IDENTIFYING RNA BIOMARKERS OF CEREBROVASCULAR DISEASE

IDENTIFYING RNA BIOMARKERS OF CEREBROVASCULAR DISEASE

By KRIPA RAMAN, B.Sc. (Hons.)

A Thesis Submitted to the School of Graduate Studies In Partial Fulfillment of the Requirements For the Degree Doctor of Philosophy

DOCTOR OF PHILOSOPHY (2015) McMaster University
(Department of Medical Sciences) Hamilton, Ontario

TITLE: Identifying RNA biomarkers of cerebrovascular disease

AUTHOR: Kripa Raman
B.Sc. (Honours) Molecular Biology and Genetics
(McMaster University)

SUPERVISOR: Dr Guillaume Paré, MD, M.Sc., FRCPC.

NUMBER OF PAGES: xv – 174

ABSTRACT

Stroke is an acute neurological deficit that results from abnormal blood flow to the brain. The term stroke encompasses two primary subgroups: hemorrhagic stroke that is due to extravasation of blood and ischemic stroke that is due to vessel obstruction. Determining stroke type and underlying etiology is a crucial step in patient management as it influences treatment strategies. Currently diagnosis of stroke relies on clinical examination and neuroimaging, but there is a lack of rapid diagnostic and prognostic testing. Using microarray technology we identified a novel association between elevated peripheral blood expression of *MCEMP1* and stroke. We have also shown that *MCEMP1* discriminates between primary stroke types and predicts one-month post-stroke prognosis. Since genetic mechanisms underlying stroke remain incompletely understood we next conducted a global gene network analysis. Network analysis identified four large groups of co-expressed genes associated with ischemic stroke. *NLRC4*, *CKLF*, and *HS.546375* were the most interconnected genes within unique modules and each was also independently associated with ischemic stroke. We show that multi-gene models have greater discriminative capacity for stroke and stroke prognosis, than single gene models. In addition to stroke biomarkers we also identified biomarkers of atrial fibrillation (AF), a known risk factor of stroke. Currently our understanding of the molecular mechanisms underlying AF remains incompletely understood. Thus we conducted whole blood expression profiling in patients with persistent AF before and after successful electrical cardioversion, a procedure that aims to restore sinus rhythm to the heart. We identified elevated expression of *SLC25A20* and *PDK4* during AF as compared with sinus rhythm.

Furthermore we show that *SLC25A20, PDK4* and NT-proBNP have incremental utility to discriminate AF from sinus rhythm. Taken together, the thesis implicates new genes with stroke and AF, and also indicates that whole blood RNA biomarkers may have clinical utility.

ACKNOWLEDGEMENTS

First, and foremost I would like to express my gratitude towards my supervisor Dr. Guillaume Paré. I never thought that a one-year undergraduate thesis project would inspire me to undertake a PhD, but here I am six years later. It was Dr. Paré who gave me the first opportunity to prove myself as a scientist. His mentorship, support and confidence in my ability have in turn built my confidence and taught me to be a critical researcher.

I am also fortunate to have many mentors at PHRI and I extend my gratitude towards my PhD committee members, Dr. Robert Hart and Dr. Matthew McQueen, for their guidance and support.

I would also like to thank all of the past and present members of the Genetic and Molecular Epidemiology Lab (GMEL) for their constant support and friendship. In particular, I would like to acknowledge Dr. Stephanie Ross for her skilled editing and Reina Ditta, Amanda Hodge, Michelle Souter, and Shana Hayter for processing samples and providing me with plentiful data to analyze. You have all made the past few years at PHRI memorable and I wish you all the best.

Finally I would like to thank all of my friends and family for their unwavering support. Without all of you this great accomplishment would not have been possible. A special thank you also goes to Ryan Tice for the endless encouragement.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

CHAPTER 3

CHAPTER 4

CHAPTER 5

LIST OF TABLES

CHAPTER 1

CHAPTER 3

CHAPTER 4

CHAPTER 5

LIST OF ABBREVIATIONS

ACTB	Beta-Actin
AF	Atrial Fibrillation
ASSERT	ASymptomatic atrial fibrillation and Stroke Evaluation in pacemaker patients and the atrial fibrillation Reduction atrial pacing Trial
AUC	Area Under the Receiver Operating Curve
BBB	Blood Brain Barrier
BMI	Body Mass Index
cDNA	Complementary Deoxyribonucleic Acid
CI	Confidence Interval
CKLF	Chemokine-like factor
CLPA	Chemical Ligation dependent Probe Amplification chemistry
CRP	C-Reactive Protein
CT	Computed Tomography
CT	Cycle Threshold
DAMP	Damage Assocaited Molecular Pattern
ECG	Electrocardiogram
ED	Emergency Department
FC	Fold Change
GS	Gene Significance
HS.546375	Gene with unknown function
Hub Gene	Highly connected gene
IDVMIA	In Vitro Diagnostic Multivariate Assay
ITGAM	Integrin, alpha M
ITGB5	Integrin, beta 5
Log2	Logarithm base 2

MAQC	MicroArray Quality Control
MCEMP1	Mast-cell expressed membrane protein 1
MDS	Multi-Dimensional Scaling
MM	Module Membership
MRI	Magnetic Resonance Imaging
mRNA	Messenger Ribonucleic Acid
mRS	Modified Rankin Score
MS	Module Summary
neqc	NormExp background correction using control probes
NIHSS	National Institutes of Health Stroke Scale
NLR family	Nucleotide-binding domain and leucine-rich repeat containing receptor family
NLRC4	NLR family, CARD domain containing 4
NLRP3	NLR family, pyrin domain containing 3
NPV	Negative Predictive Value
NRI	Net Reclassification Index
NT-proBNP	N-Terminal pro B-type Natriuretic Peptide
PCA	Principle Component Analysis
PDK4	Pyruvate dehydrogenase kinase, isozyme 4
PFO	Patent Foramen Ovale
PLAN Score	Preadmission comorbidities, Level of consciousness, Age, and Neurologic deficit Score
PPV	Positive Predictive Value
proBNP	Pro B-type Natriuretic Peptide
QC	Quality Control
qPCR	Quantitative Real-Time Polymerase Chain
RCT	Randomized Control Trial

RNA	Ribonucleic Acid
ROC	Receiver Operating Curve
SLC25A20	Solute carrier family 25 (carnitine/acylcarnitine translocase), member 20
SPIA	Signalling Pathway Impact Analysis
tPA	Tissue Plasminogen Activator
WGCNA	Weighted Gene Co-expression Network Analysis

CHAPTER 1: GENERAL INTRODUCTION

1.1 BURDEN OF STROKE

Stroke is the underlying cause of 11.1% of all deaths worldwide (Lozano *et al.*, 2012) and is the third leading cause of disability (Murray *et al.*, 2012). Stroke is classically defined as a neurological deficit that results from acute focal injury to the central nervous system due to a vascular cause (Donnan *et al.*, 2008; Sacco *et al.*, 2013). Globally, stroke consumes approximately 2 to 7% of total health-care costs (Evers *et al.*, 2004). In Canada alone, there are approximately 38, 000 stroke admissions each year, which costs the health care system over $2.8 billion CAD annually (Mittmann *et al.*, 2012).

Age, gender (Appelros *et al.*, 2009) and ethnicity (Stewart *et al.*, 1999; Sacco *et al.*, 2001; Stansbury *et al.*, 2005) are non-modifiable risk factors for stroke. The INTERSTROKE study further characterized ten additional stroke risk factors; hypertension, current smoking, abdominal obesity, diet, physical activity, diabetes mellitus, alcohol intake, psychosocial stress, cardiac causes and ratio of apolioprotein B to A1 (Martin J O'Donnell *et al.*, 2010). Stroke incidence increases with age and as global life expectancy increases in high-income counties, the prevalence and cost of stroke in Canada is projected to escalate in the future.

1.2 CLINICAL DIAGNOSIS OF STROKE

Stroke is a broad disease category encompassing two primary subgroups, ischemic stroke and hemorrhagic stroke. Ischemic stroke results from vessel obstruction and can be further subdivided based on the underlying stroke etiology (Adams *et al.*, 1993; Ay *et al.*, 2007). Hemorrhagic stroke results from extravasation of blood within or on the brain following vessel rupture. Approximately 72% of first-time stroke cases are due to ischemia, and 28% are due to hemorrhage (Krishnamurthi *et al.*, 2013). The diagnosis of stroke and distinction between primary stroke types is currently based on clinical examination and neuroimaging. However, clinical symptoms of stroke are heterogeneous. Both ischemia and hemorrhage cause damage to neurologic tissue leading to symptoms such as: hemiparesis, hemisensory loss, aphasia, impaired vision and/ or headaches. Non-vascular disease can also mimic the symptoms of stroke. Approximately 5% of patients with stroke-mimics, such as seizures, migraines and sepsis, are misdiagnosed with ischemic stroke (Scott and Silbergleit, 2003). Since primary stroke type cannot be identified solely based on clinical assessment, neurologic imaging with computed tomography (CT) or magnetic resonance imaging (MRI) is central for diagnosis. Though MRI is superior to CT for stroke diagnosis (Fiebach *et al.*, 2002; Kidwell *et al.*, 2004; Chalela *et al.*, 2007), CT is generally the modality of choice due to its lower cost and greater availability. However, CT scans are not sensitive for early ischemic stroke and studies report only moderate agreement between physicians interpreting early CT images (Grotta *et al.*, 1999; Wardlaw and Mielke, 2005). Since stroke diagnosis relies on neuroimaging, ambiguous CT results can hinder patient management.

1.3 THERAPEUTIC TREATMENT OF STROKE

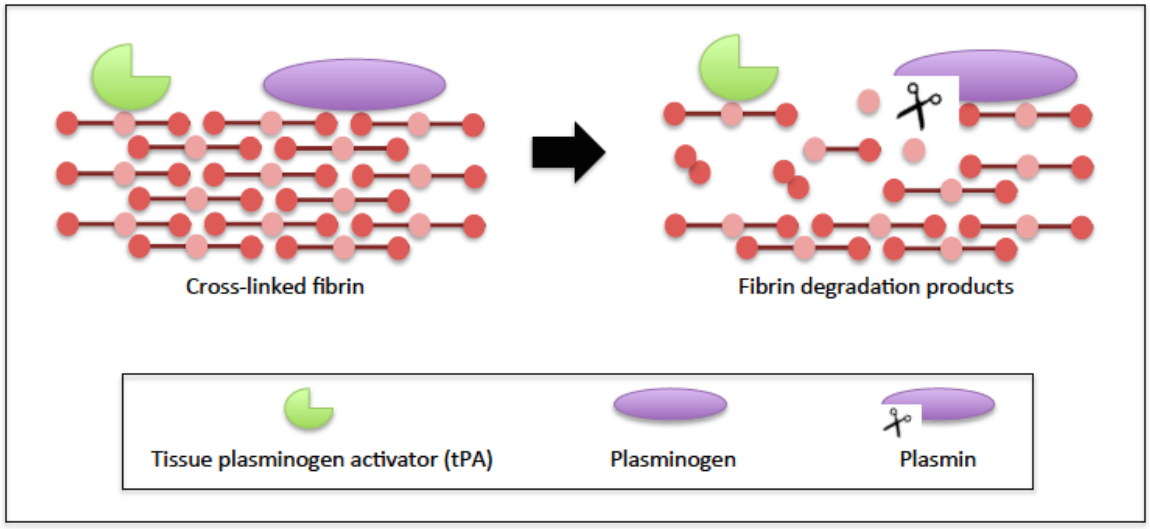
Multiple studies report that thrombolytic therapy with tissue plasminogen activator (tPA) is a cost-effective and safe treatment for ischemic stroke, (Troke and Roup, 1995; Yip and Demaerschalk, 2007; Hacke *et al.*, 2008) but tPA is currently underused. Indeed only 2 to 9% of patients with ischemic stroke are treated with tPA (Hill and Buchan, 2005; Nadeau *et al.*, 2005; Adeoye *et al.*, 2011). tPA is a serine protease that specifically bind to fibrin to promote clot dissolution by facilitating the conversion of plasminogen to plasmin (Hoylaerts *et al.*, 1982) (*Figure 1.1*). Therefore hemorrhagic stroke is a direct contraindication for tPA therapy; clot dissolution therapy worsens hemorrhage. Since treatment differs significantly between ischemic and hemorrhagic stroke, distinction between these primary subtypes is a crucial step in rapid stroke management.

The efficacy of tPA is also time dependent. Studies indicate that tPA offers the most benefit and least harm when prescribed to ischemic stroke patients within 4.5 hours of symptom onset (Lansberg *et al.*, 2009); patients given tPA after more than 4.5 hours of symptom onset had greater occurrence of intracranial hemorrhage and mortality (Emberson *et al.*, 2014). Rapid stroke diagnosis is, in part, hindered by the lack of confirmatory diagnostic testing but also by lack of rapid pre-hospital diagnostic tests for stroke. Experienced neurologists excel at diagnosing stroke, but stroke diagnosis may also be made by less experienced practitioners such as emergency department (ED) and rural physicians. Due to the associated risks, some non-neurologists are hesitant to prescribe tPA. For instance, 65% of ED physicians report feeling uncomfortable using tPA without

a neurology consult (Scott *et al.*, 2010). Furthermore, rural physicians report that the biggest barriers to the use of tPA are risk of hemorrhage and diagnostic uncertainty (Williams *et al.*, 2013). As a result, there would be clinical utility in identifying additional diagnostic tools for stroke that supplement current clinical practice.

Figure 1.1 Stages of clot dissolution using tPA.

Ischemic stroke results from vessel obstruction. Blood clots can form within one or more arties that supply blood to the brain, or can embolize from secondary locations, and inhibit blood flow. These clots are commonly composed of cross-linked fibrin strands. Both tPA and plasminogen bind to fibrin. Since the proteins are localized together, tPA converts plasminogen to plasmin. Plasmin then cleaves the cross-linked fibrin clot generating soluble degradation products. Clot dissolution facilitates recanalization of the blood vessel.



1.4 PREDICTING STROKE PROGNOSIS

Predicting a patient's risk for mortality or severe disability may be used to inform clinical decision-making and optimize stroke resource allocation. Modified Rankin Scale (mRS) score is a commonly used end point to assess disability and independence after stroke. The scale consists of simple, well-defined grades that describe global disability (Rankin, 1957; UK-TIA Study Group, 1988) (Table 1.1). A mRS score of 0 is indicative of no disability, 5 is disability requiring constant care and 6 is death.

Although stroke is a leading cause of death and functional disability, tools to predict patient outcome or mRS are infrequently used by clinicians. Currently physicians predominantly use clinical judgment to estimate stroke prognosis, yet several clinical scores have been proposed to predict functional outcome (Adams Jr. *et al.*, 1999; Weimar *et al.*, 2004; Smith *et al.*, 2010; Saposnik *et al.*, 2012; G. *et al.*, 2013; Saposnik *et al.*, 2013; Flint *et al.*, 2013). For instance, the National Institutes of Health Stroke Scale (NIHSS) has been shown to predict stroke outcome and mortality (Adams Jr. *et al.*, 1999; Weimar *et al.*, 2004; Saposnik *et al.*, 2013). NIHSS takes into account: level of consciousness, eye movement, motor skills, language, speech and attention. However, clinicians infrequently record NIHSS scores. For example in a stroke related study conducted in 1036 hospitals, NIHSS was recorded for only 40% of the 274,988 stroke patients (Smith *et al.*, 2010). NIHSS score may not be recorded due to the time required to conduct the assessment or unfamiliarity of the score by non-neurologists. Therefore prognosis scores involving subscales, including NIHSS, (Adams Jr. *et al.*, 1999; Weimar *et al.*, 2004; Smith *et al.*, 2010; Saposnik *et al.*, 2012; G. *et al.*, 2013; Saposnik *et al.*,

2013; Flint *et al.*, 2013) would also be infrequently used in the clinic. Some prediction scores are not used because they lack precision (Weimar *et al.*, 2002; König *et al.*, 2008; Lewis *et al.*, 2008) or involve complex scoring systems that are challenging to recall (Smith *et al.*, 2010; Saposnik *et al.*, 2012). Alternatively, the PLAN score has been developed. A patients PLAN score is determined based on preadmission comorbidities (P), level of consciousness (L), age (A) and neurologic deficit (N) (O'Donnell *et al.*, 2012). The PLAN score was able to predict 30-day mortality and 1-year mortality in the study population that was used to derive the score. However validation of the score is required in an independent population to confirm significance. Since neither NIHSS nor PLAN scores are regularly used in the clinic, there is still a need to identify simple tools to predict stroke prognosis.

Table 1.1 Modified Rankin Scale (mRS) score.

SCORE	DESCRIPTION
0	No symptoms at all
1	No significant disability despite symptoms; able to carry out all usual duties and activities
2	Slight disability; unable to carry out all previous activities but able to look after own affairs without assistance
3	Moderate disability; requiring some help but able to walk without assistance
4	Moderately severe disability; unable to walk without assistance and unable to attend to own bodily needs without assistance
5	Severe disability; bedridden, incontinent and requiring constant nursing care and attention
6	Dead

1.5 PERIPHERAL BLOOD BIOMARKERS

Biomarkers may be used to facilitate stroke diagnosis and prediction of prognosis. A biomarker is a substance measureable in the body or in bodily fluid that is an indicator of disease presence or outcome. Gold-standard biomarker tests are rapid, relatively non-invasive, highly sensitive, and disease specific. Good biomarker should also add independent information beyond the current standard of care. Identifying biomarkers of stroke is challenging due to the heterogeneity of the disease with regards to infarct size, location, cell population and underlying cause. Previous studies have assessed the utility of protein biomarkers that were selected based on their known functional association with stroke, such as markers of brain injury or inflammation. However detection of brain injury biomarkers is limited by the breakdown of the blood brain barrier (BBB), which normally prevents large molecules from leaving the brain capillaries and crossing into the systemic circulatory system. In addition, brain injury and inflammatory markers may be elevated in other neurologic disorders thus lacking specificity for stroke. Our currently incomplete understanding of stroke pathophysiology and protein function also hampers this knowledge-driven candidate biomarker approach. As a result more recent biomarker studies aim to use an agnostic discovery approach, such as proteome and gene expression profiling.

Currently, proteome profiling is costly and labour intensive because highly abundant proteins must be removed in order to detect low abundant proteins that would potentially have biomarker utility (Jacobs *et al.*, 2005). There are also few bioinformatics tools for proteomic data analysis, which may be due to the lack of high quality data and

lack of standardized methodologies. On the other hand, gene expression profiling is relatively cheap, reproducible, simultaneous measurement of transcriptome-wide expression is feasible, and numerous computation tools are available to facilitate analysis. In addition, mRNA expression changes are generally induced prior to protein level changes, which could allow earlier detection. For instance, Hamaouri *et al*., demonstrated that whole blood rhodopsin mRNA was elevated in patients with diabetic retinopathy as compared with healthy individuals (Hamaoui *et al.*, 2004), but elevated levels of the rhodopsin protein in these patients has not been reported.

The utility of RNA biomarkers to guide patient treatment is currently being evaluated. For instance, expression of 21-genes was recently assessed for prognostic utility in breast cancer patients (Sparano *et al.*, 2015). The study observed that patients with favourable, low expression of the 21-genes had lower rates of tumour recurrence when treated with only endocrine therapy, as compared with individuals with unfavourable, high expression that underwent both endocrine therapy and chemotherapy. Thus the gene expression data is able to provide valuable clinical information beyond currently available means. Similarly, agnostic transcriptome-wide RNA expression profiling may lead to the discovery of novel, clinically useful biomarkers for stroke.

1.6 RNA BIOMARKERS FOR STROKE DISCRIMINATION

Both animal (Tang *et al.*, 2001) and human (Moore *et al.*, 2005; Tang *et al.*, 2006; Barr *et al.*, 2010) studies have observed unique RNA expression changes in peripheral blood following stroke. The majority of RNA in whole blood is derived from leukocytes.

Gene expression within leukocytes may be influenced by factors associated with stroke pathogenesis such as blood clots or atherosclerotic plaques, circulating inflammatory molecules and/or interaction with necrotic cells. In the first proof of principle study, experimental neurologic injury was performed on rats and peripheral blood leukocyte expression was assessed using microarray after 24-hours (Tang *et al.*, 2001). The neurologic injuries were representative of ischemic stroke, hemorrhagic stroke, seizure, hypoglycemia or a sham surgery. The following results were observed: (1) many RNA transcripts were differentially expressed in each injury model, (2) a single gene did not differentiate between the injury models, and (3) a group of specific genes were differentially expressed in each injury model and together could be used to distinguish between the disease models.

The first leukocyte expression study in humans identified 22 unique genes that could distinguish between ischemic stroke cases and control participants (Moore *et al.*, 2005). Two additional studies using whole blood respectively reported 18 (Tang *et al.*, 2006) and 9 (Barr *et al.*, 2010) unique genes associated with ischemic stroke. Concordance between genes identified through the three human RNA biomarker studies (Moore *et al.*, 2005; Tang *et al.*, 2006; Barr *et al.*, 2010) was low, ranging from 0-45% overlap. The lack of concordance may be due to the heterogeneity of stroke between individuals and the small sample size used for biomarker discovery; at most 39 ischemic cases and 24 controls were used for RNA biomarker discovery. Alternatively, the low concordance may be due to differences in cell population or confounders. For instance, two of the study results may have been confounded by stroke risk factors since

established risk factors, age and hypertension, were more prevalent in stroke cases as compared with controls (Moore *et al.*, 2005; Tang *et al.*, 2006).

Limited verification of microarray results and validation in external populations was conducted by the three previous human studies. However, a targeted quantitative real-time polymerase chain reaction (qPCR) test to identify differential gene expression is currently the most simple, cost-effective and rapid test currently available for use in the clinical setting. qPCR verification was conducted for 0-44% of genes in the previous studies and qPCR was not utilized for independent validation studies. Thus, larger studies with adequate verification and external validation are required to identify biomarkers of stroke and to confirm previous reports.

1.7 INTERSTROKE COHORT

INTERSTROKE was an international, multicenter case-control study for stroke. The objective of the study was to systematically evaluate the association between stroke and traditional or emerging risk factors among different ethnic groups and geographic regions (M O'Donnell *et al.*, 2010). Between March 1, 2007 and April 23, 2010, participants were recruited from 84 centers in 22 countries worldwide: Argentina, Australia, Brazil, Canada, Chile, China, Colombia, Croatia, Denmark, Ecuador, Germany, India, Iran, Malaysia, Mozambique, Nigeria, Peru, Philippines, Poland, South Africa, Sudan, and Uganda (Martin J O'Donnell *et al.*, 2010). Stroke cases were patients with acute first-time stroke, enrolled within 5 days of symptom onset and 72-hours of hospital admission that also had neuroimaging (CT or MRI). Controls were recruited in hospital or

in the community and had no history of stroke. Hospital-based controls were patients admitted to hospital or attending outpatient clinics for disorders or procedures unrelated to stroke or transient ischemic attack, or visitors or relatives of inpatients. Controls were matched to cases based on age (within 5 years), sex, site location and ethnic origin (Martin J O'Donnell *et al.*, 2010).

A gene expression sub-study was conducted on consenting INTERSTROKE participants. Between March 2007 to April 2010, 375 INTERSTROKE participants (164 cases and 211 controls) consented to the gene expression sub-study from 6 centers in 4 countries: Canada, Columbia, Poland and Ecuador. Questionnaires were administered and physical examinations performed for consenting cases and controls. Whole blood samples were drawn from cases (within 5 days of symptom onset) and controls (at the time of interview) and frozen immediately after processing.

1.8 RNA EXPRESSION ANALYSIS WITH MICROARRAYS

Microarrays facilitate high-throughput gene expression profiling. A microarray is a small glass wafer on which thousands of oligonucleotide sequences, known as probes, are attached. Hybridization between a sample and corresponding probe results in fluorescence, which can be quantified. Analysis of fluorescent intensity values provides a snapshot of RNA transcript abundance within a sample. Comparison of fluorescence between samples can provide insight into molecular pathways underlying the biology of interest. The concept of gene expression microarrays is simple and successful, leading to

the production of many different array platforms; the two most common being the Affymetrix GeneChip® and Illumina BeadArray™.

The Affymetrix GeneChip® is a stationary piece of glass to which thousands of probe pairs are attached at specific locations (Affymetrix, n.d.). The probe pairs consist of a perfect match oligonucleotide and a mismatch oligonucleotide. The perfect match probe complements a known RNA sequence, while the mismatch probe differs by a single base substitution in the middle of the probe. As a result any hybridization to mismatch probes is indicative of non-specific binding or background fluorescence. Each probe is 25 bases in length and corresponds to unique regions near the 3' end of a gene.

In contrast, probes for Illumina BeadArrays™ are 79 oligonucleotides in length (Illumina, n.d.). Each probe is composed of a 50-oligonucleotide gene-specific sequence with a unique 29-oligonucleotide identifier sequence. Illumina probes are bound to silica beads, rather than dotted on a stationary chip. Each bead is ~3 microns in diameter and has ~700,000 copies of the same probe sequence covalently attached. Each bead type is replicated ~30 times on every BeadArray™. In addition to gene-specific sequences, there are also over 1000 non-genomic control bead types. Any hybridization to the control probes is indicative of non-specific binding. Both the gene-specific and control bead types are randomly arranged into etched wells on a stationary glass chip. As a result the location and identity of each bead is determined using the identifier sequence after the BeadArray™ has been assembled.

The MicroArray Quality Control (MAQC) project has reported that the Affymetrix and Illumina platforms can both yield high quality data with comparable

differential gene expression results (Shi *et al.*, 2006). However the Illumina BeadArray™ has greater intra-array redundancy of probes and is less costly than the Affymetrix GeneChip®.

1.8.1 PRE-PROCESING OF ILLUMINA MICROARRAY DATA

In order to obtain biologically relevant data from microarrays, data 'cleaning' is required. The general 'cleaning' or pre-processing steps include: background correction, normalization, transformation, and filtering. The fluorescent intensity from each gene probe measures the abundance of a specific RNA sequence but is also affected by non-specific sources such as, auto-fluorescence on the chip surface. The purpose of background correction is to remove the non-specific signal from the total signal. Normalization adjusts the individual fluorescent intensity values to compensate for effects of variation in the technology that could hide the true biologic variation. Intensity values may be impacted by unequal quantities of starting RNA, or difference in labeling and hybridization efficiency. Normalization aims to make the overall expression distribution similar between all of the arrays in an experiment. Next, data transformation is applied to produce a continuous spectrum of values so that up- and down regulated gene expression can be interpreted similarly (*Figure 1.2*). The most commonly used transformation method for all microarrays is logarithm base 2 (log2). The final data pre-processing step is probe filtering, which involves removal of intensity data from low quality probes. Although many methods of background correction and normalization have been

developed (Quackenbush, 2002; Xie *et al.*, 2009; Shi *et al.*, 2010), there is little agreement in the literature regarding how best to pre-process Illumina BeadArray™ data.

