakai, R. Fässler, R. O. Hynes, D. D. Wagner, Plasma fibronectin promotes thrombus growth and stability in injured arterioles. *Proc Natl Acad Sci U S A* **100**, 2415–2419 (2003).

68. J. Cho, D. F. Mosher, Enhancement of thrombogenesis by plasma fibronectin cross-linked to fibrin and assembled in platelet thrombi. *Blood* **107**, 3555–3563 (2006).

69. S. Yoshizaki, T. Tamaru, M. Hara, K. Kijima, M. Tanaka, D.-J. Konno, Y. Matsumoto, Y. Nakashima, S. Okada, Microglial inflammation after chronic spinal cord injury is enhanced by reactive astrocytes via the fibronectin/β1 integrin pathway. *J Neuroinflammation* **18**, 12 (2021).

70. J. M. J. Stoffels, J. C. de Jonge, M. Stancic, A. Nomden, M. E. van Strien, D. Ma, Z. Sisková, O. Maier, C. Ffrench-Constant, R. J. M. Franklin, D. Hoekstra, C. Zhao, W. Baron, Fibronectin aggregation in multiple sclerosis lesions impairs remyelination. *Brain* **136**, 116–131 (2013).

71. H. Saga, A. Ohhata, A. Hayashi, M. Katoh, T. Maeda, H. Mizuno, Y. Takada, Y. Komichi, H. Ota, N. Matsumura, M. Shibaya, T. Sugiyama, S. Nakade, K. Kishikawa, A Novel Highly Potent Autotaxin/ENPP2 Inhibitor Produces Prolonged Decreases in Plasma Lysophosphatidic Acid Formation In Vivo and Regulates Urethral Tension. *PLoS One* **9**, e93230 (2014).

72. L. Bitar, T. Uphaus, C. Thalman, M. Muthuraman, L. Gyr, H. Ji, M. Domingues, H. Endle, S. Groppa, F. Steffen, N. Koirala, W. Fan, L. Ibanez, L. Heitsch, C. Cruchaga, J.-M. Lee, F. Kloss, S. Bittner, R. Nitsch, F. Zipp, J. Vogt, Inhibition of the enzyme autotaxin reduces cortical excitability and ameliorates the outcome in stroke. *Science Translational Medicine* **14**, eabk0135 (2022).

73. L. W. Yuelling, C. T. Waggener, F. S. Afshari, J. A. Lister, B. Fuss, Autotaxin/ENPP2 Regulates Oligodendrocyte Differentiation in vivo in the Developing Zebrafish Hindbrain. *Glia* **60**, 1605–1618 (2012).

74. G. Fang, Y. Shen, D. Liao, ENPP2 alleviates hypoxia/reoxygenation injury and ferroptosis by regulating oxidative stress and mitochondrial function in human cardiac microvascular

endothelial cells. *Cell Stress and Chaperones* **28**, 253–263 (2023).

75. E. Matas-Rico, E. Frijlink, I. van der H. Àvila, A. Menegakis, M. van Zon, A. J. Morris, J. Koster, F. Salgado-Polo, S. de Kivit, T. Lança, A. Mazzocca, Z. Johnson, J. Haanen, T. N. Schumacher, A. Perrakis, I. Verbrugge, J. H. van den Berg, J. Borst, W. H. Moolenaar, Autotaxin impedes anti-tumor immunity by suppressing chemotaxis and tumor infiltration of CD8+ T cells. *Cell Reports* **37** (2021), doi:10.1016/j.celrep.2021.110013.

76. N. Karvelas, B. Oh, E. Wang, Y. Cobigo, T. Tsuei, S. Fitzsimons, A. Ehrenberg, M. Geschwind, D. Schwartz, J. Kramer, A. R. Ferguson, B. L. Miller, L. Silbert, H. Rosen, F. M. Elahi, Enlarged Perivascular Spaces are Associated with White Matter Injury, Brain Atrophy, Cognitive Decline and Markers of Inflammation in an Autosomal Dominant Vascular Neurodegenerative Disease (CADASIL), 2023.08.17.553732 (2023).

77. W. Khan, L. Kong, S. M. Noman, B. Brekhna, A novel feature selection method via mining Markov blanket. *Appl Intell* **53**, 8232–8255 (2023).

78. Y. Tan, Z. Liu, Feature selection and prediction with a Markov blanket structure learning algorithm. *BMC Bioinformatics* **14**, A3 (2013).

79. A. K. Jain, R. P. W. Duin, J. Mao, Statistical pattern recognition: a review. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22**, 4–37 (2000).

80. Y. Guo, T. Hastie, R. Tibshirani, Regularized linear discriminant analysis and its application in microarrays. *Biostatistics* **8**, 86–100 (2007).

81. M. B. Kursa, W. R. Rudnicki, Feature Selection with the Boruta Package. *Journal of Statistical Software* **36**, 1–13 (2010).

82. C. Ding, H. Peng, Minimum redundancy feature selection from microarray gene expression data. *J Bioinform Comput Biol* **3**, 185–205 (2005).

83. J. Y.-L. Chan, S. M. H. Leow, K. T. Bea, W. K. Cheng, S. W. Phoong, Z.-W. Hong, Y.-L. Chen, Mitigating the Multicollinearity Problem and Its Machine Learning Approach: A Review. *Mathematics* **10**, 1283 (2022).

84. I. T. Jolliffe, Ed., in *Principal Component Analysis*, Springer Series in Statistics. (Springer, New York, NY, 2002), pp. 78–110.

