Classification Models using Circulating Neutrophil Transcripts Can Detect Unruptured Intracranial Aneurysm

Kerry E Poppenberg
Canon Stroke and Vascular Research Center; Department of Neurosurgery, Jacobs School of Medicine and Biomedical Sciences https://orcid.org/0000-0002-1748-887X

Vincent M Tutino
Canon Stroke and Vascular Research Center; Department of Biomedical Engineering;Department of Neurosurgery, Jacobs School of Medicine and Biomedical Sciences;Department of Pathology and Anatomical Sciences, Jacobs School of Medicine and Biomedical Sciences

Lu Li
Department of Computer Science and Engineering

Muhammad Waqas
Department of Neurosurgery, Jacobs School of Medicine and Biomedical Sciences; Department of Neurology, Jacobs School of Medicine and Biomedical Sciences

Armond June
Department of Pathology and Anatomical Sciences, Jacobs School of Medicine and Biomedical Sciences

Lee Chaves
Department of Internal Medicine, Jacobs School of Medicine and Biomedical Sciences

Kaiyu Jiang
Genetics, Genomics, and Bioinformatics Program

James N Jarvis
Genetics, Genomics, and Bioinformatics Program;Department of Pediatrics, Jacobs School of Medicine and Biomedical Sciences

Yijun Sun
Genetics, Genomics, and Bioinformatics Program; Department of Microbiology and Immunology

Kenneth V Snyder
Canon Stroke and Vascular Research Center;Department of Neurosurgery, Jacobs School of Medicine and Biomedical Sciences;Department of Radiology, Jacobs School of Medicine and Biomedical Sciences;Department of Neurology, Jacobs School of Medicine and Biomed

Elad I Levy
Canon Stroke and Vascular Research Center;Department of Neurosurgery, Jacobs School of Medicine and Biomedical Sciences;Department of Radiology, Jacobs School of Medicine and Biomedical Sciences

Adnan H Siddiqui
Canon Stroke and Vascular Research Center;Department of Neurosurgery, Jacobs School of Medicine and Biomedical Sciences;Department of Radiology, Jacobs School of Medicine and Biomedical Sciences
John Kolega
Canon Stroke and Vascular Research Center;Department of Pathology and Anatomical Sciences, Jacobs School of Medicine and Biomedical Sciences
Hui Meng (huimeng@buffalo.edu)
Canon Stroke and Vascular Research Center;Department of Biomedical Engineering;Department of Neurosurgery, Jacobs School of Medicine and Biomedical Sciences;Department of Mechanical and Aerospace Engineering https://orcid.org/0000-0003-3884-499X

Research

Keywords: intracranial aneurysm, neutrophil, transcriptomics, machine learning, inflammation, prediction model

Posted Date: March 13th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-17161/v1

Version of Record: A version of this preprint was published on October 15th, 2020. See the published version at https://doi.org/10.1186/s12967-020-02550-2.

Abstract

Background Intracranial aneurysms (IAs) are dangerous because of their potential to rupture. We previously found significant RNA expression differences in circulating neutrophils between patients with and without unruptured IAs and trained machine learning models to predict presence of IA using 40 neutrophil transcriptomes. Here, we aim to develop a predictive model for unruptured IA using neutrophil transcriptomes from a larger population and more robust machine learning methods.

Methods Neutrophil RNA extracted from the blood of 134 patients (55 with IA, 79 IA-free controls) was subjected to next-generation RNA sequencing. In a randomly-selected training cohort (n=94), the Least Absolute Shrinkage and Selection Operator (LASSO) selected transcripts, from which we constructed prediction models via 4 well-established supervised machine-learning algorithms (K-Nearest Neighbors, Random Forest, and Support Vector Machines with Gaussian and cubic kernels). We tested the models in the remaining samples (n=40) and assessed model performance by receiver-operating-characteristic (ROC) curves. Real-time quantitative polymerase chain reaction (RT-qPCR) of 10 IA-associated genes was used to verify gene expression in a subset of 50 neutrophil RNA samples. We also examined the potential influence of demographics and comorbidities on model prediction.

Results Feature selection using LASSO in the training cohort identified 37 IA-associated transcripts. Models trained using these transcripts had a maximum accuracy of 90% in the testing cohort. The testing performance across all methods had an average area under ROC curve (AUC)=0.97, an improvement over our previous models. The Random Forest model performed best across both training and testing cohorts. RT-qPCR confirmed expression differences in 8 of 10 genes tested. Gene ontology and IPA network analyses performed on the 37 model genes reflected dysregulated inflammation, cell signaling, and apoptosis processes. In our data, demographics and comorbidities did not affect model performance.

Conclusions We improved upon our previous IA prediction models based on circulating neutrophil transcriptomes by increasing sample size and by implementing LASSO and more robust machine learning methods. Future studies are needed to validate these models in larger cohorts and further investigate effect of covariates.

Background

Intracranial aneurysms (IAs) are present in up to 6% of the general population, but only about 10% show any symptoms prior to rupture.[1] The rupture of an IA is the leading cause of nontraumatic subarachnoid hemorrhage, which has a mortality rate up to 50%.[2–4] Clinical studies have shown that the incidence of future rupture can be decreased with elective endovascular or surgical treatment.[5, 6] However, because most IAs are asymptomatic, unruptured aneurysms are usually only detected incidentally on cerebral imaging performed for other medical reasons. Imaging by magnetic resonance imaging/angiography (MRI/MRA), computed tomography angiography (CTA), and digital subtraction angiography (DSA) are not routinely used for IA screening because they are expensive and carry the risk for serious

complications due to their invasive nature (e.g., DSA) or exposure to radiation (e.g., DSA, CTA). Thus, a blood test to identify individuals with unruptured IAs could facilitate a paradigm shift to proactive IA management by enabling routine screening and preventive treatment.

We hypothesized that changes in circulating neutrophil gene expression are correlated with the presence of IA in the cerebral vasculature. Recently, several studies have shed light on the emerging role of peripheral blood neutrophils in IA pathophysiology. Investigations of resected aneurysms have found elevated levels of key proteins released from neutrophils, namely myeloperoxidase (MPO) and neutrophil gelatinase-associated lipocalin (NGAL) in the IA wall.[7,8] These proteins have also been found to be elevated in plasma levels of blood from patients with IA.[8, 9] These proteins may play a key role in extracellular matrix (ECM) degradation and neutrophil activation[10, 11] during aneurysm natural history via their production of reactive oxygen species (ROS) and protection of MMP-9 degradation, respectively.[7, 8] Additionally, neutrophils in IA produce chemokines/cytokines necessary for monocyte infiltration and, based on their role in progression and rupture of abdominal aortic aneurysms, may release neutrophil extracellular traps.[12–15]

In a small, proof-of-concept study, we previously performed differential expression analysis in case-controlled cohorts (n = 11 IA, n = 11 control) and found an 82-gene signature that distinguishes IAs from controls.[16] Bioinformatics analyses broadly reflected peripheral neutrophil activation in patients with IA, as genes with elevated expression in the IA group were associated with leukocyte activation, cell activation, and defense response.[16] In a follow-up study, we next utilized machine learning to determine whether an algorithm could predict the presence of unruptured IA using differential gene expression.[17] In an unmatched cohort (n = 30), 26 highly-informative neutrophil transcripts (FDR < 0.05, abs[fold-change] ≥ 1.5) were used to construct a diagonal Linear Discriminant Analysis model, which predicted the presence of IA with an accuracy of 90% in a small independent cohort (n = 10).[17]

While these results were exciting, due to the small sample size, it was difficult to generalize our findings to a broader population. Therefore, in this study, we aimed at confirming these results in a larger cohort of patients. Importantly, this increased sample size would enable us to: A) implement more advanced feature selection methods (in place of basic thresholding) and machine learning techniques (i.e. Random Forest) to improve prediction accuracy, and B) examine potential effect of demographics and comorbidities on model prediction.

Methods

Study enrollment

This study was approved by University at Buffalo Institutional Review Board (study no. 030-474433). All methods followed the approved protocol. Written informed consent was obtained from all subjects prior to sample collection. Patients receiving cerebral DSA at Gates Vascular Institute, Buffalo, NY with and

without IA diagnosis were enrolled in this study. Most indications for DSA included confirmation of noninvasive imaging results of presence of IAs or other cerebral vascular conditions, follow-up of non-invasive imaging for headache or visual disturbance, or follow-up of previously identified IAs. All patients who consented to participate in this study were over 18 years, English speaking, and had not previously been treated for IA. Patients with potentially altered immune systems were excluded, including, for example, patients who had recent invasive surgery, were receiving chemotherapy, had a fever (> 100^{o}F), had received solid organ transplants, had autoimmune diseases, or were taking prednisone or other immunomodulating drugs.

Between December 2013 and September 2018, we collected 232 blood samples from cerebral DSA patients at Gates Vascular Institute (103 from patients with IA, and 129 from IA-free controls). Forty-three of these samples had been sequenced as a part of our previous studies.[16, 17] In all cases, IA diagnosis was confirmed by DSA images. Patient medical record data was also collected to study demographics and comorbidities.

Neutrophil isolation

During DSA, 16 mL of blood was drawn from the access catheter in the femoral artery and transferred into two 8 mL, citrated, cell preparation tubes (BD, Franklin Lakes, NJ). Neutrophils were isolated within 1 hour of blood collection, as described elsewhere.[18] Briefly, cell preparation tubes were centrifuged at 1,700 × g for 25 minutes to separate erythrocytes and neutrophils from mononuclear cells and plasma in the peripheral blood samples via a Ficoll density gradient. Erythrocytes and neutrophils were collected into a 3 mL syringe. Following hypotonic lysis of red blood cells, neutrophils were isolated by centrifugation at 400 × g for 10 min, disrupted and stored in TRIzol reagent (Life Technologies, Carlsbad, CA) at -80^{o}C until further processing. Neutrophils isolated in this fashion are more than 98% CD66b + by flow cytometry and contain no contaminating CD14 + monocytes.[19]

RNA preparation

Neutrophil RNA was extracted as described previously[16] using TRIzol, according to the manufacturer's instructions. Trace DNA was removed by DNase I (Life Technologies, Carlsbad, CA) treatment. RNA was purified using the RNeasy MinElute Cleanup Kit (Qiagen, Venlo, Limburg, Netherlands) and suspended in RNase-free water. The purity and concentration of RNA in each sample were measured by absorbance at 260 nm and 280 nm on a NanoDrop 2000 spectrophotometer (Thermo Scientific, Waltham, MA), and 200 ng to 400 ng of RNA was reserved for sequencing. Precise RNA concentration was measured via the Quant-iT RiboGreen Assay (Invitrogen, Carlsbad, CA) with a TBS-380 Fluorometer (Promega, Madison, WI), and the quality of the RNA samples was measured with an Agilent 2100 BioAnalyzer RNA 6000 Pico Chip (Agilent, Las Vegas, NV). RNA samples to be sequenced had acceptable purity (260/280 ratio of approximately 1.9 or greater) and integrity (RIN of approximately 5 or greater) prior to RNA sequencing.