Illumina has created analytical software, BeadStudio, for the analysis of BeadArray™ data. BeadStudio provides a simple background correction method whereby the average signal from the control beads is subtracted from the gene-specific probe signals. However Dunning *et al.,* (Dunning *et al.*, 2008) have reported that Illumina's simple method introduces variability into the data, increases the number of false positives and causes a large number of low expression values to become negative, which is inappropriate for normalization and downstream analysis. Shi *et al.,* (Shi *et al.*, 2010) conducted a systematic analysis of five BeadArray™ background correction and normalization methods and determined that Normexp Background Correction Using Control Probes (neqc) was the superior method.

Regarding normalization, multiple studies report that quantile normalization in combination with log2 transformation results in good quality data that corresponds with quantitative real-time polymerase chain reaction (qPCR) data (Dunning *et al.*, 2008; Schmid *et al.*, 2010). Quantile normalization is a technique for making two distributions identical and involves replacement of probe intensity values with the ranked-mean probe intensity value (*Figure 1.3*). A limitation of quantile normalization is that genes with very high intensity are forced into the same distribution, thus reducing biological variation as well as technical variation.

Finally, probe filtering of Illumina BeadArray™ data is determined using the detection p-value. Each probe has a corresponding detection p-value describing the

confidence that the measure fluorescent intensity value of a given transcript is expressed above background noise, where background is determined based on the fluorescence observed from non-genomic, negative control bead types (Abed *et al.*, 2013). A detection p-value threshold less than either 0.01 or 0.05 is sufficient to reject the null hypothesis of no transcript detection.

Based on the literature a valid protocol for BeadArray pre-processing consists of: neqc background correction, quantile normalization, log2 transformation, and probe filtering based on the detection P-values.

Figure 1.2 Microarray data transformation.

Let the blue and red bars represent the true expression for a test and control group respectively. Suppose we are interested in the ratio between test and control [R=T/C]. Without transformation, A) genes up-regulated by a factor of 2 have an expression ratio of 2/1=2 whereas, B) genes down-regulated by the same factor would have an expression ratio of 1/2=0.5. Changing the scale to log2 makes the data more interpretable.

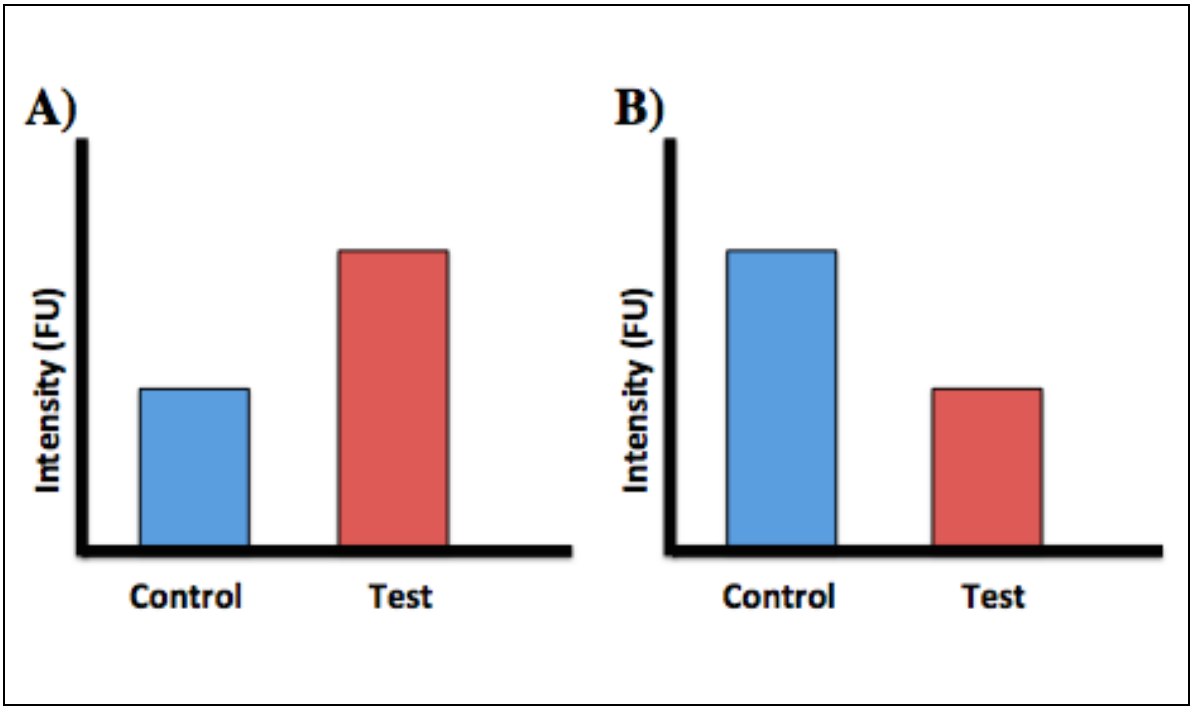
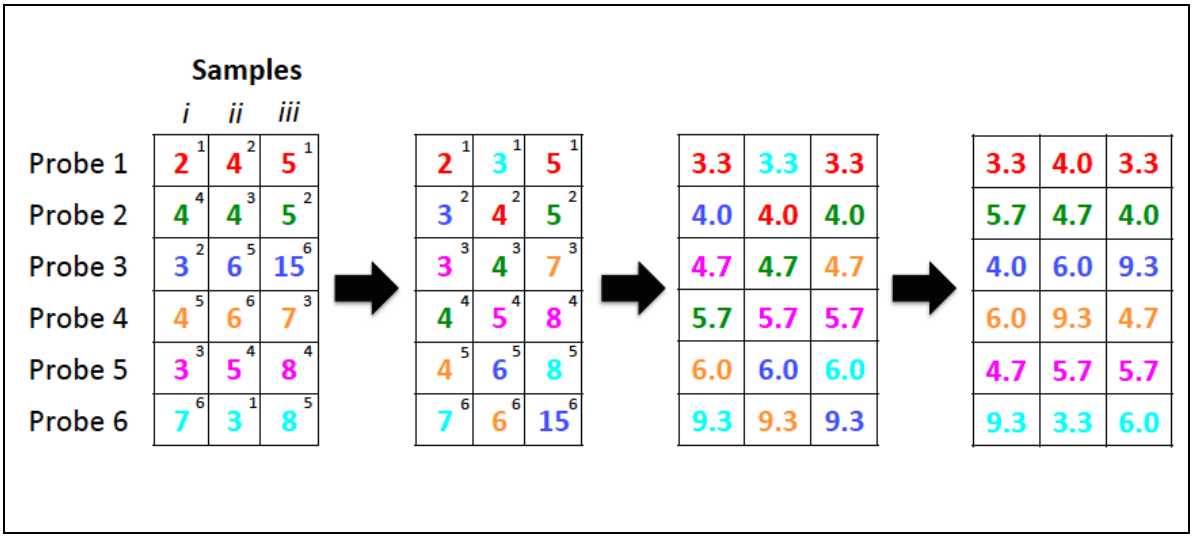


Figure 1.3 Steps of quantile normalization.

Quantile normalization is a technique to make two distributions identical. Each row of microarray data contains fluorescent intensity values from different probes while the columns represent individual samples. Initially raw intensity values are ranked. Next intensity values are ordered within each sample according to rank. The average intensity across each probe, or row, is determined. The individual intensity values are then substituted for the average intensity value. Finally the average intensity values are re-ordered according to their initial order, prior to ranking.



1.8.2 QUALITY CONTROL FOR MICORARRAY DATA

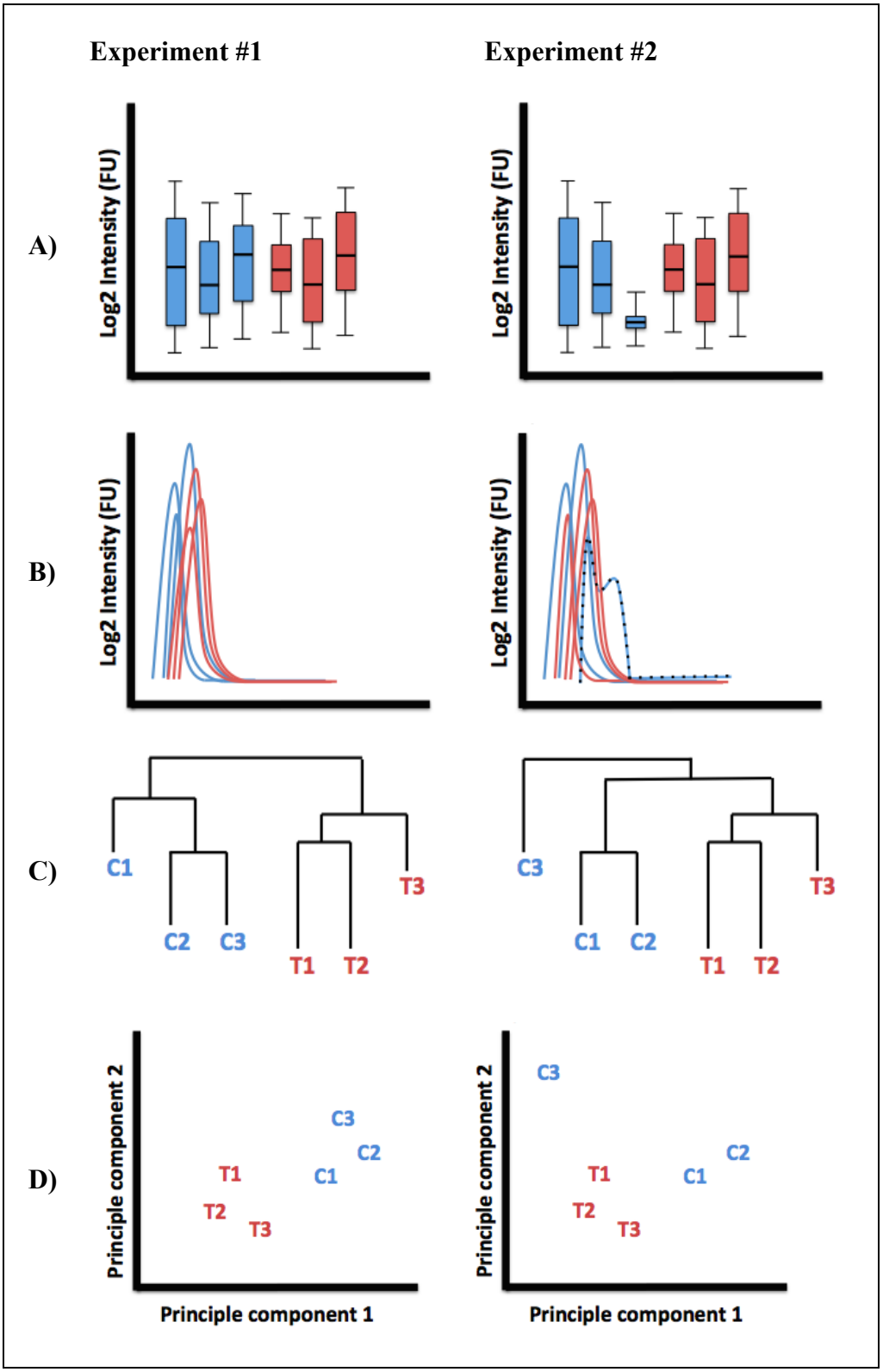
Quality control (QC) is also an essential step to obtain biologically applicable results from microarray data. QC is conducted before and after pre-processing to identify samples with suspicious intensity values relative to the majority. These atypical samples are referred to as outliers. Poor performing chips, or experimental failure may influence data quality and results in outlier samples. Since outliers can negatively impact data normalization and downstream analysis, it is preferable to identify and remove them prior to data pre-processing. However data pre-processing can also make outliers more apparent thus providing support for sample exclusion. There are no standardized protocols for microarray QC. However, relative QC metrics are commonly used to compare each array's intensity values against other arrays within the dataset. Samples that fail multiple QC matrices are more likely to be true outliers. Removal of outliers and subsequent data pre-processing will more likely result in high quality, biologically relevant microarray data. Some QC diagnostics include:

1. Boxplots of the negative and gene-specific fluorescent intensity values for each individual (*Figure 1.4A*). Since median expression and inter quartile range are expected be similar between samples, those samples with an unusual intensity distribution may be outliers.
2. Density plots of the gene-specific intensity values for each individual (*Figure 1.4B*). Samples with a bimodal intensity distribution or relatively atypical distribution may be outliers.

3. Pearson correlation of transcriptome-wide expression between samples. The majority of genes are not differentially expressed therefore high correlation is expected between the total microarray intensity values between individuals. As such, samples that have low correlation with other arrays in the dataset may be outliers.
4. Hierarchical clustering (*Figure 1.4C*) to identify samples that do not conform to the expected patterns in a dataset. Clustering methods can be used to identify samples that are similar to one another. The more similar samples will be clustered together within the same branch. Samples that group 'far away' from the majority of the dataset may be outliers.
5. Principle component analysis (PCA) (*Figure 1.4D*). PCA summarize the total expression variance within each array into a finite set of values known as principle component eigenvalues. Commonly, the first two principle components summarize the majority of the expression variance. By plotting the eigenvalues of principle component 1 against principle component 2 we can visualize the variation between samples. Samples that group 'far away' from the majority of the dataset may be outliers.

Figure 1.4 Diagnostic plots to evaluate microarray data quality before normalization.

Useful plots include: A) Boxplots B) Density plots, C) Hierarchical clustering, and D) scatter plot of first two principle components from PCA. The two columns contain QC plots for two separate experiments. Cumulatively QC metrics from experiment #1 suggest reasonable data quality where as an outlier is apparent in experiment #2; sample C3 is likely an outlier.


Experiment #1
Experiment #2
A)
B)
C)
D)
Log2 Intensity (FU)
Log2 Intensity (FU)
Log2 Intensity (FU)
Log2 Intensity (FU)
C1
C2
C3
T1
T2
T3
C3
C1
C2
T1
T2
T3
Principle component 2
Principle component 1
C3
C2
T1
C1
T2
T3
Principle component 2
Principle component 1
C3
C2
T1
C1
T2
T3

1.8.3 DIFFERENTIAL EXPRESSION ANALYSIS

Due to the normalization step, microarrays measure relative, rather than absolute, gene expression. As a result, differential gene expression is reported as relative fold change (FC). A simple microarray experiment may assess gene expression changes between a treatment and control group. There are many methods to analyze differential expression, but two simple univariate methods are Student t-test and regression. In univariate analyses each of the microarray gene-specific probes is individually assessed for association with a phenotype. Regression is the more powerful method since additional variables can be incorporated into the model such as age, body mass index (BMI), and smoking status. Since thousands of probes are being tested simultaneously, there is high probability that a false significant result would be observed by chance. Increasing the sample size increases the statistical power to detect expression difference and can also decrease the false positive rate. However the common p-value threshold of $\alpha=0.05$ or $\alpha=0.01$ would not be appropriate due to the high probability of identifying one or more false positive. P-value correction methods may be applied to control the false positive rate. The most simple and conservative method is Bonferroni adjustment. If a significance threshold of α is used, but n independent tests are performed, then the Bonferroni adjustment deems a score significant only if the corresponding p-value is $\leq \alpha/n$ (Noble, 2009). Bonferroni adjustment ensures that for a given threshold, one or more large values would be expected in the null distribution with a probability of α.

However, a limitation with univariate analysis is that it assumes each test is independent of one another, but in reality gene expression values are highly correlated.

For instance, genes may be co-expressed when they are functionally related (Eisen *et al.*, 1998) or involved in a similar regulatory system. Nonetheless, univariate analysis with Bonferroni adjustment is an effective method to identify differentially expressed genes.

1.8.4 NETWORK ANALYSIS

Traditional univariate analyses of microarray data assess the relationship between gene expression and a phenotype, however addition biomarkers may be discovered using network analysis. Univariate methods assume that each biomarker is expressed independently of one another. However, this assumption may not be appropriate for agnostic biomarker discovery. For instance, genes may be co-expressed when they are functionally related or involved in a similar regulatory system. Although univariate methods are able to identify genes associated with a clinical trait, they disregard co-expression patterns and thus provide little insight into the molecular biology underlying the phenotype or disease

Weighted gene co-expression network analysis (WGCNA) (Langfelder and Horvath, 2008) is an agnostic, computational technique to identify groups of genes with correlated expression. WGCNA has been previously used for the study of complex diseases such as atrial fibrillation (Tan *et al.*, 2013) and Alzheimer's (Lunnon *et al.*, 2012). WGCNA identifies groups of genes with correlated expression through hierarchical clustering and a tree-cutting algorithm. The large groups of genes with related expression are referred to as a module. A benefit of grouping genes into modules is that it can drastically reduce multiple hypotheses testing. Rather than test thousands of

individual genes for association with a phenotype, a small number of modules can be tested for association. Next, genes within phenotype-associated modules may be characterized. For instance, biologically significant patterns can be identified using pathway analysis. In addition, central genes to each module can be isolated. Genes with high connectivity within the module network are referred to as hub genes. Hub genes are identified based on high correlation with a clinical trait, and high correlation to the overall module variance.

Hub genes from different modules are unlikely to have correlated expression. Thus network analysis can facilitate the discovery of biomarker panels. A gene panel consisting of hub genes from independent modules may summarize more of the underlying phenotype biology and thus have greater discriminative specificity and sensitivity as compared with a single gene biomarker.

1.9 ATRIAL FIBRILLATION

Atrial fibrillation (AF) is the most common sustained cardiac arrhythmia. Common risk factors associated with AF include: age, diabetes, hypertension, congestive heart failure and valvular heart disease (Benjamin *et al.,* 1994). In addition, men are at 1.5 times greater risk of developing AF as compared with women (Benjamin *et al.*, 1994). Currently AF affects approximately 33 million people worldwide (Rahman *et al.*, 2014), but this number is projected to increase rapidly as the global population ages (Go *et al.*, 2001; Chugh *et al.*, 2014; Rahman *et al.*, 2014). Although AF itself is not life threatening, it negatively influences quality of life and is associated with increased risk of stroke

(Wolf *et al.*, 1991; Conen *et al.*, 2011) and heart failure (Benjamin *et al.*, 1998; Stewart *et al.*, 2002; Conen *et al.*, 2011). The specific molecular mechanisms underlying AF have yet to be unravelled. However structural changes and electrical remodelling play an important role in the perpetuation of AF (Iwasaki *et al.*, 2011). AF occurs when disorganized electrical signals cause the atrial chambers to contract quickly and irregularly. Due to the irregular contractions, blood is not completely pushed into the ventricles. As a result, blood pools in the atria and is prone to thrombosis.

1.10 ATRIAL FIBRILLATION AND STROKE

The Framingham study has reported that AF is associated with 5-fold increased risk for stroke (Wolf *et al.*, 1978; Wolf *et al.*, 1991; Conen *et al.*, 2011). Indeed, approximately 15% of stroke cases are likely due to documented AF (Wolf *et al.*, 1987). In order to reduce stroke risk AF patients are commonly prescribed oral anticoagulants (Ezekowitz *et al.*, 1992; Connolly *et al.*, 2009; Granger *et al.*, 2011; Patel *et al.*, 2011). Oral anticoagulants inhibit blood coagulation therefore blood pooling in the atria has less propensity to clot. Another treatment for AF is cardioversion, which aims to restore sinus rhythm to the heart using medication or an electrical procedure (Lown *et al.*, 1962). However, electrical cardioversion has limited long-term success rates and significant risks. Indeed, 40-60% of patients who undergo cardioversion experience AF recurrence (Raitt *et al.*, 2006). Developing a more detailed understanding of AF pathophysiology may allow optimization of AF therapy.

Standard methods for AF detection involve electrocardiogram (ECG). Recent studies have suggested that long-term rhythm assessment can further improve diagnosis of AF. For instance, a Swedish study observed that a single ECG was able to detect AF in 1.2% of participants while 24-hour Holter monitoring detected AF in 7.4% of participants (Engdahl *et al.*, 2013). The ASSERT study further increased monitoring time by studying implanted pacemaker data for three months in patients that had no previous history of AF and identified subclinical AF, short episodes of AF that spontaneously revert to sinus rhythm, in 10.1% of participants (Healey *et al.*, 2012). ASSERT also reported an association between subclinical AF and a 2.5-fold increased risk for ischemic stroke (Healey *et al.*, 2012). Specifically, it has been suggested that subclinical AF may be an underlying cause of cryptogenic stroke, a subtype of ischemic stroke for which cause cannot be determined (Gladstone *et al.*, 2014; Sanna *et al.*, 2014).

Stroke prognosis, recurrence risk and secondary prevention differ based on the underlying stroke etiology. Since cause cannot be determined for cryptogenic stroke patients, there are no standardized management guidelines. However detecting subclinical AF in cryptogenic stroke patients has the potential to impact stroke management. For example, risk of stroke recurrence may be reduced in these patients by prescribing new oral anticoagulants, similar to cardioembolic stroke patients. Although long term cardiac monitoring can identify AF in patients with no AF history, these methods can be moderately invasive, costly, and cumbersome to patients.

Cumulatively there is a need to improve our understanding of AF pathophysiology in order to optimize treatment strategies. In addition, development of new diagnostic tools for AF may supplement current clinical practice and improve detection of subclinical AF.

1.11 PROTEIN BIOMARKERS FOR ATRIAL FIBRILLATION

Peripheral blood biomarkers of AF may have clinical utility and improve our understanding of AF pathophysiology. Several biomarkers associated with neurohumoral responses and inflammation have been assessed for association with AF (Hijazi, Oldgren, *et al.*, 2013; Vílchez *et al.*, 2013). Current research has focused on N-terminal B-type natriuretic peptide (NT-proBNP). NT-proBNP is the stable N-terminal cleavage product of proBNP. proBNP is synthesized and released by cardiomyocytes in response to elevated cardiac wall stress (Nakagawa *et al.*, 1995; Molkentin *et al.*, 1998; Wiese *et al.*, 2000). Studies report that elevated NT-proBNP concentrations independently predict the risk of developing AF (Patton *et al.*, 2009) and having a thromboembolic event (Hijazi *et al.*, 2012; Hijazi, Wallentin, *et al.*, 2013). Studies have also shown that NT-proBNP levels rapidly decrease in AF patients that undergo successful cardioversion (Vinch *et al.*, 2004; Wozakowska-Kapłon, 2004; Buob *et al.*, 2006) suggesting that NT-proBNP may be an AF specific biomarker.

C-reactive protein (CRP), an established marker of systemic inflammation, is another biomarker under investigation for association with AF. However reports are inconsistent regarding: the relationship between CRP and risk of developing AF (Aviles *et al.*, 2003; Nyrnes *et al.*, 2012), association between CRP and stroke risk (Conway *et*

al., 2004; Lip *et al.*, 2007) and change in CRP levels following cardioversion (Buob *et al.*, 2006; Liu *et al.*, 2007; Henningsen *et al.*, 2009). As such additional studies are required to determine the utility of CRP for AF.

1.12 RNA BIOMARKERS FOR ATRIAL FIBRILLATION

Agnostic biomarker discovery may lead to the identification of new clinical tools and further improve our understanding of AF. Unique gene expression change have been detected in atrial tissue from patients with AF as compared with patients that had no history of AF (Ohki *et al.*, 2005; Barth *et al.*, 2005; Deshmukh *et al.*, 2014). The Framingham Heart Study recently conducted whole blood gene expression profiling in patients with prevalent AF and a large reference population (Lin *et al.*, 2014). Seven differentially expressed genes were identified. Many of the genes had relevant biological significance, but the microarray results have yet to be verified with qPCR or validated in an independent population. Additional studies are also required to evaluate the added benefit of gene expression biomarkers as compared with promising AF biomarkers, such as NT-proBNP.

1.12 REFERENCES

Abed, H.S., Wittert, G. a, Leong, D.P., Shirazi, M.G., Bahrami, B., Middeldorp, M.E., *et al.* (2013) Effect of weight reduction and cardiometabolic risk factor management on symptom burden and severity in patients with atrial fibrillation: a randomized clinical trial. *Jama* **310**: 2050–60.

Adams, H.P., Bendixen, B.H., Kappelle, L.J., Biller, J., Love, B.B., Gordon, D.L., and Marsh, E.E. (1993) Classification of subtype of acute ischemic stroke. Definitions for use in a multicenter clinical trial. TOAST. Trial of Org 10172 in Acute Stroke Treatment. *Stroke* **24**: 35–41.

Adams Jr., H.P., Davis, P.H., Leira, E.C., Chang, K.C., Bendixen, B.H., Clarke, W.R., *et al.* (1999) Baseline NIH Stroke Scale score strongly predicts outcome after stroke - A report of the Trial of Org 10172 in Acute Stroke Treatment (TOAST). *Neurology* **53**: 126–131.

Adeoye, O., Hornung, R., Khatri, P., and Kleindorfer, D. (2011) Recombinant tissue-type plasminogen activator use for ischemic stroke in the united states: A doubling of treatment rates over the course of 5 years. *Stroke* **42**: 1952–1955.

Affymetrix http://www.affymetrix.com/.

Appelros, P., Stegmayr, B., and Terent, A. (2009) Sex differences in stroke epidemiology: A systematic review. *Stroke* **40**: 1082–1090.

Aviles, R.J., Martin, D.O., Apperson-Hansen, C., Houghtaling, P.L., Rautaharju, P., Kronmal, R. a., *et al.* (2003) Inflammation as a Risk Factor for Atrial Fibrillation. *Circulation* **108**: 3006–3010.

Ay, H., Benner, T., Arsava, E.M., Furie, K.L., Singhal, A.B., Jensen, M.B., *et al.* (2007) A computerized algorithm for etiologic classification of ischemic stroke: the Causative Classification of Stroke System. *Stroke* **38**: 2979–84.

Barr, T.L., Conley, Y., Ding, J., Dillman, A., Warach, S., Singleton, A., and Matarin, M. (2010) Genomic biomarkers and cellular pathways of ischemic stroke by RNA gene expression profiling. *Neurology* **75**: 1009–14.

Barth, A.S., Merk, S., Arnoldi, E., Zwermann, L., Kloos, P., Gebauer, M., *et al.* (2005) Reprogramming of the human atrial transcriptome in permanent atrial fibrillation: expression of a ventricular-like genomic signature. *Circ Res* **96**: 1022–9.