85. A. Krämer, J. Green, J. Pollard, S. Tugendreich, Causal analysis approaches in Ingenuity Pathway Analysis. *Bioinformatics* **30**, 523–530 (2014).

86. A. M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).

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

88. Y. Liao, G. K. Smyth, W. Shi, featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).

89. M. I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15**, 550 (2014).

Acknowledgments

This work was funded by the National Institute on Aging and Department of Veterans Affairs (IK2CX002180), Rainwater Charitable Foundation (2790954), and Chan Zuckerberg Initiative (2022-316712) (FME). High power computing at the Icahn School of Medicine at Mount Sinai is supported in part through the computational and data resources and staff expertise provided by Scientific Computing and Data and the Clinical and Translational Science Awards (CTSA) grant UL1TR004419.

Author Contributions

JNK, HR, NK, and FME developed the concept, planned and analyzed data. JNK, HR, and NK performed the biostatistical analyses and generated figures. JNK and FME drafted, reviewed, and edited the final version of the manuscript. MMW provided brain tissue and SF and LM extracted protein from brain tissue. JFM shared plasma samples and JFAV, YTQ and FGL shared data. All authors critically read and provided edits. FME supervised the project, provided the resources, and acquired funding.

Competing Interests

All authors report no conflicts of interest relevant to this study.

Data and materials availability

All data associated with this study are presented in the article and or the Supplementary Materials. All de-identified patient data is presented in the manuscript and will be shared upon request.

Figures and Tables

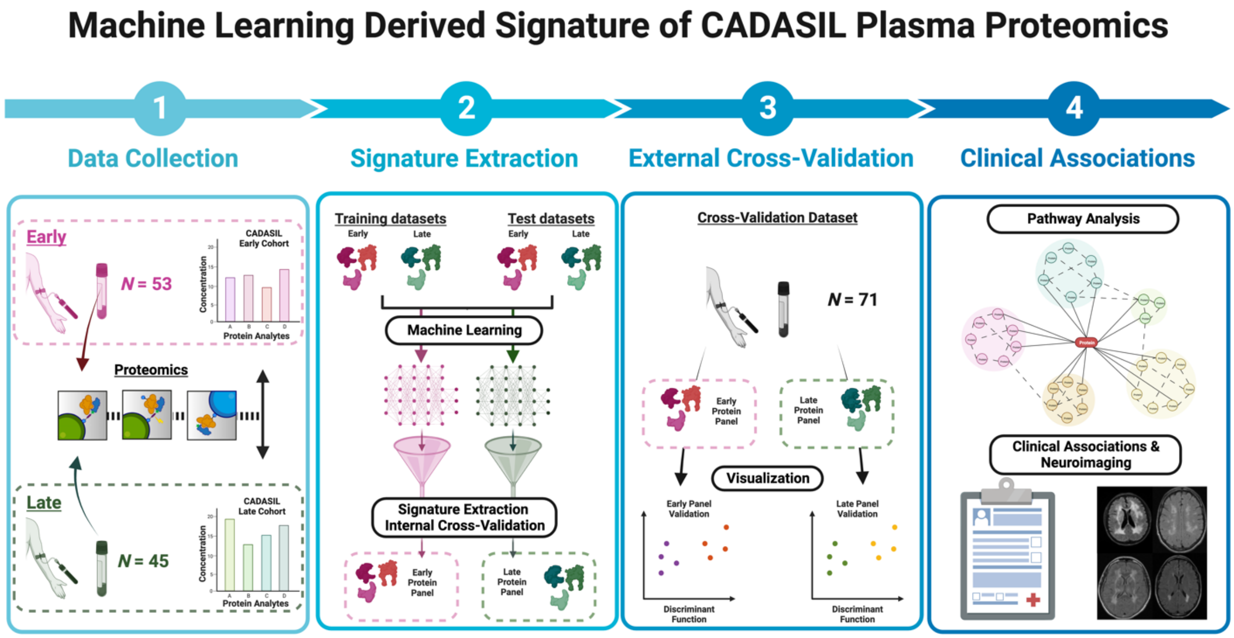

Fig. 1. Illustration of data collection and analytical workflow used in this study. Plasma samples from three distinct CADASIL cohorts were collected and analyzed using the SomaSCAN proteomics platform. Protein signatures were determined by applying a machine learning feature extraction methodology, and findings were validated both internally and externally, as well as on post-mortem brain tissue. Protein panels were interrogated using interactive enrichment analysis, and we investigated the associations of molecular measures with quantitative measures of clinical abnormalities.