RNA sequencing

For newly processed samples, the Illumina TruSeq Stranded Total RNA Gold Library Preparation Kit (Illumina, San Diego, CA) was used for library preparation. Samples were subjected to 50-cycle, single-read sequencing in a HiSeq2500 system (Illumina) and demultiplexed using Bcl2Fastq. To increase sample size, we combined reads from these new samples with reads from our previous samples[16, 17] that were sequenced in the same manner, but for which libraries were constructed using the Illumina TruSeq RNA library Prep Kit V2 (Illumina, San Diego, CA). For all data, per-cycle base-call (BCL) files generated by the Illumina HiSeq2500 were converted to per-read FASTQ files using bcl2fastq version 2.20.0.422 using default parameters. The quality of the sequencing was reviewed using FastQC version 0.11.5. Detection of potential contamination was done using FastQ Screen version 0.11.1. FastQC and FastQ Screen quality reports were summarized using MultiQC version 1.5. No adapter sequences were detected, so no trimming was performed. Genomic alignments were performed using HISAT2 version 2.1.0 using default parameters. NCBI reference GRCh38 was used for the reference genome and gene annotation set. Sequence alignments were compressed and sorted into binary alignment map (BAM) files using samtools version 1.3. Counting of mapped reads for genomic features was performed using Subread featureCounts version 1.6.2 using the parameters -s 2 –g gene_id –t exon –Q 60, the annotation file specified with –a was the NCBI GRCh38 reference from Illumina iGenomes. Aggregate quality control data (i.e. alignment statistics and feature assignment statistics) were again summarized using MultiQC.

Differential expression analysis and data exploration

Before implementing our machine learning pipeline, we performed differential expression analysis on the whole dataset to identify transcripts that were significantly differentially expressed in IA using Bioconductor package edgeR version 3.24.0. After estimating dispersion, edgeR identified differentially expressed genes by using a negative binomial distribution with generalized linear models and a quasi-likelihood F-test to identify differentially expressed genes.[20, 21] We incorporated the two sequencing batches into the design matrix to correct for any potential batch effects due to different library preparation kits. Genes with a counts sum > 0 across all samples were used as input. Multiple hypothesis testing correction was performed using Benjamini-Hochberg false discovery rate (FDR) correction.[22] Transcripts with an FDR-corrected p-value (q-value) < 0.05 were considered significantly differentially expressed. To explore how transcriptomes separated patients with and without IA on a broad scale, we performed hierarchical clustering, using the hclust package in R under default settings (complete linkage).

Verification of expression differences by qPCR in a sub-cohort

To verify expression differences in differentially expressed genes, quantitative polymerase chain reaction (qPCR) was performed. Due to limitations in RNA quantity, qPCR was performed on 10 transcripts in a subset of 50 of the 134 samples (20 IA and 30 control). We followed the protocol described previously.[16] In brief, oligonucleotide primers were designed using Primer3 software (Primer3Web 0.4.0) and Primer BLAST (NCBI, Bethesda, MD) to have a 60 °C melting temperature, a length of 15–25 nucleotides, and a

product of 50–250 base pairs (with at least one primer that spans an exon-exon junction), as well as an estimated efficiency greater than 0.8.[23] Primer sequences, annealing temperatures, efficiencies, and product lengths are reported in Supplemental Table 1. For reverse transcription, first-strand cDNA was generated from total RNA using qScript cDNA Synthesis kit (Quantabio, Beverly, MA, USA) according to the manufacturer's directions. qPCR was run with 5 ng of cDNA in 20 µL reactions in triplicate in Bio-Rad CFX Connect (Bio-Rad, Hercules, California) using the qScript One-Step SYBR Green Master Mix kit (Quantabio, Beverly, MA, USA) and gene-specific primers at a concentration of 0.02 µM each. The temperature profile consisted of an initial step of 95 °C for 1 min, followed by 40 cycles of 95 °C for 15 seconds and 60 °C for 1 min, and then a final melting curve analysis from 60 °C to 95 °C over 20 min.

Gene-specific amplification was demonstrated by a single peak using the Bio-Rad dissociation melt curve. Samples were normalized based on HPRT1 (a housekeeping gene[24]) expression, which was run in parallel reactions to the genes of interest. These values were used to calculate average fold-change between the two groups using the $2^{-\Delta\Delta Ct}$ method for RNAseq and qPCR data. Fold-change in gene expression measured by qPCR data was compared to the fold-change calculated from RNA sequencing on the same samples to determine if fold-change in expression was in the same direction and statistically different (Student's t-test, significance at p-value < 0.05).

Feature selection for classification model development

To build predictive models for IA, we began with raw counts to eliminate bias and uncertainty associated with distribution modeling incorporated in edgeR. Raw counts were then normalized to transcript per million (TPM) values to facilitate comparison of expression between samples by normalizing by both gene length and sequencing depth. Then, we applied an abundance filtering by only selecting protein coding genes with average TPM > 1 across all samples, reducing the set of potential transcripts to 18,833. To account for the two sequencing batches in our study design (as done in edgeR analyses), we performed batch effect correction using ComBat under the default settings in R.[25, 26] Then, 70% of samples were randomly allocated to a training cohort (n = 39 IA and n = 55 control) and 30% to a testing cohort (n = 16 IA and n = 24 control), maintaining the proportion of IA and controls.

For feature selection in the training cohort, we performed a supervised feature selection by using the Hilbert-Schmidt Independence Criterion Least Absolute Shrinkage and Selection Operator (HSIC LASSO) method. HSIC LASSO was implemented in the ComBat corrected dataset to select features for the model. To visualize how those selected transcripts separated samples from patients with and without IAs, we performed principal component analysis (PCA) using the prcomp package under the default settings.[27]

Model training

We used MATLAB Statistics and Machine Learning Toolbox (MathWorks, Natick, MA) to train 4 popular algorithms (K-Nearest Neighbor, Random Forest, Support Vector Machine with Gaussian and cubic kernels) on 2 different gene panels – the 37 transcripts identified by LASSO in this study and the 26 transcripts identified by filtering in our previous study. While these algorithms have been used in other

disease classification applications[28 – 33], we implemented all 4 algorithms to determine which best suited our data. Specific parameters for each algorithm are as follows:

- For K-Nearest Neighbors, we used a Euclidean metric and 10 neighbors (k). The resulting model classified test samples by calculating their distance to each training sample and the test sample labels were predicted by choosing the class that was most common among their 10 nearest neighbors.
- For Random Forest, which constructs a multitude of decision trees in training and outputs the mode of the classes as the predicted class,[34] we set the number of trees as 1000. The Random Forest was built by contrasting a multitude of decision trees based on subsets of the training data generated by random sampling with replacement and the resulting model classified testing samples by the majority vote of the decision trees.
- For Support Vector Machines, we used two different kernels[35], Gaussian and cubic. To separate a binary-labeled sample, the Support Vector Machine transforms them into a multidimensional space using the kernel, and then a hyper-plane, which maximizes the distance to samples of either class, is established. The resulting model classified testing samples by transforming them into a higher dimensional space with the corresponding kernel and making decisions based on their signed distance to the hyper-plane.

Model assessment

We estimated the performance of each model for the new LASSO-identified features as well as our previously-identified 26 features by a leave-one-out (LOO) cross-validation within the training cohort. Model predictions were compared to each patient's clinical diagnosis from imaging, and the true positives, true negatives, false positives, and false negatives were tallied. We then calculated each model's sensitivity, specificity, and accuracy, as described elsewhere.[17] Based on model predictions, we created receiver operating characteristic (ROC) curves and calculated the area under the ROC curve (AUC) to assess model performance. Additionally, to gauge predictive value of models, we determined positive predictive value (PPV) and negative predictive value (NPV). PPV and NPV were estimated using formulas based on Bayes' theorem as previously described[17] with 5% aneurysm prevalence, which is within the range of IA prevalence reported in the literature (3.2-7%[36–39]). The classification models were then independently evaluated in the testing cohort (n = 40), and classification results were compared to clinical diagnoses to calculate the true sensitivity, specificity, and accuracy for each model. ROC curves were constructed and AUCs, along with PPV and NPV, were used to assess the performance of each classifier. This was performed for algorithms trained on LASSO-selected features and the previous 26 features.

Testing influence of covariates on gene expression differences

While we randomly assigned samples to training or testing cohorts, this study was not cohort-controlled and used a large, heterogeneous population. Consequently, it is possible that factors other than IA status,

such as demographics or comorbidities, could be affecting differential expression and model performance. To determine if patient characteristics influenced model performance, we first performed a chi-square test to determine if there were different rates in the aneurysm and control populations. We examined gender, hypertension, heart disease, stroke, high cholesterol, cancer, diabetes, arthritis, asthma, smoking status, and age. Additionally, we performed covariate matching in the MatchIt program in R to create 6 subclasses under default settings with similar distribution of covariates (age [60 and under vs over 60], sex, smoking status [non-smoker vs current smoker], hypertension, heart disease, stroke history, high cholesterol, cancer, diabetes, arthritis, asthma, and IA family history) for aneurysm and control populations.[40, 41] To create subclasses, we used a distance measure determined by a logistic regression model to estimate the propensity score. We then examined misclassification rate in each of the subclasses to determine if any group with a specific “covariate profile” was associated with greater misclassification.

Bioinformatics

Gene ontology enrichment analysis was performed using Gene Ontology enRIchment anaLysis and visuaLizAtion tool (GORILLA).[42] A background list of neutrophil expression from 3 healthy individuals (average fragments per kilobase million > 0.5) was used to compute hypergeometric statistics and assign significance to GO terms.[18] GO functions and processes with a p-value < 0.001 were reported. Ingenuity Pathway Analysis (IPA) software (Qiagen Inc., https://www.qiagenbioinformatics.com/products/ingenuity-pathway- analysis) was used to investigate networks associated with the differentially expressed genes identified by edgeR ($q < 0.05$, fold-change > 2) and those selected by LASSO during feature selection. Each gene identifier was mapped to its corresponding gene object in the Ingenuity Knowledge Base and overlaid onto a molecular network derived from information accumulated in the Knowledge Base. Gene networks were algorithmically generated based on their “connectivity” derived from known interactions between the products of these genes. Networks were considered significant if their p-scores were ≥ 15. Network score is calculated as p-score=-log10(p-value), so a score of 15 corresponds to a p-value of 1E-15.[43]

Results

Study Population

We obtained and analyzed an additional 91 samples from individuals undergoing cerebral DSA that met data and RNA quality criteria. Combined with the 43 samples we previously analyzed, our total dataset was 134 neutrophil transcriptomes – 55 from patients with IA and 79 from control patients. The characteristics of study population are presented in Table 1; detailed aneurysm characteristics in Supplemental Table 2. Overall, the 134 samples had an average 260/280 ratio of 2.04 and an average RIN of 6.7 (quality data reported in Supplemental Table 3). Patients with IA had 73 aneurysms (as 12 individuals had multiple IAs), which ranged in size from 1 mm to 19 mm measured by largest diameter on 2D images.

Table 1

Clinical characteristics of training and testing cohorts.*

	Training Cohort		Testing Cohort	
	Control (n = 55)	Aneurysm (n = 39)	Control (n = 24)	Aneurysm (n = 16)
Age (Mean ± SE)	62 ± 2.0	61 ± 1.7	59 ± 2.9	57 ± 3.3
Age [Median (Q1/Q3)]	66 (54/72)	60 (54/68)	59 (54/68)	58.5 (49.25/63.25)
Sex (% of patients)				
Female	56.36%	69.23%	50%	75%
Smoker (% of patients)				
Yes	10.91%	26.64%	20.83%	43.75%
Comorbidities (% of patients)				
Hypertension	61.82%	53.85%	54.17%	50%
Heart Disease	30.91%	23.08%	25%	18.75%
High Cholesterol	52.73%	48.72%	62.50%	50%
Stroke History	12.73%	10.26%	25%	0%
Diabetes	29.09%	17.95%	8.33%	31.25%
Arthritis	16.36%	30.77%	16.67%	18.75%

*Clinical characteristics of the randomly-created training and testing cohorts. With the exception of age, these factors were quantified as binary data points. The clinical factors were retrieved from the patients' medical records via the latest "Patient Medical History" form administered prior to imaging.

