ANNALS of Clinical and Translational Neurology
Open Access
ANA American Neurological Association since 1875

RESEARCH ARTICLE

Blood transcriptome changes after stroke in an African American population

Robert Meller[1], Andrea N. Pearson[1], Jimmaline J. Hardy[1], Casey L. Hall[2,3], Dawn McGuire[1], Michael R. Frankel[2,3] & Roger P. Simon[1,2]

[1]Translational Stroke Program, Neuroscience Institute, Morehouse School of Medicine, Atlanta, Georgia
[2]Grady Memorial Hospital, Atlanta, Georgia
[3]Department of Neurology, Emory University, Atlanta, Georgia

Correspondence
Robert Meller, Neuroscience Institute, Morehouse School of Medicine, 720 Westview Dr SW, Atlanta, GA 30310.
Tel: 4047565789; Fax: 404-752-1041;
E-mail: rmeller@msm.edu

Funding Information
R. P. S. (NINDS) R56NS073714-01; Supported by the National Center for Advancing Translational Sciences of the National Institutes of Health UL1TR000454; R. M. (NIMHD) Grant 8 U54 MD007588, (NINDS) R01NS59588; J. H. (NINDS) U54 NS083932. C. L. H. (NINDS) R25NS065739. The contents are solely the responsibility of the authors and do not necessarily represent the official views of NIMHD, NINDS, or National Institutes of Health.

Received: 6 July 2015; Revised: 7 October 2015; Accepted: 28 October 2015

***Annals of Clinical and Translational Neurology* 2016; 3(2): 70–81**

doi: 10.1002/acn3.272

Abstract

Objective: Molecular diagnostic medicine holds much promise to change point of care treatment. An area where additional diagnostic tools are needed is in acute stroke care, to assist in diagnosis and prognosis. Previous studies using microarray-based gene expression analysis of peripheral blood following stroke suggests this approach may be effective. Next-generation sequencing (NGS) approaches have expanded genomic analysis and are not limited to previously identified genes on a microarray chip. Here, we report on a pilot NGS study to identify gene expression and exon expression patterns for the prediction of stroke diagnosis and prognosis. **Methods**: We recruited 28 stroke patients and 28 age- and sex-matched hypertensive controls. RNA was extracted from 3 mL blood samples, and RNA-Seq libraries were assembled and sequenced. **Results**: Bioinformatical analysis of the aligned RNA data reveal exonic (30%), intronic (36%), and novel RNA components (not currently annotated: 33%). We focused our study on patients with confirmed middle cerebral artery occlusion ischemic stroke (n = 17). On the basis of our observation of differential splicing of gene transcripts, we used all exonic RNA expression rather than gene expression (combined exons) to build prediction models using support vector machine algorithms. Based on model building, these models have a high predicted accuracy rate >90% (spec. 88% sen. 92%). We further stratified outcome based on the improvement in NIHss scores at discharge; based on model building we observe a predicted 100% accuracy rate. **Interpretation**: NGS-based exon expression analysis approaches have a high potential for patient diagnosis and outcome prediction, with clear utility to aid in clinical patient care.

Introduction

Stroke diagnosis and assessment is essential prior to the administration of thrombolytic therapy (recombinant tissue plasminogen activator: rt-PA).[1] While imaging is a prerequisite for determination of hemorrhage and infarction volume, the utility of imaging as a prognostic marker is more limited.[2] In part, since many stroke patients do not have access to specialist imaging and neurology services, rt-PA administration rates are frequently found to be low (~5%).[3] Studies of potential biomarkers for stroke diagnosis and prognosis have identified a number of candidates for brain injury,[4] however, few single protein studies show specificity for stroke subtype. As an alternative, microarray-based gene expression analysis offers a novel approach to identify stroke subtype, based on the transcriptome response in circulating blood immune cells.[5] This approach enables the subtyping of acute stroke (ischemic vs. hemorrhagic) as well as the prediction of ischemic stroke subtype (atherosclerotic vs. cardioembolic).[6,7] The power of this approach is exemplified by the observation that a prediction of cause of cryptogenic stroke is possible.[8] However, microarray technology is being superseded by high throughput sequencing technologies, such as RNA sequencing (RNA-Seq).[9,10] To date the power of RNA-Seq for identifying

gene panels for stroke diagnosis and prognosis is not established.

It is well documented that African American populations suffer from a higher burden of stroke compared to Caucasians (294 vs. 174/100,000).[3] Traditional risk factors for stroke, such as obesity, hypertension (uncontrolled), and lifestyle only partially account for the higher risk.[11] Furthermore, race-associated differences in response to rt-PA have been recently described.[12,13] Gene expression studies of vascular disease reveal a difference in the transcriptome of African Americans versus Caucasians.[14] However, an assessment of gene expression in blood following stroke did not report African Americans, or had insufficient power for separate racial analysis.[6] Therefore, in this pilot study, we used next-generation RNA-Seq approaches to determine whether gene expression profiles in blood have diagnostic as well as prognostic potential, with a focus on an African American patient cohort.

Subjects/Materials and Methods

Participant recruitment

We recruited 28 African American stroke patients (Fig. 1). Matched healthy controls (28) with similar hypertensive profiles, but without history of stroke were recruited from an outpatient clinic. Admission stroke assessment included NIH stroke scale ($NIHss_0$). A second measure of stroke assessment was performed on discharge ($NIHss_1$), and the change was calculated as follows: $(100\% \times((NIHss_0 - NIHss_1)/NIHss_0))$. Final stroke diagnosis was determined after reviewing the medical records by the neurologist panel (Dr's Simon, Hall, and Frankel). While all 28 patient samples were subjected to sequencing, only patients with a confirmed stroke in the territory of the middle cerebral artery (MCA) were subjected to further analysis (17 in total, of which 11 received rt-PA).

Blood was drawn into PAXgene™ vacutainer tubes from stroke patients the first morning following admission to Grady Memorial Hospital. The 24-h time point was validated in previous studies reporting the blood transcriptome to be stable between 4 and 24 h following a brain injury event.[15] The average time between stroke and study blood draw was 22.9 ± 4.5 (mean ± SE) hour.

RNA library assembly

RNA was extracted from whole blood collected in PAXgene™ tubes using PreAnalytiX™ RNA isolation procedures (Qiagen, Valencia, CA, USA) (see Fig. S1 and details for more information). The RNA concentration was determined spectrometrically (A_{260}). RNA libraries, assembled blind to the clinical diagnosis, were created using the Total