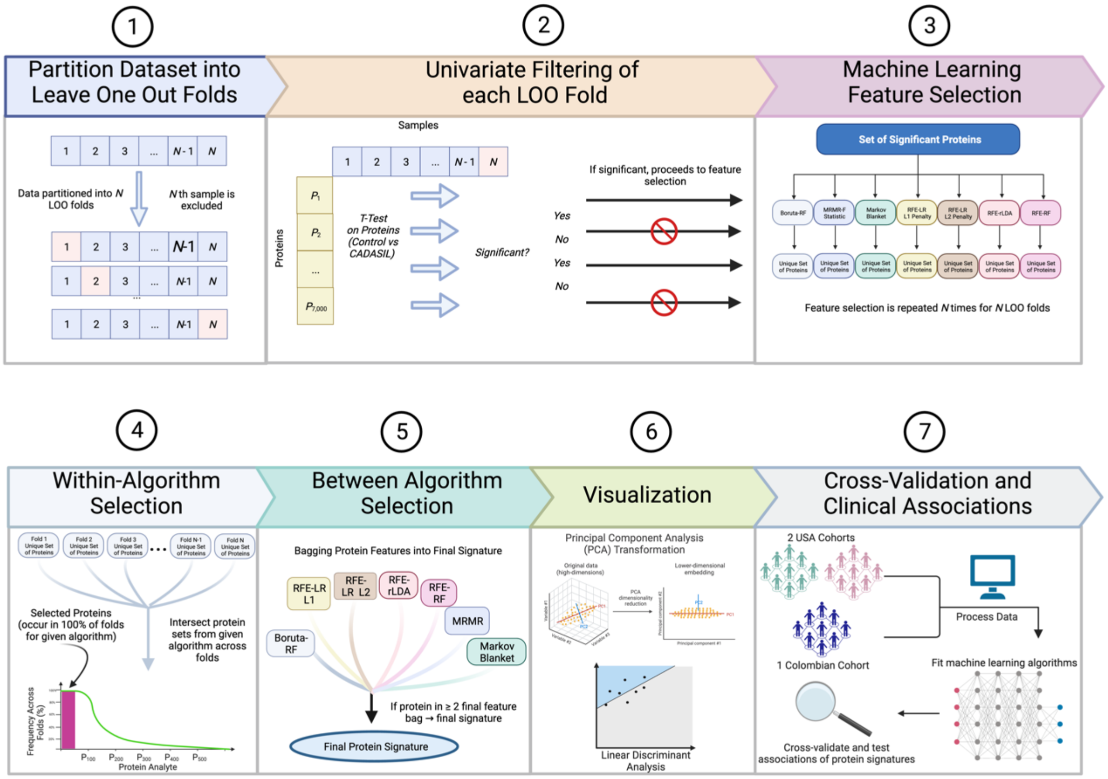
1
Partition Dataset into Leave One Out Folds
1 2 3 ... N - 1 N
Data partitioned into N LOO folds
N th sample is excluded
2
Univariate Filtering of each LOO Fold
Samples
Proteins
P_1
P_2
...
$P_{7,000}$
T-Test on Proteins (Control vs CADASIL)
Significant?
Yes
No
Yes
No
If significant, proceeds to feature selection
3
Machine Learning Feature Selection
Set of Significant Proteins
Boruta-RF
MRMR-F Statistic
Markov Blanket
RFE-LR L1 Penalty
RFE-LR L2 Penalty
RFE-rLDA
RFE-RF
Unique Set of Proteins
Feature selection is repeated N times for N LOO folds
4
Within-Algorithm Selection
Fold 1 Unique Set of Proteins
Fold 2 Unique Set of Proteins
Fold 3 Unique Set of Proteins
Fold N-1 Unique Set of Proteins
Fold N Unique Set of Proteins
Selected Proteins (occur in 100% of folds for given algorithm)
Intersect protein sets from given algorithm across folds
Frequency Across Folds (%)
P_{100} P_{200} P_{300} P_{400} P_{500}
Protein Analyte
5
Between Algorithm Selection
Bagging Protein Features into Final Signature
RFE-LR L1
RFE-LR L2
RFE-rLDA
RFE-RF
MRMR
Boruta-RF
Markov Blanket
If protein in ≥ 2 final feature bag → final signature
Final Protein Signature
6
Visualization
Principal Component Analysis (PCA) Transformation
Original data (high-dimensions)
PCA dimensionality reduction
Lower-dimensional embedding
PC1
PC2
Linear Discriminant Analysis
7
Cross-Validation and Clinical Associations
2 USA Cohorts
1 Colombian Cohort
Process Data
Fit machine learning algorithms
Cross-validate and test associations of protein signatures

Fig. 2. Illustration of machine learning methodology. Multivariate analytical workflow for CADASIL proteomic signature identification, featuring robust statistical validation, dimensionality reduction from 7,000 to 1,300 proteins, LOO method for bias minimization, diverse feature selection algorithms, and a stringent ensemble aggregation technique leading to two definitive protein signatures.