Benjamin, E.J., Levy, D., Vaziri, S.M., D'Agostino, R.B., Belanger, A.J., and Wolf, P.A. (1994) Independent risk factors for atrial fibrillation in a population-based cohort. The Framingham Heart Study. *JAMA* **271**: 840–4.

Benjamin, E.J., Wolf, P.A., D'Agostino, R.B., Silbershatz, H., Kannel, W.B., and Levy, D. (1998) Impact of Atrial Fibrillation on the Risk of Death: The Framingham Heart Study. *Circulation* **98**: 946–952.

Buob, A., Jung, J., Siaplaouras, S., Neuberger, H.R., and Mewis, C. (2006) Discordant regulation of CRP and NT-proBNP plasma levels after electrical cardioversion of persistent atrial fibrillation. *PACE - Pacing Clin Electrophysiol* **29**: 559–563.

Chalela, J. a, Kidwell, C.S., Nentwich, L.M., Luby, M., Butman, J. a, Demchuk, A.M., *et al.* (2007) Magnetic resonance imaging and computed tomography in emergency assessment of patients with suspected acute stroke: a prospective comparison. *Lancet* **369**: 293–8.

Chugh, S.S., Havmoeller, R., Narayanan, K., Singh, D., Rienstra, M., Benjamin, E.J., *et al.* (2014) Worldwide epidemiology of atrial fibrillation: A global burden of disease 2010 study. *Circulation* **129**: 837–847.

Conen, D., Chae, C.U., Glynn, R.J., Tedrow, U.B., Everett, B.M., Buring, J.E., and Albert, C.M. (2011) Risk of death and cardiovascular events in initially healthy women with new-onset atrial fibrillation. *JAMA* **305**: 2080–2087.

Connolly, S.J., Ezekowitz, M.D., Yusuf, S., Eikelboom, J., Oldgren, J., Parekh, A., *et al.* (2009) Dabigatran versus warfarin in patients with atrial fibrillation. *N Engl J Med* **361**: 1139–1151.

Conway, D.S.G., Buggins, P., Hughes, E., and Lip, G.Y.H. (2004) Prognostic significance of raised plasma levels of interleukin-6 and C-reactive protein in atrial fibrillation. *Am Heart J* **148**: 462–466.

Deshmukh, A., Barnard, J., Sun, H., Newton, D., Castel, L., Pettersson, G., *et al.* (2014) Left Atrial Transcriptional Changes Associated with Atrial Fibrillation Susceptibility and Persistence. *Circ Arrhythm Electrophysiol* **8**: CIRCEP.114.001632–.

Donnan, G.A., Fisher, M., Macleod, M., and Davis, S.M. (2008) Stroke. *Lancet (London, England)* **371**: 1612–23.

Dunning, M.J., Barbosa-Morais, N.L., Lynch, A.G., Tavaré, S., and Ritchie, M.E. (2008) Statistical issues in the analysis of Illumina data. *BMC Bioinformatics* **9**: 85.

Eisen, M.B., Spellman, P.T., Brown, P.O., and Botstein, D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* **95**: 14863–8.

Emberson, J., Lees, K.R., Lyden, P., Blackwell, L., Albers, G., Bluhmki, E., *et al.* (2014) Effect of treatment delay, age, and stroke severity on the effects of intravenous thrombolysis with alteplase for acute ischaemic stroke: a meta-analysis of individual patient data from randomised trials. *Lancet (London, England)* **384**: 1929–35.

Engdahl, J., Andersson, L., Mirskaya, M., and Rosenqvist, M. (2013) Stepwise screening of atrial fibrillation in a 75-year-old population: Implications for stroke prevention. *Circulation* **127**: 930–937.

Evers, S.M. a a, Struijs, J.N., Ament, A.J.H. a, Genugten, M.L.L. van, Jager, J.H.C., and Bos, G. a M. van den (2004) International comparison of stroke cost studies. *Stroke* **35**: 1209–15.

Ezekowitz, M.D., Bridgers, S.L., James, K.E., Carliner, N.H., Colling, C.L., Gornick, C.C., *et al.* (1992) Warfarin in the prevention of stroke associated with nonrheumatic atrial fibrillation. Veterans Affairs Stroke Prevention in Nonrheumatic Atrial Fibrillation Investigators. .

Fiebach, J.B., Schellinger, P.D., Jansen, O., Meyer, M., Wilde, P., Bender, J., *et al.* (2002) CT and Diffusion-Weighted MR Imaging in Randomized Order: Diffusion-Weighted Imaging Results in Higher Accuracy and Lower Interrater Variability in the Diagnosis of Hyperacute Ischemic Stroke. *Stroke* **33**: 2206–2210.

Flint, A.C., Faigeles, B.S., Cullen, S.P., Kamel, H., Rao, V. a, Gupta, R., *et al.* (2013) THRIVE score predicts ischemic stroke outcomes and thrombolytic hemorrhage risk in VISTA. *Stroke* **44**: 3365–9.

G., L., G., N., H., Z., P., M., D.Z., W., J., F., *et al.* (2013) External validation of the ASTRAL score to predict 3- and 12-month functional outcome in the China National Stroke Registry. *Stroke* **44**: 1443–1445.

Gladstone, D.J., Spring, M., Dorian, P., Panzov, V., Thorpe, K.E., Hall, J., *et al.* (2014) Atrial fibrillation in patients with cryptogenic stroke. *N Engl J Med* **370**: 2467–77.

Go, A.S., Hylek, E.M., Phillips, K.A., Chang, Y., Henault, L.E., Selby, J. V, and Singer, D.E. (2001) Prevalence of diagnosed atrial fibrillation in adults: national implications for rhythm management and stroke prevention: the AnTicoagulation and Risk Factors in Atrial Fibrillation (ATRIA) Study. *JAMA* **285**: 2370–2375.

Granger, C.B., Alexander, J.H., McMurray, J.J. V, Lopes, R.D., Hylek, E.M., Hanna, M., *et al.* (2011) Apixaban versus warfarin in patients with atrial fibrillation. *N Engl J Med* **365**: 981–992.

Grotta, J.C., Chiu, D., Lu, M., Patel, S., Levine, S.R., Tilley, B.C., *et al.* (1999) Agreement and Variability in the Interpretation of Early CT Changes in Stroke Patients Qualifying for Intravenous rtPA Therapy. *Stroke* **30**: 1528–1533.

Hacke, W., Kaste, M., Bluhmki, E., Brozman, M., Dávalos, A., Guidetti, D., *et al.* (2008) Thrombolysis with alteplase 3 to 4.5 hours after acute ischemic stroke. *N Engl J Med* **359**: 1317–29.

Hamaoui, K., Butt, A., Powrie, J., and Swaminathan, R. (2004) Concentration of Circulation Rhodopsin mRNA in Diabetic Retinopathy. *Clin Chem* **50**: 2150–2.

Healey, J.S., Connolly, S.J., Gold, M.R., Israel, C.W., Gelder, I.C. Van, Capucci, A., *et al.* (2012) Subclinical Atrial Fibrillation and the Risk of Stroke. *N Engl J Med* **366**: 120–129.

Henningsen, K.M.A., Therkelsen, S.K., Bruunsgaard, H., Krabbe, K.S., Pedersen, B.K., and Svendsen, J.H. (2009) Prognostic impact of hs-CRP and IL-6 in patients with persistent atrial fibrillation treated with electrical cardioversion. *Scand J Clin Lab Invest* **69**: 425–432.

Hijazi, Z., Oldgren, J., Andersson, U., Connolly, S.J., Ezekowitz, M.D., Hohnloser, S.H., *et al.* (2012) Cardiac biomarkers are associated with an increased risk of stroke and death in patients with atrial fibrillation: a Randomized Evaluation of Long-term Anticoagulation Therapy (RE-LY) substudy. *Circulation* **125**: 1605–16.

Hijazi, Z., Oldgren, J., Siegbahn, A., Granger, C.B., and Wallentin, L. (2013) Biomarkers in atrial fibrillation: a clinical review. *Eur Heart J* **34**: 1475–80.

Hijazi, Z., Wallentin, L., Siegbahn, A., Andersson, U., Christersson, C., Ezekowitz, J., *et al.* (2013) N-terminal pro-B-type natriuretic peptide for risk assessment in patients with atrial fibrillation: Insights from the aristotle trial (apixaban for the prevention of stroke in subjects with atrial fibrillation). *J Am Coll Cardiol* **61**: 2274–2284.

Hill, M.D., and Buchan, A.M. (2005) Thrombolysis for acute ischemic stroke: results of the Canadian Alteplase for Stroke Effectiveness Study. *C Can Med Assoc J J lAssociation medicale Can* **172**: 1307–1312.

Hoylaerts, M., Rijken, D.C., Lijnen, H.R., and Collen, D. (1982) Kinetics of the activation of plasminogen by human tissue plasminogen activator. Role of fibrin. *J Biol Chem* **257**: 2912–9.

Illumina https://www.illumina.com/.

Iwasaki, Y., Nishida, K., Kato, T., and Nattel, S. (2011) Atrial fibrillation pathophysiology: implications for management. *Circulation* **124**: 2264–74.

Jacobs, J.M., Adkins, J.N., Qian, W.J., Liu, T., Shen, Y., Camp, D.G., and Smith, R.D. (2005) Utilizing human blood plasma for proteomic biomarker discovery. *J Proteome Res* **4**: 1073–1085.

Kidwell, C.S., Chalela, J. a, Saver, J.L., Starkman, S., Hill, M.D., Demchuk, A.M., *et al.* (2004) Comparison of MRI and CT for detection of acute intracerebral hemorrhage. *JAMA* **292**: 1823–30.

König, I.R., Ziegler, A., Bluhmki, E., Hacke, W., Bath, P.M.W., Sacco, R.L., *et al.* (2008)

Predicting long-term outcome after acute ischemic stroke: A simple index works in patients from controlled clinical trials. *Stroke* **39**: 1821–1826.

Krishnamurthi, R. V., Feigin, V.L., Forouzanfar, M.H., Mensah, G. a., Connor, M., Bennett, D. a., *et al.* (2013) Global and regional burden of first-ever ischaemic and haemorrhagic stroke during 1990-2010: Findings from the Global Burden of Disease Study 2010. *Lancet Glob Heal* **1**: e259–e281.

Langfelder, P., and Horvath, S. (2008) WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**: 559.

Lansberg, M.G., Bluhmki, E., and Thijs, V.N. (2009) Efficacy and safety of tissue plasminogen activator 3 to 4.5 hours after acute ischemic stroke: a metaanalysis. *Stroke* **40**: 2438–41.

Lewis, S.C., Sandercock, P. a G., and Dennis, M.S. (2008) Predicting outcome in hyper-acute stroke: validation of a prognostic model in the Third International Stroke Trial (IST3). *J Neurol Neurosurg Psychiatry* **79**: 397–400.

Lin, H., Yin, X., Lunetta, K.L., Dupuis, J., McManus, D.D., Lubitz, S. a., *et al.* (2014) Whole blood gene expression and atrial fibrillation: the framingham heart study. *PLoS One* **9**: e96794.

Lip, G.Y.H., Patel, J. V., Hughes, E., and Hart, R.G. (2007) High-sensitivity C-reactive protein and soluble CD40 ligand as indices of inflammation and platelet activation in 880 patients with nonvalvular atrial fibrillation: Relationship to stroke risk factors, stroke risk stratification schema, and prognosis. *Stroke* **38**: 1229–1237.

Liu, T., Li, G., Li, L., and Korantzopoulos, P. (2007) Association Between C-Reactive Protein and Recurrence of Atrial Fibrillation After Successful Electrical Cardioversion. A Meta-Analysis. *J Am Coll Cardiol* **49**: 1642–1648.

Lown, B., Amarasingham, R., and Neuman, J. (1962) New method for terminating cardiac arrhythmias. Use of synchronized capacitor discharge. *JAMA* **182**: 548–555.

Lozano, R., Naghavi, M., Foreman, K., Lim, S., Shibuya, K., Aboyans, V., *et al.* (2012) Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: A systematic analysis for the Global Burden of Disease Study 2010. *Lancet* **380**: 2095–2128.

Lunnon, K., Ibrahima, Z., Proitsi, P., Lourdusamy, A., Newhouse, S., Sattlecker, M., *et al.* (2012) Mitochondrial dysfunction and immune activation are detectable in early alzheimer's disease blood. *J Alzheimer's Dis* **30**: 685–710.

Mittmann, N., Seung, S.J., Hill, M.D., Phillips, S.J., Hachinski, V., Coté, R., *et al.* (2012) Impact of disability status on ischemic stroke costs in Canada in the first year. *Can J*

Neurol Sci **39**: 793–800.

Molkentin, J.D., Lu, J.R., Antos, C.L., Markham, B., Richardson, J., Robbins, J., *et al.* (1998) A calcineurin-dependent transcriptional pathway for cardiac hypertrophy. *Cell* **93**: 215–228.

Moore, D.F., Li, H., Jeffries, N., Wright, V., Cooper, R. a, Elkahloun, A., *et al.* (2005) Using peripheral blood mononuclear cells to determine a gene expression profile of acute ischemic stroke: a pilot investigation. *Circulation* **111**: 212–21.

Murray, C.J.L., Vos, T., Lozano, R., Naghavi, M., Flaxman, A.D., Michaud, C., *et al.* (2012) Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990-2010: A systematic analysis for the Global Burden of Disease Study 2010. *Lancet* **380**: 2197–2223.

Nadeau, J.O., Shi, S., Fang, J., Kapral, M.K., Richards, J.A., Silver, F.L., and Hill, M.D. (2005) TPA use for stroke in the Registry of the Canadian Stroke Network. *Can J Neurol Sci* **32**: 433–439.

Nakagawa, O., Ogawa, Y., Itoh, H., Suga, S.I., Komatsu, Y., Kishimoto, I., *et al.* (1995) Rapid transcriptional activation and early mRNA turnover of brain natriuretic peptide in cardiocyte hypertrophy: Evidence for brain natriuretic peptide as an "emergency" cardiac hormone against ventricular overload. *J Clin Invest* **96**: 1280–1287.

Noble, W.S. (2009) How does multiple testing correction work? *Nat Biotechnol* **27**: 1135–7.

Nyrnes, A., Njolstad, I., Mathiesen, E.B., Wilsgaard, T., Hansen, J.B., Skjelbakken, T., *et al.* (2012) Inflammatory biomarkers as risk factors for future atrial fibrillation. an eleven-year follow-up of 6315 men and women: The Troms?? study. *Gend Med* **9**: 536–547.e2.

O'Donnell, M., Xavier, D., Diener, C., Sacco, R., Lisheng, L., Zhang, H., *et al.* (2010) Rationale and design of INTERSTROKE: a global case-control study of risk factors for stroke. *Neuroepidemiology* **35**: 36–44.

O'Donnell, M.J., Fang, J., D'Uva, C., Saposnik, G., Gould, L., McGrath, E., and Kapral, M.K. (2012) The PLAN score: a bedside prediction rule for death and severe disability following acute ischemic stroke. *Arch Intern Med* **172**: 1548–56.

O'Donnell, M.J., Xavier, D., Liu, L., Zhang, H., Chin, S.L., Rao-Melacini, P., *et al.* (2010) Risk factors for ischaemic and intracerebral haemorrhagic stroke in 22 countries (the INTERSTROKE study): a case-control study. *Lancet* **376**: 112–23.

Ohki, R., Yamamoto, K., Ueno, S., Mano, H., Misawa, Y., Fuse, K., *et al.* (2005) Gene expression profiling of human atrial myocardium with atrial fibrillation by DNA microarray analysis. **102**: 233–238.

Patel, M.R., Mahaffey, K.W., Garg, J., Pan, G., Singer, D.E., Hacke, W., *et al.* (2011) Rivaroxaban versus warfarin in nonvalvular atrial fibrillation. *N Engl J Med* **365**: 883–891.

Patton, K.K., Ellinor, P.T., Heckbert, S.R., Christenson, R.H., DeFilippi, C., Gottdiener, J.S., and Kronmal, R. a (2009) N-terminal pro-B-type natriuretic peptide is a major predictor of the development of atrial fibrillation: the Cardiovascular Health Study. *Circulation* **120**: 1768–74.

Quackenbush, J. (2002) Microarray data normalization and transformation. *Nat Genet* **32 Suppl**: 496–501.

Rahman, F., Kwan, G.F., and Benjamin, E.J. (2014) Global epidemiology of atrial fibrillation. *Nat Rev Cardiol* **11**: 639–54.

Raitt, M.H., Volgman, A.S., Zoble, R.G., Charbonneau, L., Padder, F. a., O'Hara, G.E., and Kerr, D. (2006) Prediction of the recurrence of atrial fibrillation after cardioversion in the Atrial Fibrillation Follow-up Investigation of Rhythm Management (AFFIRM) study. *Am Heart J* **151**: 390–396.

Rankin, J. (1957) Cerebral vascular accidents in patients over the age of 60. II. Prognosis. *Scott Med J* **2**: 200–15.

Sacco, R.L., Boden-Albala, B., Abel, G., Lin, I.F., Elkind, M., Hauser, W. a, *et al.* (2001) Race-ethnic disparities in the impact of stroke risk factors: the northern Manhattan stroke study. *Stroke* **32**: 1725–1731.

Sacco, R.L., Kasner, S.E., Broderick, J.P., Caplan, L.R., Connors, J.J., Culebras, A., *et al.* (2013) An updated definition of stroke for the 21st century: A statement for healthcare professionals from the American heart association/American stroke association. *Stroke* **44**: 2064–2089.

Sanna, T., Diener, H.-C., Passman, R.S., Lazzaro, V. Di, Bernstein, R. a, Morillo, C. a, *et al.* (2014) Cryptogenic stroke and underlying atrial fibrillation. *N Engl J Med* **370**: 2478–86.

Saposnik, G., Fang, J., Kapral, M.K., Tu, J. V., Mamdani, M., Austin, P., and Johnston, S.C. (2012) The iScore predicts effectiveness of thrombolytic therapy for acute ischemic stroke. *Stroke* **43**: 1315–1322.

Saposnik, G., Guzik, A.K., Reeves, M., Ovbiagele, B., and Johnston, S.C. (2013) Stroke Prognostication using Age and NIH Stroke Scale: SPAN-100. *Neurology* **80**: 21–28.

Schmid, R., Baum, P., Ittrich, C., Fundel-Clemens, K., Huber, W., Brors, B., *et al.* (2010) Comparison of normalization methods for Illumina BeadChip HumanHT-12 v3. *BMC Genomics* **11**: 349.

Scott, P.A., and Silbergleit, R. (2003) M isdiagnosis of Stroke in Tissue Plasminogen Activator – Treated Patients : Characteristics and Outcomes. 1–8.

Scott, P.A., Xu, Z., Meurer, W.J., Frederiksen, S.M., Haan, M.N., Westfall, M.W., *et al.* (2010) Attitudes and beliefs of Michigan emergency physicians toward tissue plasminogen activator use in stroke: baseline survey results from the INcreasing Stroke Treatment through INteractive behavioral Change Tactic (INSTINCT) trial hospitals. *Stroke* **41**: 2026–32.

Shi, L., Shi, L., Reid, L.H., Jones, W.D., Shippy, R., Warrington, J. a, *et al.* (2006) The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat Biotechnol* **24**: 1151–1161.

Shi, W., Oshlack, A., and Smyth, G.K. (2010) Optimizing the noise versus bias trade-off for Illumina whole genome expression BeadChips. *Nucleic Acids Res* **38**: e204.

Smith, E.E., Shobha, N., Dai, D., Olson, D.M., Reeves, M.J., Saver, J.L., *et al.* (2010) Risk score for in-hospital ischemic stroke mortality derived and validated within the get with the guidelines-stroke program. *Circulation* **122**: 1496–1504.

Sparano, J.A., Gray, R.J., Makower, D.F., Pritchard, K.I., Albain, K.S., Hayes, D.F., *et al.* (2015) Prospective Validation of a 21-Gene Expression Assay in Breast Cancer. *N Engl J Med* **9**: 960–6.

Stansbury, J.P., Jia, H., Williams, L.S., Vogel, W.B., and Duncan, P.W. (2005) Ethnic disparities in stroke: Epidemiology, acute care, and postacute outcomes. *Stroke* **36**: 374–386.

Stewart, J.A., Dundas, R., Howard, R.S., Rudd, A.G., and Wolfe, C.D.A. (1999) Ethnic differences in incidence of stroke: Prospective study with stroke register. *Br Med J* **318**: 967–971.

Stewart, S., Hart, C.L., Hole, D.J., and McMurray, J.J. V (2002) A population-based study of the long-term risks associated with atrial fibrillation: 20-Year follow-up of the Renfrew/Paisley study. *Am J Med* **113**: 359–364.

Tan, N., Chung, M.K., Smith, J.D., Hsu, J., Serre, D., Newton, D.W., *et al.* (2013) Weighted gene coexpression network analysis of human left atrial tissue identifies gene modules associated with atrial fibrillation. *Circ Cardiovasc Genet* **6**: 362–71.

Tang, Y., Lu, A., Aronow, B.J., and Sharp, F.R. (2001) Blood genomic responses differ after stroke, seizures, hypoglycemia, and hypoxia: Blood genomic fingerprints of disease. *Ann Neurol* **50**: 699–707.

Tang, Y., Xu, H., Du, X., Lit, L., Walker, W., Lu, A., *et al.* (2006) Gene expression in blood changes rapidly in neutrophils and monocytes after ischemic stroke in humans: a

microarray study. *J Cereb Blood Flow Metab* **26**: 1089–102.

Troke, S.T.S., and Roup, S.T.G. (1995) Tissue plasminogen activator for acute ischemic stroke. **333**.

UK-TIA Study Group (1988) United Kingdom transient ischaemic attack (UK-TIA) aspirin trial: interim results. *Bmj* **296**: 316–320.

Vílchez, J. a., Roldán, V., Hernández-Romero, D., Valdés, M., Lip, G.Y.H., and Marín, F. (2013) Biomarkers in atrial fibrillation: An overview. *Int J Clin Pract* **68**: 434–443.

Vinch, C.S., Rashkin, J., Logsetty, G., Tighe, D. a., Hill, J.C., Meyer, T.E., *et al.* (2004) Brain natriuretic peptide levels fall rapidly after cardioversion of atrial fibrillation to sinus rhythm. *Cardiology* **102**: 188–193.

Wardlaw, J.M., and Mielke, O. (2005) Early signs of brain infarction at CT: observer reliability and outcome after thrombolytic treatment--systematic review. *Radiology* **235**: 444–453.

Weimar, C., König, I.R., Kraywinkel, K., Ziegler, a., and Diener, H.C. (2004) Age and National Institutes of Health Stroke Scale Score Within 6 Hours after Onset Are Accurate Predictors of Outcome after Cerebral Ischemia: Development and External Validation of Prognostic Models. *Stroke* **35**: 158–162.

Weimar, C., Ziegler, A., König, I.R., and Diener, H.-C. (2002) Predicting functional outcome and survival after acute ischemic stroke. *J Neurol* **249**: 888–95.

Wiese, S., Breyer, T., Dragu, a, Wakili, R., Burkard, T., Schmidt-Schweda, S., *et al.* (2000) Gene expression of brain natriuretic peptide in isolated atrial and ventricular human myocardium: influence of angiotensin II and diastolic fiber length. *Circulation* **102**: 3074–3079.

Williams, J.M., Jude, M.R., and Levi, C.R. (2013) Recombinant tissue plasminogen activator (rt-PA) utilisation by rural clinicians in acute ischaemic stroke: a survey of barriers and enablers. *Aust J Rural Health* **21**: 262–7.

Wolf, P. a, Abbott, R.D., and Kannel, W.B. (1987) Atrial fibrillation: a major contributor to stroke in the elderly. The Framingham Study. *Arch Intern Med* **147**: 1561–1564.

Wolf, P.A., Abbott, R.D., and Kannel, W.B. (1991) Atrial fibrillation as an independent risk factor for stroke: the Framingham Study. *Stroke* **22**: 983–8.

Wolf, P.A., Dawber, T.R., Thomas, H.E., and Kannel, W.B. (1978) Epidemiologic assessment of chronic atrial fibrillation and risk of stroke: the Framingham study. *Neurology* **28**: 973–977.

Wozakowska-Kapłon, B. (2004) Effect of sinus rhythm restoration on plasma brain

natriuretic peptide in patients with atrial fibrillation. *Am J Cardiol* **93**: 1555–1558.

Xie, Y., Wang, X., and Story, M. (2009) Statistical methods of background correction for Illumina BeadArray data. *Bioinformatics* **25**: 751–7.

Yip, T.R., and Demaerschalk, B.M. (2007) Estimated cost savings of increased use of intravenous tissue plasminogen activator for acute ischemic stroke in Canada. *Stroke* **38**: 1952–5.

CHAPTER 2: GENERAL HYPOTHESIS, OBJECTIVE & APPROACH

2.1 GENERAL HYPOTHESIS

We hypothesize that peripheral blood RNA expression can be used to discriminate between stroke cases and controls, predict stroke outcome, and/or identify patients currently in AF. We also believe RNA biomarkers will also provide new insight into the underlying pathophysiology of cerebrovascular disease.

2.2 GENERAL OBJECTIVE

The overall objective of this PhD thesis is to identify novel RNA biomarkers of cerebrovascular disease.

2.2 RATIONALE AND APPROACH

Both animal and human studies have detected unique RNA expression changes in peripheral blood following stroke. However previous human studies have not had complementary results. In addition, limited qPCR verification and/or validation have been conducted. Furthermore, studies have yet to evaluate RNA biomarkers for stroke prognosis and AF discrimination.

We will conduct agnostic RNA biomarker discovery using low-cost Illumina microarrays that measure transcriptome-wide gene expression simultaneously. Since stroke is a heterogeneous disease we take advantage of the large INTERSTROKE cohort

to identify novel biomarkers. The discriminative capacity of biomarkers for stroke, primary stroke type and stroke prognosis will be evaluated. Significant results will be verified using qPCR and validated in a small group of samples independent from the discovery cohort (Chapter 3).

Since previous stroke expression studies have disregarded global gene networks, we will apply new statistical methods to identify groups of genes with correlated expression. We will further characterize these large groups of genes by conducting pathway analysis and isolating central, highly interconnected genes within each group. The most significant gene within each group will be assessed for discrimination of stroke and stroke prognosis (Chapter 4).

Finally, we will compare whole blood between participants with and without AF to identify novel AF biomarkers. Blood samples will be collected at two time points for each participant: pre-electrical cardioversion (ECV) while patients are in AF and post-ECV when participants are in sinus rhythm. We believe that the case-cross over study design, assessing the same individuals at two time points, will minimizes the impact of inter-individual heterogeneity and co-morbidities. Novel biomarkers identified through microarray analysis will be verified using qPCR and validated in an independent cohort (Chapter 5).