Differential RNA expression in neutrophils from patients with IA vs. controls

RNA sequencing data were used to identify differentially expressed neutrophil transcripts between IA and control groups. Overall, our sequencing experiments had an average of 55.06 million reads per sample and a 95% read mapping rate (or % aligned), as reported in Supplemental Table 3. The scatter plot in Fig. 1A shows neutrophil expression differences between IA patients and controls in terms of average fold-change in expression and significance level. Differential expression analysis in edgeR identified 65 transcripts that were significantly differentially expressed ($q < 0.05$, fold-change > 2) (red and blue points in Fig. 1B). Twenty-three genes showed lower expression in the IA group, and 42 showed higher expression. Using all transcriptome data, we performed supervised hierarchical clustering to determine if gene expression in general could also discriminate patients with IAs from controls. On the dendrogram in

Fig. 1C, samples from IA and control groups are separated. The dendrogram shows 7 clusters of primarily IA or control samples (highlighted sections). Overall, hierarchical clustering congregated 73% of the samples with their respective groups.

To gain biological insight into the observed neutrophil RNA expression differences between IA and control groups, we performed bioinformatics analyses using gene set enrichment analysis and physiological pathway modeling. We used GORILLA to analyze the ontologies associated with the edgeR genes with increased and decreased expression in IA as compared to background of healthy individuals. Genes with increased expression in IA were associated with cell migration, cell motility, T cell migration, and lymphocyte migration processes. On the other hand, genes with decreased expression in IA had functions related to sodium channel activity, ion channel activity, and gated channel activity, as well as signaling and regulation of membrane potential processes. A full list of ontologies associated with edgeR genes is reported in Supplemental Table 4. IPA gene network analysis identified 3 significant networks with p-scores of 21, 21, and 15, respectively (Fig. 5). The first network was associated with cell morphology, cell-to-cell signaling and interaction, nervous system development and function, with hubs around IRS1, GRIK1, GRIN2A, and L-glutamic acid. The second network is associated with connective tissue development and function, dermatological diseases and conditions, organismal injury and abnormalities, with hubs of ELAVL1, CCND1, and FMOD. Finally, the last network was enriched for cell death and survival, connective tissue disorders, and inflammatory disease functions, reflected by a predominant hub being TNF. Associated molecules and diseases/functions for each network are listed in Supplemental Table 5.

Verification of expression differences by RT-qPCR

We confirmed expression differences of 10 prominent IA-associated transcripts (C1QL1, GPR15, PDE9A, HES4, PVRL2, CD163, CYP1B1, CDH2, ZBTB16, PTGDS) using RT-qPCR. These genes were selected because they were prominently differentially expressed transcripts, i.e. were in the models we trained, were highly abundant in at least one cohort, or were significantly differentially expressed. This confirmation was performed in a subset of 50 patient samples; however, one IA sample did not provide sufficient data for analysis across all genes and so that data is not included here. Figure 3 demonstrates that the expression differences between patients with and without IA were of the same direction and of similar magnitudes when calculated by both RNA sequencing and RT-qPCR, with the exception of PDE9A and PVRL2. In the case of PVRL2, this may be in part due to poor primer efficiency (see Supplemental Table 1).

Selected transcripts for model training

Feature selection using LASSO identified 37 IA-associated transcripts with significant expression in the training cohort, which were used to create models with 4 machine learning algorithms. Table 2 reports gene-specific accuracy, sensitivity, and specificity of the 37 model genes. The PCA in Fig. 4A illustrates these transcripts' ability to clearly separate aneurysm samples from control in training cohort. Compared

to the PCA using the 26 previously identified genes in Fig. 4D, it is visually evident that the transcripts identified by LASSO were able to better separate IA and control groups.

Table 2

The 37 transcripts selected for classification model training.*

			Training Cohort			Testing Cohort		
Gene	Gene ID	Accession #	Acc.	Sen.	Spe.	Acc.	Sen.	Spe.
AC011380.1	–	AC011380	0.56	0.64	0.51	0.58	0.69	0.50
C1QL1†	10882	NM_006688.5	0.80	0.56	0.96	0.93	0.88	0.96
CCDC42B	387885	NM_001144872.2	0.59	0.05	0.96	0.65	0.19	0.96
CEP295NL†	100653515	NM_001243541.1	0.71	0.59	0.80	0.80	0.56	0.96
CERS4†	79603	NM_024552.3	0.79	0.72	0.84	0.85	0.88	0.83
CLP1	10978	NM_006831.3	0.57	0.00	0.98	0.58	0.00	0.96
DCUN1D1	54165	NM_020640.3	0.57	0.00	0.98	0.60	0.00	1.00
EIF4EBP3	8637	NM_003732.3	0.47	0.26	0.62	0.68	0.69	0.67
FLT1	2321	NM_002019.4	0.52	0.00	0.89	0.50	0.00	0.83
GBGT1†	26301	NM_021996.6	0.71	0.74	0.69	0.70	0.88	0.58
GPR15†	2838	NM_005290	0.79	0.79	0.78	0.93	1.00	0.88
GPR157	80045	NM_024980.5	0.57	0.00	0.98	0.60	0.00	1.00
GTF2B	2959	NM_001514.6	0.55	0.03	0.93	0.55	0.00	0.92
HBB	3043	NM_000518.5	0.79	1.00	0.64	0.58	0.94	0.33
HIST1H4E	8367	NM_003545.3	0.57	0.00	0.98	0.60	0.00	1.00
HIST2H2AB	317772	NM_175065.2	0.57	0.00	0.98	0.60	0.00	1.00
ISY1	57461	NM_001199469.1	0.60	0.05	0.98	0.65	0.13	1.00
KIAA1324	57535	NM_020775.5	0.59	0.05	0.96	0.55	0.06	0.88
KIAA1614	57710	NM_020950.2	0.63	0.15	0.96	0.63	0.06	1.00
LOC100129697	100129697	NM_001290330.2	0.65	0.26	0.93	0.58	0.06	0.92
LOC105377284	105377284	XR_938891.2	0.59	0.05	0.96	0.65	0.13	1.00
LRRN3†	54674	NM_001099658.2	0.78	0.79	0.76	0.85	1.00	0.75
MFSD6L	162387	NM_152599.3	0.67	0.59	0.73	0.65	0.50	0.75

*We show their per-transcript performance in the training and testing dataset. Transcripts with high accuracy (> 0.70) in both training and testing cohorts are denoted by †.

			Training Cohort			Testing Cohort		
MORC3	23515	NM_015358.3	0.57	0.00	0.98	0.60	0.00	1.00
MTRNR2L1	100462977	NM_001190452.1	0.57	0.03	0.96	0.60	0.00	1.00
NECAB1	64168	NM_022351.5	0.62	0.15	0.95	0.58	0.06	0.92
NEIL3	55247	NM_018248.3	0.57	0.00	0.98	0.60	0.00	1.00
PDCD10	11235	NM_007217.4	0.57	0.00	0.98	0.60	0.00	1.00
PGM5	5239	NM_021965.4	0.55	0.00	0.95	0.50	0.00	0.83
RFFL	117584	NR_037713.1	0.57	0.00	0.98	0.58	0.00	0.96
SDCBP2	27111	NM_080489.5	0.71	0.33	0.98	0.65	0.31	0.88
SMIM8	57150	NM_001042493.3	0.57	0.03	0.96	0.50	0.00	0.83
SYP	6855	NM_003179.2	0.59	0.05	0.96	0.63	0.13	0.96
TGS1[†]	96764	NM_024831.7	0.80	0.79	0.80	0.78	0.94	0.67
TMC4	147798	NM_001145303.2	0.64	0.49	0.75	0.63	0.44	0.75
USF1	7391	NM_007122.5	0.59	0.03	0.98	0.60	0.06	0.96
UTY	7404	NM_182660.1	0.57	0.00	0.98	0.60	0.00	1.00

*We show their per-transcript performance in the training and testing dataset. Transcripts with high accuracy (> 0.70) in both training and testing cohorts are denoted by [†].

To investigate how the biological underpinnings of IA specifically influence the genes selected by LASSO, we performed bioinformatics analyses using only the 37 new panel genes (n = 23 with decreased expression in IA group, n = 14 with increased expression in IA group). Model genes with decreased expression in the IA group were associated with negative regulation of execution phase of apoptosis, negative regulation of endothelial cell proliferation, and regulation of execution phase of apoptosis (Supplemental Table 6). However, model genes with increased expression in the IA group did not return any significant functions or processes. Two networks using all 37 genes produced by IPA had significant p-scores (47, 25). The first network showed hubs around TNF and MMP3 and was associated with cancer, cellular movement, and connective tissue disorders. The second had hubs around HBB and MAPK and was associated with cell cycle, cellular assembly and organization, DNA replication, recombination, and repair. We note TNF was incorporated in networks generated using both edgeR and LASSO gene sets. See Fig. 5 and Supplemental Table 7 for details on these networks, including associated molecules and top diseases and functions.

Predictive models of IA have high performance and high NPV in training and testing

Sensitivity, specificity, accuracy, NPV, and PPV estimated by LOO cross-validation in the training cohort are reported in Fig. 4B for models using the new 37-transcript panel. Each classification method achieved high performance, with accuracies that ranged from 0.85 to 0.91. Evaluation by ROC curve analysis showed a range in AUCs from 0.95 to 0.98 (Fig. 4C) across all methods. All models had high NPV of approximately 1 (0.98-1). Random Forest outperformed K-Nearest Neighbor and both Support Vector Machine algorithms, with a sensitivity of 0.87, specificity of 0.95, accuracy of 0.92, AUC of 0.98, NPV of 0.99, and PPV of 0.46. Figure 4E reports performances of the 4 classification models trained with the 26 previously-identified genes. Sensitivity, specificity, accuracy, NPV, and PPV were estimated by LOO cross-validation in the training cohort. AUCs ranged from 0.71–0.92, as shown in Fig. 4F. Overall performance in the training cohort was superior using the transcripts selected by LASSO; all metrics (accuracy, sensitivity, specificity, AUC, 5% PPV, 5% NPV) when averaged across the 4 models were greater using the new 37-gene panel.

As in the training cohort, PCA in the testing cohort (see Fig. 6A) shows that the 37 transcripts could discriminate patients with IAs from controls. The separation between classes was more obvious using the 37 newly-identified transcripts than the 26 previously-identified transcripts (Fig. 6D). Using the 37 features selected by LASSO, the models predicted aneurysm status in the testing cohort with accuracies ranging from 0.83 to 0.90 (Fig. 6B). The ROC analysis in Fig. 6C shows that model AUCs ranged from 0.95 to 0.99. In the testing cohort, the Random Forest model again performed well, with a sensitivity of 1.0, specificity of 0.75, accuracy of 0.85, and AUC of 0.99. The performance of the previously identified 26-gene panel (Fig. 6E, F) was similar to that of the 37-gene panel in the testing cohort with accuracies ranging from 0.83 to 0.93 and AUCs of 0.84–0.97. While average accuracy for the 4 models was the same (86%) using the 37-gene panel identified by LASSO and the 26-gene panel previously identified, the models using the 37 LASSO features had greater average AUC (0.97 vs 0.91).

Presence of covariates and effect on model performance

Table 3 shows the rates of demographics and comorbidities in aneurysm and control populations. Only smoking was significantly higher in the IA population ($p = 0.017$), which can be expected as it is a well-known risk factor.[44] We created 5 subclasses using MatchIt as there were too few samples in the 6th subclass. Misclassification by the 37-gene prediction model for each subclass ranged from 8–19%, indicating no one subclass could be driving misclassification.