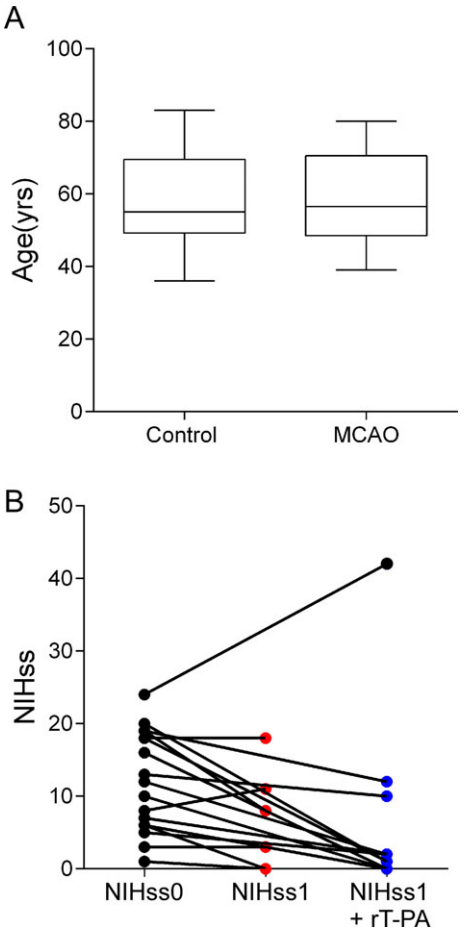


Figure 1. Confirmed MCA patients ($n = 17$) and control data ($n = 28$) for preliminary studies. (A) Age Matching of stroke patients ($n = 17$) with controls ($n = 28$) data are not statistically significantly different (Student's *t*-test). (B) Stroke severity was rated using the NIH stroke scale at admittance ($NIHss_0$) and discharge ($NIHss_1$). Patients receiving rt-PA are grouped separately (blue; $n = 11$) compared to those who did not receive rt-PA (red; $n = 6$). Overall, we observe a statistically significant decrease in NIHss rating from admission to discharge for all patients ($P < 0.01$ Wilcoxon signed matched test). Data shown are mean ± SEM. Note one deceased patient who received rt-PA is denoted by the filled black circle and assigned an arbitrary $NIHss_1$ of 42. MCA, middle cerebral artery; rt-PA, recombinant tissue plasminogen activator.

RNA workflow (Life Technologies, Foster City, CA, USA). RNA (800 ng) was fragmented using RNase III and fragmented RNA (200 ng) was hybridized to adapters, subjected to reverse transcription, and amplified using polymerase chain reaction (PCR) (AmpliTaq) with barcoded primers. Libraries were assessed using a Bioanalyzer DNA High Sensitivity chip (Agilent, Santa Clara, CA, USA) and quantified with qPCR using a known standard. The library (1.2 pmol/L from 8 × 0.15 pmol/L of each library in the seeding reaction) was cloned onto sequencing beads (E80 reaction) and deposited on three lanes of a sequencing

flowchip. The libraries were sequenced on a SOLiD 5500XL sequencer using an F50 kit (50 base single end reads). Resulting xsq data files were transferred to a Penguin Cluster and aligned to the Hg19 human reference genome using LifeScope Software with Whole Transcriptome Analysis default settings (Life Technologies). Transcripts were annotated using the RefSeq Hg19 annotation (version 09/2013, UCSC). In addition, a novel annotation guide was created following alignment of Bam file reads to the hg19 Refseq annotation guide using Cuffmerge (part of the Cufflinks suite).[16,17] The resultant .gtf annotation file merges the Hg19 Refseq database with the novel, un-annotated RNAs we discovered using whole transcriptome analysis (see Fig. 2D). This annotation guide is available upon request.

Technical replicates are typically not required for next-generation sequencing (NGS) experiments.[18] To control for potential technical issues, we prepared samples in batches of eight enabling color balancing of the barcoded primers and sequencing at ~50–100 million reads/sample. Direct comparison of multiple libraries created from the same

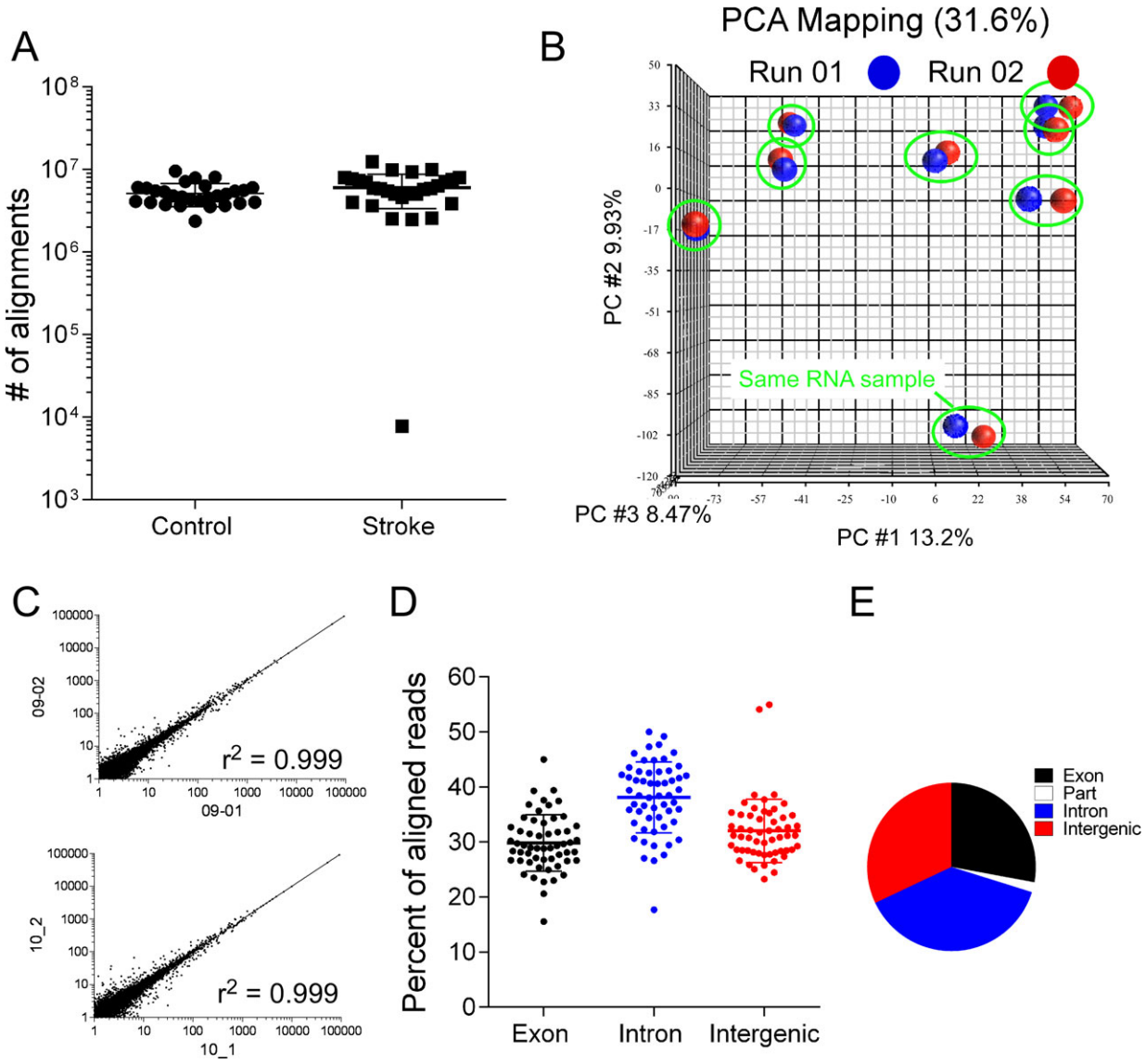


Figure 2. Assessment of RNA-Seq library preparation from controls and recruited patients. (A) Alignment of RNA-Seq libraries to hg19 reference genome, following removal (filtering) of ribosomal RNA reads. Data shown are number of reads from each sample, one library of 56 samples failed library building (one patient sample not analyzed further). (B) Assessment of reproducibility of RNA-Seq library preparation. Two sets of eight libraries (four controls and four patients) were constructed 6 months apart and then sequenced. Note the close overlap of blue and red samples (01 and 02 denote run). (C) Individual regression analysis of two libraries yields a fit of $r^2 = 0.999$ (09 – female middle cerebral artery sample, and 10 – female control). (D and E), Mapping summary of all 56 samples reveals a large proportion of RNA reads that align to intronic and intergenic regions.