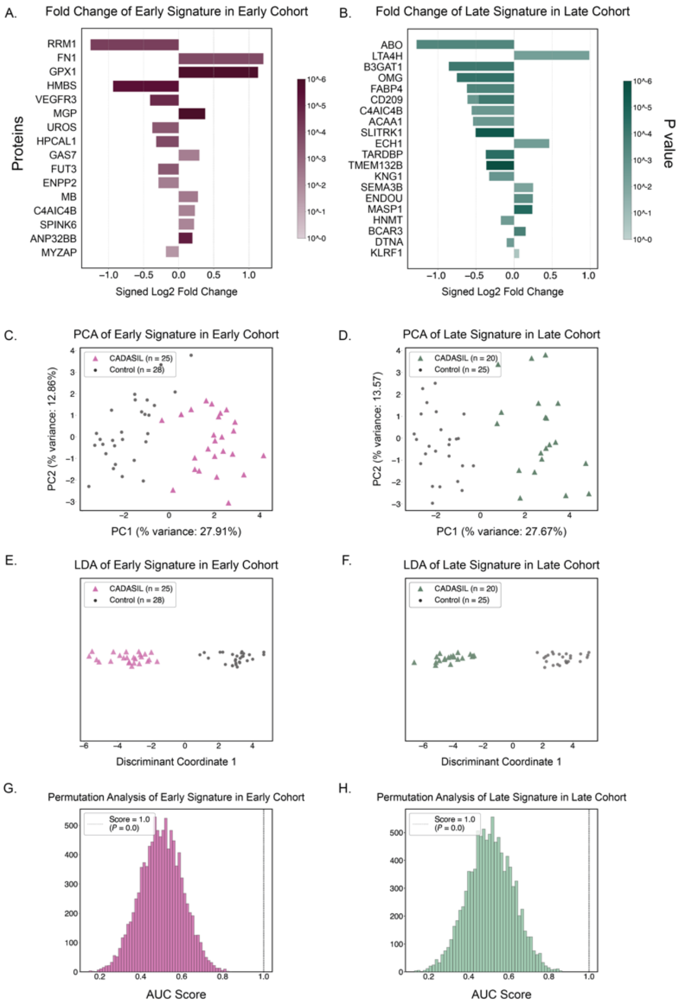

Fig. 3. Machine learning identifies highly accurate and precise plasma proteomic signatures that discriminate CADASIL from control groups in early and late disease stages. **(A)** Fold changes ($\log_2$FC) of Early signature protein expression between CADASIL and Control groups. **(B)** Fold changes ($\log_2$FC) of Late signature protein expression between CADASIL and Control groups. **(C)** Principal component analysis of CADASIL-Early cohort separated from Control group, using Early signature proteins as features. **(D)** Principal component analysis of CADASIL-Late cohort using Late signature proteins as features. **(E)** Regularized linear discriminant analysis of CADASIL-Early cohort using Early signature proteins as features. **(F)** Regularized linear discriminant analysis of CADASIL-Late cohort using Late signature proteins as features. **(G)** Histogram of permutation results of random feature selection performance on Early cohort compared to derived Early signature. **(H)** Histogram of permutation analysis on Late cohort compared to derived Late signature.

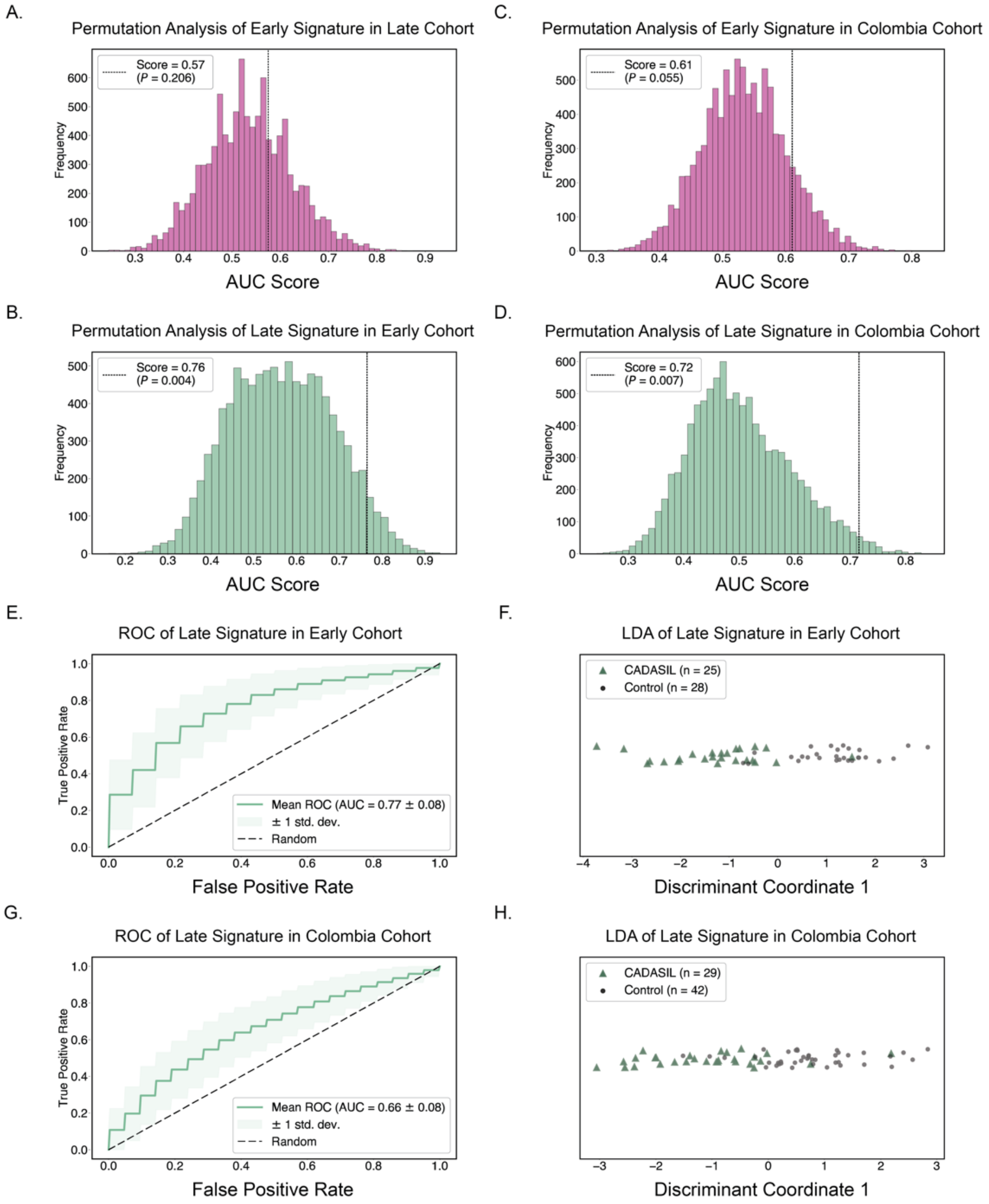

Fig. 4. Validation of machine learning-derived Early and Late signatures. **(A)** Permutation analysis of Early signature performance in Late Cohort. **(B)** Permutation analysis of Late signature performance in Early Cohort. **(C)** Permutation analysis of Early signature in the external Colombia Cohort. **(D)** Permutation analysis of Late signature in the external Colombia Cohort. **(E, F)** Biplot combining ROC curve and LDA plot for Late signature in Early Cohort. **(G, H)** Biplot of ROC curve and LDA plot for Late signature in Colombia Cohort.