Table 3
Clinical characteristic differences in entire population.*

	Control (n = 79)	Aneurysm (n = 55)	Chi-square test
Age (Mean ± SE)	61 ± 1.7	60 ± 1.5	(age 60 cutoff) 0.243
Age [Median (Q1/Q3)]	65 (54/72)	60 (54/67)	
Sex (% of patients)			
Female	54.43%	70.91%	0.054
Smoking (% of patients)			
Current	13.92%	30.91%	0.017†
Comorbidities (% of patients)			
Arthritis	16.46%	27.27%	0.130
Asthma	7.59%	18.18%	0.063
Cancer	11.39%	9.09%	0.668
Diabetes	22.78%	21.82%	0.895
Heart Disease	29.11%	21.82%	0.344
High Cholesterol	55.70%	49.09%	0.451
Hypertension	59.49%	52.73%	0.437
IA Family History	7.69%	12.73%	0.336
Stroke History	18.99%	9.09%	0.114
*None of the reported covariates were significantly different in either group (chi-square test < 0.05) except for smoking†.			

Discussion

More robust machine learning strategy improves biomarker performance

In this study, we implemented a new machine learning strategy for IA biomarker discovery, which consisted of a larger dataset (94 training, 40 testing), LASSO for feature selection, and more robust algorithms, K-Nearest Neighbor, Random Forest, and Support Vector Machine with cubic and Gaussian kernels. Our larger dataset and LASSO feature selection led to a new panel of 37 genes to use in IA predictive models. Two genes of these 37 genes, C1QL1 and TGS1, were also in our previously-discovered

26-gene panel. The new learning algorithms trained using the 37 genes all performed very well in the testing cohort with accuracies of 0.83–0.90 and AUCs of 0.95–0.99, a marked increase over our previous algorithms. Interestingly, all 4 new models had an NPV of 1, indicating that in the testing dataset there were no false negatives. This may be important for future applications of these biomarkers as a prescreen, since false negatives would be particularly deleterious.

To examine how the increased sample size and improved algorithms affected model performance, we retrained the previous 26-gene panel using the new algorithms in the current, larger dataset. The performance of the retrained models in the testing set (n = 40) using the 26-gene panel improved from our previous study with accuracies ranging from 0.83–0.93 and AUCs of 0.84–0.97. Despite this increase in performance using the new algorithms, models using the previously identified 26 genes still fell short of those using the newly identified 37 genes; the average testing AUC using 26 genes was 0.91 compared to 0.97 when using 37 genes. This suggests that the 37 features identified by LASSO are more reliable for IA prediction than the 26 selected by filtering in our last study.

We believe that improved IA prediction can be attributed to our increased sample size, which afforded several advantages. First, it allowed us to use LASSO to identify features instead of simple filtering methods. Thresholding filters like we used in our last study consider each gene independently, which can neglect groups of genes that function together in pathophysiologic mechanisms and could be useful as a biomarker. Filtering methods can also select highly correlated, redundant genes, which can increase the number of features required to make accurate predictions. HSIC LASSO, a nonlinear feature selection method, overcomes these issues and identifies combinations of non-redundant genes with strong dependence on disease status. Implementing LASSO in the training dataset identified 37 unique IA-associated genes, two of which (C1QL1, TGS1) had also been identified as part of the 26-gene panel in our past expression profiling study.[17] The identification of non-redundant features may be one reason why the biomarkers created in this study outperform our past efforts, as some of the 26 features (with the exception of C1QL1 and TGS1) may have ultimately been uninformative for classification.

Secondly, a larger sample size also enabled us to leverage more complex machine learning models, namely Support Vector Machine and Random Forests which perform better in larger datasets.[45] In our previous effort we did implement Support Vector Machine, but only achieved a testing accuracy of 0.70, possibly because the training dataset contained only 30 patients.[17] In this this larger study we were able to achieve an accuracy of 0.85 for Support Vector Machine (Gaussian kernel). Nevertheless, we found that in our data Random Forest consistently performed the best, with a testing accuracy of 0.85 and AUC of 0.99. Both Random Forests and K-Nearest Neighbors are weighted neighbors schemes. However, the K-Nearest Neighbors algorithm may have had poorer performance because this classifier simply uses the training data for prediction instead of learning a discriminative rule. The performance of the K-Nearest Neighbors classifier is reliant on the quality of the training data, which in the case of transcriptomes derived from human samples may be noisy. However, this problem is well-solved in Random Forest. Through the random sampling process, Random Forest handles outliers by binning them. Also, by averaging the decision trees, the Random Forest method provides a low bias and moderate variance

model, which improves the generalizability of the output model. In other words, Random Forest not only attains a good performance in the training data but also performs well in unknown (testing) data. And while Support Vector Machine performed well here, Random Forest likely surpassed Support Vector Machine by avoiding overfitting and achieving better predictive power.

We note that increasing sample size may have introduced more variability in our data due to a larger, heterogeneous population that was not cohort-controlled. For example, in our entire population we found smoking was significantly higher in patients with IA ($\chi = 0.017$), which may be because smoking is a well-known risk factor for IA formation and rupture.[46–48] Indeed two genes in our model, LRRN3 and GPR15, are among the top differentially expressed genes in blood between current and never smokers according to a meta-analysis by Huan et al.[49] Their presence in our predictive model may be because of the higher proportion of smokers in IA group or because these genes are capturing biological mechanisms related to smoking that are important in IA pathogenesis, such as endothelial dysfunction.[50–52] Still, when we performed covariate analysis using MatchIt to create subgroups with similar distributions of covariates between IA and control groups, we found that no one subgroup had significantly higher misclassification rates. For instance, 61% of all subjects in “Subclass 5” were smokers, and this subgroup had a misclassification rate of 13%. Yet, “Subclass 1”, which had 0% smokers, had a misclassification rate of 14%. These results suggest that our prediction models may not be affected greatly by covariate imbalance, albeit testing this in even larger cohorts will be needed to confirm these results.

Complex role of circulating neutrophils in intracranial aneurysm

Inflammation is widely-recognized to play a central role in the pathophysiology of IA.[53–55] It is commonly thought that in IA neutrophils are recruited to the sac, where they infiltrate the wall and coordinate the inflammatory responses.[53, 56, 57] In this study, gene ontology enrichment analysis showed that genes with higher expression in IA identified by edgeR in the entire dataset were related to cell migration and lymphocyte migration ontologies. These processes, which were also observed in neutrophils from patients with IAs in our previous studies[16, 17], increase upon peripheral activation and prompt inflammatory cell migration and infiltration of diseased tissue.[53, 58] IPA analysis mirrored these results, showing 3 significant networks, 2 of which were involved in activation-related processes: cell-to-cell signaling and interaction, and inflammatory disease function. Interestingly, one of the largest nodes of gene connectivity in all the networks was TNF, a proinflammatory cytokine with many functions including regulation of cell proliferation and apoptosis. TNF has been shown to have a mechanistic role in IA formation in animal models,[59] and an increased presence in human IA tissue compared to superficial temporal artery control tissue.[60] In this network, TNF has a predicted connection to DEFA1, which was significantly elevated in neutrophils of IA patients. Higher levels of this cytotoxic defensin protein that is contained within neutrophil granules have been reported in IA tissue, suggesting that production of this protein may occur peripherally before neutrophils enter the IA wall.[61] Here the molecules CCR4 and CCR6

(receptors of MIP-3 alpha and MIP-1, RANTES, CCL17, and MCP-1, respectively) were also related to the TNF node. We suspect that these receptors, which play a role in dendritic and T cell migration and recruitment during inflammation,[62] may coordinate inflammatory cell migration once expressed in aneurysm tissue.

We also observed the dysregulation of inflammation and a potential role of TNF in our bioinformatics analyses of model genes selected by LASSO in the training dataset. TNF was a hub of connectivity in networks created using the LASSO genes. In these networks, we observed an indirect relationship between TNF and the complement system (i.e. C1QTNF1), which is also associated with C1QL1 (one of 37 model genes). This may be because complement activation plays a critical role in the inflammatory response,[63] has been implicated in IA wall degradation and rupture,[64] and involves proteins that are increased in human IA tissue (including CFB, CFH, C1Q, and C3AR1[65]). We suspect that the complement alternative pathway may be one mechanism through which neutrophils become activated as it can amplify activation through a positive feedback mechanism.[66] In addition to complement members, the TNF node was also related to CD44, a cell surface glycoprotein critical to neutrophil recruitment during inflammation. Because neutrophils interact with CD44, PSGL-1, and E-selectin ligand 1 as they roll along activated endothelial cells, this result may reflect neutrophils transmigrating into inflamed endothelium.[67] Our data shows TNF may also interact with the transcription factor TP53, a node with connections to numerous molecules, many of which have decreased expression. TP53 plays a variety of roles in inflammation, such as acting on the NF- κB pathway.[68] Overall, our bioinformatics analyses of genes selected by LASSO, while not overlapping greatly with the differentially expressed genes selected by edgeR in the entire dataset (with the exception of C1QL1, GRP15), show that the biology of neutrophil activation and inflammation responses are captured by the IA prediction model gene panel.

In addition to neutrophil activation and heightened inflammatory signaling, we observed other aberrant neutrophil functions not specifically characterized in IA, including our previous studies.[16, 17] In genes identified in the whole dataset by edgeR, gene ontology enrichment analysis showed that the differentially expressed genes with decreased expression in IA had functions related to sodium channel activity, ion channel activity, and gated channel activity, as well as signaling and regulation of membrane potential processes (ASIC2, GRIK3, SCN5A). GRIK3, glutamate receptor 7, is particularly interesting as glutamate is a chemotactic factor for neutrophils after injury or infection.[69] Glutamate binding to its receptors can trigger release of cytokines and MMPs and can activate immune responses, all critical processes in IA.[70, 71] Future studies are needed to better understand how these channel activities impact IA pathogenesis.

New ontologies were also captured using the genes identified by LASSO in the training dataset. Using the LASSO genes with lower expression in IA, we found dysregulation of apoptosis as gene ontology enrichment analysis reported both negative regulation of execution phase of apoptosis and regulation of execution phase of apoptosis. These were associated with MTRNR2L1, a neuroprotective and antiapoptotic factor[72], and RFFL, which is related to TNF signaling.[73] Dysregulated MTRNR2L1

expression may be responsible for increasing the lifespan of neutrophils. Increased lifespan provides further evidence of neutrophil activation in IA. We note that TP53, previously discussed, also induces apoptosis.[74, 75] These results are echoed in the blood profiling study of IA published by Jin et al.[76] They reported hsa-miR-21, an upregulated miRNA in IA serum, induces apoptosis by extracellular signals, potentially triggering more apoptotic reactions to facilitate the medial thinning and destructive remodeling, a hallmark of IA pathogenesis.[77–80] Overall, we suspect that capturing neutrophil activation and inflammation responses involved in IA is the reason why the 37-gene panel was able to detect IA.

Limitations

In this study, we increased sample size from our previous study by adding 94 samples to 40 samples we previously analyzed. However, these two batches used different versions of the Illumina kit for library preparation, which necessitated the implementation of batch effect correction that could potentially have introduced bias or skewed our dataset.[81] Secondly, all samples were recruited from patients receiving cerebral imaging at a single center, which may introduce selection bias. Future studies are needed to validate our predictive models using broader patient populations from multiple centers. Thirdly, inflammatory or vascular diseases other than IA could affect model prediction. Larger studies with multiple control groups of individuals with other vascular and inflammatory conditions are needed to refine our model.

Conclusions

We improved IA predictive model performance from circulating neutrophil transcripts by using LASSO for feature selection and powerful machine learning techniques in a large dataset. The Random Forest algorithm performed the best with a testing AUC of 0.99. Bioinformatics using all 134 samples implicated inflammation through TNF and neutrophil activation as key processes in IA. IPA networks using the 37 LASSO-selected genes also reflected these increased inflammatory and signaling pathways. Comorbidities and demographics did not significantly affect IA prediction. Future studies are needed to validate these predictive models.