RNA reveal a correlation r^2 of 0.999. In addition, Partek analysis to identify batch effects showed no significant effect of library batch numbers. Therefore, we did not transform the data to remove batch effects, as has been performed on other transcriptome studies using microarrays.[19]

Statistics

BAM (binary alignment map) output files were analyzed using Partek Genomics Studio Software v 6.6 (Partek Inc., St Louis, MO, USA). Gene expression and exon expression data were indicated as reads per kilobase per million aligned reads (RPKM).[20] Expression values were filtered to remove low expression genes (<10 reads/gene) and low occurrence (present in <50% of samples) quartile normalized, Log2 (+1 offset) transformed and subjected to analysis of variance (ANOVA) with diagnosis (control or MCA as the discriminating factor). Resultant data were used for hierarchical clustering and other representations. Additional analysis used stratification of NIHss scores (0–5 minor, 6–15 moderate and >15 severe).

Gene and exon expression data were subjected to modeling in order to identify a panel of genes that best fits/discriminates between stroke (MCA occlusion) patients and controls, as well as the NIHSS outcome data of the MCA patients. Genes and exons identified as significantly different with ANOVA were used to create a model for prediction analysis. We used the data to train an algorithm to predict the diagnosis of patients based on exon expression values. Data were modeled using support vector machine (SVM) models, with shrinking centroids variable selection method and different cross-validation strategies (one-level cross-validation – full leave one out, two-level cross validation (inner partition – full leave one out, outer partition 10 partitions), and Bootstrap cross validation). Models were deemed "best" based on their normalized correct rate, sensitivity and specificity in training, and the lowest number of variables used to discriminate the samples.

Study approval

All procedures were approved by the Morehouse School of Medicine and Grady Memorial Hospital Institutional Review Boards. Written informed consent was received from all participants prior to their inclusion in the study. Patient blood samples were deidentified and assigned random number identifiers prior to analysis.

Results

In order to determine the utility of transcriptome analysis for the diagnosis of stroke, we recruited and consented patients admitted to the ER at Grady Memorial Hospital based upon their presentation of stroke symptoms: Acute hemiparesis. In total 28 patients gave consent for the study, and we obtained a blood sample from 28 age and sex-matched hypertensive controls (Fig. 1 and Table 1). The RNA was extracted from 3 mL blood samples and RNA-Seq libraries were assembled by researchers blinded to the condition. The yields of extracted RNA from the blood of controls and stroke patients were 5.3 ± 0.4 and 3.4 ± 0.4 μg, respectively ($P < 0.01$). Following sequencing, and filtering for ribosomal RNA we obtained between 4 and 9 million RNA sequences (reads) aligned to the human genome (hg19) (Fig. 2A). There was no significant difference between numbers of reads obtained from the control and stroke patient cohorts (5.16 ± 0.3 million reads vs. 6.0 ± 0.5 million aligned reads, $n = 56$, $P = 0.14$; unpaired Student's t-test). Reproducibility testing of the transcriptomes generated from the same RNA

Table 1. Baseline characteristics of patients.

	Stroke ($n = 17$)	Control ($n = 28$)
Age, y, mean ± SD	57 ± 13	59 ± 13
Women, n (%)	7 (41)	9 (32)
Risk factors, n (%)		
Prior stroke or TIA	0	
Hypertension	9 (53)	28 (100)
Coronary artery disease	3 (18)	0
Congestive heart failure	4 (24)	0
Diabetes mellitus	4 (24)	1 (4)
Atrial fibrillation	1 (6)	0
Other	10 (43)[1]	0
Vascular territory involved, n (%)		
MCA	17 (100)	–
ACA	2 (12)	–
PCA	1 (6)	–
Stroke subtype[2], n (%)		
Atherogenic	12 (71)	–
Cardioembolic	5 (29)	–
Thrombolytic therapy, n (%)	11 (65)	–
NIHSS on admission, median, IQR	12 (6–18)	–
NIHSS on hospital discharge, median, IQR	2 (0–9.5)[3]	–

Baseline characteristics of confirmed MCA territory stroke patients and hypertensive controls. All samples were obtained from self-determined African Americans. Severity of stroke is determined by National Institutes of Health Stroke Scale (NIHss). MCA, middle cerebral artery; ACA, anterior cerebral artery; PCA, posterior cerebral artery; IQR, interquartile range; TIA, transient ischemic attack.

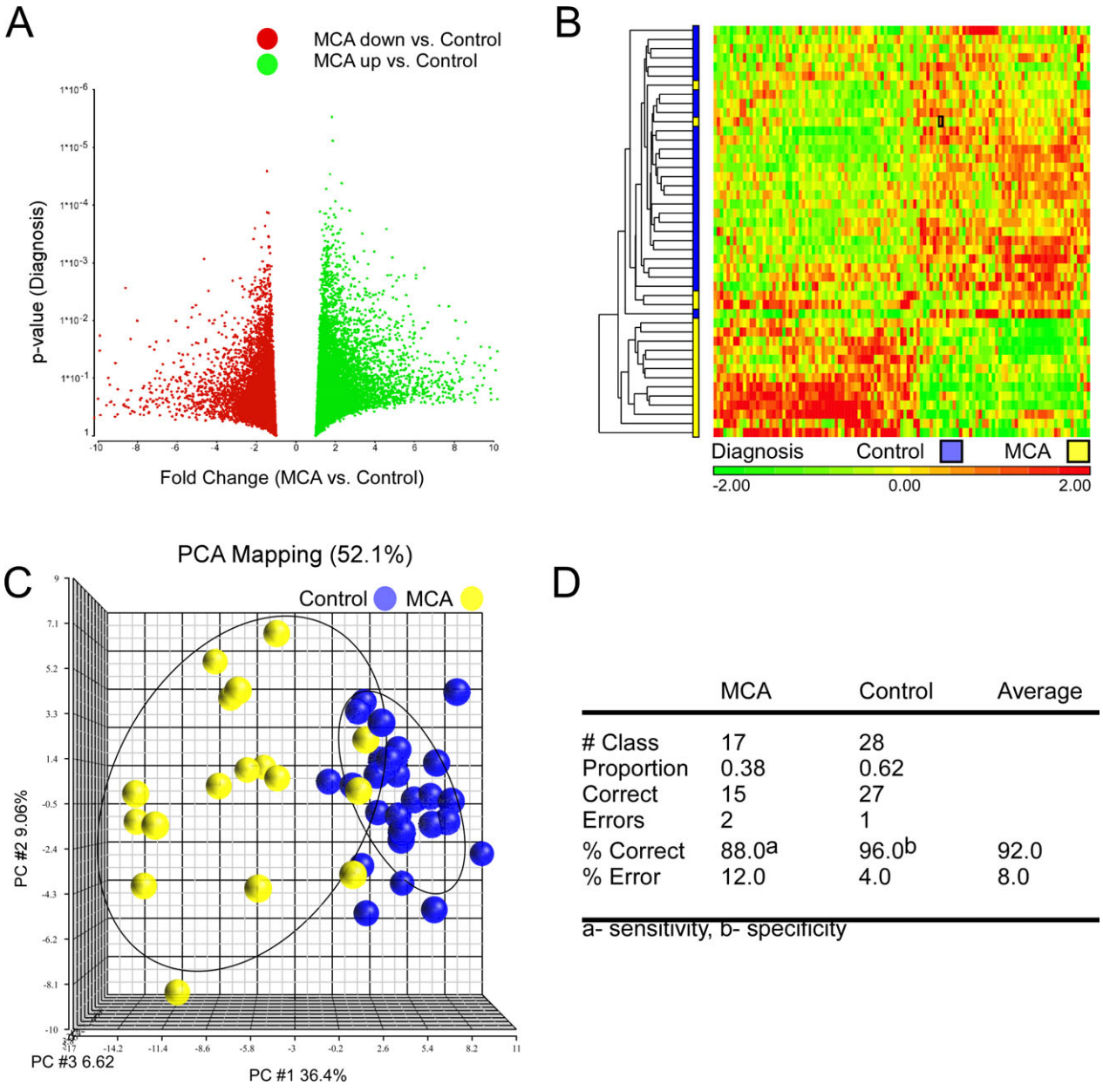
[1]Other risk factors in stroke group (Asthma 1, lupus 2, malignancy 2, normal pressure hydrocephalus 1, polysubstance abuse, cocaine and alcohol 3, syphilis history 1).