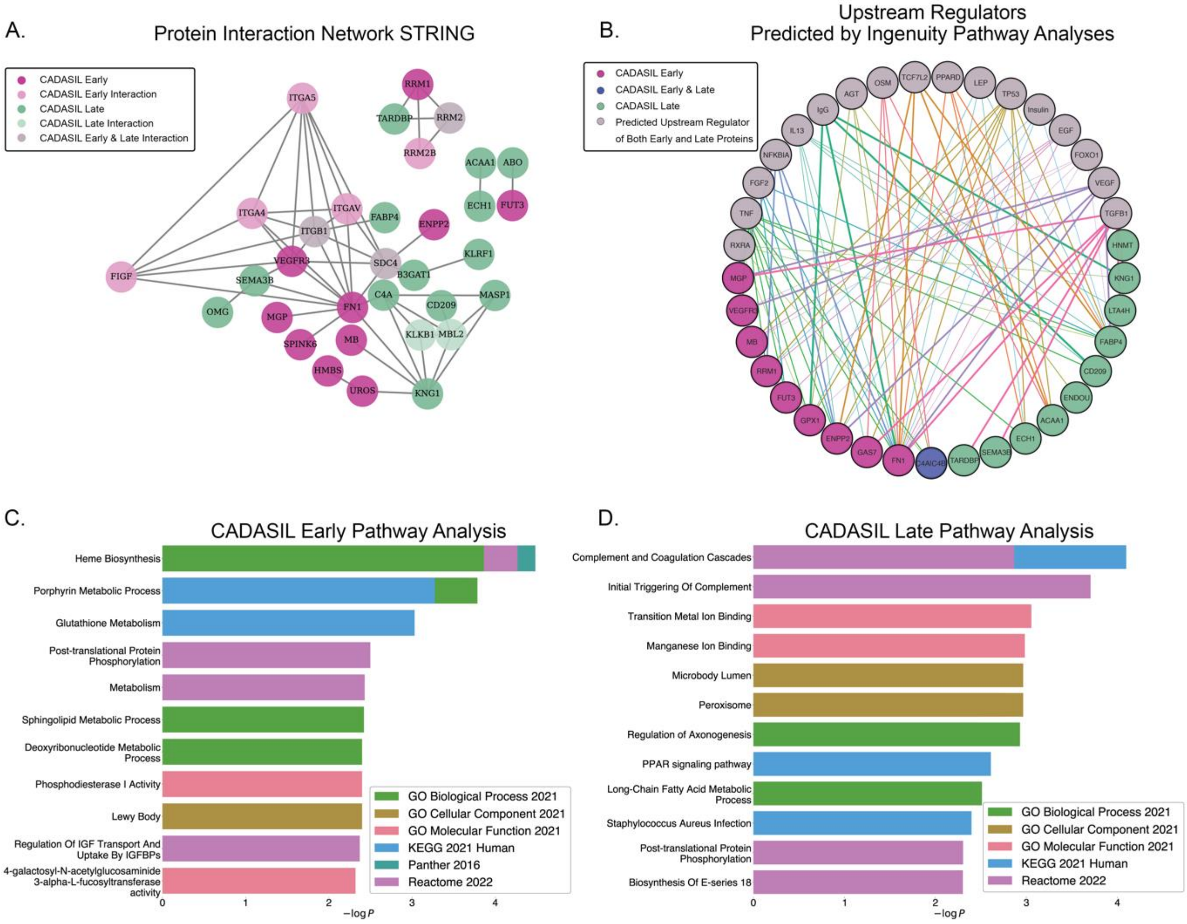

Fig. 5. Pathway and network analyses of Early and Late signatures. **(A)** STRING network of Early and Late signature proteins with predicted protein-protein physical and functional interactions. **(B)** IPA network of Early and Late signature proteins as well as predicted upstream regulators. Edges are colored according to the corresponding upstream regulator and edge width was assigned based on predicted activation Z-scores (see also Table S4). **(C)** Collated EnrichR pathway analysis from several libraries using Early signature. **(D)** Collated EnrichR pathway analysis from several libraries using Late signature.