Abbreviations

AUC: area under ROC curve, BAM: binary alignment map, BCL: per-cycle base-call, CTA: computed tomography angiography, DSA: digital subtraction angiography, ECM: extracellular matrix, FDR: false discovery rate, GO: gene ontology, GORILLA: Gene Ontology enRIchment anaLysis and visuaLizAtion, HSIC: Hilbert-Schmidt independence criterion, IA: intracranial aneurysm, IPA: Ingenuity pathway analysis, LASSO: least absolute shrinkage and selection operator, LOO: leave one-out, MPO: myeloperoxidase, MRA: magnetic resonance angiography, MRI: magnetic resonance imaging, NGAL: neutrophil gelatinase-associated lipocalin, NPV: negative predictive value, PCA: principal component analysis, PPV: positive predictive value, qPCR: quantitative polymerase chain reaction, RIN: RNA integrity, ROC: receiver-

operating-characteristic, ROS: reactive oxygen species, RT-qPCR: real-time quantitative polymerase chain reaction, TPM: transcript per million

Declarations

Ethics approval and consent to participate: This study was approved by the University at Buffalo Health Sciences Institutional Review Board (study no. 030-474433). All methods were carried out in accordance with approved protocol and informed consent was obtained from all subjects.

Consent for publication: No specific patient information or images are given in the manuscript.

Availability of data and materials: The datasets used in the current study are available from the corresponding author on reasonable request.

Competing interests: KEP, LL, MW, AJ, LC, KJ, YS, JK—None.

VMT— Principal investigator: National Science Foundation Award No. 1746694, and the Brain Aneurysm Foundation grant, Center for Advanced Technology grant, and Cummings Foundation grant mentioned above. Co-founder: Neurovascular Diagnostics, Inc.

JNJ—Principal Investigator: NIH Grant R01-AR-060604.

KVS—Consulting and teaching for Canon Medical Systems Corporation, Penumbra Inc., Medtronic, and Jacobs Institute. Co-Founder: Neurovascular Diagnostics, Inc.

EIL—Intratech Medical Ltd., NeXtGen Biologics. Principal investigator: Medtronic US SWIFT PRIME Trials. Honoraria–Medtronic. Consultant–Pulsar Vascular. Advisory Board-Stryker, NeXtGen Biologics, MEDX. Cognition Medical. Other financial support—Abbott Vascular for carotid training sessions.

AHS—Financial Interest/Investor/Stock Options/Ownership: Amnis Therapeutics, Apama Medi- cal, BlinkTBI, Inc, Buffalo Technology Partners, Inc., Cardinal Health, Cerebrotech Medical Systems, Inc, Claret Medical, Cognition Medical, Endostream Medical, Ltd, Imperative Care, International Medical Distribution Partners, Rebound Therapeutics Corp., Silk Road Medical, StimMed, Synchron, Three Rivers Medi- cal, Inc., Viseon Spine, Inc. Consultant/Advisory Board: Amnis Therapeutics, Boston Scientific, Canon Medical Systems USA, Inc., Cerebrotech Medical Systems, Inc., Cerenovus, Claret Medical, Corindus, Inc., Endostream Medical, Ltd, Guidepoint Global Consulting, Imperative Care, Integra, Medtronic, Micro-Vention, Northwest University—DSMB Chair for HEAT Trial, Penumbra, Rapid Medical, Rebound Therapeutics Corp., Silk Road Medical, StimMed, Stryker, Three Rivers Medical, Inc., VasSol, W.L. Gore & Associates. National PI/Steering Committees: Cerenovus LARGE Trial and ARISE II Trial, Medtronic SWIFT PRIME and SWIFT DIRECT Trials, MicroVention FRED Trial & CONFIDENCE Study, MUSC POSITIVE Trial, Penumbra 3D Separator Trial, COMPASS Trial, INVEST Trial.

HM— Principal investigator: NIH Grants R01-NS-091075 and R01-NS-064592. Grant support: Canon Medical Systems. Co-founder: Neurovascular Diagnostics, Inc.

Funding: This work was funded by National Science Foundation Award No. 1746694 - through Neurovascular Diagnostics, Inc., the Brain Aneurysm Foundation, the New York State Center for Advanced Technology in Big Data and Health Sciences, and the Cummings Foundation.

Authors' contributions: Study conception and design: KEP, VMT, JNJ, JK, HM. Acquisition of data: KEP, VMT, MW, AJ, LC, KVS, EIL, AHS. Analysis and interpretation of data: KEP, VMT, LL, LC, KJ, JNJ, YS, JK, HM. Drafting of manuscript: KEP, VMT, LL, JNJ, YS, JK, HM. Critical revision: KEP, VMT, LL, MW, AJ, LC, KJ, JNJ, YS, KVS, EIL, AHS, KJ, HM. All authors read and approved the final manuscript.

Acknowledgements: We thank the patients who participated in this study, Jonathan Bard MA and Brandon Marzullo MS for RNA sequencing data analysis assistance, and Jennifer L. Gay CCRP for study protocol management. This work was performed in part at the New York State Center of Excellence in Bioinformatics and Life Sciences' Genomics and Bioinformatics Core.

References

1. Vega, C., J.V. Kwoon, and S.D. Lavine, *Intracranial aneurysms: current evidence and clinical practice.* Am Fam Physician, 2002. **66**(4): p. 601-8.
2. Olafsson, E., G. Hauser Wa Fau - Gudmundsson, and G. Gudmundsson, *A population-based study of prognosis of ruptured cerebral aneurysm: mortality and recurrence of subarachnoid hemorrhage.* Neurology, 1997(0028-3878 (Print)).
3. Hop, J.W., et al., *Case-fatality rates and functional outcome after subarachnoid hemorrhage: a systematic review.* Stroke, 1997. **38**(0039-2499 (Print)).
4. Nieuwkamp, D.J., et al., *Changes in case fatality of aneurysmal subarachnoid haemorrhage over time, according to age, sex, and region: a meta-analysis.* The Lancet Neurology, 2009. **8**(7): p. 635-642.
5. Greving, J.P., et al., *Cost-effectiveness of preventive treatment of intracranial aneurysms: new data and uncertainties.* Neurology, 2009. **73**(4): p. 258-65.
6. Juvela, S., *Treatment options of unruptured intracranial aneurysms.* Stroke, 2004. **35**(2): p. 372-4.
7. Gounis, M.J., et al., *Myeloperoxidase in human intracranial aneurysms: preliminary evidence.* Stroke, 2014. **45**(5): p. 1474-7.
8. Serra, R., et al., *Metalloproteinase-9 and neutrophil gelatinase-associated lipocalin plasma and tissue levels evaluation in middle cerebral artery aneurysms.* Br J Neurosurg, 2014.
9. Chu, Y., et al., *Myeloperoxidase is increased in human cerebral aneurysms and increases formation and rupture of cerebral aneurysms in mice.* Stroke, 2015. **46**(6): p. 1651-6.
10. Stapleton, P.P., H.P. Redmond, and D.J. Bouchier-Hayes, *Myeloperoxidase (MPO) may mediate neutrophil adherence to the endothelium through upregulation of CD11B expression--an effect*

downregulated by taurine. Advances in experimental medicine and biology, 1998. **442**: p. 183.

11. Leopold, J.A., *The Central Role of Neutrophil Gelatinase-Associated Lipocalin in Cardiovascular Fibrosis.* Hypertension (Dallas, Tex. : 1979), 2015. **66**(1): p. 20-22.
12. Korai, M., et al., *Abstract 197: Roles of Neutrophil Extracellular Trap in the Rupture of Intracranial Aneurysm.* Stroke, 2017. **48**(suppl_1): p. A197-A197.
13. Meher, A.K., et al., *Novel Role of IL (Interleukin)-1β in Neutrophil Extracellular Trap Formation and Abdominal Aortic Aneurysms.* Arteriosclerosis, thrombosis, and vascular biology, 2018. **38**(4): p. 843-853.
14. Spinosa, M., et al., *Resolvin D1 decreases abdominal aortic aneurysm formation by inhibiting NETosis in a mouse model.* Journal of Vascular Surgery, 2018. **68**(6, Supplement): p. 93S-103S.
15. Yan, H., et al., *Neutrophil Proteases Promote Experimental Abdominal Aortic Aneurysm via Extracellular Trap Release and Plasmacytoid Dendritic Cell Activation.* Arteriosclerosis, Thrombosis, and Vascular Biology, 2016. **36**(8): p. 1660-1669.
16. Tutino, V.M., et al., *Circulating neutrophil transcriptome may reveal intracranial aneurysm signature.* PLOS ONE, 2018. **13**(1): p. e0191407.
17. Tutino, V.M., et al., *Biomarkers from circulating neutrophil transcriptomes have potential to detect unruptured intracranial aneurysms.* Journal of translational medicine, 2018. **16**(1): p. 373-373.
18. Jiang, K., et al., *RNA sequencing from human neutrophils reveals distinct transcriptional differences associated with chronic inflammatory states.* BMC Med Genomics, 2015. **8**: p. 55.
19. Jarvis, J.N., et al., *Novel approaches to gene expression analysis of active polyarticular juvenile rheumatoid arthritis.* Arthritis Res Ther, 2004. **6**(1): p. R15-r32.
20. Robinson, M.D., D.J. McCarthy, and G.K. Smyth, *edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.* Bioinformatics, 2009. **26**(1): p. 139-140.
21. McCarthy, D.J., Y. Chen, and G.K. Smyth, *Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation.* Nucleic Acids Research, 2012. **40**(10): p. 4288-4297.
22. Benjamini, Y. and Y. Hochberg, *Controlling the false discovery rate: a practical and powerful approach to multiple testing.* Journal of the Royal statistical society: series B (Methodological), 1995. **57**(1): p. 289-300.
23. Mallona, I., J. Weiss, and M. Egea-Cortines, *pcrEfficiency: a Web tool for PCR amplification efficiency prediction.* BMC Bioinformatics, 2011. **12**(1): p. 404.
24. de Kok, J.B., et al., *Normalization of gene expression measurements in tumor tissues: comparison of 13 endogenous control genes.* Laboratory Investigation, 2005. **85**(1): p. 154-159.
25. Leek, J.T., et al., *The sva package for removing batch effects and other unwanted variation in high-throughput experiments.* Bioinformatics, 2012. **28**(6): p. 882-883.
26. Johnson, W.E., C. Li, and A. Rabinovic, *Adjusting batch effects in microarray expression data using empirical Bayes methods.* Biostatistics, 2006. **8**(1): p. 118-127.