[2]Stroke subtypes lacunar and cryptogenic are not included given focus on large vessel, cortical-type ischemic strokes.

[3]One patient deceased.

in two independent library builds show the pipeline for library building has high accuracy and reproducibility. Reproducibility is shown by principle component analyses of the libraries, and representative individual library expression correlations of samples collected from our two recruitment sites (09 and 10) ($r^2 > 0.999$) (Fig. 2B and C).

We first investigated from where in the human genome the RNA is transcribed. The RNA-Seq reads were aligned using the hg19 reference genome and the RefSeq annotation guide (09-2013) (see Fig. S1 for a diagram of the workflow). Of the known RNA transcripts in the human genome annotation guide, 62% were



	MCA	Control	Average
# Class	17	28	
Proportion	0.38	0.62	
Correct	15	27	
Errors	2	1	
% Correct	88.0[a]	96.0[b]	92.0
% Error	12.0	4.0	8.0

a- sensitivity, b- specificity

Figure 3. Analysis of Gene Expression in blood following MCA stroke using RNA seq. Gene expression was quantified in RNA-Seq libraries and differences between MCA patients (17) and controls (28) were analyzed. (A) Volcano plot representing changes in expression of genes and *P* values. Of note most absolute changes are small <2.0-fold change. (B). Hierarchical clustering of genes whose expression change (1.2-fold, unadjusted $P < 0.001$). (C) Principle component analysis of 115 genes which change between control and MCA patients. (D) Results matrix from support vector machine model to predict diagnosis based on gene expression values (Best model – 35 variables, SVM, Shrinking centroids, cost 701, nu 0.5, tol 0.01, kern, rbf deg 3, gamma 1e -05, coeff 0). MCA, middle cerebral artery; SVM, Support Vector Machine.

detected in the whole blood RNA libraries. Approximately 30% of the reads aligned with known gene encoding exons. In contrast, 36% aligned with intronic regions of annotated genes, whereas 33% aligned with novel nongene encoding regions (intergenic) (Fig. 2D and E). These data show the discovery potential of RNA-Seq. In order to quantify the expression of these novel RNAs, not present in the Ref-Seq annotation guide, we created a new annotation guide using Cuffmerge (part of the Cufflinks suites of RNA-Seq tools; available on request from the authors).

The goal of our study was to identify panels of genes from which we had the ability to discriminate between control and stroke patients. Because we had a clinically homogeneous cohort of hemi-paretic patients, we further focused our analysis on stroke patients with a confirmed MCA territory infarct. Analysis of gene expression was performed on read data (normalized to RPKM values[18,20]). We define a gene as consisting of multiple exons that define differential splicing and isoforms of the gene. For this analysis multiple exon expression values are aggregated to yield a single gene expression value. Common in blood transcriptomic profiling,[15] most gene expression changes were small, <2-fold (Fig. 3A). Accordingly, we relaxed our *P*-value to an unadjusted $P < 0.001$, yielding 115 differentially expressed genes, of which 36 were "novel RNAs". The gene list was subjected to hierarchical cluster analysis and principle component analysis (PCA) (Fig. 3B and C). The stroke (yellow) and control (blue) data show clear separation in both analyses (note the ellipse), indicating a distinct pattern of gene expression in blood following ischemic stroke that might have diagnostic potential. To test whether a panel of gene expression profiles has the ability to discriminate between stroke and control patients, we trained our dataset using a SVM mathematical model. Modeling suggested that 35 gene expression values could yield a normalized accuracy rate of 92%, with a potential sensitivity of 88% and specificity of 96% (Fig. 3D). This result supports the conclusion that gene expression analysis in blood, as determined by RNA-Seq, has the potential to identify patients who have suffered a stroke.

In order to improve the performance of our test, we investigated whether exon-specific expression values would improve the accuracy of the model. This concept was based on the observation that statistically significant changes in the differential splicing of 17 genes were observed following stroke ($\pm$1.2-fold change, post hoc false detection rate (FDR) of 0.1). For example, *SIGMAR1* does not show significant increase in gene expression following stroke, but significant changes in differential transcript expression were observed; following stroke, the alignment of reads to isoform NM_001282205 versus other isoforms increased (Fig. 4A). These changes in transcript isoforms may be detected by differential levels of exon expression. The expression of reads aligning to individual exons, which make up genes, and the novel RNAs identified with our novel annotation guide were assembled. Exon expression data were filtered and subjected to ANOVA to identify candidate modeling genes; 345 potential exons were identified (unadjusted $P < 0.001$), of which 53 were "novel RNA". SVM modeling was again employed to identify the smallest set of genes capable of discriminating between the control and MCA stroke datasets. The best model identified utilized 90 exons and had an estimated normalized accuracy of 94% (88% sensitivity and 100% specificity). As such this model only had two false negatives (patients who had suffered a stroke being called control), and it is of note that these patients had received rt-PA and showed an improvement in NIHSS scores upon discharge from hospital. Further testing of this model using two-level and bootstrap cross validation suggest that the achievable accuracy was 96% (see Supporting Information).

We investigated the exon RNA-Seq profiles from the 17 patients who suffered a confirmed MCA stroke. We used the following criteria to stratify the severity of stroke deficits based on the entry NIH Stroke scale rating of neurological deficits: 0–5 – Minor, 6–15 – moderate > 15 severe. MCA data were subjected to ANOVA (categorical stroke severity as the discriminating factor). When we perform this, we observe six genes which pass the FDR test. However, since all other datasets are reported as those genes with an unadjusted *P* value difference of 0.001, we will consider this larger dataset of 174 exon fragments (For gene exon lists see Tables S1–S8). Both hierarchical cluster and PCA analysis on the regulated exon data show a clear discrimination between the stoke patients $NIHss_0$ score (Fig. 4E and F). Together these data show that RNA-Seq analysis of whole blood transcriptomes has diagnostic potential, with respect to determining the severity of neurological deficit.