CHAPTER 3: Peripheral blood *MCEMP1* gene expression as a biomarker for stroke prognosis

Kripa Raman, BSc; Martin J. O'Donnell, MB, MRCPI; Anna Czlonkowska, MD, PhD; Yan Carlos Duarte, MD; Patricio Lopez-Jaramillo, MD, PhD, FACP; Ernesto Peñaherrera, MD; Mike Sharma, MD, MSc; Ashkan Shoamanesh, MD, FRCPC; Marta Skowronska, MD; Salim Yusuf, MD, DPhil, MRCP; Guillaume Paré, MD, MSc, FRCPC

Population Health Research Institute, and Hamilton Health Sciences (KR, MOD, MS, AS, SY, GP)

Department of Medical Science, McMaster University, Hamilton, Ontario, Canada (KR)

Department of Medicine, McMaster University, Hamilton, Ontario, Canada and HRB-Clinical Research Facility, National University of Ireland, Galway, Ireland (MOD)

Department of Neurology, McMaster University, Hamilton, Ontario, Canada (AS, MS)

Department of Pathology and Molecular Medicine, McMaster University, Hamilton, Ontario, Canada (GP)

Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario, Canada (SY)

Second Department of Neurology, Institute of Psychiatry and Neurology, Warsaw Medical University, Warsaw, Poland (AC, MS)

Luis Vernaza Hospital, Guayaquil, Ecuador (YCD, EP)

Ophthalmological Foundation of Santander (FOSCAL), Floridablanca, Santander, Colombia; Instituto Masira, School of Health Sciences, University of Santander (UDES), Bucaramanga, Santander, Colombia (PLJ)

3.1 FORWARD

The lack of rapid diagnostic and/or prognostic testing hinders rapid treatment of stroke patients. Previous stroke biomarker studies have used a candidate biomarker approach or involved few participants. This manuscript conducts agnostic RNA biomarker discovery on a portion of participants recruited for the INTERSTROKE study. We identify a novel association between elevated expression of *MCEMP1* in stroke cases as compared with controls. We also demonstrate that *MCEMP1* discriminates between ischemic stroke from hemorrhagic stroke, one-month disability from no disability, and mortality from survival. Moreover, *MCEMP1* expression improves discrimination of one-month prognosis as compared with available clinical characteristics, such as stroke type and baseline mRS.

This manuscript was published in the journal Stroke on March 2016, Volume 47, Issue 3 (PMID: 26846866). In addition, the results of the manuscript have been highlighted in Nature Reviews Neurology (PMID: 26891770) and EBioMedicine (PMID: 26288825).Salim Yusuf and Guillaume Paré conceptualized and designed the study. Kripa Raman designed the analysis plan, conducted all data analysis, qPCR-related laboratory work, and wrote the manuscript. Critical revisions to the manuscript were made by: Martin O'Donnell, Anna Czlonkowska, Yan Carlos Duarte, Patricio Lopez-Jaramillo, Ernesto Peñaherrera, Mike Sharma, Ashkan Shoamanesh, Marta Skowronska, Salim Yusuf and Guillaume Paré.

3.2 ABSTRACT

BACKGROUND AND PURPOSE: A limitation when making early decisions regarding stroke management is the lack of rapid diagnostic and prognostic testing. Our study sought to identify peripheral blood RNA biomarkers associated with stroke. The secondary aims were to assess the discriminative capacity of RNA biomarkers for primary stroke type and stroke prognosis at one-month.

METHODS: Whole blood gene expression profiling was conducted on the discovery cohort, 129 first-time stroke cases that had blood sampling within five days of symptom onset and 170 control participants with no history of stroke.

RESULTS: Through multiple regression analysis we determined that expression of the gene *MCEMP1* had the strongest association with stroke, out of 11,181 genes tested. *MCEMP1* increased by 2.4 fold in stroke as compared with controls (CI 2.0-2.8, p=8.2x10^{-22}). In addition, expression was elevated in intracerebral hemorrhage as compared with ischemic stroke cases (p=3.9x10^{-4}). *MCEMP1* was also highest soon after symptom onset and had no association with stroke risk factors. Furthermore, *MCEMP1* expression independently improved discrimination of one-month outcome. Indeed, discrimination models for disability and mortality that included *MCEMP1* expression, baseline modified Rankin score, and primary stroke type improved discrimination as compared with a model without *MCEMP1* (disability Net Reclassification Index=0.76, p=3.0x10^{-6} and mortality NRI=1.3, p=1.1x10^{-9}). Significant associations with *MCEMP1* were confirmed in an independent validation cohort of 28 stroke cases and 34 controls.

CONCLUSION: This study demonstrates that peripheral blood expression of *MCEMP1* may have utility for stroke diagnosis and as a prognostic biomarker of stroke outcome at one-month.

KEYWORDS: Biomarker; Blood; Gene expression profiling; Stroke; Prognosis

3.3 INTRODUCTION

Stroke is the second leading cause of death worldwide, and a major cause of disability.[1,2] Stroke diagnosis is dependent on clinical assessment and neuroimaging. However the lack of rapid diagnostic testing hinders patient management. Although an effective ischemic stroke treatment is available, tissue plasminogen activator (tPA),[3] studies have observed an underuse by rural[4] and ER physicians[5,6] owing in part to diagnostic uncertainty, risk of hemorrhage and the short therapeutic time window. However a biomarker that establishes diagnosis of stroke and distinguishes hemorrhage from ischemia has the potential to minimize the time from symptom onset to treatment, and improve patient outcomes. Determining a patient's risk for disability or mortality may also be used to inform clinical decision-making, evaluate risk-benefit and optimize allocation of healthcare resources. Indeed, although one third of stroke patients die or experience disability within the first month,[2] clinical risk scores to predict patient outcome are infrequently used by clinicians due to lack of precision, validation and complexity. Identification of biomarkers that quickly distinguish stroke cases from controls, ischemia from hemorrhage and predict prognosis could improve patient management.

The advent of high-throughput genomic technology provides a novel, agnostic approach for biomarker discovery. RNA gene expression levels vary rapidly in response to physiologic changes. Rapid point-of-care RNA tests are currently in development,[7] therefore peripheral blood RNA may be used in the clinic in the future. Both animal[8] and human[9,10] studies have observed unique RNA expression changes in whole blood

following ischemic stroke. However these clinical studies were conducted on a relatively small sample size, consisting of a maximum of 39 ischemic stroke cases and 25 controls,[9,10] and still require validation. In addition, previous studies have assessed RNA biomarkers for stroke diagnosis, but not stroke prognosis. Due to the clinical need for stroke biomarkers and heterogeneity in stroke pathophysiology, large studies are crucial to robustly identify novel biomarkers and to assess clinical value.

In this report, we used a large discovery population, 299 INTERSTROKE participants (129 stroke cases and 170 controls), to identify novel RNA biomarkers of stroke. We then determined whether RNA biomarkers distinguished between primary stroke types and stroke outcome. Significant results were validated in an independent group of participants (28 stroke cases and 34 controls).

3.4 METHODS

3.4.1 Patient population

The INTERSTROKE study has been described in detail elsewhere.[11] Briefly, INTERSTROKE was a large, international, standardized case-control study consisting of stroke cases and control participants from 22 countries. Stroke cases were patients admitted to hospital with first-time acute stroke that presented within five days of symptom onset and within 72-hours of hospital admission. Distinction between stroke subtypes was confirmed with neuroimaging (CT or MRI). Control participants were recruited from the hospital or within the community, and had no history of stroke. 375

INTERSTROKE participants recruited from six centers consented to the expression profiling sub-study. Our expression study benefitted from the international recruitment by the increased ethnic diversity, and the greater prevalence of ICH in South America[12] that increased the number of samples available for analysis. Peripheral whole blood was collected into PAXgene Blood RNA tubes (PreAnalytiX) and stored at -80^{0}C prior to sample processing.

Only two patients with subarachnoid hemorrhage had blood samples collected, and so were excluded from the analysis. 99% of participants were either Latin American or Caucasian, so we excluded participants of other ethnicities (N=9) to reduce potential population stratification. As a result, our study consisted of 364 participants. Initially 302 participants (131 cases and 171 controls) were recruited and consecutively assigned for biomarker discovery. An additional 62 participants (28 cases and 34 controls) were recruited for independent validation.

3.4.2 Sample processing and array hybridization

Total RNA was isolated from the discovery cohort using the QIAsymphony PAXgene Blood RNA kit (Qiagen) according to the manufacturer's protocol. RNA was isolated from the validation cohort using the MagMAX Stabilized Blood Tube RNA Isolation kit (LifeTech). RNA quality was assessed with Nanodrop2000 (Nanodrop) and 2100 Bioanalyzer (Agilent) then quantified using Quant-IT RiboGreen (LifeTech). Total RNA was amplified and biotinylated using the Illumina TotalPrep RNA Amplification

Kit (LifeTech). Samples were then hybridized to Illumina HumanRef-8v4 BeadChips (Illumina) and scanned on the iScan System (Illumina) as per manufacturer protocol.

3.4.3 Microarray data pre-processing

The Illumina HumanRef-8v4 BeadChip interrogates expression of 34,694 unique genes using 47,323 probes. The raw sample probe profile and control probe profile were exported from GenomeStudio version1.9.0 (Illumina). All analysis was performed in R (http://r-project.org). In the discovery cohort, three samples did not pass quality control metrics and were excluded from further analysis. Data pre-processing involved background correction using the non-genomic control probes,[13] quantile normalization and log2 transformation.[14] Probes with detection $p<0.01$ in >50% of the samples were considered expressed.

3.4.4 Statistical analysis

Microarrays (and quantitative PCR) measure relative rather than absolute gene expression, or in other words the relative increase or decrease in expression of a gene as compared with global expression (or housekeeping genes). Differential gene expression was thus reported as fold change (FC), with 95% confidence intervals (CI). Regression models were used to identify RNA transcripts associated with stroke in the discovery cohort. Each model tested a single gene's association with stroke while adjusting for gender, age, body mass index (BMI), ethnicity, and hybridization chip. The hybridization chip variable acted as a surrogate for batch effect and other unwanted technical

variation.[15] To correct for multiple hypotheses testing a conservative Bonferroni correction was applied, setting the significance threshold at $0.05/11{,}181=4.5\text{x}10^{-6}$. As external validation, we assessed the significance of genes reported to be associated with stroke by Tang *et al.*,[9] and Barr *et al.*[10]

Further analysis was conducted on the most significant transcript associated with stroke in the discovery cohort. Regression was used to assess the association between expression and stroke risk factors. The relationship between expression and hours from symptom onset was assessed using regression and T-tests. Comparison of expression between controls and primary stroke types, hemorrhagic and ischemic stroke, and between ischemic subtypes was performed with T-tests. We used ordinal logistic regression to evaluate the association between functional disability, measured as modified Rankin Scale score (mRS) and gene expression. Our analysis utilized mRS recorded soon after the stroke (at baseline) and at the one-month follow-up. One-month mRS was also dichotomized to represent either functional disability (mRS 0-2 vs mRS>2) or mortality (mRS 0-5 vs mRS 6). Using pROC,[16] receiver operator curves (ROC) were constructed from logistic regression models for the dichotomized outcomes. Area under the ROC (AUC) was determined as a measure of sensitivity and specificity. The odds ratio (OR), positive predictive value (PPV) and negative predicative value (NPV) were determined based on the optimal univariate ROC expression threshold. The continuous Net Reclassification Index (NRI)[17] was calculated using Hmisc[18] to compare multiple discrimination models. An NRI >0.6 was considered a strong improvement in discriminative capacity, 0.4 was intermediate, and 0.2 was considered weak.

3.4.6 Quantitative PCR validation and replication

For quantitative real-time PCR (qPCR), complementary DNA was synthesized using the QuantiTect Reverse Transcription Kit (Qiagen). TaqMan qPCR was performed on a Viia7 Real-Time System (LifeTech) where *MCEMP1* was monitored with Hs00545333_g1 (LifeTech) and normalized to *ACTB*, monitored with Hs01060665_g1 (LifeTech). Cycle threshold values were calculated automatically with default parameters and FC was calculated using the **δ**CT method.[19] qPCR confirmed microarray results if Pearson correlation>0.8 and regression $p<0.05$. The independent validation cohort was analyzed using one-sided T-tests and ordinal logistic regression.

3.5 RESULTS

3.5.1 Patient Characteristics

Between March 2007 to April 2010, 364 INTERSTROKE participants consented to the gene expression sub-study. Biomarker discovery was conducted on 299 samples (129 stroke cases and 170 controls) that passed quality control. 62 additional participants (28 stroke cases and 34 controls) were recruited as an independent validation cohort. Patient demographics for the discovery and validation cohorts are presented in *Supplemental Table I*. Among stroke cases in the discovery cohort, 19.4% (N=25) were intracerebral hemorrhage (ICH) and 80.6% (N=104) were ischemic. Based on TOAST criteria,[20] 21.7% of the ischemic strokes were classified as cardioembolic, 7.8% large

vessel, 15.5% small vessel, 24% cryptogenic and 11.6% other. Clinical features of stroke cases were similar to controls in the discovery cohort, except for presence of hypertension (p=0.02), migraine (p=4.5x10^{-3}) and smoking (p=0.01), all of which were more common among stroke cases.

Supplemental Table I. Participant demographics

	DISCOVERY (N=299)			VALIDATION (N=62)		
	Stroke, N = 129	**Control**, N = 170	**P-value**	**Stroke**, N = 28	**Control**, N = 34	**P-value**
Gender (% female)	54 (41.9)	78 (45.9)	0.49	12 (42.9)	14 (41.2)	0.90
Age, mean ± SD	67.6 ± 13.3	66.0 ± 14	0.34	67.5 ± 12.5	67.9 ± 12.1	0.92
BMI, mean ± SD	26.56 ± 4.6	27.16 ± 4.8	0.27	25.84 ± 3.8	26.8 ± 4.0	0.33
Hyperlipidemia (N)	46 (35.7)	62 (36.5)	0.89	14 (50.0)	15 (44.1)	0.65
Hypertension	95 (73.6)	103 (60.6)	0.02*	21 (75.0)	18 (52.9)	0.07
Diabetes	23 (17.8)	21 (12.4)	0.20	5 (17.9)	8 (23.5)	0.59
Atrial fibrillation	9 (7.0)	6 (3.5)	0.20	0	1 (2.9)	0.32
Migraine	33 (25.6)	21 (12.4)	0.0045*	5 (17.9)	2 (5.9)	0.16
Current smoker	26 (20.2)	15 (8.8)	0.01*	3 (10.7)	0	0.08
Race, (%)						
European	43 (33.3)	53 (31.2)		2 (7.1)	13 (38.2)	
Latin American	86 (66.7)	117 (68.8)		26 (92.9)	21 (61.8)	
Stroke type, (%)						
Intracranial hemorrhage	25 (19.4)	NA		4 (14.3)	NA	
Subarachnoid hemorrhage	0	NA		0	NA	
Ischemic stroke	104 (80.6)	NA		24 (85.7)	NA	
Cardioembolic	28 (21.7)	NA		8 (28.6)	NA	
Large vessel	10 (7.8)	NA		1 (3.6)	NA	
Small vessel	20 (15.5)	NA		3 (10.7)	NA	
Cryptogenic	31 (24.0)	NA		8 (28.6)	NA	
Other	15 (11.6)	NA		4 (14.3)	NA	
Stroke severity (N)						
Baseline mRS 0-2	53	NA		10	NA	
Baseline mRS >2	76	NA		18	NA	
One-month mRS 0-2	65	NA		11	NA	
One-month mRS >2	64	NA		17	NA	
Time from symptom onset, mean ± SD	52.1 ± 23.7	NA		63.2 ± 26.5	NA	
<24hrs	17	NA		1	NA	
24-48	38	NA		8	NA	
48-72	41	NA		8	NA	
72-96	29	NA		7	NA	
96+	4	NA		4	NA	

3.5.2 Association between gene expression and stroke

Microarray expression profiling was conducted in the discovery cohort (129 stroke cases and 170 controls). Each of the 11,181 RNA probes were tested for association with stroke and 19% were significantly associated after Bonferroni correction ($p<4.5x10^{-6}$, *Supplemental Figure IA*). As external validation, we compared our significant associations to the 18 genes previously associated with stroke by Tang *et al.*[9] and 9 genes by Barr *et al*.[10] 81.2% of genes identified by Tang *et al.* and 77.8% of genes identified by *Barr et al.* had genome-wide significant association with stroke in our data (*Supplemental Table II and III*). The direction of effect was consistent between our study and previous reports for all replicated genes.

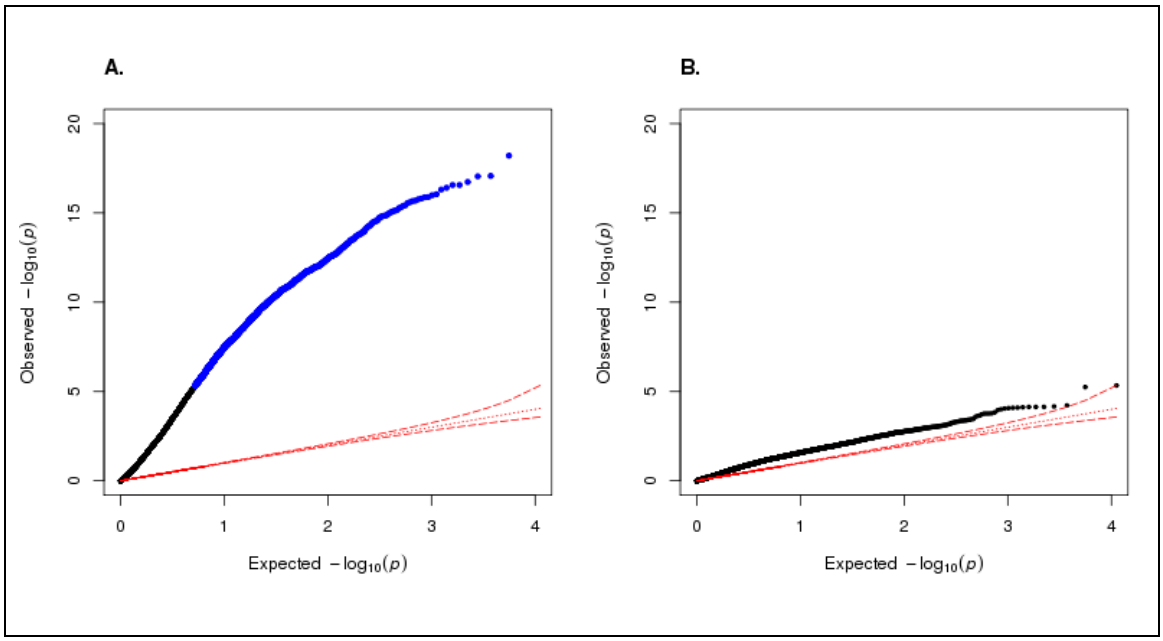
Table 1 presents the ten most significant genes associated with stroke in our discovery cohort. The most significant gene identified was *MCEMP1,* which had a 2.4 fold expression increase in stroke cases compared with controls (CI 2.0-2.8, $p=8.2x10^{-22}$, *Figure 1A*). The AUC of *MCEMP1* for stroke was 0.81 (CI 0.76 - 0.86, *Supplemental Figure IIA*). To test whether multiple probes were non-redundantly associated with stroke, we included *MCEMP1* expression in the initial association models as a co-variable and tested all 11,180 remaining probes. Under this model, the most significant gene was *MSRA* ($p=4.6x10^{-6}$), which did not reach our threshold for statistical significance after Bonferroni correction (*Supplemental Figure IB*).

Differential expression of *MCEMP1* was verified using qPCR in a subset of the discovery cohort (76 stroke cases and 66 controls). We observed high correlation between *MCEMP1* expression levels measured by qPCR and microarray ($r^2=0.88$ and

p=4.8x10^{-48}). Using qPCR a 2.4 fold increase in *MCEMP1* was detected in stroke cases as compared with controls (CI 1.8-3.2, p=1.6x10^{-8}, *Supplemental Figure III*).

Supplemental Figure I. Quantile-quantile plots of P-values from the association between microarray gene expression and stroke

Each point represents one of the RNA transcript probes tested. The x-axis represents expected P-value under the Null hypothesis of no association. The y-axis represents the observed P-value based on association with stroke. Points in blue represent probes that are differentially expressed after Bonferroni correction ($p<4.5x10^{-6}$). The solid red lines represent the expected distribution under the Null (i.e. uniform distribution) with dashed red lines highlight the 95%CI. Panel **(A)** illustrates the P-value distribution from the association between each of the 11, 181 probes and stroke, after adjustment for gender, age, BMI, ethnicity and hybridization chip, while panel **(B)** illustrates a similar association, but further adjusts for *MCEMP1* gene expression. As such this plot visualizes the P-values distribution from 11, 180 probes.



Supplemental Table II. Significance of genes identified by Tang *et al.*, in our INTERSTROKE dataset

Tang *et al.*, identified 18 genes associated with stroke.[9] Only 16 of these genes are represented on the Illumina microarray. In our data set, 13 of the genes had significant association with stroke ($p < 4.5 \times 10^{-6}$).

Gene	P-value	Fold Change
S100A12	1.7×10^{-16}	1.8
BCL6	7.9×10^{-16}	1.6
ARG1	2.7×10^{-14}	2.5
ETS2	4.2×10^{-13}	1.5
PYGL	7.0×10^{-13}	1.4
F5	1.1×10^{-12}	1.5
LY96	2.0×10^{-12}	1.6
CKAP4	2.0×10^{-10}	1.4
SLC16A6	2.4×10^{-10}	1.3
MMP9	6.3×10^{-10}	1.8
S100P	1.1×10^{-9}	2.0
CA4	2.4×10^{-9}	1.5
FPR1	1.6×10^{-6}	1.3
RNASE2	8.2×10^{-6}	1.4
NPL	1.8×10^{-5}	1.2
S100A9	0.01	1.0
HOXA2	NA	NA
HIS2H2AA3	NA	NA

Supplemental Table III. Significance of genes identified by Barr *et al.*, in our INTERSTROKE dataset

Barr *et al*., identified 9 genes associated with stroke.[10] In our data set, 7 of these genes surpassed the pre-processing criteria and had significant association with stroke ($p < 4.5 \times 10^{-6}$).

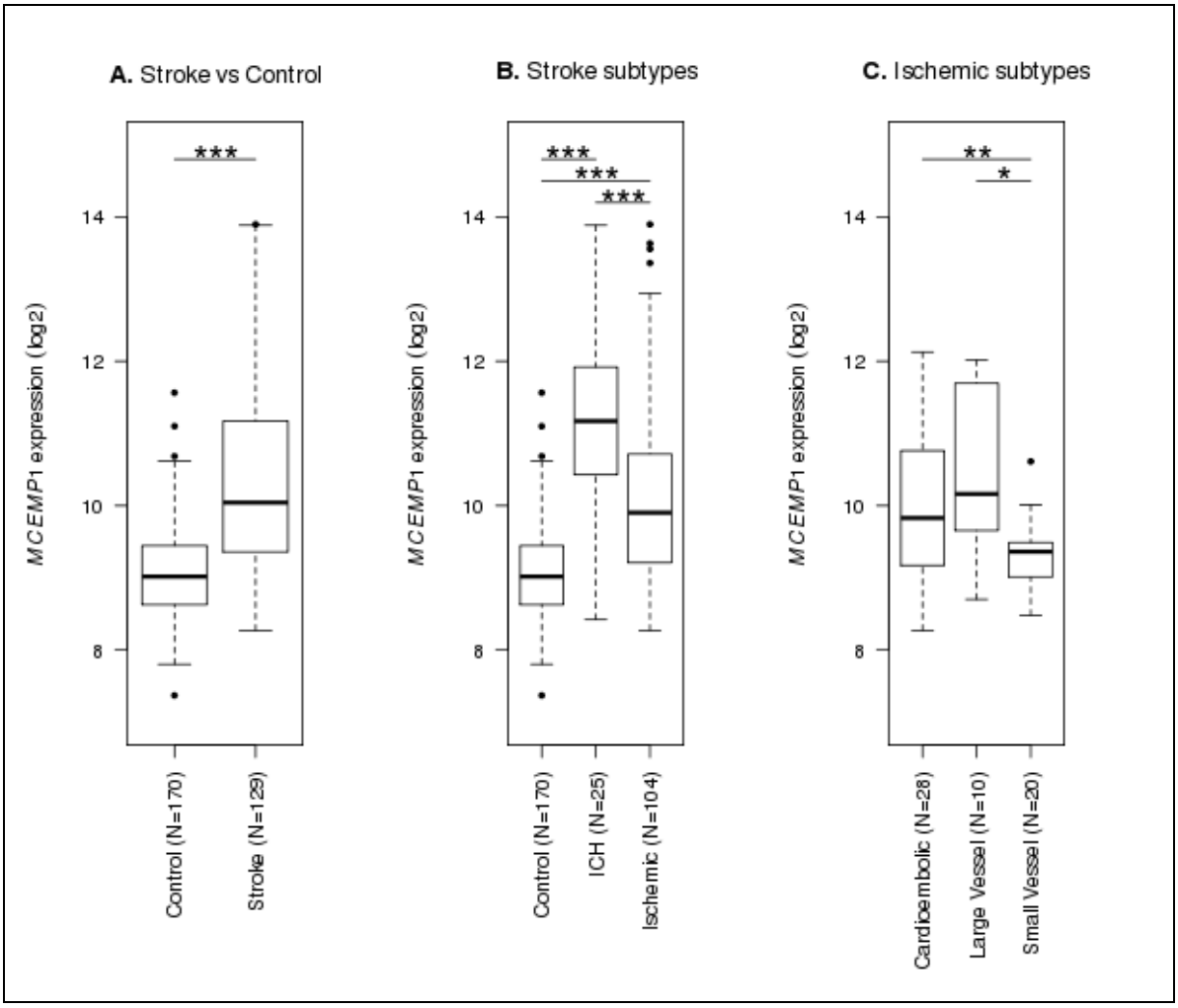
Gene	P-value	Fold Change
S100A12	1.7×10^{-16}	1.8
ARG1	2.7×10^{-14}	2.5
IQGAP1	1.8×10^{-12}	1.3
LY96	2.0×10^{-12}	1.6
CA4	2.4×10^{-9}	1.5
CCR7	1.8×10^{-8}	0.7
ORM1	2.7×10^{-6}	1.7
CSPG2	NA	NA
MMP9	NA	NA