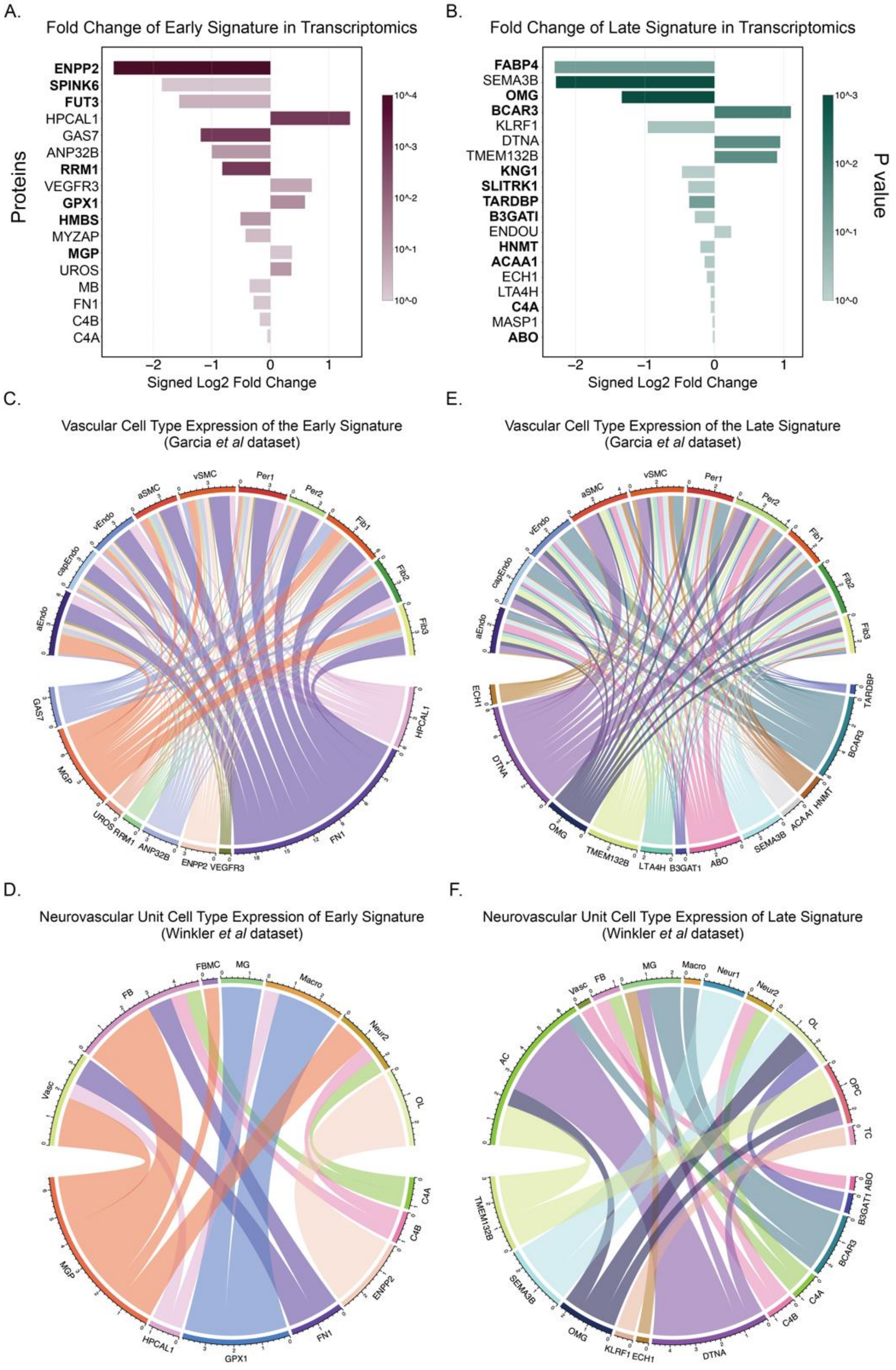

Fig. 6. Brain transcriptomic analysis of Early and Late signatures in CADASIL. **(A-B)** Bulk RNASeq analysis depicting fold changes (log2FC) of the **(A)** Early signature and **(B)** Late signature in CADASIL versus control brain tissue. Bolded proteins are co-directional in plasma and brain tissue. **(C-D)** Chord diagrams illustrating vascular cell type expression of the Early signature based on Garcia et al. dataset **(C)**, and Winkler et al. dataset **(D)**. **(E-F)** Chord diagrams showcasing vascular cell type expression of the Late signature based on Garcia et al. dataset **(E)**, and Winkler et al. dataset **(F)**. **(C-F)** The width of the band represents the degree of upregulation

of the protein in a specific cell type. **(C, E)** aEndo, arterial endothelial cell; aSMC, arterial smooth muscle cell; capEndo, capillary endothelial cell; Fib1, fibroblast cluster 1; Fib2, fibroblast cluster 2; Fib3, fibroblast clutter 3; Per1, pericyte cluster 1; Per2, pericyte cluster 2; vEndo, venous endothelial cell; vSMC, vascular smooth muscle cell. **(D, F)** AC, astrocyte; Vasc, vascular cells; FB, perivascular fibroblast; FBMC, fibromyocyte; Macro, macrophage; TC, T cell; Neu1, Neuron Cluster 2; Neu1, Neuron Cluster 2; MG, microglia; OL, oligodendrocyte; and OPC, oligodendrocyte precursor cell.