Of the 17 MCA patients, 11 were treated with the clot dissolving agent, rt-PA. We plotted the improvement in NIH stroke severity rating from entry ($NIHss_0$) to discharge ($NIHss_1$), and the percent improvement (Fig. 5A and B). When considered as a whole there was a significant overall reduction in NIHss scores from admission to discharge (Wilcoxon ranked sign test, $P < 0.01$), however, analysis of the rt-PA treated versus non-RT-PA treated groups did not yield a significant difference (Fig. 1B). Since two patients with a strong improvement in NIHss rating clustered with controls, we asked whether the gene expression datasets had prognostic ability to predict those patients who would have good versus poor outcomes. We

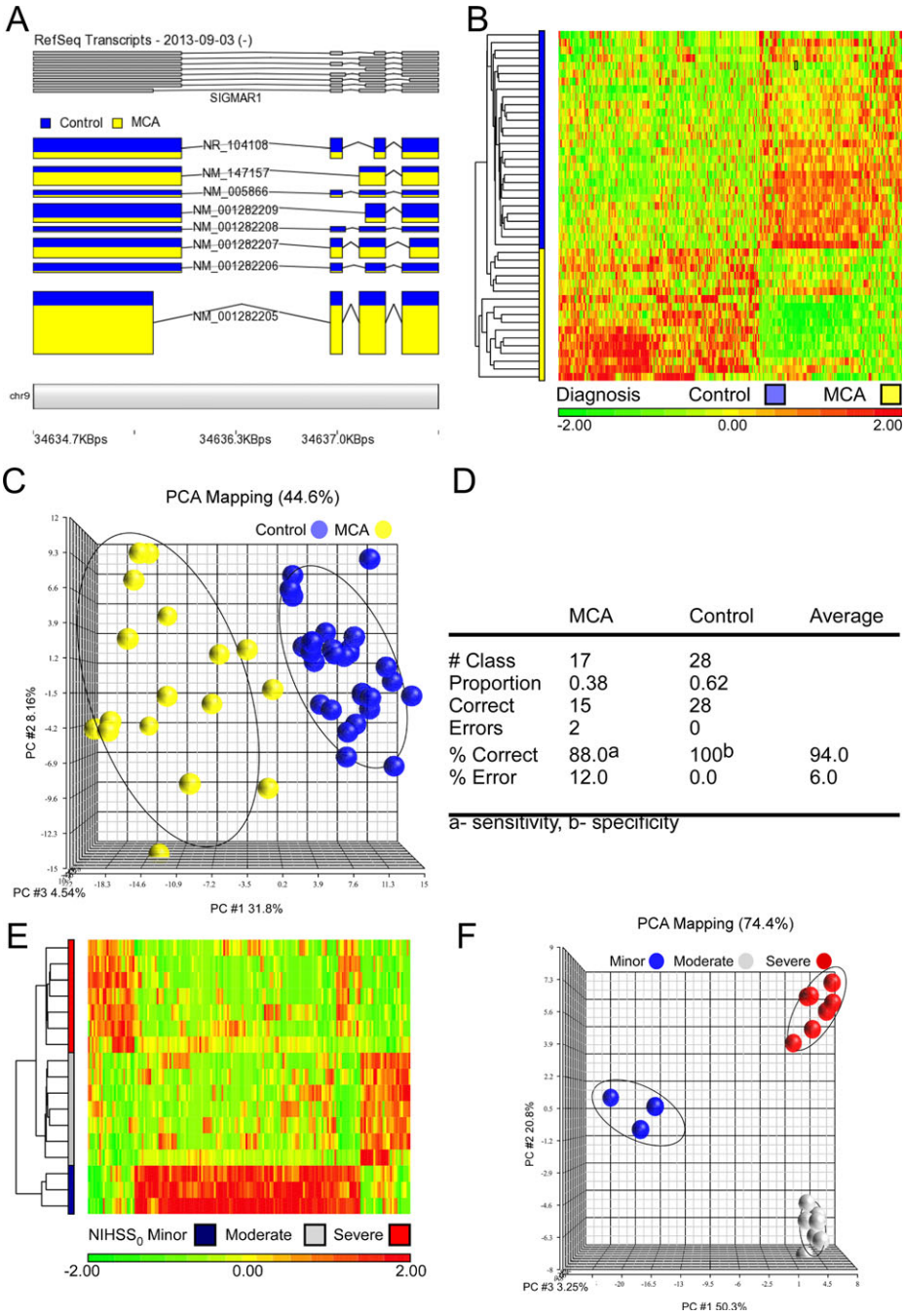


Figure 4. Prediction of stroke diagnosis using analysis of exon expression in blood. Exon expression was quantified in RNA-Seq libraries and differences between MCA patients (17) and controls (28) were analyzed. (A) Genome viewer image of alternative transcript splicing of *SIGMAR*1 in controls (blue) and MCA patients (yellow). Note the increase in nm_001282205 isoform following MCA. (B) Hierarchical clustering of 345 exons whose expression change (1.2-fold, unadjusted $P < 0.001$). (C) Principle component analysis of 345 exons which change between control and MCA patients. (D) Results matrix from support vector machine model to predict diagnosis based on exon expression values (Best model- 90 variables, SVM, Shrinking centroids, cost 201, nu 0.5, tol 0.001, kern, rbf deg 3, gamma 0.0001, coeff 0). (E) Hierarchical clustering of 174 exons showing a significant change (unadj. $P < 0.001$) with respect to the severity of their stroke based on admission $NIHss_0$ score in 17 confirmed MCA patients (stratified as 0–4 – minor [blue], 5–15 – moderate [gray] 15 < severe [red]). (F) Principle component analysis of 174 exons with respect to severity of admission $NIHss_0$ score (stratified as 0–4 – minor [blue], 5–15 – moderate [gray] 15 < severe [red]). MCA, middle cerebral artery; SVM, Support Vector Machine.

stratified the MCA stroke dataset based on the mean improvement in NIHss score from admission to discharge ($60 \pm 11.05\%$) (Fig. 5A): a prognosis of good was attributed to a mean improvement of $\geq$60% in NIHss. The exon expression data were filtered and subjected to ANOVA to identify candidate exons for modeling, revealing 144 exons differentially expressed (unadjusted $P < 0.001$). Hierarchical cluster and PCA analysis revealed strong clustering of the datasets (Fig. 5B and C). Preliminary SVM modeling (one-level cross validation based on 17 partitions [full leave one out]) revealed a model whereby a set of 15, 20, 25, and 30 exon expression patterns had the estimated normalized correct rate of 100% (sensitivity and specificity of 100%) (Fig. 5D). We performed a two-level cross-validation test of the model that also had a 100% normalized accuracy rate for 15–30 variables (inner cross-validation partition – 10 random partitions) Subsequent bootstrap cross-validation suggest a normalized correct rate of 99%, 97%, 95%, and 100%, respectively, therefore a model of 30 exon expression variables would appear to be our best model to predict prognosis (sup data contains the exon list).

In addition, we subjected the MCA data to ANOVA using the stratification of the discharge NIH stroke scale ratings as above ($NIHss_1$). We observe 31 genes that pass the FDR test ($P < 0.05$). However, since all other datasets are reported as those genes with an unadjusted P value difference of 0.001, we will considered the larger dataset of 359 exon fragments (For gene exon lists see Tables S1–S8). Both hierarchical cluster and PCA analysis on the regulated exon data show a clear discrimination between the severity of the patients discharge stoke $NIHss_1$ score (Fig. 5E and F). We annotated the patients who received rt-PA on the hierarchical cluster to show there was no clear clustering effect of rt-PA in this analysis (Fig. 5F). Together these data strongly support the further investigation of gene and exon expression analysis to identify discriminant panels for stroke diagnosis and prognosis.