Table 1. Ten most significant genes associated with stroke in the discovery cohort

Gene	P-value	Fold Change	Upper CI	Lower CI	Description
MCEMP1	$8.2x10^{-22}$	2.4	2.8	2.0	Mast cell-expressed membrane protein 1
SPOCK2	$6.3x10^{-19}$	0.6	0.7	0.6	Sparc/osteonectin, cwcv and kazal-like domains proteoglycan
SEPT9	$8.6x10^{-18}$	0.8	0.8	0.7	Septin 9
IRAK3	$9.0x10^{-18}$	1.7	2.0	1.6	Interleukin-1 receptor- associated kinase 3
ANXA3	$1.9x10^{-17}$	2.3	2.8	1.9	Annexin A3
RBM47	$2.8x10^{-17}$	1.6	1.7	1.4	RNA binding motif protein 47
IL18BP	$2.8x10^{-17}$	0.7	0.8	0.7	Interleukin 18 binding protein
TLR5	$3.9x10^{-17}$	1.8	2.0	1.6	Toll-like receptor 5
PCED1B	$5.0x10^{-17}$	0.7	0.8	0.6	PC-esterase domain containing 1B
HS.407903	$9.0x10^{-17}$	1.6	1.8	1.5	-

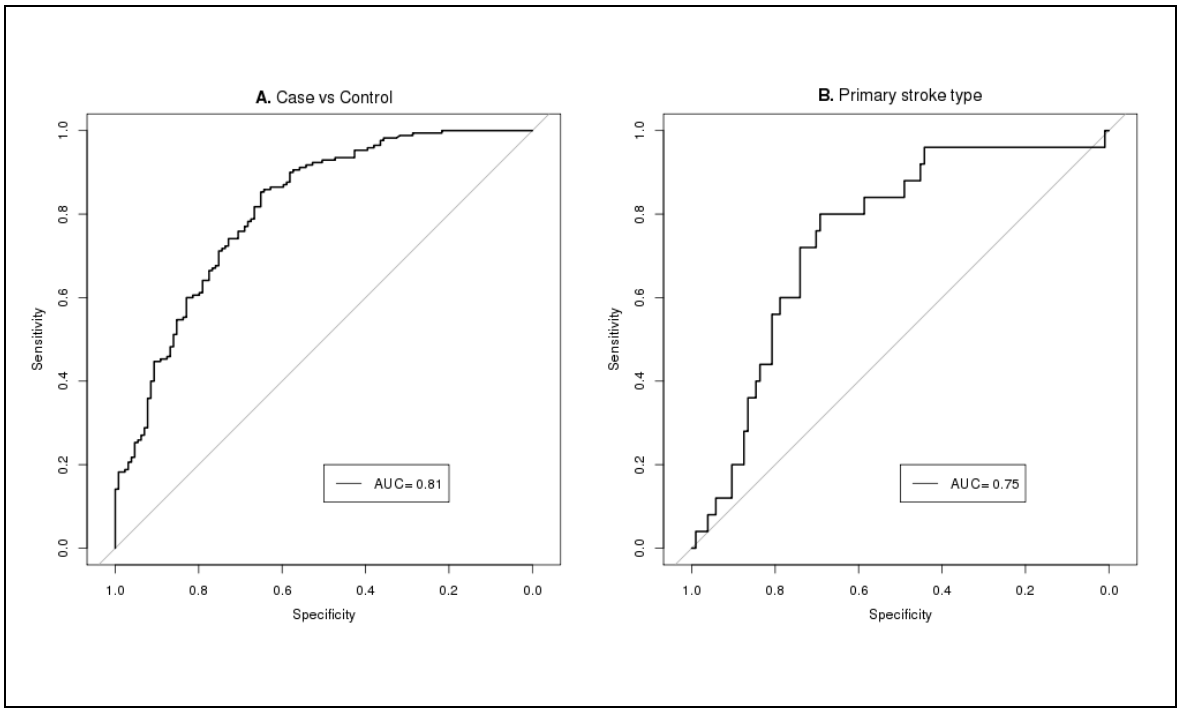
Figure 1. Box-plots of *MCEMP1* microarray expression

Box-plots of *MCEMP1* expression grouped by **(A)** stroke cases and controls, **(B)** controls, ischemic and hemorrhagic cases and **(C)** ischemic stroke subtypes. A symbol directly above a bar indicates a significant difference between groups using Student t-test; $p<0.0005$ (***), $p<0.005$ (**), $p<0.05$ (*).



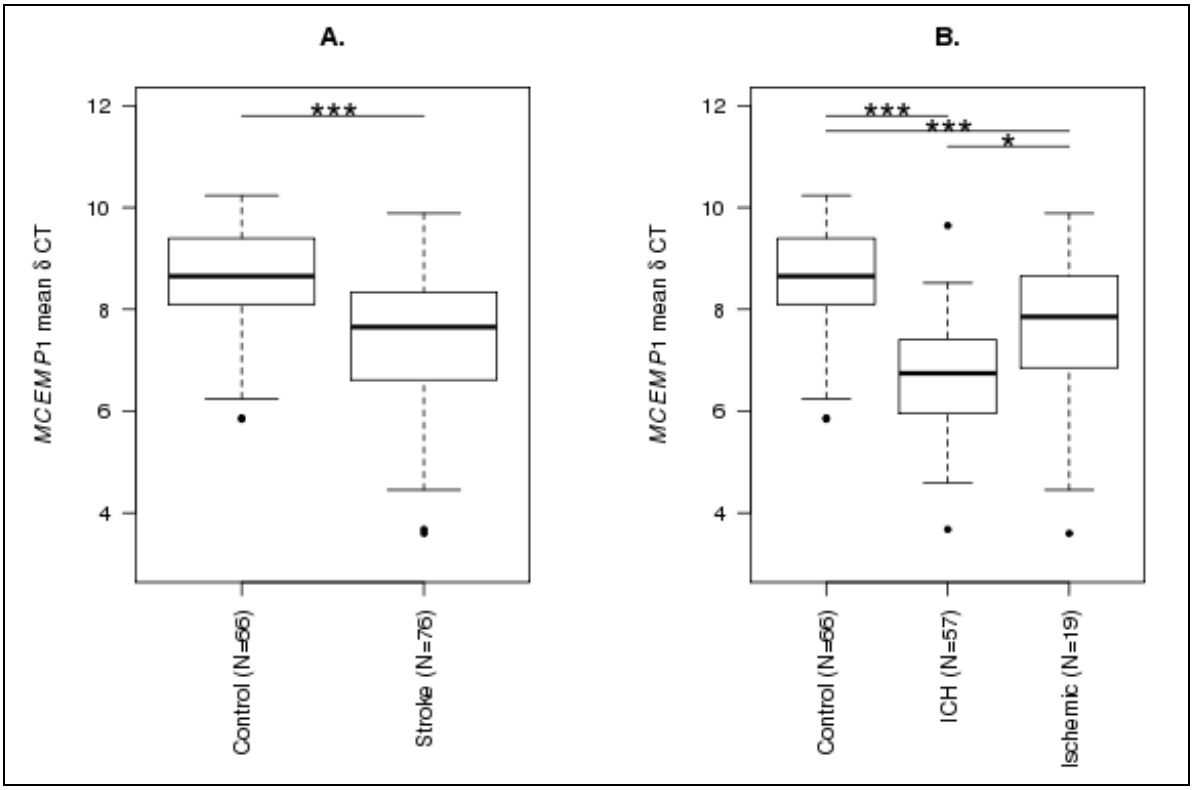
Supplemental Figure II. Receiver-operating-characteristic curves for *MCEMP1* expression discrimination of stroke

(A) Depicts the discriminative capacity of *MCEMP1* for stroke cases vs controls, and **(B)** depicts discrimination between the primary stroke types, hemorrhagic stroke vs ischemic stroke.



Supplemental Figure III. Box-plots of *MCEMP1* expression in a subset of the discovery cohort (N=142)

Boxes extend from the 25th to the 75th percentile, with the horizontal line representing the median. Outliers are identified as samples with an expression value 1.5 times more or less than the interquartile range. The CT (cycle threshold) is the number of PCR cycles required for the fluorescent signal to exceed background levels. Unlike microarray values, CT values are inversely proportional to the amount of target nucleic acid in a sample. **(A)** Stroke cases and controls, **(B)** controls, ischemic stroke cases and ICH stroke cases. Mean ΔCT values are inversely proportional to the amount of target nucleic acid in a sample. A symbol directly above a bar indicates a significant difference between groups; p <0.0005 (***), p<0.005 (**), p<0.05 (*).



3.5.3 *MCEMP1* expression is not associated with stroke risk factors

Restricting the analysis to healthy controls (N=170), we tested *MCEMP1* for association with stroke risk factors including age, gender, BMI, ethnicity, hyperlipidemia, diabetes, atrial fibrillation, hypertension, migraine and smoking status. After adjustment for multiple hypothesis testing, we observed no association between *MCEMP1* and stroke risk factors (p>0.05/9=0.0056, *Supplemental Table IV*). A modest association between elevated *MCEMP1* and hypertension was identified (FC=1.2, CI 1.1-1.4, p=8.8x10^{-3}). However, adjusting the initial stroke association model for all available risk factors did not attenuate the association with *MCEMP1* (FC=3.4, CI 2.4-4.9, p=2.6x10^{-11}).

Supplemental Table IV. Association between *MCEMP1* expression and available stroke risk factors in controls (N=170)

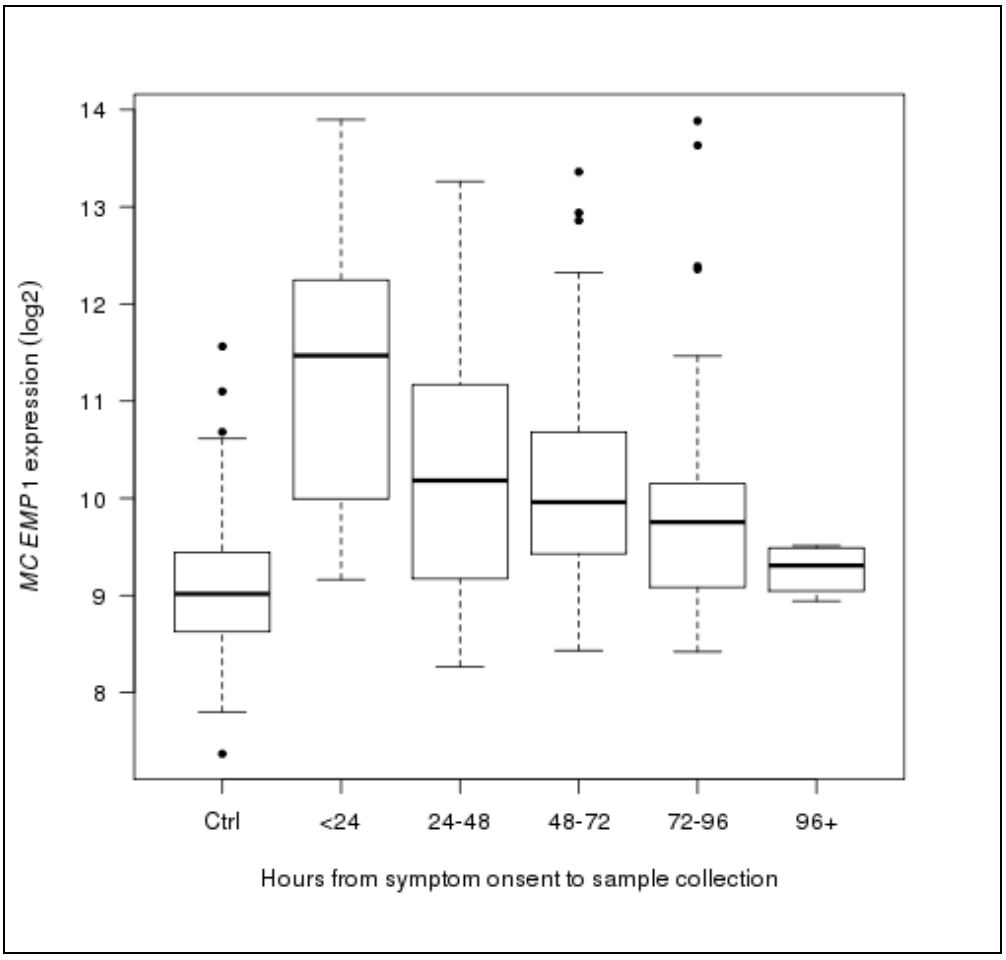
Risk Factor	P-value
Gender	0.05
Age	0.06
BMI	0.38
Ethnicity	0.97
Hyperlipidemia	0.82
Hypertension	0.0088*
Diabetes	0.58
Atrial fibrillation	0.88
Migraine	0.29
Current smoker	0.12

3.5.4 *MCEMP1* expression is associated with time from symptom onset

INTERSTROKE case participants were recruited at varying times after symptom onset. Since the time from symptom onset to blood sampling varied between individuals, we were able to assess the temporal profile of *MCEMP1* change after stroke. We identified a significant relationship where *MCEMP1* decreased by 1% per hour from symptom onset (CI 0.98-1.0, p=3.7x10^{-3}, *Figure 2*), even after adjustment for stroke risk factors. Separating stroke cases by primary stroke type, we also observed that *MCEMP1* decreased as time from symptom onset increased (*Supplemental Figure IV*). Furthermore, *MCEMP1* was highest in samples collected <24-hours of symptoms onset as compared with controls (FC=5.3, CI 3.2-8.5, p=1.7x10^{-6}) or stroke cases collected >24-hours (FC=1.9, CI 1.4-3.9, p=1.9x10^{-3}).

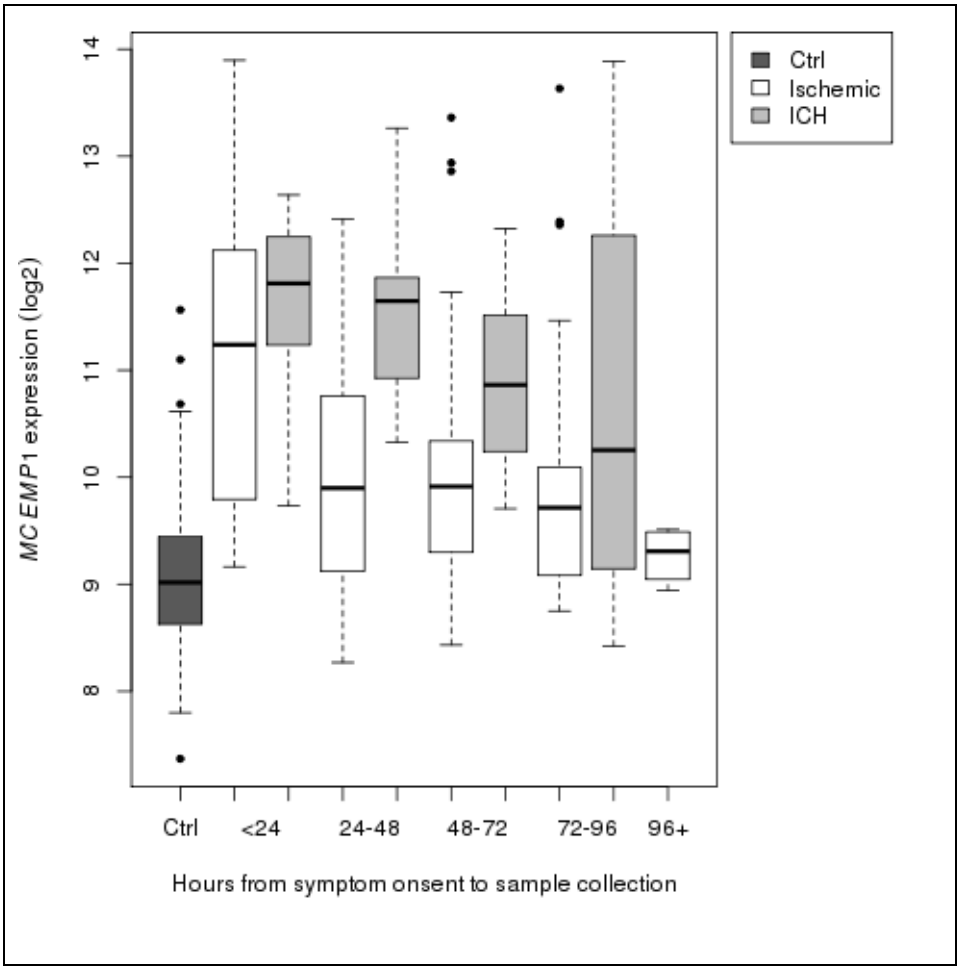
Figure 2. Box-plot of *MCEMP1* expression according to hours from symptom onset

Box-plot of *MCEMP1* expression in controls (N=170) and stroke samples according to hours from symptom onset; <24 (N=17), 24-48 (N=38), 48-72 (N=41), 72-96 (N=29), and 96+ (N=4). A symbol directly above a bar indicates a significant difference between groups using Student t-test; $p<0.0005$ (***).



Supplemental Figure IV. Boxplots of *MCEMP1* expression according to hours from symptom onset and primary stroke type

Boxes extend from the 25th to the 75th percentile, with the horizontal line representing the median. Outliers are identified as samples with an expression value 1.5 times more or less than the interquartile range. Dark grey boxes are controls (N=170), white are ischemic cases (N=104) and light grey are ICH cases (N=25). Of the ischemic cases, 12 samples were collected within 24hrs, 31 within 24-48hrs, 35 within 48-72hrs, 22 within 72-96hrs and 4 after 96+hrs. Of the ICH cases, 6 samples were collected within 24hrs, 7 within 24-48hrs, 8 within 48-72hrs, 4 within 72-96hrs, and 0 after 96+hrs.



3.5.5 *MCEMP1* expression differs between stroke types

MCEMP1 was increased by 4.5 fold in ICH cases as compared with controls (CI 3.1-6.4, p=3.4x10^{-9}) and by 2.1 fold in ischemic cases compared with controls (CI 1.8-2.6, p=3.4x10^{-13}). Accordingly, a 2.1 fold increase in *MCEMP1* was observed in ICH cases as compared with ischemic (CI 1.4-3.1, p=3.9x10^{-4}, *Figure 1B*). The area under the ROC for primary stroke type discrimination by *MCEMP1* was 0.75 (CI 0.65-0.85, *Supplemental Figure IIB*). Expression differences were also detected between ischemic stroke subtypes. *MCEMP1* was elevated in cardioembolic (FC=1.5, CI 1.1-2.1, p=8.1x10^{-3}) and large vessel (FC=2.3, CI 1.2-4.1, p=0.012) stroke as compared with small vessel stroke (*Figure 1C*).

3.5.6 Baseline and one-month mRS associated with *MCEMP1* expression

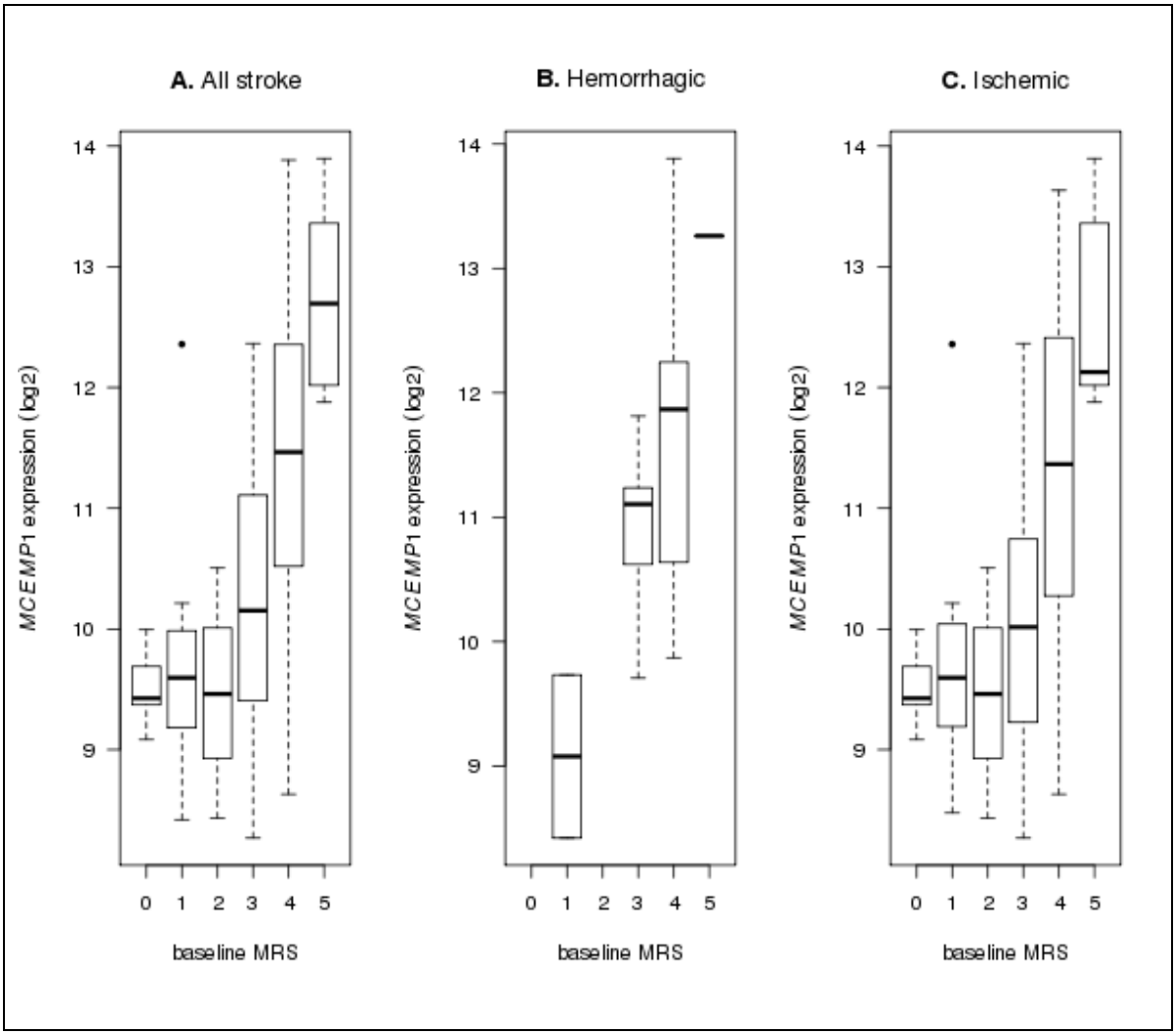
Baseline mRS, measured soon after stroke, was associated with *MCEMP1* expression (p=4.0x10^{-13}, *Supplemental Figure V*). One unit *MCEMP1* expression increase was associated with a 3.3 odds (CI 2.4-4.5) increase in baseline mRS. The association remained significant after adjustment for stroke risk factors, stroke type, tPA treatment and hours from symptom onset (OR=3.1, CI=2.4-4.5, p=1.8x10^{-9}).

One-month mRS was also associated with *MCEMP1* expression (p=1.3x10^{-14}, *Supplemental Figure VI*). One unit *MCEMP1* expression increase was associated with a 3.4 odds (CI 2.5-4.6) increase in one-month mRS and the association remained significant after adjustment for stroke risk factors, primary stroke type, tPA treatment, hours from symptom onset and baseline mRS as a categorical variable (OR=1.8, CI=1.2-2.8,

p=6.6x10^{-3}). In fact, only *MCEMP1* expression (p=6.6x10^{-3}), baseline mRS (p=3.2x10^{-3}-6.4x10^{-10}), and primary stroke type (p=1.0x10^{-3}) were independently associated with one-month mRS.

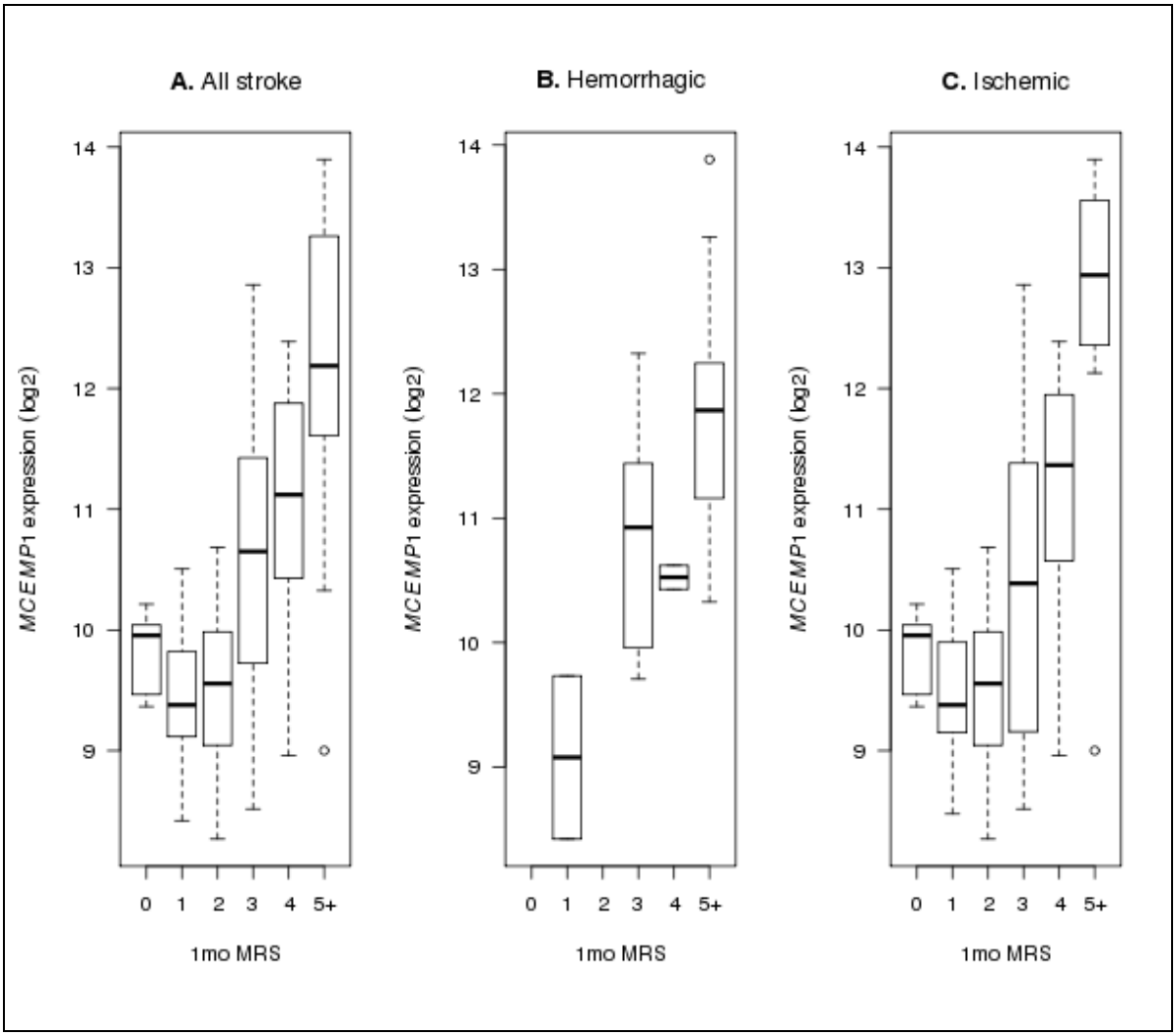
Supplemental Figure V. Boxplots of *MCEMP1* expression according to baseline modified Rankin Score (mRS)

Boxes extend from the 25th to the 75th percentile, with the horizontal line representing the median. Outliers are identified as samples with an expression value 1.5 times more or less than the interquartile range. (A) Includes all stroke cases (N=129), (B) ICH stroke cases (N=25), and (C) ischemic stroke cases (N=104).