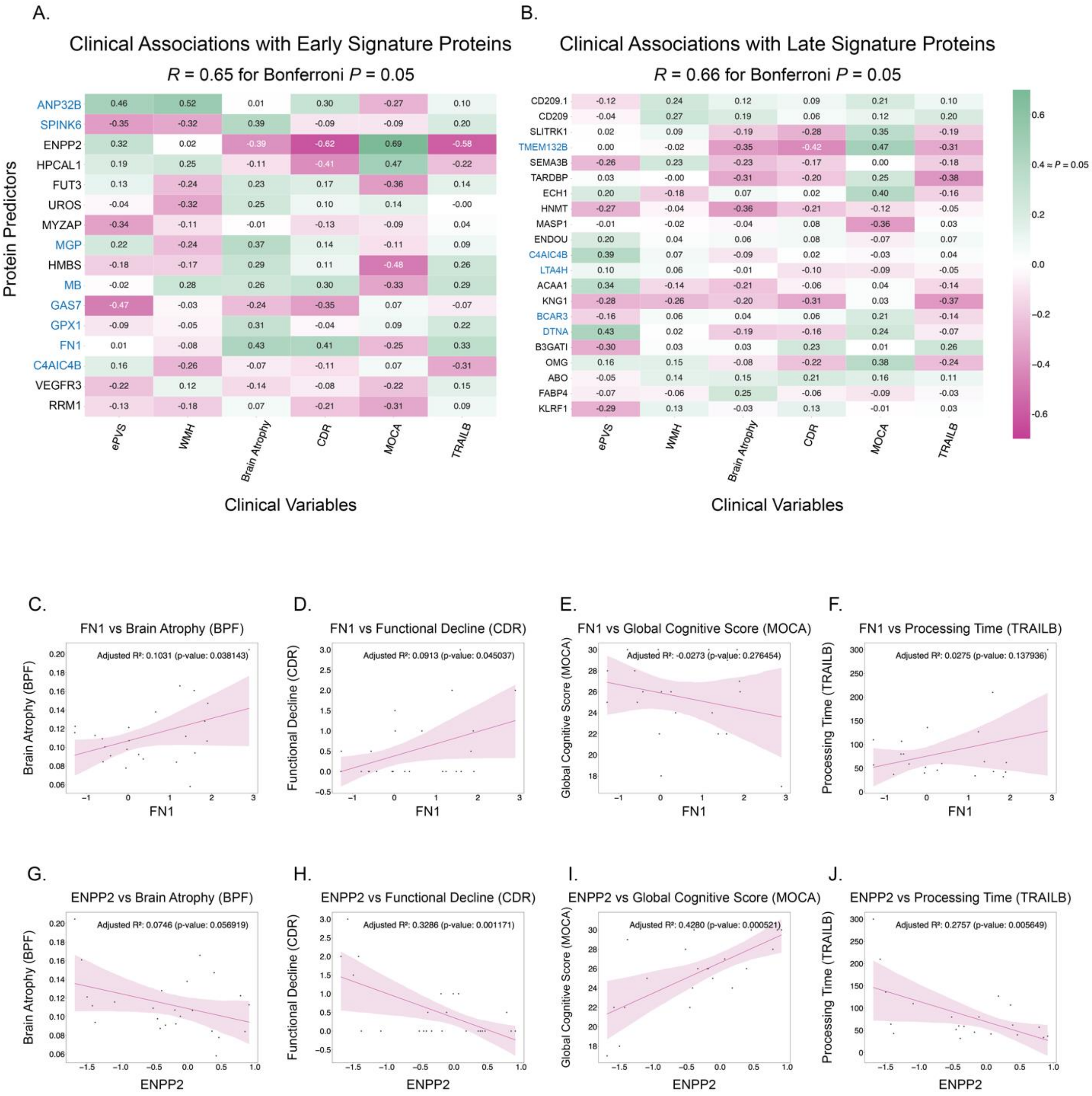

Fig. 7. Plasma proteomic signatures in CADASIL patients: association and regression analyses. **(A)** Association analysis of the Early signature. **(B)** Association analysis of the Late signature. Proteins labeled in blue are upregulated in plasma; proteins labeled in black are downregulated in plasma. **(C, D, E, F)** Regression analyses for protein FN1 against various imaging and clinical metrics: **(C)** brain atrophy, **(D)** functional decline, **(E)** global cognition, and **(F)** processing time. **(G, H, I, J)** Regression analysis of protein ENPP2 against various imaging and clinical metrics: (G) brain atrophy (-log(BPF)), **(H)** functional decline (CDR), **(I)** global cognition (MOCA), and **(J)** processing time (TRAILB). Brain Atrophy, BPF, brain parenchymal fraction; Functional Decline, CDR, clinical dementia rating score; ePVS, Enlarged Perivascular Space Volume; Global Cognition, MOCA, Montreal Cognitive Assessment Score; Processing Time, TRAILB,

Trail Making Test Part B Completion; Time; WMH, White Matter Hyperintensities. Protein labels colored blue indicate upregulation in CADASIL plasma compared to controls. Proteins colored black indicate downregulation in CADASIL plasma compared to controls.

Cohort	Status	Sample Size (N)	Female (%)	Age (yrs.) (mean ± SD)
Early	Control	28	50	52 (± 11)
	CADASIL	25	36	51 (± 12)
Late	Control	25	20	64 (± 11)
	CADASIL	20	50	57 (± 12)
Colombia	Control	42	40	41 (± 15)
	CADASIL	29	28	39 (± 10)

Table 1. Summary Demographic Data of CADASIL Cohorts. Control, healthy control; CADASIL, CADASIL cases.