Discussion

In this study, we investigate gene expression and exon expression data from peripheral blood, and use these data to build models with discriminant power to identify stroke patients from controls without stroke, and to categorize stroke prognosis. Attempts at biomarker identification for neurocritical care have generated many approaches. The majority of these have focused on single protein biomarkers in a biofluid (usually blood or c.s.f.). For example, efficacy (surrogate) biomarkers, such as beta amyloid have been investigated for Alzheimer's disease therapies, and the release of intracellular proteins from neurons has been a focus as a biomarker for brain injury.[21] While proteomic approaches to identify panels of proteins are becoming more common,[22] few single protein biomarkers have the sophistication to subtype neurological injuries. As such the application of gene expression analysis, using unbiased application of mathematical cluster analysis to identify stroke subtypes offers promise. Studies of microarrays and NGS show the diagnostic potential of such an approach.[5–8,10,23–25] Using molecular technology, preliminary diagnostic discrimination following stroke is over 90% accurate, with similar high sensitivity and specificity. Our data show a clear cluster stratification to enable the prediction of admission NIHss severity ratings. The novel application of exon expression and utilization of the dataset for prognostic stratification (both discharge NIHss and improvement in NIHss) show the clear utility of this approach to aid in clinical patient care.

Here, we report the feasibility of NGS to identify panels of genes for stroke discrimination. We focused on a homogeneous subset of ischemic hemiparesis patients due to MCA occlusion. We recruited 28 stroke patients, and focused on 17 confirmed MCA territory stroke for this study. In this study, we chose to collect blood during the morning blood draw, to reduce circadian effects.[26] As such this generated a series of samples with a variety of durations since stroke onset. However, we did not observe a significant effect of this factor in our data (also see[5]). The goal of this study was to identify optimum methodology for obtaining candidate gene sets for mathematical modeling rather than a biological interpretation. We note that gene ontology analysis identified candidate biological signaling pathways associated with nuclear events and inflammatory processes to be regulated following MCA (Tables S1–S8); however, these were not investigated further in this study. Preliminary modeling studies suggest that the data are sufficiently heterogeneous to enable a discriminant model to be identified and built. Our preliminary data based on gene expression levels suggest an accuracy of 92% (88% sensitivity and >90% specificity). These values are similar to the accuracy obtained by previous microarray studies.[6,8,23] Using a unique approach of evaluating exon-specific expression patterns (due to alternative transcript splicing) we obtained a higher normalized accuracy rate of 93%, with 100% specificity. Notable is the observation of novel-regulated RNA discovered by the whole transcriptome analysis approach. Only 30% of the identified RNA transcripts align with known protein coding exons, 36% align with annotated introns, and 33% align to novel, not previously reported genomic regions (per RefSeq reference: Fig. 2A). The sensitivity of gene isoform and gene exon expression to ischemia and other cardiovascular diseases has recently been reported.[10,19] We further stratified the data based on the average improve-

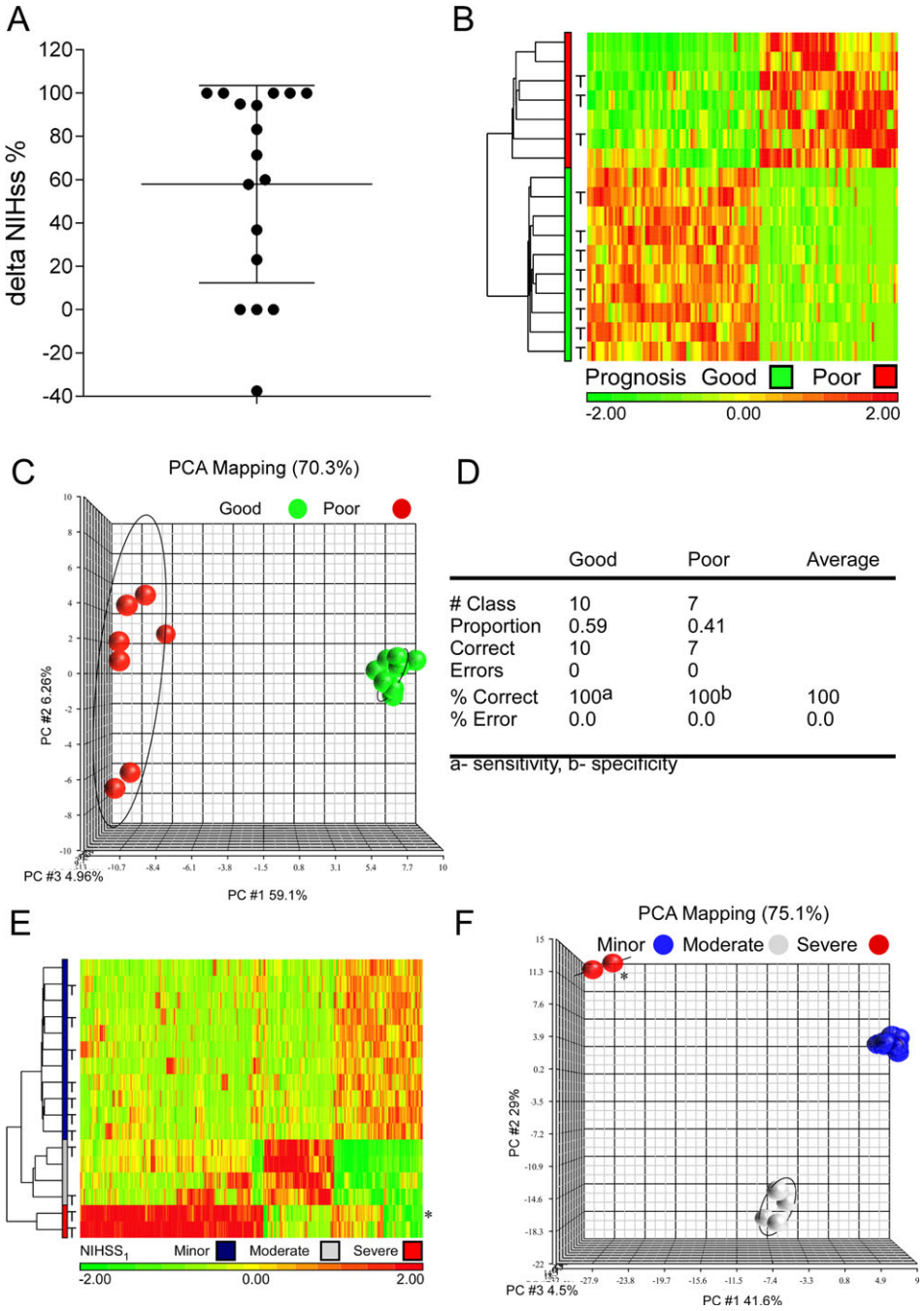


Figure 5. Prediction of stroke prognosis following mathematical modeling of RNA-Seq data. (A). The improvement of NIHSS rating from admission to discharge of 17 MCA patients subjected to sequencing. From this data, a mean improvement was the level for our stratification of good versus poor. (B) Hierarchical clustering of 144 exons whose expression change (1.2-fold, unadjusted $P < 0.001$) between MCA patients with good or poor outcome (defined as mean NIHss improvement or better). (C) Principle component analysis of 144 exons which change between MCA with better than mean improvement of NIHss. (D) Results matrix from support vector machine model to predict prognosis based on exon expression values (Best model – 30 variables, SVM, Shrinking centroids, cost 1, nu 0.5, tol 0.001, kern, rbf deg 3, gamma 0.01, coeff0 0). (E) Hierarchical clustering of 359 exons showing a significant change (unadj. $P < 0.001$) with respect to the severity of their stroke based on discharge $NIHss_1$ score in 17 confirmed MCA patients (stratified as 0–5 – minor [blue], 6–15 – moderate [gray] > 15 severe [red]; asterisk (*)-denotes deceased patient). (F) Principle component analysis of 359 exons with respect to severity of admission $NIHss_1$ score (stratified as 0–5 – minor [blue], 6–15 – moderate [gray] 15 < severe [red]). Patients treated with rt-PA denoted by T. NIHSS, NIH stroke scale; MCA, middle cerebral artery; SVM, Support Vector Machine; rt-PA, recombinant tissue plasminogen activator.

ment in NIHss score and found that a set of 15 genes identified in admission blood samples, have the power to discriminate between the outcome subgroups at discharge with 100% sensitivity and specificity.