Supplemental Figure VI. Boxplots of *MCEMP1* expression according to one-month modified Rankin Score (mRS)

Boxes extend from the 25th to the 75th percentile, with the horizontal line representing the median. Outliers are identified as samples with an expression value 1.5 times more or less than the interquartile range. (A) Includes all stroke (N=129), (B) hemorrhagic stroke cases (N=25), and (C) ischemic stroke cases (N=104).

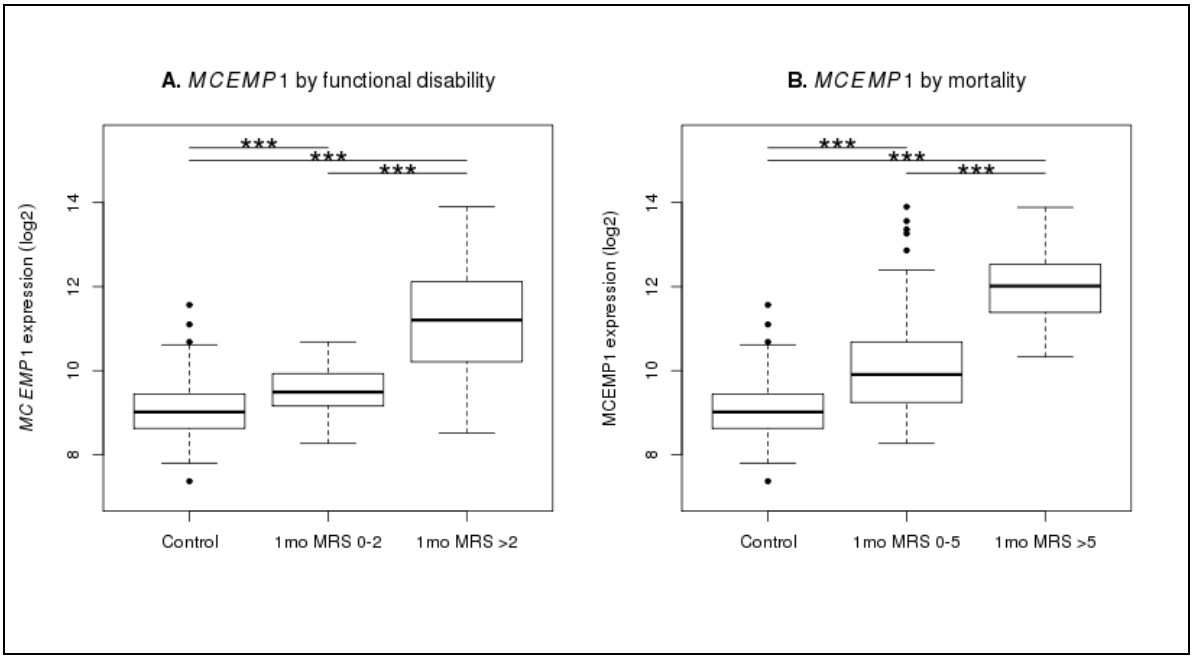


3.5.7 *MCEMP1* expression is associated with disability at one-month

To further explore the association, we dichotomized one-month mRS into two groups, mRS of 0, 1 or 2 and mRS >2, representing the ability to live autonomously or not post-stroke. Individuals with disability at one-month had elevated baseline *MCEMP1* as compared with controls (FC=4.7, CI 3.5-5.7, p=1.6x10^{-19}) or individuals without disability (FC=3.2, CI 2.5-4.2, p=1.8x10^{-14}, *Figure 3A*). A disability discrimination model including *MCEMP1* expression, primary stroke type, and baseline mRS as a categorical variable, strongly improved discrimination as compared with a model including only primary stroke type and baseline mRS (AUC with *MCEMP1*=0.96, AUC without *MCEMP1*= 0.93, NRI=0.76, p=3.0x10^{-6}). The optimal *MCEMP1* threshold had a specificity of 80.3%, sensitivity of 86.2%, with corresponding PPV of 78.1% and NPV of 87.7%, for disability (*Supplemental Table V*). The odds ratio for disability was 6.6 (CI 1.9-22.7) in individuals with high versus low *MCEMP1* expression after adjustment for stroke type and baseline mRS.

Figure 3. Box-plots of *MCEMP1* expression according to dichotomized one-month mRS

Box-plots of *MCEMP1* expression in controls (N=170) and in cases according to one-month mRS representing (A) functional disability, mRS 0-2 (N=65) and mRS >2 (N=64), or (B) mortality, mRS 0-5 (N=114) and mRS >5 (N=15). A symbol directly above a bar indicates a significant difference between groups using Student t-test; $p<0.0005$ (***).



Supplemental Table V. Two-way contingency table of disability at one-month and baseline *MCEMP1* expression

One-month disability was determined by dichotomizing one-month mRS to represent no disability (mRS < 2) or disability (mRS > 2).

		One-month outcome	
		No disability (mRS 0-2)	**Disability (mRS >2)**
expression	**High**	8	50
	Low	57	14

3.5.8 *MCEMP1* expression is associated with mortality at one-month

MCEMP1 was also associated with one-month mortality after adjustment for stroke risk factors, baseline mRS, primary stroke type, tPA treatment and hours from symptom onset (FC=3.8, CI 1.4-11.0, p=9.9x10^{-3}, *Figure 3B*). Comparing univariate one-month mortality discrimination models, *MCEMP1* appeared more informative (AUC=0.88) than primary stroke type (AUC=0.80). Moreover a model including *MCEMP1*, primary stroke type and baseline mRS, as a categorical variable, strongly improved mortality discrimination as compared with a model without *MCEMP1* (AUC with *MCEMP1*=0.97, AUC without *MCEMP1*= 0.92, NRI=1.3, p=1.1x10^{-9}). Selecting the optimal discrimination threshold, *MCEMP1* had a specificity of 97.8%, sensitivity of 35.1%, PPV of 86.7%, and NPV of 78.9% for mortality (*Supplemental Table VI*). The odds ratio for mortality was 20.7 (CI 2.5-174.6) in individuals with high as compared with low expression, after adjustment for baseline mRS and stroke type.

Supplemental Table VI. Two-way contingency table of one-month mortality and baseline *MCEMP1* expression

One-month mortality was determined by dichotomizing one-month mRS to represent survival (mRS < 5) or death (mRS = 6).

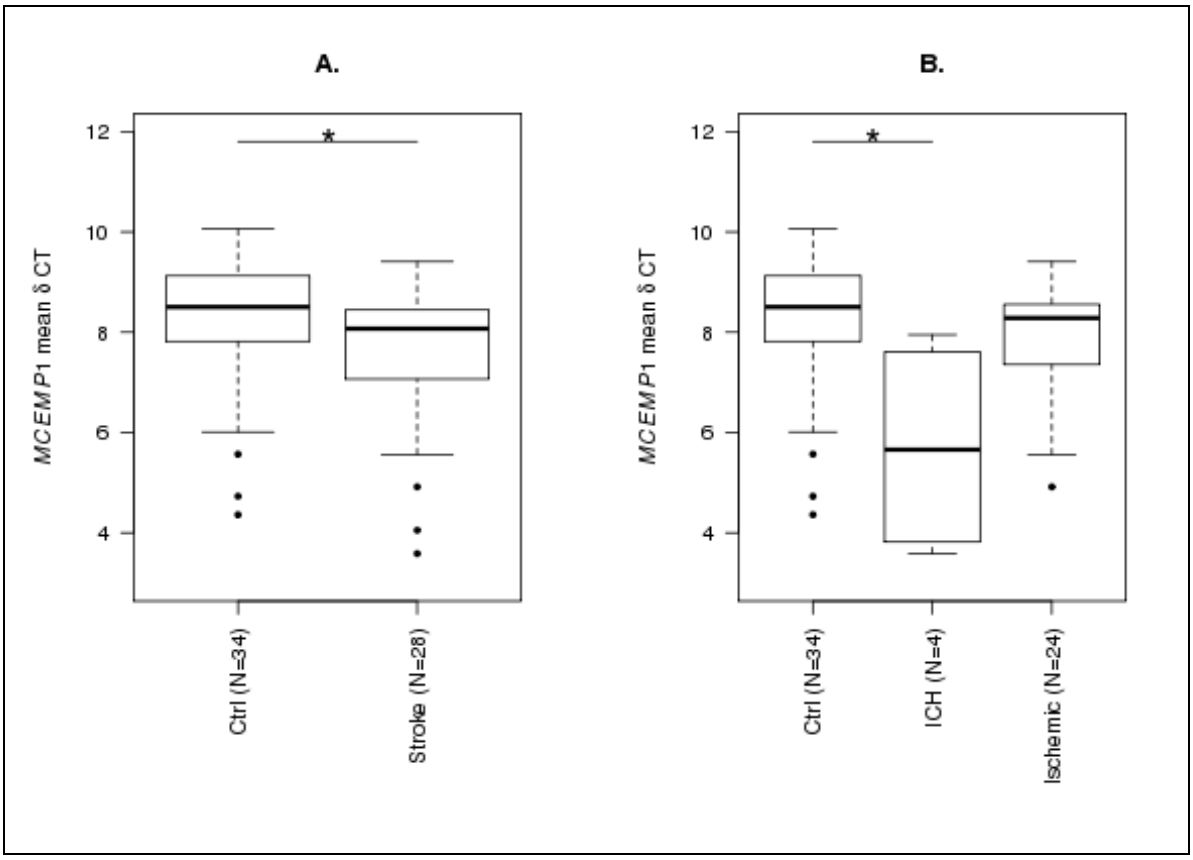
		One-month outcome	
		Alive (mRS 0-5)	**Dead (mRS 6)**
expression	**High**	24	13
	Low	90	2

3.5.9 Replication of *MCEMP1* associations in validation cohort

The significance of *MCEMP1* was validated in a small independent cohort (28 stroke cases and 34 controls) using qPCR. Power calculations indicated that we had sufficient power (>99%) to detect expression difference at a significance of $p<0.05$. We detected increased *MCEMP1* in stroke cases as compared with controls (FC=1.6, $p=0.039$, *Supplemental Figure VII*). We also observed trends towards higher *MCEMP1* in ICH cases compared with controls (FC=5.6, $p=0.05$), higher expression in ischemic cases than controls (FC=1.3, $p=0.14$), and higher expression in ICH cases than ischemic (FC=4.4, $p=0.074$). Finally, both baseline mRS and one-month mRS were associated with *MCEMP1* expression (baseline $p=0.049$, one-month $p=3.3\text{x}10^{-3}$).

Supplemental Figure VII. Box-plots of *MCEMP1* qPCR values in the validation cohort (N=62)

Boxes extend from the 25th to the 75th percentile, with the horizontal line representing the median. Outliers are identified as samples with an expression value 1.5 times more or less than the interquartile range. The CT (cycle threshold) is the number of PCR cycles required for the fluorescent signal to exceed background levels. Unlike microarray values, CT values are inversely proportional to the amount of target nucleic acid in a sample. **(A)** Stroke cases and controls, **(B)** controls, ischemic stroke cases and ICH stroke cases. Mean ΔCT values are inversely proportional to the amount of target nucleic acid in a sample. A symbol directly above a bar indicates a significant difference between groups; $p<0.05$ (*).



3.6 DISCUSSION

The present study evaluated peripheral blood gene expression in a sub-group of INTERSTROKE participants. We identified elevated expression of a novel gene, *MCEMP1,* in stroke cases as compared with controls. *MCEMP1* decreased as time from symptom onset increased and expression was increased in hemorrhagic stroke cases as compared with ischemic. We also identified an association between one-month mRS and *MCEMP1*, where increased functional disability and mortality were associated with increased *MCEMP1*. One-month prognosis discrimination models that included *MCEMP1*, primary stroke type and baseline mRS improved discrimination as compared with similar models without *MCEMP1*. The associations between *MCEMP1* with stroke, primary stroke type and stroke prognosis were independently confirmed in the validation cohort.

Our results demonstrate that non-invasive measurement of *MCEMP1* soon after stroke provides additional prognostic information to clinical characteristics. We observed that *MCEMP1* decreased as time from symptom onset increased, suggesting that *MCEMP1* may have utility for estimating time from symptom onset. In addition, identifying patients with poor prognosis may be beneficial for informed clinical decision making and assessing the risk-benefit ratio for acute therapies. Although several clinical scores have been proposed to predict outcome and mortality,[21–23] these scores are not routinely used in the clinic in part due to their complexity. We have shown that a simple model including only baseline mRS, primary stroke type and *MCEMP1* expression may

predict one-month disability and mortality. Indeed, inclusion of *MCEMP1* strongly improved discrimination of one-month prognosis.

Mast-cell expressed membrane protein 1 (MCEMP1), also known as C19ORF59, is a transmembrane protein first identified in mast cells,[24] but also expressed by macrophages and other tissue.[25] The exact function of MCEMP1 has yet to be determined, however the gene's promoter region contains NF-kB and NF-AT binding motifs, similar to many immune receptor genes.[24] Although limited research has been conducted on MCEMP1, recent findings indicate an emerging role for brain resident mast cells in acute stroke. Experimental stroke studies have reported that mast cells are first responders to cerebral ischemia and act as early regulators of blood-brain barrier (BBB) permeability.[26-28] The increase in *MCEMP1* expression observed in stroke cases may be the result of cerebral mast cell activation and mast cell mediated BBB disruption. Furthermore the expression difference detected between the primary stroke types and ischemic stroke subtypes may indicate an association between *MCEMP1* expression and infarct size.

Whole genome expression following ischemic stroke has been previously assessed[9,10] in small studies including 15 to 39 stroke cases. We used a larger discovery cohort of 129 stroke cases and 170 controls, and verified expression of 77.8% to 81.2% of previously reported genes. To our knowledge, our study shows the largest proportion of overlap with previous reports, providing confidence in both novel and known results. In addition, a recent study reported significantly elevated *MCEMP1* in 12 ischemic stroke cases as compared with 12 controls.[29] Though the study had a small sample size and

lacked replication, stroke samples were collected within 24-hours of symptom onset thus further positioning *MCEMP1* as a marker of acute stroke. Our study robustly identified *MCEMP1* as a stroke biomarker in a significantly larger, multi-ethnic population and confirmed our findings in an independent validation cohort. The overlap and concordance between our study and previous works add credence to our findings and the significance of *MCEMP1* as an acute stroke biomarker.

There are a few study limitations that warrant further discussion. First, our study lacks stroke mimics and non-strokes such as transient ischemic attacks, migraines, seizure and other neurological or inflammatory disorders. Thus we were unable to assess the specificity of *MCEMP1*. Second, an acute stroke biomarker would have the greatest clinical utility if increased concentrations were detected very shortly after symptom onset. In our discovery cohort only 17 samples were collected within 24-hours of symptom onset, but we observed elevated *MCEMP1* expression in these samples, as compared with controls or stroke samples collected after 24-hours. Nonetheless, very early sampling (<6 hours) will be required to assess the utility of *MCEMP1* in guiding acute stroke treatment. Third, due to the nature of the INTERSTROKE study design, there was limited neurological imaging data available, and consequently, the effect of infarct size on expression could not be assessed. However, *MCEMP1* may provide useful information early after stroke onset at times where stroke volume is difficult to assess with plain CT-scans. Future studies including neurological imaging and NIHSS score will be useful to further characterize the association between *MCEMP1* and stroke severity. Finally our

study focused on a single gene to differentiate between the various stroke groups, but there may be other genes with diagnostic potential.

The results of the study demonstrate that *MCEMP1* expression has prognostic capacity beyond baseline mRS and stroke type. *MCEMP1* may also have diagnostic capacity. These observations are promising given the currently limited number of simple clinical tools available to predict outcome and mortality, and lack of an established non-invasive stroke biomarker. The results also point towards an important role for mast cells in stroke and unraveling the biological mechanisms may lead to the identification of new therapeutic targets. Future clinical studies including a stroke mimic cohort, very early blood sampling and measurement of stroke severity will help determine the diagnostic capacity and clinical utility of *MCEMP1*.

3.7 REFERENCES

1. Go AS, Mozaffarian D, Roger VL, Benjamin EJ, Berry JD, Borden WB, et al. Heart disease and stroke statistics--2013 update: a report from the American Heart Association. *Circulation*. 2013;127:e6–e245.
2. Donnan GA, Fisher M, Macleod M, Davis SM. Stroke. *Lancet (London, England)*. 2008;371:1612–23.
3. Emberson J, Lees KR, Lyden P, Blackwell L, Albers G, Bluhmki E, et al. Effect of treatment delay, age, and stroke severity on the effects of intravenous thrombolysis with alteplase for acute ischaemic stroke: a meta-analysis of individual patient data from randomised trials. *Lancet (London, England)*. 2014;384:1929–35.
4. Williams JM, Jude MR, Levi CR. Recombinant tissue plasminogen activator (rt-PA) utilisation by rural clinicians in acute ischaemic stroke: a survey of barriers and enablers. *Aust. J. Rural Health*. 2013;21:262–7.
5. Brown DL, Barsan WG, Lisabeth LD, Gallery ME, Morgenstern LB. Survey of emergency physicians about recombinant tissue plasminogen activator for acute ischemic stroke. *Ann. Emerg. Med.* 2005;46:56–60.
6. Scott PA, Xu Z, Meurer WJ, Frederiksen SM, Haan MN, Westfall MW, et al. Attitudes and beliefs of Michigan emergency physicians toward tissue plasminogen activator use in stroke: baseline survey results from the INcreasing Stroke Treatment through INteractive behavioral Change Tactic (INSTINCT) trial hospitals. *Stroke.* 2010;41:2026–32.
7. Dobson MG, Galvin P, Barton DE. Emerging technologies for point-of-care genetic testing. *Expert Rev. Mol. Diagn.* 2007;7:359–370.
8. Tang Y, Lu A, Aronow BJ, Sharp FR. Blood genomic responses differ after stroke, seizures, hypoglycemia, and hypoxia: Blood genomic fingerprints of disease. *Ann. Neurol.* 2001;50:699–707.
9. Tang Y, Xu H, Du X, Lit L, Walker W, Lu A, et al. Gene expression in blood changes rapidly in neutrophils and monocytes after ischemic stroke in humans: a microarray study. *J. Cereb. Blood Flow Metab.* 2006;26:1089–102.
10. Barr TL, Conley Y, Ding J, Dillman A, Warach S, Singleton A, et al. Genomic biomarkers and cellular pathways of ischemic stroke by RNA gene expression profiling. *Neurology*. 2010;75:1009–14.
11. O'Donnell M, Xavier D, Diener C, Sacco R, Lisheng L, Zhang H, et al. Rationale and design of INTERSTROKE: a global case-control study of risk factors for

stroke. *Neuroepidemiology*. 2010;35:36–44.

12. O'Donnell MJ, Xavier D, Liu L, Zhang H, Chin SL, Rao-Melacini P, et al. Risk factors for ischaemic and intracerebral haemorrhagic stroke in 22 countries (the INTERSTROKE study): a case-control study. *Lancet*. 2010;376:112–23.
13. Shi W, Oshlack A, Smyth GK. Optimizing the noise versus bias trade-off for Illumina whole genome expression BeadChips. *Nucleic Acids Res.* 2010;38:e204.
14. Schmid R, Baum P, Ittrich C, Fundel-Clemens K, Huber W, Brors B, et al. Comparison of normalization methods for Illumina BeadChip HumanHT-12 v3. *BMC Genomics*. 2010;11:349.
15. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*. 2007;8:118–27.
16. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*. 2011;12:77.
17. Pencina MJ, D'Agostino RB, Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat. Med.* 2008;27:157–172.
18. Harrel FE. Hmisc: A package of miscellaneous R functions [Internet]. 2015 [cited 2015 Feb 1];Available from: https://cran.r-project.org/web/packages/Hmisc/index.html
19. Livak KJ, Schmittgen TD. Analysis of relative gene expression data using real-time quantitative PCR and. *Methods*. 2001;25:402–408.
20. The Publications Committee for the Trial of ORG 10172 in Acute Stroke Treatment (TOAST) Investigators. Low molecular weight heparinoid, ORG 10172 (danaparoid), and outcome after acute ischemic stroke: a randomized controlled trial. *JAMA*. 1998;279:1265–72.
21. Saposnik G, Kapral MK, Liu Y, Hall R, O'Donnell M, Raptis S, et al. IScore: A risk score to predict death early after hospitalization for an acute ischemic stroke. *Circulation*. 2011;123:739–749.
22. O'Donnell MJ, Fang J, D'Uva C, Saposnik G, Gould L, McGrath E, et al. The PLAN score: a bedside prediction rule for death and severe disability following acute ischemic stroke. *Arch. Intern. Med.* 2012;172:1548–56.
23. Adams Jr. HP, Davis PH, Leira EC, Chang KC, Bendixen BH, Clarke WR, et al. Baseline NIH Stroke Scale score strongly predicts outcome after stroke - A report of the Trial of Org 10172 in Acute Stroke Treatment (TOAST). *Neurology*.

1999;53:126–131.

24. Li K, Wang S-W, Li Y, Martin RE, Li L, Lu M, et al. Identification and expression of a new type II transmembrane protein in human mast cells. *Genomics*. 2005;86:68–75.

25. Andersson S, Nilsson K, Fagerberg L, Hallström BM, Sundström C, Danielsson A, et al. The transcriptomic and proteomic landscapes of bone marrow and secondary lymphoid tissues. *PLoS One*. 2014;9:e115911.

26. Strbian D, Karjalainen-Lindsberg M-L, Tatlisumak T, Lindsberg PJ. Cerebral mast cells regulate early ischemic brain swelling and neutrophil accumulation. *J. Cereb. Blood Flow Metab.* 2006;26:605–12.

27. Lindsberg PJ, Strbian D, Karjalainen-Lindsberg M-L. Mast cells as early responders in the regulation of acute blood-brain barrier changes after cerebral ischemia and hemorrhage. *J. Cereb. Blood Flow Metab.* 2010;30:689–702.

28. Mattila OS, Strbian D, Saksi J, Pikkarainen TO, Rantanen V, Tatlisumak T, et al. Cerebral mast cells mediate blood-brain barrier disruption in acute experimental ischemic stroke through perivascular gelatinase activation. *Stroke.* 2011;42:3600–5.

29. Oh S-H, Kim O-J, Shin D-A, Song J, Yoo H, Kim Y-K, et al. Alteration of immunologic responses on peripheral blood in the acute phase of ischemic stroke: blood genomic profiling study. *J. Neuroimmunol.* 2012;249:60–5.

CHAPTER 4: Identifying biomarkers of ischemic stroke using gene co-expression analysis

Kripa Raman[1,2,3], and Guillaume Paré[1,2,4]

[1] Population Health Research Institute, David Braley Cardiac, Vascular and Stroke Research Institute, 237 Barton Street East, Hamilton, ON L8L 2X2, Canada

[2] Thrombosis and Atherosclerosis Research Institute, David Braley Cardiac, Vascular and Stroke Research Institute, 237 Barton Street East, Hamilton, ON L8L 2X2, Canada

[3] Department of Medical Sciences, McMaster University, 1280 Main Street West, Hamilton ON L8S 4K1, Canada

[4] Department of Pathology and Molecular Medicine, McMaster University, Michael G. DeGroote School of Medicine, 1280 Main Street West, Hamilton ON L8S 4K1, Canada

4.1 FORWARD

Previous RNA biomarker studies for stroke have disregarded gene expression correlations. However, study of global gene expression networks may provide new insights into the molecular biology underlying stroke and lead to the identification of biomarker panels. Using weighted gene co-expression analysis (WGCNA) we identified groups of genes with correlated expression that were associated with ischemic stroke. These groups of genes are referred to as modules. Central, interconnected genes within each module were then identified and tested for discriminative capacity. These highly interconnected genes are referred to as hub genes. *NLRC4, CKLF*, and *HS.546375* were the top hub greens associated with three different modules. Each gene was also independently associated with ischemic stroke. We then determined that multi-gene models had greater discriminative capacity for stroke and stroke prognosis as compared with single gene models.

This project was designed, conducted and written by Kripa Raman and Guillaume Paré.

4.2 ABSTRACT

INTRODUCTION: Genetic mechanisms underlying ischemic stroke remain incompletely understood. Previous differential expression studies for ischemic stroke have been limited by small samples sizes and provided limited understanding of global gene networks, suggesting a need for large-scale, network-based analyses.

METHODS: As a sub-study of INTERSTROKE, whole blood gene expression profiling was conducted on 104 first-time ischemic stroke cases and 170 controls with no history of stroke. Weighted gene co-expression network analysis was performed to detect groups of co-expressed genes referred to as modules.

RESULTS: Of the 15 modules identified, four were associated with ischemic stroke after adjustment for clinical risk factors. Pathway analysis of genes within the four modules suggested activation of cytokine-cytokine receptor interactions, chemokine signalling and RNA transport. *NLRC4, CKLF*, and *HS.546375* were the most highly connected genes within unique modules and were independently associated with stroke. An ischemic stroke discrimination model including the three genes *(*AUC=0.83, CI 0.78-0.88*)* moderately improved performance compared with single gene models (NRI>0.54). One-month disability discrimination was also improved using the three-gene model (AUC=0.81, CI 0.73-0.90) as compared with single gene models (NRI>0.45). Disability discrimination was further improved using a multi-gene model consisting of *NLRC4, CKLF*, *HS.546375, MCEMP1* and baseline modified Rankin scale score as compared with the three-gene model (NRI=0.48, p=0.013) or single-gene *MCEMP1* model (NRI=0.52, p=0.0073).

CONCLUSION: This study demonstrates that network analysis can implicate new genes with stroke such as *NLRC4*, *CKLF* and *HS.546375*. In addition, multi-gene panels identified through network analysis can improve discrimination of stroke and stroke prognosis.