27. Hothorn, T. and B.S. Everitt, *A handbook of statistical analyses using R*. 2014: Chapman and Hall/CRC.
28. Jabbar, M.A., B.L. Deekshatulu, and P. Chandra, *Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm.* Procedia Technology, 2013. **10**: p. 85-94.
29. Chen, H.-L., et al., *An efficient diagnosis system for detection of Parkinson's disease using fuzzy k-nearest neighbor approach.* Expert Systems with Applications, 2013. **40**(1): p. 263-271.
30. Moore, P.J., et al., *Random forest prediction of Alzheimer's disease using pairwise selection from time series data.* PLOS ONE, 2019. **14**(2): p. e0211558.
31. Lebedev, A.V., et al., *Random Forest ensembles for detection and prediction of Alzheimer's disease with a good between-cohort robustness.* NeuroImage: Clinical, 2014. **6**: p. 115-125.
32. Lee, Y. and C.-K. Lee, *Classification of multiple cancer types by multicategory support vector machines using gene expression data.* Bioinformatics, 2003. **19**(9): p. 1132-1139.
33. Yu, W., et al., *Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes.* BMC medical informatics and decision making, 2010. **10**: p. 16-16.
34. Liaw, A. and M. Wiener, *Classification and regression by randomForest.* R news, 2002. **2**(3): p. 18-22.
35. Furey, T.S., et al., *Support vector machine classification and validation of cancer tissue samples using microarray expression data.* Bioinformatics, 2000. **16**(10): p. 906-914.
36. Harada, K., et al., *Prevalence of unruptured intracranial aneurysms in healthy asymptomatic Japanese adults: differences in gender and age.* Acta Neurochir, 2013(0942-0940 (Electronic)).
37. Li Mh Fau - Chen, S.-W., et al., *Prevalence of unruptured cerebral aneurysms in Chinese adults aged 35 to 75 years: a cross-sectional study.* Ann Intern Med, 2013(1539-3704 (Electronic)).
38. Rinkel, G.J., *Intracranial aneurysm screening: indications and advice for practice.* Lancet Neurol, 2005(1474-4422 (Print)).
39. Vlak, M.H., et al., *Prevalence of unruptured intracranial aneurysms, with emphasis on sex, age, comorbidity, country, and time period: a systematic review and meta-analysis.* 2011(1474-4465 (Electronic)).
40. Ho, D.E., et al., *Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference.* Political Analysis, 2007. **15**(3): p. 199-236.
41. Ho, D., et al., *MatchIt: Nonparametric Preprocessing for Parametric Causal Inference.* 2011, 2011. **42**(8): p. 28.
42. Eden, E., et al., *GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists.* BMC bioinformatics, 2009. **10**: p. 48-48.
43. Gopurappilly, R. and R. Bhonde, *Transcriptional profiling and functional network analyses of islet-like clusters (ILCs) generated from pancreatic stem cells in vitro.* Genomics, 2015. **105**(4): p. 211-219.
44. Müller Tomm, B., et al., *Risk Factors for Unruptured Intracranial Aneurysms and Subarachnoid Hemorrhage in a Prospective Population-Based Study.* Stroke, 2019. **50**(10): p. 2952-2955.

45. Kim, S.-Y., *Effects of sample size on robustness and prediction accuracy of a prognostic gene signature.* BMC bioinformatics, 2009. **10**: p. 147-147.
46. Woo, D., et al., *Smoking and family history and risk of aneurysmal subarachnoid hemorrhage.* Neurology, 2009. **72**(1): p. 69-72.
47. Juvela, S., K. Poussa, and M. Porras, *Factors Affecting Formation and Growth of Intracranial Aneurysms.* Stroke, 2001. **32**(2): p. 485-491.
48. Juvela, S., et al., *Cigarette smoking and alcohol consumption as risk factors for aneurysmal subarachnoid hemorrhage.* Stroke, 1993. **24**(5): p. 639-646.
49. Huan, T., et al., *A whole-blood transcriptome meta-analysis identifies gene expression signatures of cigarette smoking.* Human Molecular Genetics, 2016. **25**(21): p. 4611-4623.
50. Kamio, Y., et al., *Roles of Nicotine in the Development of Intracranial Aneurysm Rupture.* Stroke, 2018. **49**(10): p. 2445-2452.
51. Can, A., et al., *Association of intracranial aneurysm rupture with smoking duration, intensity, and cessation.* Neurology, 2017. **89**(13): p. 1408-1415.
52. Chalouhi, N., et al., *Cigarette Smoke and Inflammation: Role in Cerebral Aneurysm Formation and Rupture*. Vol. 2012. 2012. 12.
53. Chalouhi, N., et al., *Biology of intracranial aneurysms: role of inflammation.* Journal of cerebral blood flow and metabolism, 2012. **32**(9): p. 1659-1676.
54. Chyatte, D., et al., *Inflammation and Intracranial Aneurysms.* Neurosurgery, 1999. **45**(5): p. 1137-1147.
55. Turkmani, A.H., N.J. Edwards, and P.R. Chen, *The role of inflammation in cerebral aneurysms.* Neuroimmunology and Neuroinflammation; Vol 2, No 2 (2015), 2015.
56. Strong, M., et al., *The role of leukocytes in the formation and rupture of intracranial aneurysms.* Neuro Immunology and Inflammation, 2015. **2**(2): p. 107-114.
57. Chalouhi, N., et al., *Localized increase of chemokines in the lumen of human cerebral aneurysms.* Stroke, 2013. **44**(9): p. 2594-7.
58. Sawyer, D.M., et al., *Lymphocytes influence intracranial aneurysm formation and rupture: role of extracellular matrix remodeling and phenotypic modulation of vascular smooth muscle cells.* Journal of neuroinflammation, 2016. **13**(1): p. 185-185.
59. Starke, R.M., et al., *Critical role of TNF-α in cerebral aneurysm formation and progression to rupture.* Journal of neuroinflammation, 2014. **11**: p. 77-77.
60. Jayaraman, T., et al., *Tumor necrosis factor alpha is a key modulator of inflammation in cerebral aneurysms.* NEUROSURGERY, 2005. **57**(3): p. 558-563.
61. Wang, C., et al., *Proteomic identification of differentially expressed proteins in vascular wall of patients with ruptured intracranial aneurysms.* Atherosclerosis, 2015. **238**(2): p. 201-206.
62. Yamazaki, T., et al., *CCR6 regulates the migration of inflammatory and regulatory T cells.* Journal of immunology (Baltimore, Md. : 1950), 2008. **181**(12): p. 8391-8401.

63. Sarma, J.V. and P.A. Ward, *The complement system.* Cell and tissue research, 2011. **343**(1): p. 227-235.
64. Tulamo, R., et al., *COMPLEMENT ACTIVATION ASSOCIATES WITH SACCULARCEREBRAL ARTERY ANEURYSM WALL DEGENERATION AND RUPTURE.* Neurosurgery, 2006. **59**(5): p. 1069-1077.
65. Shi, C., et al., *Genomics of Human Intracranial Aneurysm Wall.* Stroke, 2009. **40**(4): p. 1252.
66. Camous, L., et al., *Complement alternative pathway acts as a positive feedback amplification of neutrophil activation.* Blood, 2011. **117**(4): p. 1340.
67. Yago, T., et al., *E-selectin engages PSGL-1 and CD44 through a common signaling pathway to induce integrin alphaLbeta2-mediated slow leukocyte rolling.* Blood, 2010. **116**(3): p. 485-494.
68. Cooks, T., C.C. Harris, and M. Oren, *Caught in the cross fire: p53 in inflammation.* Carcinogenesis, 2014. **35**(8): p. 1680-1690.
69. Gupta, R., S. Palchaudhuri, and D. Chattopadhyay, *Glutamate induces neutrophil cell migration by activating class I metabotropic glutamate receptors.* Amino Acids, 2013. **44**(2): p. 757-767.
70. Flood, S., et al., *Modulation of interleukin-6 and matrix metalloproteinase 2 expression in human fibroblast-like synoviocytes by functional ionotropic glutamate receptors.* Arthritis & Rheumatism, 2007. **56**(8): p. 2523-2534.
71. Lin, Y.-J., et al., *Genetic variants of glutamate receptor gene family in Taiwanese Kawasaki disease children with coronary artery aneurysms.* Cell & bioscience, 2014. **4**(1): p. 67-67.
72. Niikura, T., H. Tajima, and Y. Kita, *Neuronal cell death in Alzheimer's disease and a neuroprotective factor, humanin.* Current neuropharmacology, 2006. **4**(2): p. 139-147.
73. Liao, W., et al., *CARP-2 is an endosome-associated ubiquitin ligase for RIP and regulates TNF-induced NF-kappaB activation.* Current biology : CB, 2008. **18**(9): p. 641-649.
74. Aubrey, B.J., et al., *How does p53 induce apoptosis and how does this relate to p53-mediated tumour suppression?* Cell Death & Differentiation, 2018. **25**(1): p. 104-113.
75. Haupt, S., et al., *Apoptosis - the p53 network.* Journal of Cell Science, 2003. **116**(20): p. 4077.
76. Jin, H., et al., *Circulating microRNA: a novel potential biomarker for early diagnosis of intracranial aneurysm rupture a case control study.* J Transl Med, 2013. **11**: p. 296.
77. Metaxa, E., et al., *Characterization of Critical Hemodynamics Contributing to Aneurysmal Remodeling at the Basilar Terminus in a Rabbit Model.* Stroke, 2010. **41**(8): p. 1774-1782.
78. Frösen, J., et al., *Saccular intracranial aneurysm: pathology and mechanisms.* Acta Neuropathologica, 2012. **123**(6): p. 773-786.
79. Kolega, J., et al., *Cellular and molecular responses of the basilar terminus to hemodynamics during intracranial aneurysm initiation in a rabbit model.* Journal of vascular research, 2011. **48**(5): p. 429-442.
80. Meng, H., et al., *Progressive aneurysm development following hemodynamic insult.* 2011. **114**(4): p. 1095.

81. Nygaard, V., E.A. Rødland, and E. Hovig, *Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses.* Biostatistics, 2015. **17**(1): p. 29-39.

Figures

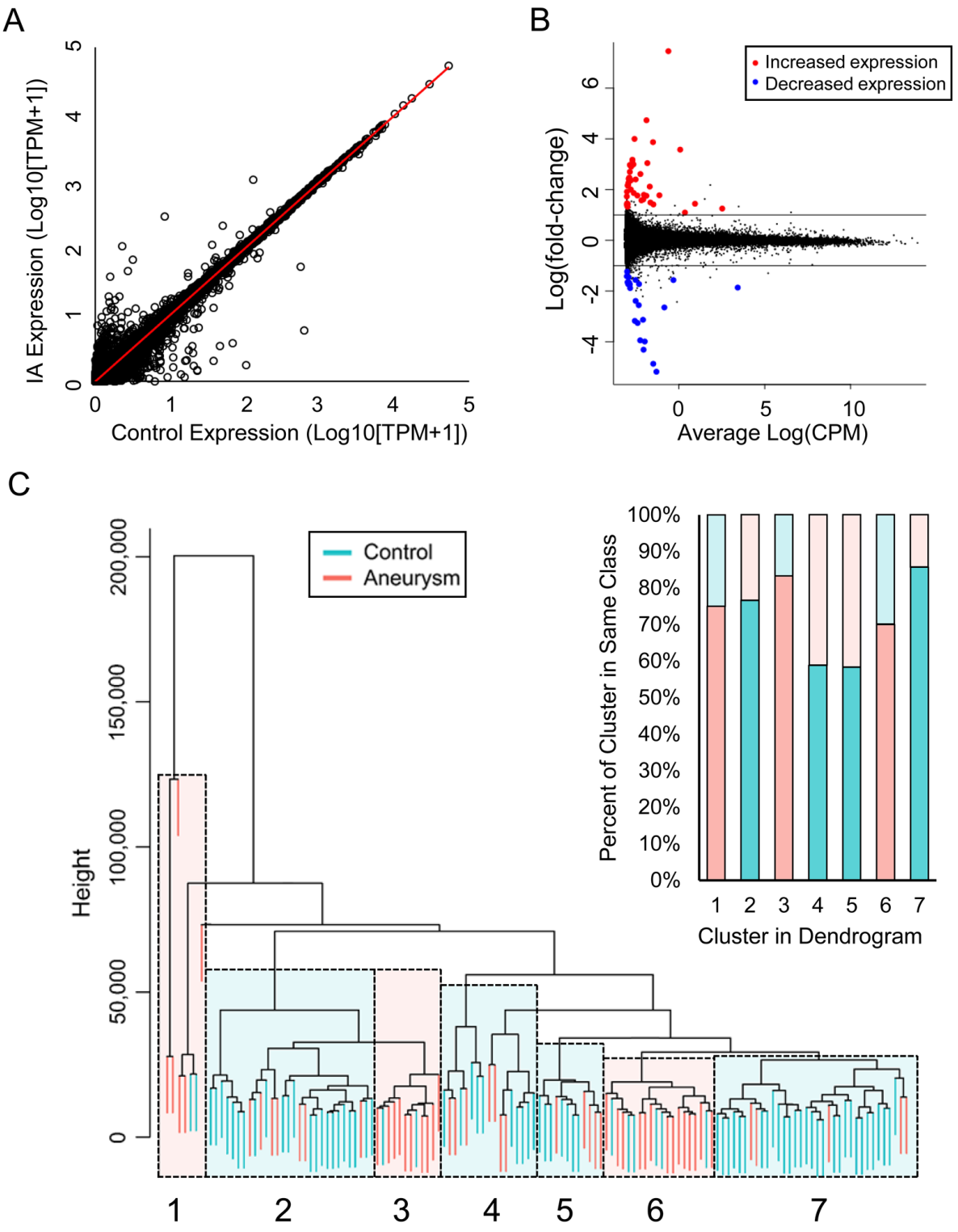

Figure 1

RNAseq data from whole dataset (n=134). A) The scatter plot demonstrates the dispersion in expression between the IA and control groups. B) The volcano plot produced by edgeR demonstrates that there are 65 differentially expressed genes. Red points are increased in IA group and blue points are decreased in IA group. C) Clustering performed on all transcriptome data demonstrates several distinct clusters of IA and control samples. Overall, 73% of samples were assigned to the correct group.