Clearly this study would not have utility for clinical diagnosis, due to the significant time it takes to build, and sequencer RNA-Seq libraries. However, the current approach may yield panels of genes that could be utilized for more rapid technological platforms, such as qPCR. Faster electrode-based sequencing systems are on the horizon[27] and show much promise for the speeding up of the transcriptome analysis process, which may make this approach suitable for point of care use in the future. An alternative will be the development of a rapid PCR-based multiplex assay, which can be performed and analyzed rapidly (>1 h). Clearly this is an area where technological advances may have a dramatic impact on point of care diagnostics in critical acute situations.

The higher risk of stroke in the African American population has been well documented, and accordingly our area for investigation.[11] While not part of this study, the expansion of these data to include Caucasians, may reveal additional gene expression profiles which are unique for racial susceptibility to stroke. It is of note that the previous RNA-Seq study of blood transcriptomics was performed on samples obtained from Caucasian patients.[10] The identification of such candidate gene expression profiles may also open the opportunity of targeting drugs to such population differences. In addition to African American populations having a higher stroke risk,[12] a recent retrospective analysis of rt-PA studies suggested that African American women show no significant benefit from rt-PA.[13] Analysis of our transcriptome data show a clear transcriptome difference by sex following rt-PA administration (unpubl. obs. Robert Meller RM), which agrees with reports of gene expression differences following stroke between males and females.[28,29] These preliminary data show the clear need for further studies to better understand both racial and sex-associated differences in response to stroke.

The possibility of a blood biomarker test to diagnose and subtype stroke appears highly feasible. Previous studies using microarrays have set the stage for this technological approach to clinical diagnosis of neurological disorder.[5,6,8,15,23] More recently, the same group published a preliminary RNA-Seq study.[10] In this study, they were able to show differential gene splicing in association with various subtypes of stroke (intracerebral hemorrhagic stroke vs. ischemic stroke). While different protocols were used between our study and the Dykstra-Aiello study, together they have similar findings. First, that RNA-Seq studies of whole blood have diagnostic potential, with respect to subtyping[10] or severity assessment (Fig. 4). Second that prognosis is feasible based on RNA-Seq data from whole blood (Fig. 5), which may have more accuracy than predictions based on entry NIHss ratings. Although the two studies investigated different racial profiles of patients, cellular immune signaling pathways were commonly regulated following stroke (Tables S1–S8). Both studies show the clear potential of RNA-Seq studies on stroke patients, and support the larger scale investigation of this approach for clinical use. Clearly clinical adoption will require the application of faster techniques than current RNA-Seq platforms to identify key genes whose expression predict stroke diagnosis and prognosis.

Limitations of the study

The following limitations of our study are noted; first we had a limited sample size, which may have reduced our ability to observe many expression changes that passed a false detection rate post hoc test. The low sample size also impacted our use of unadjusted *P* values in our cluster models. The purpose of the study was to determine the utility of the methodology and process. However, our preliminary analysis suggests that further samples will reveal more genes that are statistically significantly regulated, and some of the expression differences pass FDR post hoc analysis (e.g., NIHss1).

We tested the ability of our data to model the diagnosis based on gene expression using SVM models. These were initially run with one-level cross validation (full leave one out). Further tests using both bootstrap and two-level cross validation (sets of 10) suggest robust models. Unfortunately, because of our sample size we were unable to test these models on additional datasets, as previous studies have. This is a clear goal for the future, to refine the models to identify the minimal gene exon expression dataset that can accurately identify the clinical diagnosis of stroke. Because of our small sample size we did not perform a subtyping analysis, therefore it is not yet clear whether this panel of genes will identify other stroke subtypes, or are unique of those to MCA stroke. When we have further samples, we will perform an analysis of stroke subtyping, similar to that shown using microarray. However, we were able to show distinct exon expression patterns associated with the severity of the neurological deficits of the patients, as determined using the NIH Stroke scale (Fig. 4).

The omission of a ribosomal depletion step (to simplify the experimental pipeline and because pull-down protocols yielded variable results [not shown]) resulted in loss of 80% of usable reads. Future experiments will include an rRNA pull down step, which will increase coverage of the transcriptome.[30] The depth of sequencing was between 5 and 10 million aligned reads. While lower than ENCODE standards, this level is compatible with recent studies suggesting that a read depth of 10 million aligned

fragments is sufficient for most studies.[31,32] We chose not to perform Globin mRNA depletion for similar reasons. In addition, we find that each depletion step may change the expression of nontarget mRNAs (J. H. unpubl. obs.).

Summary

In summary, our proof-of-principal study demonstrates the ability of NGS transcriptome profiling to identify diagnostic gene expression patterns following stroke, and prognostic patterns as well. Our analysis highlights the potential of clinical diagnostic and prognostic information generated by such studies. In addition, these observations suggest that further investigations into the large noncoding RNA signature in blood following stroke will reveal valuable prognostic and diagnostic information. This pilot study strongly suggests that the RNA-Seq analysis of the blood transcriptome will yield novel and instrumental data for utilization in the development of diagnostic biomarker panels to aid in the identification and subtyping of stroke. In addition, identification of prognostic biomarkers may assist with clinical care decisions, and identify novel targets for hypothesis driven therapy.

Acknowledgments

We acknowledge assistance of the Morehouse School of Medicine Clinical Research Center for recruiting controls (J. Wainwright, E. Stanley, N. Silvestrov), Martha B Johnson Ph.D. for manuscript preparation and the technical assistance of Life Technologies (Now Thermo Fisher Scientific) Technical Assistance teams (M. Osentoski, L. Combs & Y. Wang).

Author Contribution

R. M., R. P. S., M. F., D. M., and C. H. conceptualized the project, Data collection and analysis was performed by R. M., J. H., and A. P. Stroke confirmation was by C. H., M. F., and R. P. S. Manuscript preparation was performed by R. M., J. H., A. P., and R. P. S.

Conflict of Interest

Dr. Hall reports grants from National Institute of Health NINDS, outside the submitted work. Dr. Meller reports grants from NIH, during the conduct of the study.