KEYWORDS: Biomarker; Blood; Gene expression profiling; Stroke

4.3 INTRODUCTION

Stroke is a leading cause of death and disability worldwide.[1,2] Ischemia is the underlying cause of 80% of stroke cases,[3] however treatment of ischemic stroke patients is impeded by the lack of rapid diagnostic testing. Multiple studies have sought to identify blood-based biomarker of ischemic stroke.[4–7] Biomarkers may also provide insight into the molecular mechanisms underlying stroke. With advancements in high-throughput genomic technology, such as microarrays and sequencing, it is now possible to evaluate the expression of thousands of genes simultaneously. Using microarray data and univariate analysis, we have previously reported that whole blood expression of *MCEMP1* may have utility for stroke diagnosis and predicting prognosis.[7] A limitation of our previous work was that the study focused on a single gene to differentiate between the various groups and thus provided little understanding of the global gene interactions.

Univariate analysis assumes that each gene is expressed independent of one another. However genes are often co-expressed. For instance, genes may be co-expressed when they are functionally related[8] or involved in a similar regulatory system. Weighted gene co-expression network analysis (WGCNA)[9] is a computational technique that identifies groups of genes with correlated expression. These large groups of co-expressed genes are referred to as modules. Study of gene modules may provide new insight into the molecular pathways underlying ischemic stroke. In addition, modules may lead to the identification of a biomarker panel with greater discriminative capacity than single biomarkers.

In this study we implement WGCNA to 1) gain insight into the molecular pathways and genes associated with ischemic stroke, 2) use network analysis to derive multi-gene models and 3) determine the discriminative capacity of a multi-gene score for stroke prognosis.

4.4 METHODS

4.4.1 Patient population

The INTERSTROKE study has been described in detail elsewhere.[10] Briefly, INTERSTROKE was a large, international, standardized case-control study consisting of stroke cases and controls from 22 countries. Stroke cases were patients admitted to hospital with first-time acute stroke that presented within five days of symptom onset and within 72-hours of hospital admission. Distinction between stroke subtypes was confirmed with neuroimaging (CT or MRI). Control participants were recruited from the hospital or within the community, and had no history of stroke. Peripheral whole blood was collected into PAXgene Blood RNA tubes (PreAnalytiX) and stored at -80^{0}C prior to sample processing. Only confirmed ischemic stroke cases and controls were assessed in this study.

4.4.2 Sample processing and microarray hybridization

All sample processing was conducted at the Genetic and Molecular Epidemiology Laboratory of PHRI and McMaster University. Total RNA was isolated using the

QIAsymphony PAXgene Blood RNA kit (Qiagen) on the QIASymphony (Qiagen). RNA quality was determined using Nanodrop2000™ (Nanodrop) and 2100 Bioanalyzer (Agilent), while quantity was determined using Quant-IT RiboGreen® (LifeTech). 500ng of total RNA was amplified and biotinylated using the Illumina TotalPrep RNA Amplification Kit (LifeTech). The final biotin-labeled cRNA species were then hybridized to the Illumina HumanRef-8v4.0 expression BeadChip (Illumina). BeadChips hold 12 samples at a time, so to minimize batch effect samples were randomly assigned to chips for hybridization. BeadChips were then washed, dried and scanned on the iScan System (Illumina) as per manufacturer protocol.

4.4.3 Microarray data pre-processing

The Illumina HumanRef-8v4.0 expression BeadChip interrogated expression of 34,694 unique genes using 47,323 probes. The raw sample probe profile and control probe profile were exported from GenomeStudio version 1.9.0 (Illumina). All analysis was performed in R (http://r-project.org). Data pre-processing was conducted with the LIMMA (Linear models for microarray analysis)[11] package and involved background correction using the non-genomic control probes,[12] quantile normalization and log2 transformation.[13] Probes with detection P-value < 0.01 in >50% of the samples were considered expressed. As a result the final pre-processed expression data set consisted of 11,099 RNA transcript probes for each of the 274 individuals (104 ischemic stroke cases and 170 control participants).

4.4.4 Statistical analysis

Groups of co-expressed genes, referred to as modules, were identified by constructing weighted gene co-expression networks as previously described.[14] Briefly, networks were determined using pairwise Pearson correlation between all probes across all participants. The absolute value of the Pearson correlation was raised to a power of $\beta = 6$ to emphasize large correlations at the expense of low correlations thus resulting in a weighted network. Using the default tree-cutting algorithm, modules were identified.

Principal component analysis was used to summarize the expression variance within each module. The first principle component eigenvalue for each module is referred to as the module summary (MS) value. MS values were assed for correlation with stroke and stroke risk factors using Pearson correlation. A Bonferroni corrected p-value $<0.05/15=0.0033$ was considered significant. We also constructed multivariable regression models to evaluate the association between MS values and stroke while adjusting for risk factors. As external validation, we determined module assignment for genes reported to be associated with stroke by Moore *et al.*,[4] Tang *et al.*,[5] and Barr *et al.*[6]

Signalling Pathway Impact Analysis (SPIA)[15] was conducted on ischemic stroke associated modules. SPIA identifies the number of differentially expressed genes per pathway, in comparison to the probability of identifying more pathway-associated genes by chance. A Bonferroni corrected global p-value $<3.5 \times 10^{-2}$ was indicative of significant pathway activation or inhibition.

Next we identified highly connected genes within each module, referred to as hub genes. Hub genes were defined as genes with high gene significance (GS), correlation

between gene expression and case-control status, and high module membership (MM), correlation between gene expression and respective MS value. Genes with GS>0.3 and MM>0.75 were considered hub genes. To confirm that the top hub genes were associated with ischemic stroke, we regressed on stroke status while adjusting for available risk factors. Correlations in hub gene expression were evaluated using Pearson correlation. To then identify hub genes independently associated with ischemic stroke we constructed a multivariable logistic regression model.

Finally, regression models were constructed to evaluate the association between hub gene expression and functional disability, measured as modified Rankin Scale score (mRS). Our analysis utilized mRS recorded soon after the stroke (at baseline) and at the one-month follow-up. One-month mRS was dichotomized to represent functional disability (mRS 0-2 vs mRS>2). Using pROC,[16] receiver operator curves (ROC) were constructed from logistic regression models associating one-month disability and hub gene expression. Area under the ROC (AUC) was determined as a measure of sensitivity and specificity. The strength of multiple models were compared using the Net Reclassification Index (NRI).[17,18] NRI was calculated using the Hmisc[19] package. An NRI >0.6 was considered a strong improvement in discriminative capacity, 0.4 was intermediate, and 0.2 was considered weak.

4.5 RESULTS

4.5.1 Patient characteristics

Demographic and clinical features of the 104 ischemic stroke cases and 170 controls are summarized in *Table 1*. The average age for stroke cases was 68.9 ± 12.3 and 43% were female. 26.9% of ischemic cases were classified as cardioembolic, 9.6% large vessel, 19.2% small vessel, 29.8% cryptogenic and 14.4% other. Clinical features of ischemic stroke cases were similar to controls, except for presence of smoking (p=0.01), which was more common among stroke cases.

Table 1. Participant demographics.

	Ischemic Stroke N = 104	Control N = 170	P-Value
Gender (% female)	45 (43.3)	78 (45.9)	0.67
Age, mean ± SD	68.9 ± 12.3	66.1 ± 14.2	0.08
BMI, mean ± SD	26.6 ± 4.8	27.16 ± 4.8	0.37
Hyperlipidemia	44 (42.3)	62 (36.5)	0.34
Hypertension	73 (70.2)	103 (60.6)	0.10
Diabetes	21 (20.2)	21 (12.4)	0.10
Atrial Fibrillation	9 (8.7)	6 (3.5)	0.10
Migraine	21 (20.2)	21 (12.4)	0.10
Current smoker	21 (20.2)	15 (8.8)	0.01*
Race, (%)			
European	42 (40.4)	53 (31.2)	
Latin American	62 (59.6)	117 (68.8)	
Stroke Type, (%)			
Cardioembolic	28 (26.9)	NA	
Large vessel	10 (9.6)	NA	
Small vessel	20 (19.2)	NA	
Cryptogenic	31 (29.8)	NA	
Other	15 (14.4)	NA	

4.5.2 Gene co-expression network construction and module identification

Microarray expression profiling was conducted on all samples and groups of co-expressed genes were identified. These groups of genes are referred to as modules and in total 15 were identified (*Figure 1*). Module size ranged from 86 genes to 2,758 genes. Five of the 15 modules were correlated with ischemic stroke after adjustment for multiple hypothesis testing ($p<0.05/15=0.0033$, *Supplementary Figure 2, Supplementary Table 1*). Modules were named based on the significance in association with stroke. Thus module 1 was most significantly correlated with stroke ($p=2.4x10^{-13}$), then module 2 ($p=1.6x10^{-8}$), module 3 ($p=6.1x10^{-6}$), module 4 ($p=1.1x10^{-5}$) and module 5 ($p= 8.5x10^{-4}$).

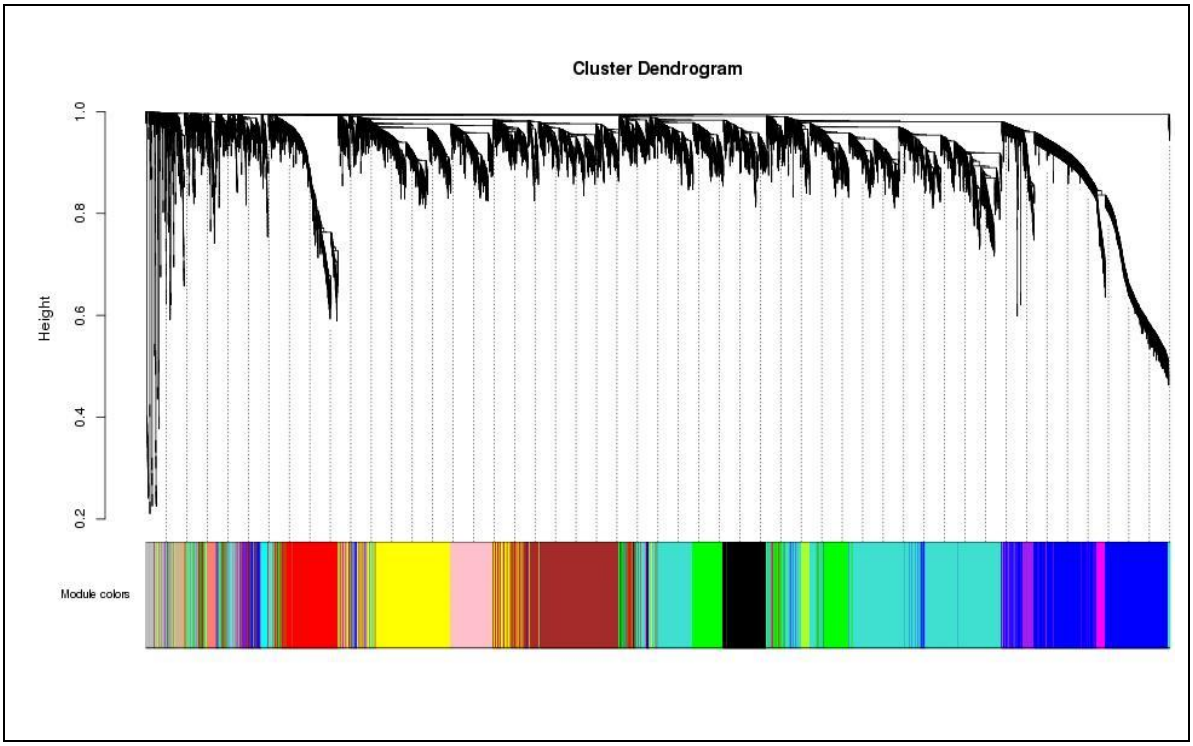
We also tested each module for correlation with stroke risk factors including: age, gender, BMI, ethnicity, hyperlipidemia, diabetes, atrial fibrillation, hypertension, migraine and smoking status. A modest association between module 5 and smoking was identified ($p=2.6\ x10^{-3}$). Each module was further tested for association with ischemic stroke while adjusting for available risk factors. Under these models, only modules 1, 2, 3 and 4 were significantly associated with stroke ($p<0.05/15=0.0033$, *Supplementary Table 2*).

As external validation, we determined module assignment for the 18 genes previously associated with stroke by Tang *et al.*[5], 9 genes by Barr *et al*.[6] and 22 genes by Moore *et al*.[4] The majority of genes identified by each group corresponded to our stroke associated modules 1 or 2. Specifically, 37-71% of genes in each list were associated with module 1 and, 5-14% of genes were associated with module 2. Module membership for

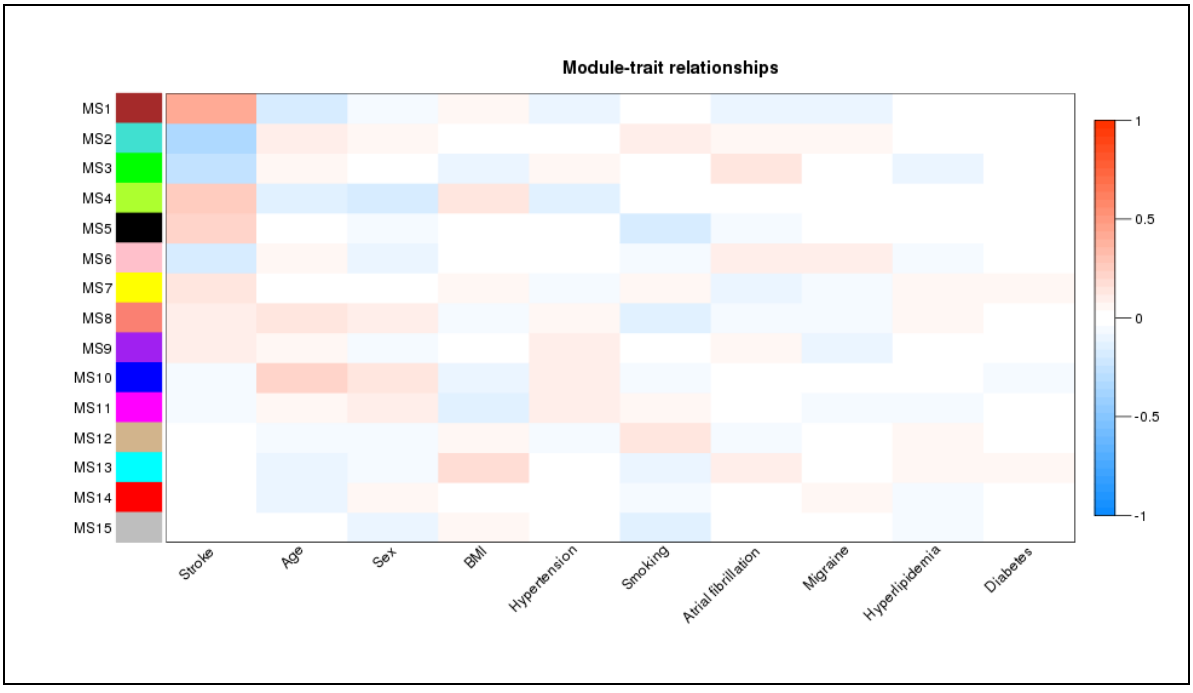
the remaining genes was inconsistent between the gene lists. The full results are presented in *Supplementary Table 3*.

Figure 1. Dendogram of gene expression and identification of modules.

The dendogram was constructed by average-link hierarchical clustering. Initially each gene was considered a unique cluster. Clusters were progressively combined based on shortest distance (ie greatest similarity); the distance between two clusters was equal to the average distance from any member in one cluster to any member of the other cluster. Modules were identified by dividing the dendogram at significant branch points. 15 modules were identified. Genes within each module were color-coded to improve visualization.



Supplementary Figure 1. Heat maps of module summary values with clinical traits. Modules summary (MS) values summarize the variation in gene expression between all genes within each module for each individual (N=274); MS values are equivalent to the eigenvector of the first principal component. Each MS value was tested for correlation with stroke and stroke risk factors using all samples. The heat map colors indicate strength and direction of correlation. Dark red indicates a strong positive correlation, white no correlation, and dark blue a strong negative correlation. MS values were named based on the significance of their association with stroke.



Supplementary Table 1. P-values from the correlation between module summary values and clinical traits.

An adjusted p-value <0.0033 was considered significant.

	Stroke	Age	Sex	BMI	Hypertension	Smoking	Atrial fibrillation	Migraine	Hyperlipidemia	Diabetes
MS1	$2.4x10^{-13}$	$5.8 x10^{-3}$	0.26	0.37	0.10	0.85	0.17	0.09	0.99	0.94
MS2	$1.6 x10^{-8}$	0.15	0.41	0.82	0.77	0.10	0.21	0.43	0.95	0.70
MS3	$6.1 x10^{-6}$	0.24	0.76	0.16	0.22	0.76	$9.1 x10^{-3}$	0.80	0.17	0.86
MS4	$1.1 x10^{-5}$	$3.4 x10^{-2}$	$7.8 x10^{-3}$	$4.6 x10^{-2}$	1.1E-02	0.93	0.56	0.80	0.83	0.76
MS5	$8.5 x10^{-4}$	0.74	0.50	0.88	0.83	$2.6 x10^{-3}$	0.46	0.61	0.83	0.74
MS6	$7.1 x10^{-3}$	0.36	0.16	0.69	0.66	0.45	0.15	0.08	0.37	0.81
MS7	$2.2 x10^{-2}$	0.57	0.90	0.21	0.43	0.49	0.15	0.46	0.28	0.48
MS8	0.10	$3.2 x10^{-2}$	0.06	0.41	0.50	$2.6 x10^{-2}$	0.43	0.49	0.25	0.93
MS9	0.10	0.28	0.20	0.78	0.13	0.79	0.25	0.17	0.96	0.90
MS10	0.44	$4.8 x10^{-4}$	$4.0 x10^{-2}$	0.13	0.08	0.22	0.69	0.61	0.78	0.45
MS11	0.46	0.37	0.13	$3.3 x10^{-2}$	0.10	0.31	0.55	0.41	0.48	0.98
MS12	0.65	0.35	0.51	0.30	0.21	$2.7 x10^{-2}$	0.43	0.89	0.42	0.59
MS13	0.70	0.10	0.24	$6.9 x10^{-3}$	0.68	0.12	0.15	0.61	0.35	0.31
MS14	0.72	0.08	0.39	0.57	0.93	0.39	0.63	0.27	0.37	0.91
MS15	0.96	0.84	0.15	0.49	0.89	$1.2 x10^{-2}$	0.76	0.68	0.36	0.95

Supplementary Table 2. Association between module summary value and stroke after adjustment for clinical risk factors.

An adjusted p-value <0.0033 was considered significant.

	Beta	P-value
MS1	0.050	1.70×10^{-11}
MS2	-0.038	3.65×10^{-7}
MS3	-0.031	3.28×10^{-5}
MS4	0.032	2.81×10^{-5}
MS5	0.022	3.61×10^{-3}
MS6	-0.018	1.48×10^{-2}
MS7	0.018	2.49×10^{-2}
MS8	0.011	0.16
MS9	0.017	0.04
MS10	-0.005	0.51
MS11	-0.005	0.52
MS12	-0.002	0.78
MS13	0.001	0.88
MS14	-0.005	0.53
MS15	-0.003	0.72

Supplementary Table 3. Module membership of genes previously associated with stroke.

Tang et al.,[6] identified 18 genes associated with stroke, Barr *et al*.,[7] identified 9 genes and Moore *et al*.,[5] identified 22 genes. Respectively, 16, 7 and 19 genes were represented in our gene expression data. Module membership for the overlapping genes was determined.

Module	Tang *et al.*[6]	Barr *et al.*[7]	Moore *et al.*[5]
1	*ARG1, BCL6, PYGL, RNASE2, S100A12, F5, S100P*	*ARG1, CCR7, IQGAP1, ORM1, S100A12*	*ADM, FCGR1A, ENTPD1, CD163, TLR2, IL13RA1, PTEN*
2	*LY96*	*LY96*	*CD36*
3	*MMP9*	NA	*PILRA*
4	*C44, CKAP4*	*CA4*	NA
5	NA	NA	*VCAN, KIAA0146*
6	*NPL, ETS2, FPR1*	NA	*BST1, FOS, NPL, ETS2*
7	NA	NA	NA
8	NA	NA	NA
9	NA	NA	NA
10	NA	NA	NA
11	NA	NA	NA
12	NA	NA	*APLP2, CD14, LTA4H, CYBA*
13	NA	NA	NA
14	*S100A9*	NA	NA
15	NA	NA	NA

4.5.3 Pathway analysis of stroke associated modules

Pathway analysis of genes in module 1 indicated increased activation of cytokine-cytokine receptor interaction pathways in ischemic stroke as compared with controls ($p=4.3x10^{-9}$). Additional pathways with significant activation or inhibition are listed in *Supplementary Table 4*. In module 2 we detected increased activation of chemokine signalling pathways ($p=9.0x10^{-6}$) and inhibition of actin cytoskeletal regulation ($p=8.5x10^{-5}$). Significant changes were not observed in module 3. In module 4, increased activation of RNA transport pathways was detected ($p=2.65x10^{-4}$).

Supplementary Table 4. Significant pathways identified in module 1.

Molecular or disease pathway	P-Value	Bonferroni P-value	Status
Cytokine-cytokine receptor interaction	4.3×10^{-9}	5.8×10^{-7}	Activated
Asthma	6.9×10^{-8}	9.3×10^{-6}	Activated
Salmonella infection	1.1×10^{-6}	1.5×10^{-4}	Inhibited
Allograft rejection	7.2×10^{-6}	9.7×10^{-4}	Activated
Leishmaniasis	1.0×10^{-5}	1.4×10^{-3}	Inhibited
Rheumatoid arthritis	1.1×10^{-5}	1.5×10^{-3}	Inhibited
Legionellosis	1.1×10^{-5}	1.5×10^{-3}	Inhibited
Autoimmune thyroid disease	3.4×10^{-5}	4.7×10^{-3}	Activated
Focal adhesion	5.4×10^{-5}	7.4×10^{-3}	Activated
Intestinal immune network for IgA production	7.8×10^{-5}	1.0×10^{-2}	Inhibited
Systemic lupus erythematosus	1.2×10^{-4}	1.6×10^{-2}	Activated
Toxoplasmosis	1.6×10^{-4}	2.1×10^{-2}	Inhibited
Type I diabetes mellitus	1.7×10^{-4}	2.3×10^{-2}	Inhibited
Graft-versus-host disease	2.6×10^{-4}	3.5×10^{-2}	Inhibited

4.5.4 Identification of hub genes within stroke associated modules

Each module consists of many genes, so we sought to pinpoint genes with the greatest impact on each network. These central genes with high connectivity are referred to as hub genes. 207 hub genes were identified in module 1, the two most significant being *NLRC4*, and *MCEMP1*. Within modules 2, 3 and 4 we respectively identified 49, 32 and 31 hub genes. The top hub gene in module 2 was *CKLF*, module 3 was *ITGAM* and module 4 was *HS.546375*. *Table 2* describes the ten most significant hub genes for each stroke-associated module.

Expression of *NLRC4, CKLF, HS.546375* and *ITGAM* were associated with stroke after adjustment for available risk factors. A 1.6 fold increase in *NLRC4* was identified in ischemic stroke cases as compared with controls (CI 1.4-1.8, $p=3.0x10^{-15}$*, Supplementary Figure 2A*). In addition, elevated expression of *CKLF* (FC=1.3, CI 1.2-1.4, $p=1.6x10^{-10}$) and *ITGAM* (FC=1.3, CI 1.2-1.4, $p=1.0x10^{-8}$) were observed in stroke cases, while *HS.546375* expression was decreased (FC= 0.6, CI 0.6-0.7, $p=2.8x10^{-13}$, *Supplementary Figure 2D*). Correlation was low between expression of *NLRC4, CKLF, HS.546375* and *ITGAM* ($r^2<0.66$, *Supplementary Table 5*). To identify genes independently associated with stroke, we constructed an association model including the four genes and available stroke risk factors. The model indicated that only *NLRC4* (p=0.0025), *CKLF* (p=0.018) and *HS.546375* ($p=3.6x10^{-4}$) were independently associated with stroke status.

Table 2. Top 10 stroke associated hub genes identified within modules 1 to 4.

Genes with high connectivity tend to have high module membership (MM) and gene significance (GS). MM is the correlation between gene expression and module summary (MS) values. GS is the correlation between gene expression and stroke or control status of a sample. Hub genes had MM>0.75 and GS>0.3.