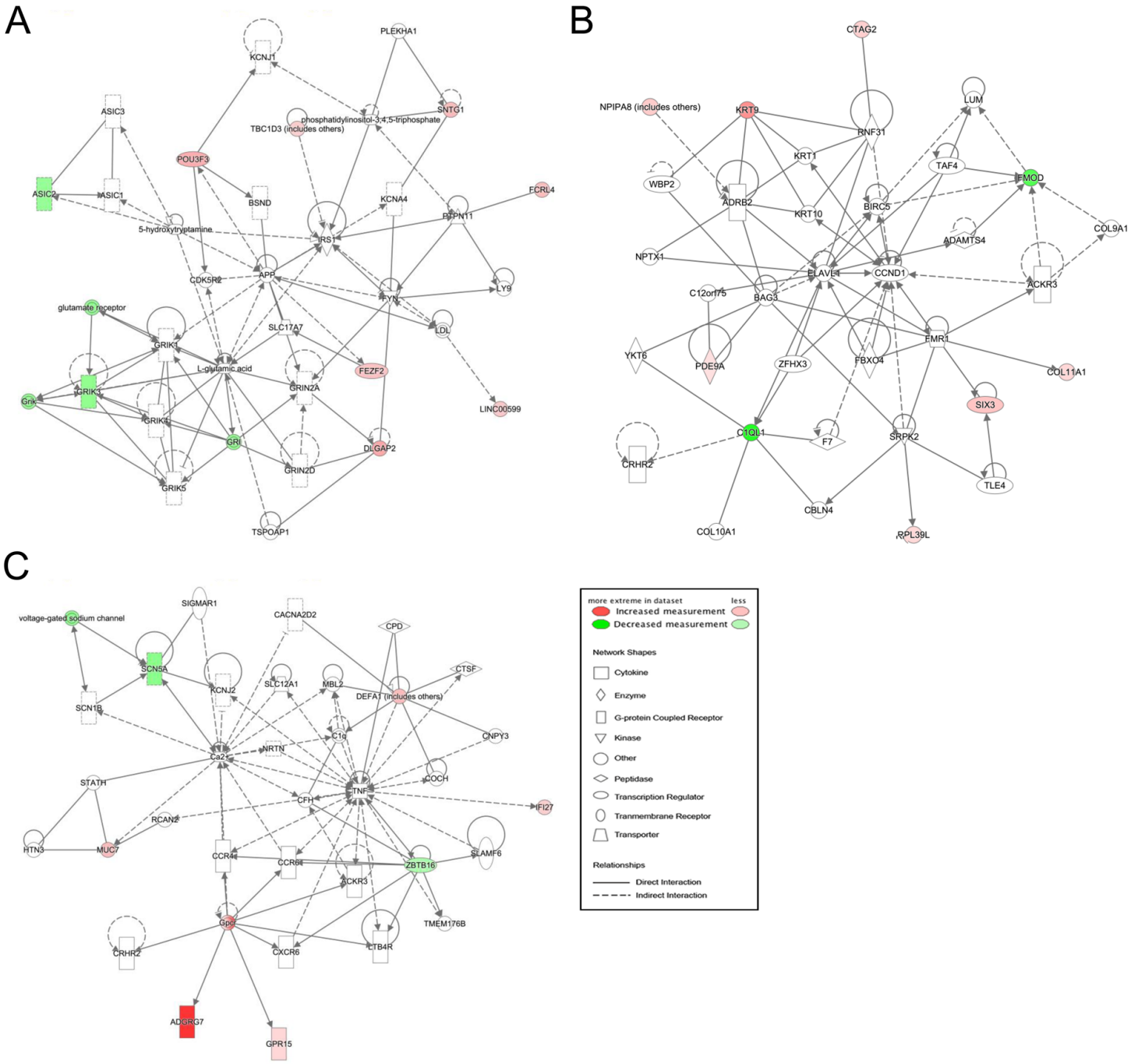

Figure 2

Networks derived from IPA of the 65 differentially expressed transcripts (q<0.05, fold-change>2). Transcripts with increased expression in IA are red; transcripts with lower expression in IA are green; fold-change is represented by intensity. A) This network (p-score=21) has related functions of cell-to-cell

signaling and interaction, nervous system development and function, and cell morphology B) This network (p-score=21) associated with dermatological diseases and conditions, organismal injury and abnormalities, and connective tissue development and function. C) This network (p-score=15) has ties to cell death and survival, connective tissue disorders, and inflammatory disease.

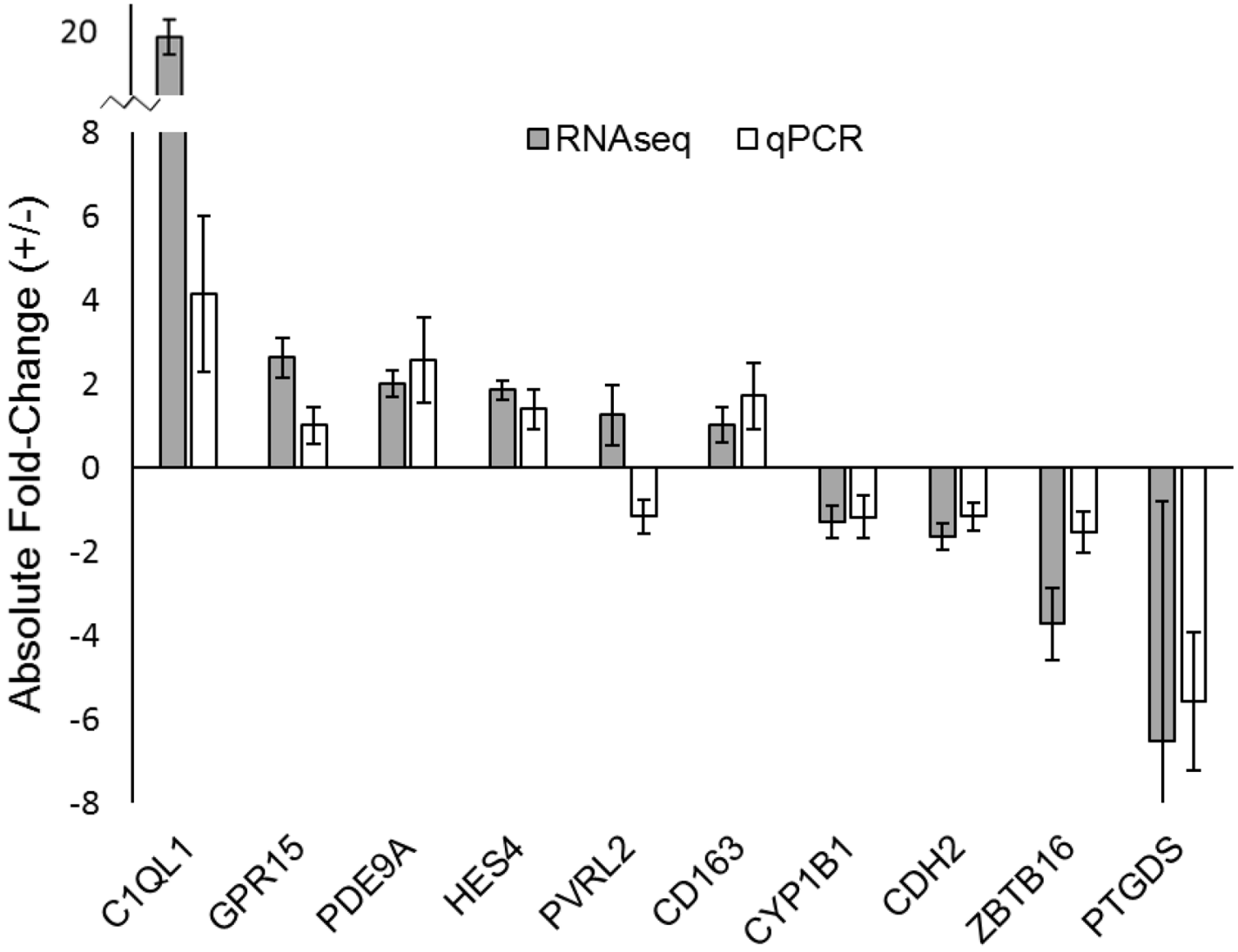

Figure 3

Verification of RNA-Sequencing data for 10 transcripts by qPCR. Fifty of the sequenced samples were analyzed by RT-qPCR, as the other samples did not have enough RNA for the additional reactions. Nine of the 10 transcripts in samples in a subset of patients had the same expression difference on qPCR. There were no statistically significant differences in fold-change in expression between RNAseq and qPCR. (Negative fold-change values calculated by negative inverse of fold-change, error bars = standard error.)