References

1. Cronin CA. Intravenous tissue plasminogen activator for stroke: a review of the ECASS III results in relation to prior clinical trials. J Emerg Med 2010;38:99–105.
2. Gonzalez RG. Clinical MRI of acute ischemic stroke. J Magn Reson Imaging 2012;36:259–271.
3. Mozaffarian D, Benjamin EJ, Go AS, et al. Heart disease and stroke statistics-2015 update: a report from the American Heart Association. Circulation 2015;131:e29–322.
4. Reynolds MA, Kirchick HJ, Dahlen JR, et al. Early biomarkers of stroke. Clin Chem 2003;49:1733–1739.
5. Sharp FR, Xu H, Lit L, et al. Genomic profiles of stroke in blood. Stroke 2007;38:691–693.
6. Stamova B, Xu H, Jickling G, et al. Gene expression profiling of blood for the prediction of ischemic stroke. Stroke 2010;41:2171–2177.
7. Jickling GC, Xu H, Stamova B, et al. Signatures of cardioembolic and large-vessel ischemic stroke. Ann Neurol 2010;68:681–692.
8. Jickling GC, Stamova B, Ander BP, et al. Prediction of cardioembolic, arterial, and lacunar causes of cryptogenic stroke by gene expression and infarct location. Stroke 2012;43:2036–2041.
9. Shendure J. The beginning of the end for microarrays? Nat Methods 2008;5:585–587.
10. Dykstra-Aiello C, Jickling GC, Ander BP, et al. Intracerebral hemorrhage and ischemic stroke of different etiologies have distinct alternatively spliced mRNA profiles in the blood: a pilot RNA-Seq study. Transl Stroke Res 2015;6:284–289.
11. Howard G, Cushman M, Kissela BM, et al. Traditional risk factors as the underlying cause of racial disparities in stroke: lessons from the half-full (empty?) glass. Stroke 2011;42:3369–3375.
12. Romano JG, Smith EE, Liang L, et al. Outcomes in mild acute ischemic stroke treated with intravenous thrombolysis: a retrospective analysis of the Get With the Guidelines-Stroke Registry. JAMA Neurol 2015;72:423–431.
13. Mandava P, Murthy SB, Munoz M, et al. Explicit consideration of baseline factors to assess recombinant tissue-type plasminogen activator response with respect to race and sex. Stroke 2013;44:1525–1531.
14. Wei P, Milbauer LC, Enenstein J, et al. Differential endothelial cell gene expression by African Americans versus Caucasian Americans: a possible contribution to health disparity in vascular disease and cancer. BMC Med 2011;9:2.
15. Sharp FR, Jickling GC, Stamova B, et al. Molecular markers and mechanisms of stroke: RNA studies of blood in animals and humans. J Cereb Blood Flow Metab 2011;31:1513–1531.
16. Roberts A, Pimentel H, Trapnell C, Pachter L. Identification of novel transcripts in annotated genomes using RNA-Seq. Bioinformatics 2011;27:2325–2329.
17. Trapnell C, Roberts A, Goff L, et al. Differential gene and transcript expression analysis of RNA-Seq experiments with TopHat and Cufflinks. Nat Protoc 2012;7:562–578.

18. Marioni JC, Mason CE, Mane SM, et al. RNA-Seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res 2008;18:1509–1517.
19. Joehanes R, Ying S, Huan T, et al. Gene expression signatures of coronary heart disease. Arterioscler Thromb Vasc Biol 2013;33:1418–1426.
20. Mortazavi A, Williams BA, McCue K, et al. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 2008;5:621–628.
21. Chou SH, Robertson CS. Monitoring biomarkers of cellular injury and death in acute brain injury. Neurocrit Care 2014;21 (Suppl 2):S187–214.
22. Bian F, Simon RP, Li Y, et al. Nascent proteomes in peripheral blood mononuclear cells as a novel source for biomarker discovery in human stroke. Stroke 2014;45:1177–1179.
23. Jickling GC, Ander BP, Stamova B, et al. RNA in blood is altered prior to hemorrhagic transformation in ischemic stroke. Ann Neurol 2013;74:232–240.
24. Jickling GC, Stamova B, Ander BP, et al. Profiles of lacunar and nonlacunar stroke. Ann Neurol 2011;70:477–485.
25. Stamova B, Ander BP, Barger N, et al. Specific regional and age-related small noncoding RNA expression patterns within superior temporal gyrus of typical human brains are less distinct in autism brains. J Child Neurol 2015;30:1930–1946.
26. Zhu J, Chen Y, Leonardson AS, et al. Characterizing dynamic changes in the human blood transcriptional network. PLoS Comput Biol 2010;6:e1000671.
27. Mikheyev AS, Tin MM. A first look at the Oxford Nanopore MinION sequencer. Mol Ecol Resour 2014;14:1097–1102.
28. Tian Y, Stamova B, Jickling GC, et al. Effects of gender on gene expression in the blood of ischemic stroke patients. J Cereb Blood Flow Metab 2012;32:780–791.
29. Stamova B, Jickling GC, Ander BP, et al. Gene expression in peripheral immune cells following cardioembolic stroke is sexually dimorphic. PLoS One 2014;9:e102550.
30. Tarazona S, Garcia-Alcalde F, Dopazo J, et al. Differential expression in RNA-Seq: a matter of depth. Genome Res 2011;21:2213–2223.
31. Hart SN, Therneau TM, Zhang Y, et al. Calculating sample size estimates for RNA sequencing data. J Comput Biol 2013;20:970–978.
32. Liu Y, Zhou J, White KP. RNA-Seq differential expression studies: more sequence or more replication? Bioinformatics 2014;30:301–304.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Figure S1. Overview of RNA Seq pipeline. This figure details the procedures employed for the collection of blood, extraction of RNA, Whole transcriptome library assembly, sequencing and bioinformatic analysis. Extraction library assembly and sequencing was performed in batches of eight samples (four controls and four patients). The library creation took ~5 lab days, and the sequencing run was 5–6 day duration.

Table S1. Genelist_Gene_MCA versus Cont. List of genes with unadjusted $P < 0.001$ between MCA and control groups. Gene ID are from RefSeq, and XLOC refers to novel RNAs identified in study. These data were used to create Figure 3C and D.

Table S2. GO_Gene_MCA versus control. Gene ontology analysis of gene IDs from Table S1.

Table S3. Genelist EXON MCA versus control. List of exon fragments significantly different between control and MCA groups (Unadjusted $P < 0.001$). Column ID are chromosomal coordinates and GeneID from RefSeq, and XLOC refers to novel RNAs identified in study. These data were used to create Figure 4C and D.

Table S4. Genelist EXON NIHss0. List of exon fragments significantly different between different NIHSss0 groups (Unadjusted $P < 0.001$). Column ID are chromosomal coordinates and GeneID from RefSeq, and XLOC refers to novel RNAs identified in study. These data were used to create Figure 4E and F.

Table S5. Genelist EXON prognosis. List of exon fragments significantly different between patients with above average (Good) or below average (Poor) improvement in NIHss score between admission and discharge (Unadjusted $P < 0.001$). Column ID are chromosomal coordinates and GeneID from RefSeq, and XLOC refers to novel RNAs identified in study. These data were used to create Figure 5C and D.

Table S6. Genelist EXON NIHss1. List of exon fragments significantly different between different NIHSss1 groups (Unadjusted $P < 0.001$). Column ID are chromosomal coordinates and GeneID from RefSeq, and XLOC refers to novel RNAs identified in study. These data were used to create Figure 5E and F.

Table S7. Model for Prognosis. List of 30 exon fragments with best model for prediction based on Support vector machine modeling and Bootstrap cross validation on groups of 10.

Table S8. Model for diagnosis. List of 90 exon fragments with best model for prediction of diagnosis based on Support vector machine modeling and Bootstrap cross validation on groups of 10.