Supplementary Material

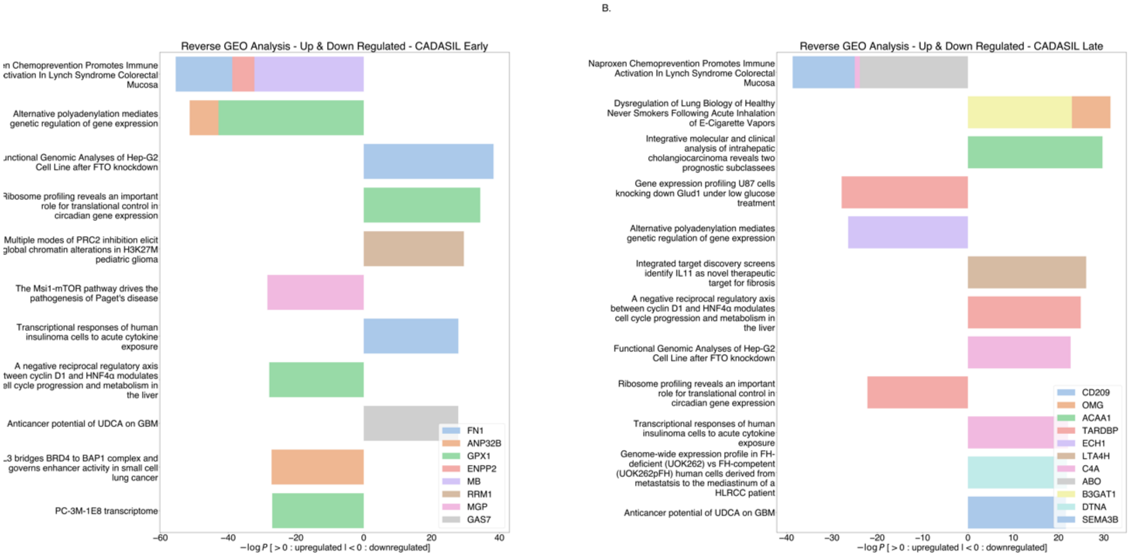

Fig. S1. Protein-centric Reverse GEO Search of **(A)** CADASIL-Early and **(B)** CADASIL-Late Signature.

Cohort	Evaluator	ROC	Accuracy	Precision	Recall	F-score	Explained Variance
CADASIL-Early	Ridge Classifier	1.000	0.991	0.990	0.994	0.991	0.969
	LR	1.000	0.979	1.000	0.956	0.975	0.926
	rLDA	1.000	0.989	0.983	0.996	0.989	0.959
	Linear SVC	0.998	0.961	0.995	0.924	0.955	0.860
	Perceptron	0.995	0.951	0.973	0.928	0.944	0.824
	Decision Trees	0.730	0.731	0.738	0.718	0.717	0.054
CADASIL-Late	LR	1.000	0.999	0.998	1.000	0.999	0.996
	Linear SVC	0.999	0.988	0.989	0.985	0.986	0.954
	Perceptron	0.998	0.969	0.959	0.988	0.969	0.896
	Ridge Classifier	0.997	0.976	0.978	0.970	0.972	0.907
	rLDA	0.992	0.951	0.948	0.953	0.946	0.818
	Decision Trees	0.849	0.857	0.867	0.835	0.827	0.480

Table S1. CADASIL-Early ML classifiers and their performance using the CADASIL-Early, CADASIL-Late, and CADASIL-Colombia protein signature.

Cohort Best ML Model ROC			*P* value of ROC	
			Based on ROC Score Distribution of Random Protein Predictor Models	Based on ROC Score Distribution of Permuted Label Models
CADASIL-Late	LDA	0.575	0.2058	0.3312
CADASIL-Colombia	SVC	0.610	0.0552	0.1308

Table S2. Best ML classifiers performance metrics using the CADASIL-Early protein signature on different cohorts.

Cohort	Best ML Model	ROC	*P* value of ROC	
			Based on ROC Score Distribution of Random Protein Predictor Models	Based on ROC Score Distribution of Permuted Label Models
CADASIL-Early	SVC	0.765	0.0042	0.049
CADASIL-Colombia	SVC	0.716	0.0066	0.0168

Table S3. Best ML classifiers performance metrics using the CADASIL-Late protein signature on different cohorts.

Upstream Regulator	Molecule Type	Activation z-score	Overlap P value	Target Molecules in Dataset
IgG	complex	1.982	0.00144	CD209,FN1,GPX1,KNG1
TGFB1	growth factor	1.691	0.00686	FN1,GAS7,HNMT,KNG1,LTA4H,MGP,SEMA3B,TARDBP
Vegf	group	1.342	0.000791	ENPP2,FABP4,FLT4,FN1,MGP
TCF7L2	transcription regulator	1.131	0.00314	ACAA1,ENPP2,FABP4,FN1
NFKBIA	transcription regulator	1	0.00377	ENPP2,FN1,MGP,RRM1
OSM	cytokine	0.254	0.00935	ACAA1,C4A/C4B,FN1,GAS7
TNF	cytokine	0.117	0.000114	C4A/C4B,CD209,ECH1,ENPP2,FABP4,FLT4,FN1,FUT3,GPX1,MGP
FGF2	growth factor	0.054	0.000707	ENPP2,FABP4,FN1,FUT3,MGP
PPARD	ligand-dependent nuclear receptor	0.036	0.000816	ACAA1,ECH1,FABP4,FN1
TP53	transcription regulator	0	0.00196	ACAA1,ECH1,ENPP2,FABP4,FN1,GAS7,GPX1,MB,RRM1
AGT	growth factor	-0.5	0.0454	C4A/C4B,ENDOU,FN1,GPX1
EGF	growth factor	-0.694	0.0184	FN1,FUT3,MB,RRM1
IL13	cytokine	-0.885	4.74E-05	CD209,ENPP2,FABP4,FN1,GAS7,LTA4H
LEP	growth factor	-1	0.00573	ECH1,FABP4,FN1,GPX1
RXRA	ligand-dependent nuclear receptor	-1.067	0.000332	C4A/C4B,FABP4,GPX1,KNG1,MGP
Insulin	group	-1.128	0.0171	ENPP2,FABP4,FN1,RRM1

FOXO1	transcription regulator	-1.342	0.000576	FABP4,FLT4,FN1,GPX1,MB

Table S4. IPA Predicted Upstream Regulators of Early and Late signature proteins.

Cohort	Status	Sample Size (*N*)	Female (%)	Age (mean ± SD)
CADASIL Brain Transcriptomics	CO	5	40	67(± 4)
	CADASIL	7	29	66 (± 6)

Table S5. Summary Demographic Data of CADASIL Brain Tissue Transcriptomics Cohort. CO, healthy control; CADASIL, CADASIL cases.