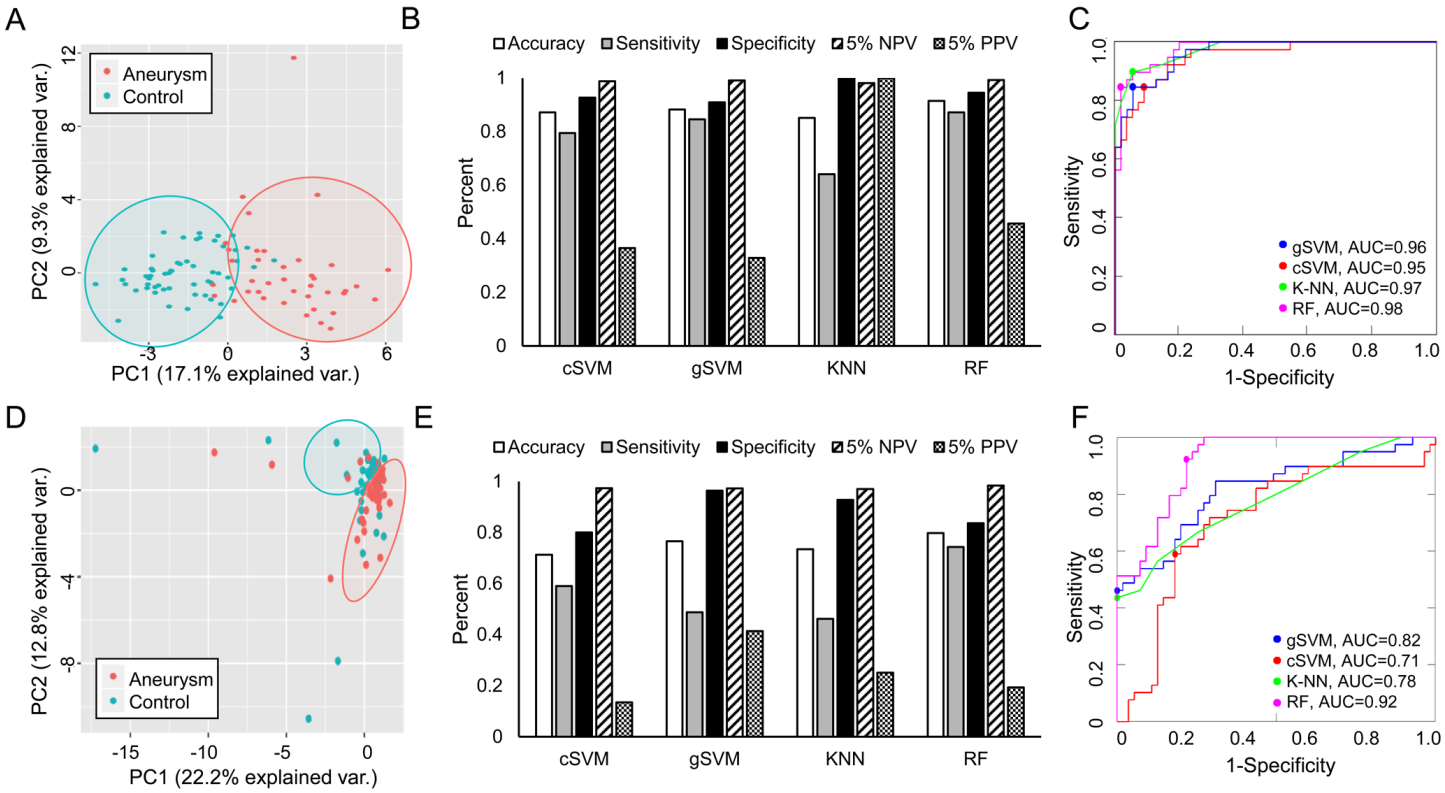

Figure 4

Models' performance in the training dataset. A) PCA using the 37 selected transcripts demonstrated clear separation between samples from patients with IA and those from controls. B) Estimation of model performance during LOO C-V in the training cohort demonstrated that models performed with an accuracy of 0.85–0.91. Considering a 5% prevalence of IA, PPV ranged from 0.33-1 and NPV ranged from 0.98-0.99 C) ROC analysis showed that all models had AUCs ≥0.95. D) PCA using the 26 previously-identified transcripts demonstrated inferior separation between IA and control cases. E) Estimation of model performance during LOO C-V in the training cohort demonstrated that models performed with an accuracy of 0.71-0.80. Considering a 5% prevalence, PPV and NPV ranged from 0.13-0.41 and 0.97-0.98 respectively. F) ROC analysis also showed subpar performance compared to newly identified transcripts (AUC range 0.71-0.92).

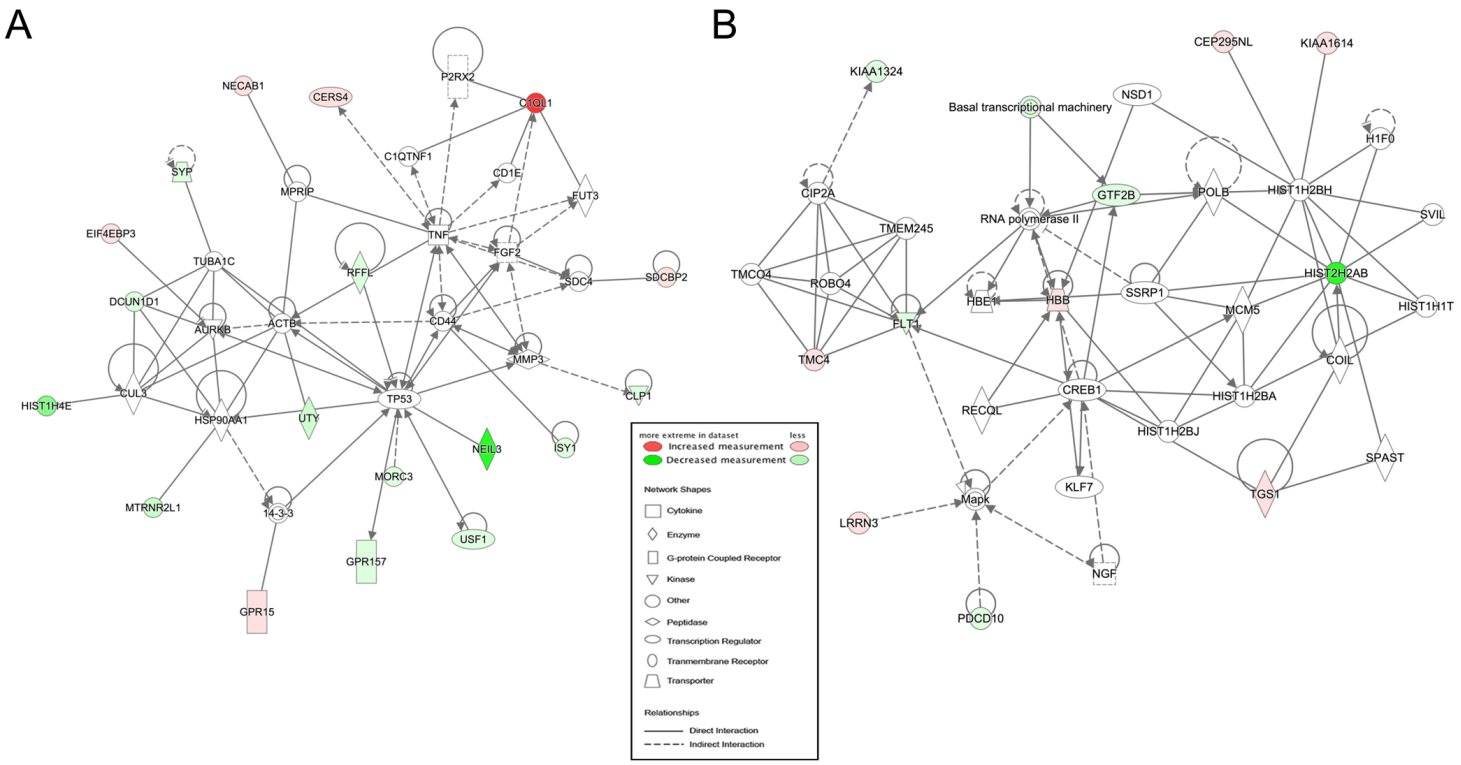

Figure 5

Networks derived from IPA of the 37 genes identified by LASSO. Transcripts with increased expression in IA are red; transcripts with lower expression in IA are green; fold-change is represented by intensity. A) This network (p-score=47) affiliated with cancer, cellular movement, and connective tissue disorders. B) This network (p-score=25) has associated functions of cell cycle, cellular assembly and organization, DNA replication, recombination, and repair.

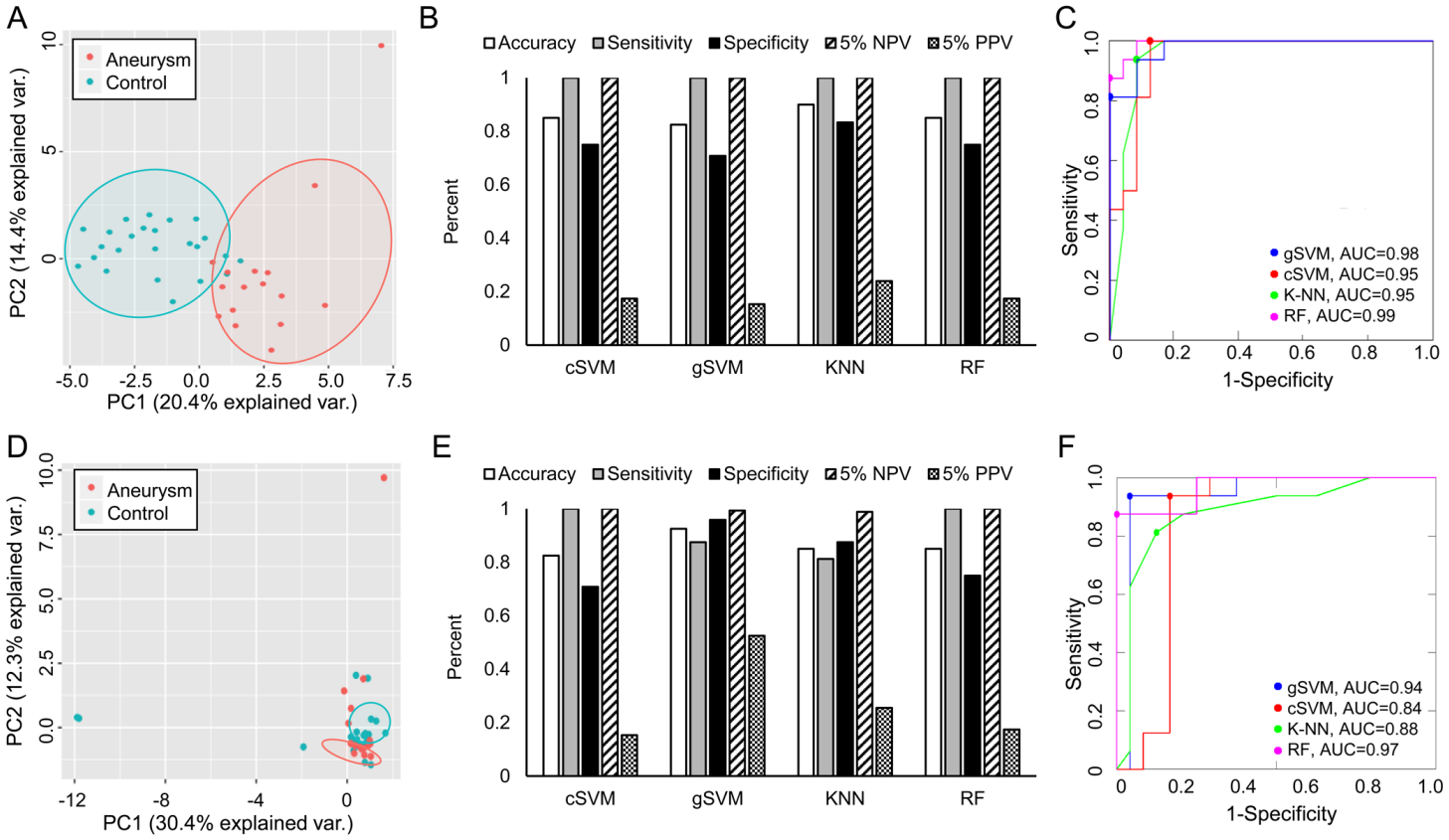

Figure 6

Models' performance in the testing dataset. A) PCA using the 37 selected transcripts in this independent dataset also demonstrated strong separation between samples from patients with IA and from controls. B) Assessment of true model performance showed that models performed with an accuracy of 0.83–0.90. In this dataset all models had a sensitivity of 1. At 5% IA prevalence, the PPV ranged from 0.15-0.24 and NPV was 1 for all models. C) ROC analysis showed that all models again had AUCs ≥0.95. D) PCA using the 26 previously-identified transcripts demonstrated mediocre separation between IA and control cases. E) Estimation of model performance in the testing cohort demonstrated that models performed with an accuracy of 0.83-0.93. Considering a 5% prevalence, PPV and NPV ranged from 0.15-0.52 and 0.99-1 respectively. F) ROC analysis also showed inferior performance compared to newly identified transcripts (AUC range 0.84-0.97).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementalTable4.docx
- SupplementalTable2.docx
- SupplementalTable1.docx
- SupplementalTable5.docx

- SupplementalTable3.docx
- SupplementalTable6.docx
- SupplementalTable7.docx