Module 1

Gene	GS	MM
NLRC4	0.49	0.79
MCEMP1	0.49	0.81
IRAK3	0.47	0.83
SPOCK2	0.46	0.84
PPP4R1	0.44	0.80
ANXA3	0.44	0.87
BCL6	0.44	0.78
EXOC6	0.44	0.85
CLEC4D	0.44	0.79
FLJ20273	0.44	0.81

Module 2

Gene	GS	MM
CKLF	0.40	0.78
CCPG1	0.40	0.80
FAM160B1	0.39	0.84
LY96	0.39	0.77
E2F3	0.38	0.85
TXN	0.38	0.77
RP5-1022P6.2	0.38	0.88
CKLF	0.37	0.80
KIAA1600	0.37	0.91
TXN	0.36	0.78

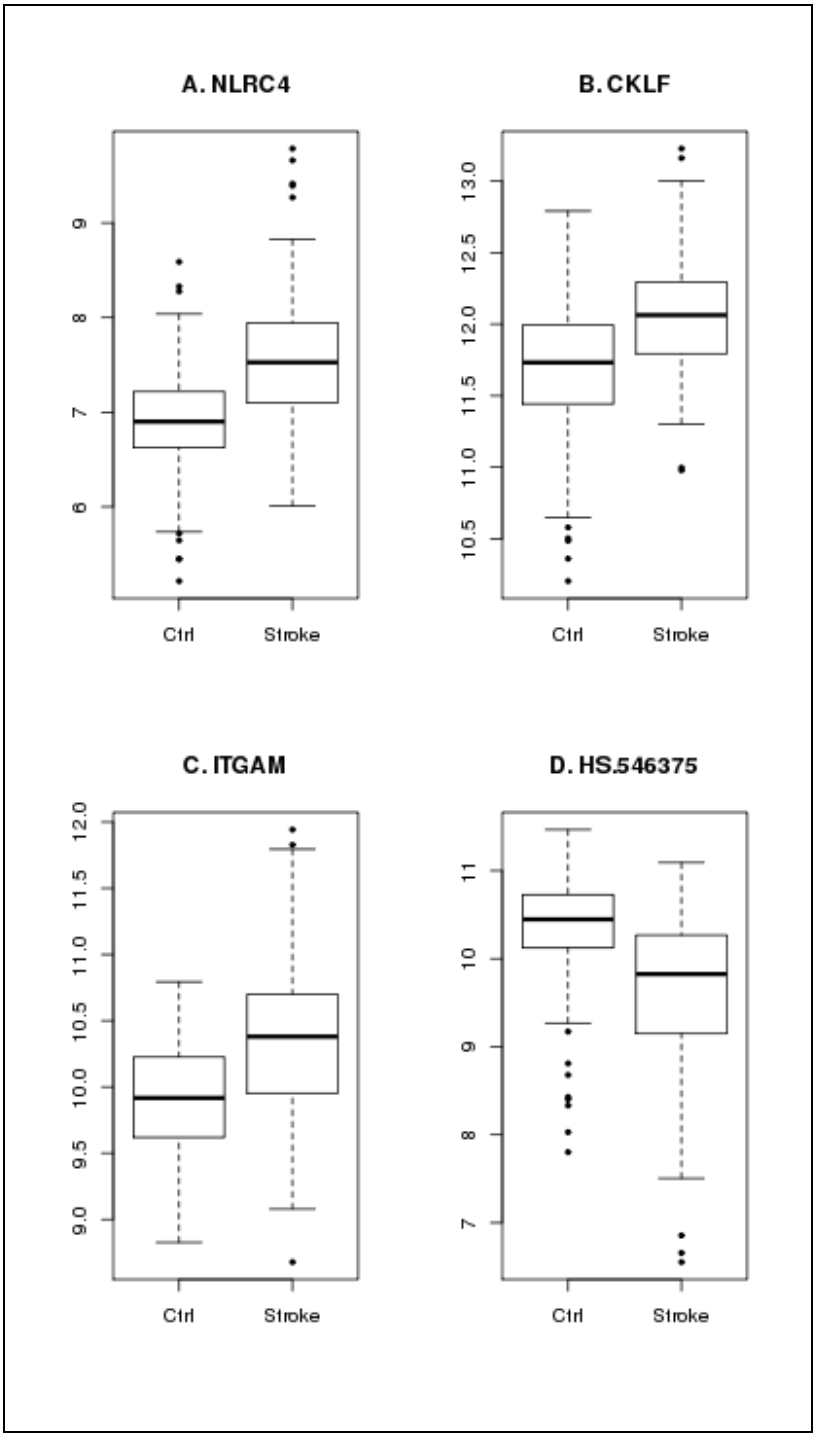
Module 3

Gene	GS	MM
ITGAM	0.38	0.78
KIAA1881	0.37	0.81
TMEM88	0.37	0.81
REM2	0.37	0.75
DOCK5	0.36	0.80
LOC100134734	0.35	0.80
MMP25	0.35	0.83
ALPL	0.35	0.77
SLC9A8	0.34	0.81
MANSC1	0.34	0.78

Module 4

Gene	GS	MM
HS.546375	0.42	0.81
LOC646294	0.41	0.80
BCL11B	0.40	0.79
ATP5G2	0.40	0.76
HS.534439	0.39	0.75
IMP3	0.37	0.86
RPS2	0.37	0.82
EIF3F	0.36	0.85
ST6GAL1	0.36	0.83
LOC347544	0.36	0.78

Supplementary Figure 2. Boxplots of gene expression for the top hub genes for stroke associated modules.



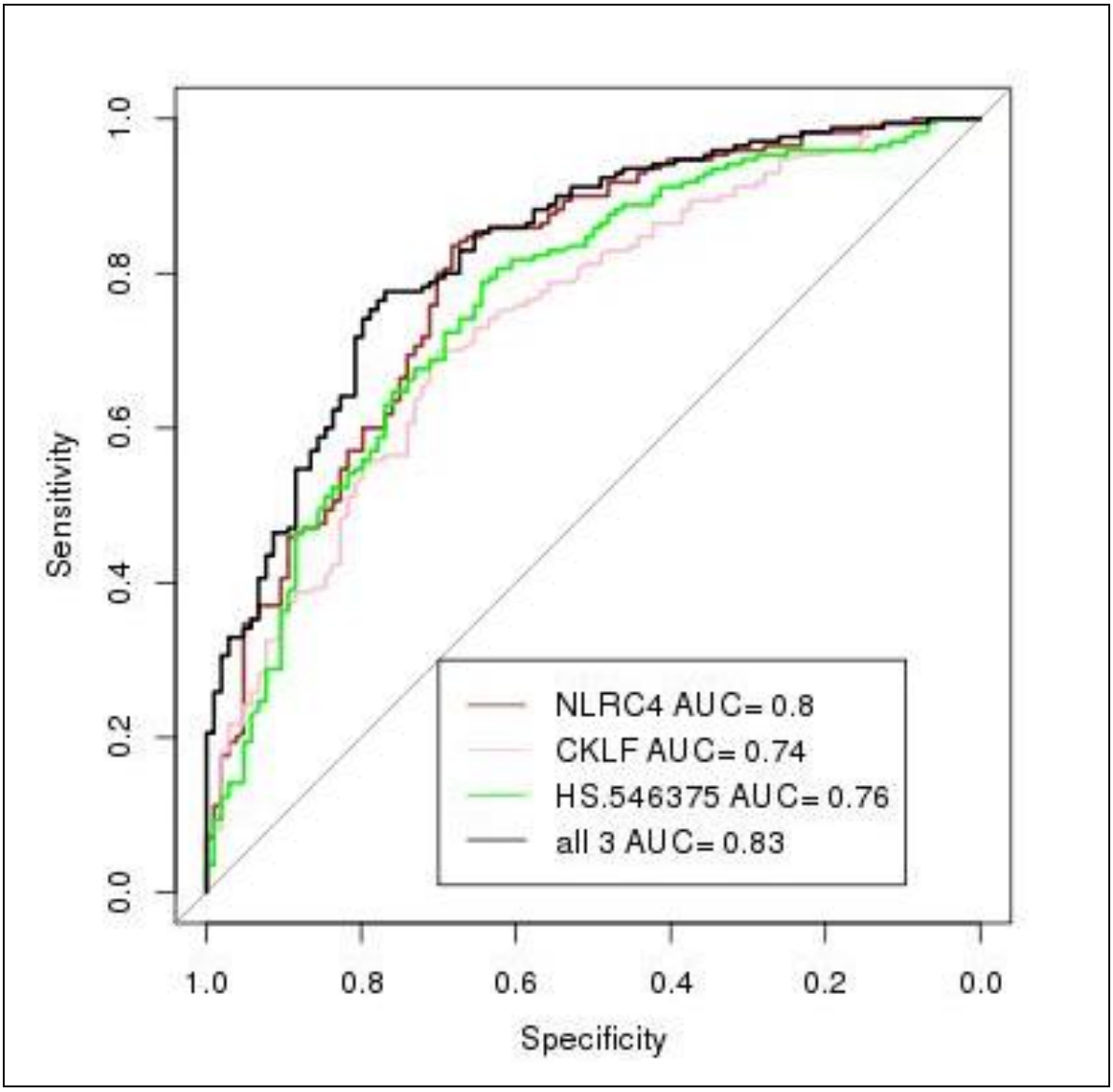
Supplementary Table 5. Absolute pair-wise Pearson correlation between the top hub genes from stroke associated modules.

	Module 1	**Module 2**	**Module 3**	**Module 4**
	NLRC4	***CKLF***	***ITGAM***	***HS.546375***
NLRC4	1	0.55	0.64	0.66
CKLF	0.55	1	0.46	0.41
ITGAM	0.64	0.46	1	0.53
HS.546375	0.66	0.41	0.53	1

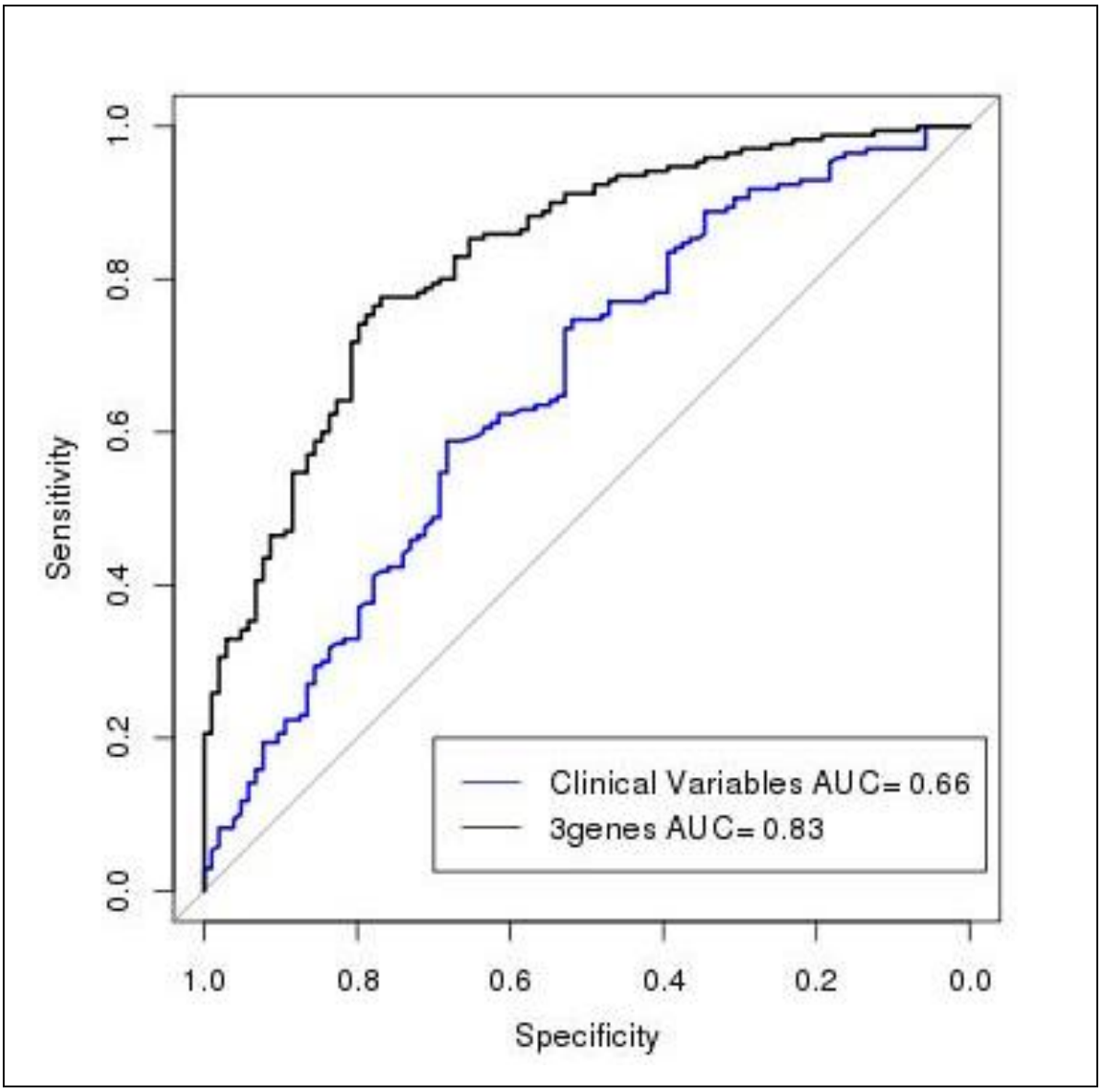
4.5.5 Multi-gene model improves discrimination of ischemic stroke

To assess the discriminative capacity of *NLRC4, CKLF* and *HS.546375* for stroke we constructed ROC curves. Based on univariate models, *NLRC4* (AUC 0.80, CI 0.74-0.85) had greater discriminative capacity as compared with either *CKLF* (AUC=0.74, CI 0.68-0.80) or *HS.546375* (AUC=0.76, CI 0.70-0.82). However, a model combining expression of *NLRC4, CKLF* and *HS.546375,* had the best discriminative capacity (AUC=0.83, CI 0.78-0.88, *Supplementary Figure 3*). The three-gene model strongly improved discrimination as compared with univariate models (*NLRC4* NRI=0.54, $p=7.8x10^{-6}$; *CKLF* NRI=0.81, $p=8.9x10^{-13}$; and *HS.546375* NRI=0.73, $p=2.3x10^{-10}$). Furthermore, the three-gene model strongly improved ischemic stroke discrimination as compared with a model including only clinical variables (AUC=0.66, CI 0.59-0.72, NRI=0.83, $p=3.1x10^{-13}$, *Supplementary Figure 4*).

Supplementary Figure 3. Discriminative capacity of hub genes for ischemic stroke.



Supplementary Figure 4. Discriminative capacity of hub gene panel and clinical variables for ischemic stroke.

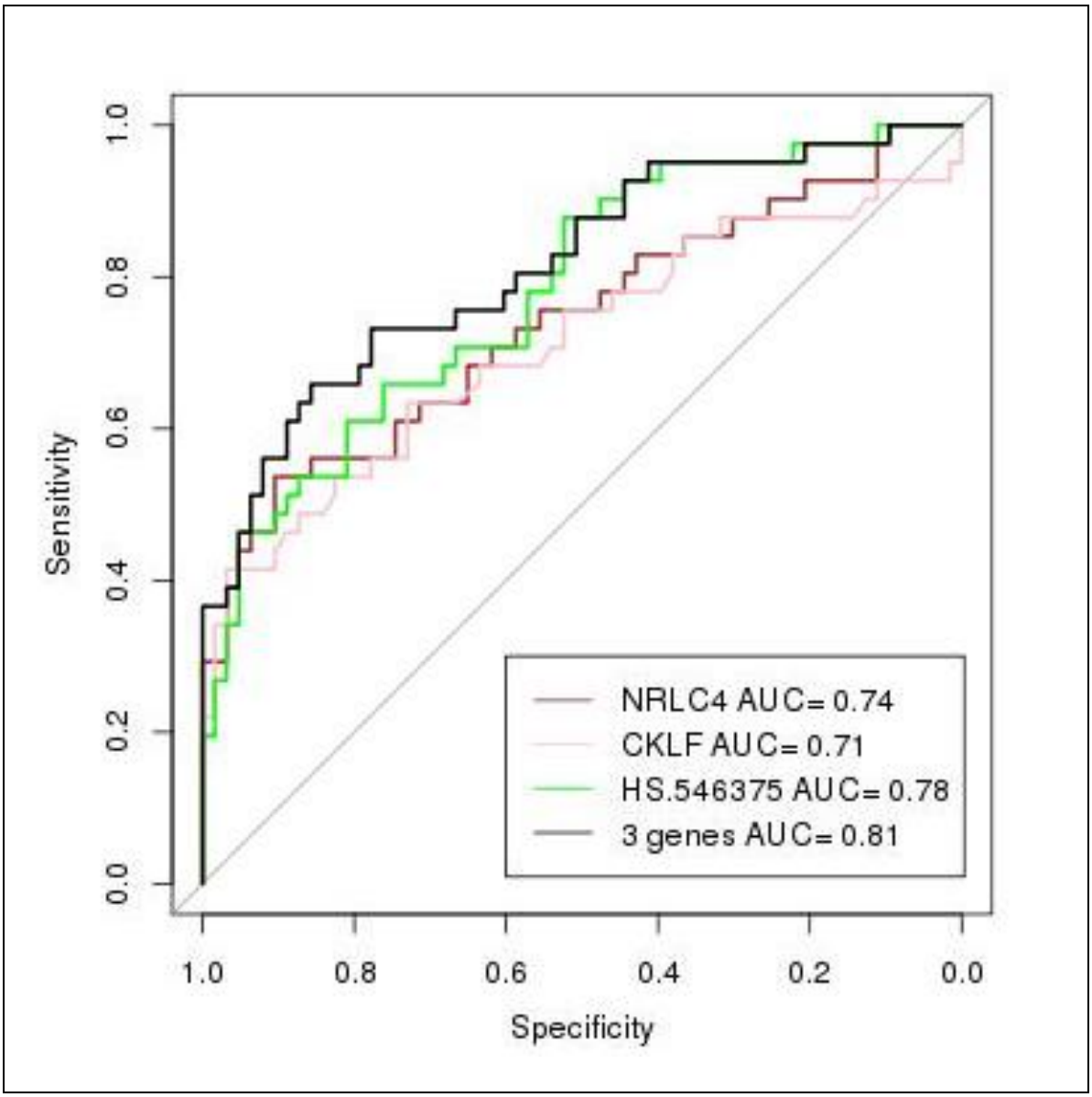


4.5.5 Multi-gene model improves discrimination of one-month disability

Modified Rankin Scale score (mRS) was recorded for each stroke patient soon after stroke, at baseline, and re-evaluated one-month after the stroke. We dichotomized one-month mRS into two groups, mRS of 0, 1 or 2 and mRS >2, which represented post-stroke disability. We observed that a model including expression of the *NLRC4, CKLF* and *HS.546375* (AUC=0.81, CI 0.73-0.90) improved disability discrimination as compared with single gene models (*NLRC4* NRI=0.62, p=0.0013; *CKLF* NRI=0.79, $p=1.7x10^{-5}$; *HS.546375* NRI=0.45, p=0.019; *Supplementary Figure 5*).

Our previous study observed that whole blood expression of *MCEMP1*, baseline mRS and primary stroke type had prognostic capacity for stroke.[7] Thus we sought to evaluate the discriminative capacity of multi-gene models as compared with a single gene model. Since all stroke cases in this study were due to ischemia the stroke type variable was disregarded, but baseline mRS was included in all models. We observed that a disability discrimination model including three genes, *NLRC4, CKLF, HS.546375* (AUC=0.93, CI 0.88-0.98) had equivalent performance to the *MCEMP1* model (AUC=0.93, CI 0.88-0.98). However a model consisting the three genes identified through network analysis and *MCEMP1* (AUC=0.94, CI 0.90-0.98) moderately improved disability discrimination as compared to the three-gene model (NRI=0.48, p=0.013) or *MCEMP1* (NRI=0.52, p=0.0073). Models for mortality discrimination could not be developed due to the limited number of fatal events one-month after ischemic stroke.

Supplementary Figure 5. Discriminative capacity of hub genes for one-month disability.



4.6 DISCUSSION

In this study we identified four gene modules associated with ischemic stroke. We further characterized the modules by identifying enriched pathways and hub genes. *NLRC4, CKLF* and *HS.546375* were the top hub genes identified in three different modules and each was independently associated with ischemic stroke. Stroke discrimination was improved using a model that included all three genes as compared with single gene models or a model with clinical variables. Similarly, the three-gene model improved discrimination of one-month disability. One-month disability discrimination was similar between the previously reported *MCEMP1* model and the three-gene model. But, disability discrimination was moderately improved by the model consisting of *NLRC4, CKLF, HS.546375, MCEMP1* and baseline mRS.

Network analysis can provide new insights into the genes and molecular pathways underlying ischemic stroke. In our previous work[7] we employed univariate analysis which implicated only one gene, *MCEMP1*, with stroke whereas network analysis identified many hub genes with possible biomarker capabilities. For instance, one of the top hub genes identified was *NLRC4*. Nucleotide-binding and oligomerization domain–like receptor (NLR) proteins are involved in immune-surveillance. Upon sensing pathogenic bacteria, it has been reported that NLRC4 oligomerizes and leads to the formation of the NLRC4 inflammasome.[20] Inflammasomes are multi-protein complexes that play a critical role in mediating innate immune responses. The inflammatory response in cerebral tissue following stroke contributes to the progression of brain injury and exacerbation of neurological deficits. Previous studies have implicated NLRP1 and NLRP3 with

stroke,[21,22] but a recent study suggests that NLRC4 may be the driver of sterile inflammatory responses in the brain.[23] Denes *et al.*,[23] observed that *NLRC4*$^{-/-}$ mice that underwent experimental stroke had reduced ischemic brain injury and improved neurological outcomes as compared with wild type or *NLRP3*$^{-/-}$ mice. In concordance, we have identified elevated *NLRC4* expression in ischemic stroke patients as compared with controls. Together these studies are the first studies to implicate *NLRC4* with stroke.

Since network analysis implicates many genes it is possible to conduct pathway analysis. It is well recognized that inflammatory mechanisms are triggered by stroke.[24,25] In agreement, we detected activation of cytokine-cytokine receptor interactions and chemokine signalling in ischemic stroke. We also observed activation of RNA transport pathways. Although the role of RNA transport is rather non-specific it could reflect activation of leukocytes and other inflammatory cells. Pathway analysis of stroke-associated modules did not provide novel insights, but they were able to confirm known aspects of stroke biology. Indeed, the ability to identify new pathways is dependent on annotation of protein functions and interactions, which is expected to improve as our understanding of biological pathways increases.

Our results also demonstrate that multi-gene panels can improve discrimination as compared with single genes. For instance, a stroke discrimination model including three genes that were identified through network analysis had greater discriminative capacity as compared with single gene models or clinical variables. Thus network analysis can improve the derivation of discriminative gene scores. Our three-gene model included *NLRC4, CKLF* and *HS.54637,* which represented three different modules. NLRC4 is

involved in inflammasome formation; the chemokine-like receptor (CKLF) protein is a potent chemoattractant for neutrophils, monocytes and lymphocytes[26] and *HS.54637* encodes a protein of unknown function. Together the three genes may describe three distinct and significant mechanisms underlying ischemic stroke, which facilitates improved discriminative capacity.

In our previous work we demonstrate that *MCEMP1* expression may have utility as a non-invasive biomarker for stroke prognosis. The study focused on the discriminative capacity of a single gene since multiple genes independently associated with stroke could not be identified. In the current study we observed that the discriminative performance of *MCEMP1* was similar to the three-gene model identified through network analysis. But, we determined that a one-month disability score comprised of *NLRC4, CKLF*, *HS.54637*, *MCEMP1*, and baseline mRS, had improved performance as compared with other model, albeit a moderate improvement. Although there is greater complexity and cost associated with multi gene scores, the increased test sensitivity may have clinical benefits. Future studies including stroke mimics and neurologic inflammatory conditions are required to determine the specificity of a multi gene models for stroke as compared with single genes. With additional research we can determine the clinical utility of multi gene panels.

The results of the present study demonstrate that network analysis can implicate new genes and molecular pathways with stroke. To our knowledge, ours is the first human study to implicate *NLRC4* with ischemic stroke. Network analysis also improves detection of multi-gene scores as compared to univariate gene expression analysis. We demonstrate that multi-gene scores for ischemic stroke diagnosis and one-month

prognosis have greater discriminative capacity as compared with single-gene models. The multi-gene scores may describe more of the expression changes associated with stroke, which increases complexity but also improves discriminative capacity. Additional research will be required to determine the cost-benefit ratio of multi-gene models. However, our results suggest that network analysis and identification of modules may be a superior method for RNA biomarker discovery. Further analysis of genes within each module may improve our understanding of the molecular mechanisms of stroke.

4.7 REFERENCES

1. Go AS, Mozaffarian D, Roger VL, Benjamin EJ, Berry JD, Borden WB, et al. Heart disease and stroke statistics--2013 update: a report from the American Heart Association. *Circulation*. 2013;127:e6–e245.
2. Donnan GA, Fisher M, Macleod M, Davis SM. Stroke. *Lancet (London, England)*. 2008;371:1612–23.
3. O'Donnell MJ, Xavier D, Liu L, Zhang H, Chin SL, Rao-Melacini P, et al. Risk factors for ischaemic and intracerebral haemorrhagic stroke in 22 countries (the INTERSTROKE study): a case-control study. *Lancet*. 2010;376:112–23.
4. Moore DF, Li H, Jeffries N, Wright V, Cooper R a, Elkahloun A, et al. Using peripheral blood mononuclear cells to determine a gene expression profile of acute ischemic stroke: a pilot investigation. *Circulation*. 2005;111:212–21.
5. Tang Y, Xu H, Du X, Lit L, Walker W, Lu A, et al. Gene expression in blood changes rapidly in neutrophils and monocytes after ischemic stroke in humans: a microarray study. *J. Cereb. Blood Flow Metab.* 2006;26:1089–102.
6. Barr TL, Conley Y, Ding J, Dillman A, Warach S, Singleton A, et al. Genomic biomarkers and cellular pathways of ischemic stroke by RNA gene expression profiling. *Neurology*. 2010;75:1009–14.
7. Raman K, O'Donnell MJ, Czlonkowska A, Duarte YC, Lopez-Jaramillo P, Peñaherrera E, et al. Peripheral Blood *MCEMP1* Gene Expression as a Biomarker for Stroke Prognosis. *Stroke*. 2016;STROKEAHA.115.011854.
8. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U. S. A.* 1998;95:14863–8.
9. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*. 2008;9:559.
10. O'Donnell M, Xavier D, Diener C, Sacco R, Lisheng L, Zhang H, et al. Rationale and design of INTERSTROKE: a global case-control study of risk factors for stroke. *Neuroepidemiology*. 2010;35:36–44.
11. Smyth GK. Limma : Linear Models for Microarray Data. *Springer*. 2005;397–420.
12. Plancade S, Rozenholc Y, Lund E. Generalization of the normal-exponential model: exploration of a more accurate parametrisation for the signal distribution on Illumina BeadArrays. *BMC Bioinformatics*. 2012;13:329.

13. Schmid R, Baum P, Ittrich C, Fundel-Clemens K, Huber W, Brors B, et al. Comparison of normalization methods for Illumina BeadChip HumanHT-12 v3. *BMC Genomics*. 2010;11:349.

14. Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* 2005;4:Article17.

15. Tarca AL, Draghici S, Khatri P, Hassan SS, Mittal P, Kim J-S, et al. A novel signaling pathway impact analysis. *Bioinformatics*. 2009;25:75–82.

16. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*. 2011;12:77.

17. Pencina MJ, D'Agostino RB, Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat. Med.* 2008;27:157–172.

18. Pencina MJ, D'Agostino RB, Steyerberg EW. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. *Stat. Med.* 2011;30:11–21.

19. Harrel FE. Hmisc: A package of miscellaneous R functions [Internet]. 2015 [cited 2015 Feb 1];Available from: https://cran.r-project.org/web/packages/Hmisc/index.html

20. von Moltke J, Ayres JS, Kofoed EM, Chavarría-Smith J, Vance RE. Recognition of bacteria by inflammasomes. 2013.

21. Abulafia DP, de Rivero Vaccari JP, Lozano JD, Lotocki G, Keane RW, Dietrich WD. Inhibition of the inflammasome complex reduces the inflammatory response after thromboembolic stroke in mice. *J. Cereb. Blood Flow Metab.* 2009;29:534–544.

22. Savage CD, Lopez-Castejon G, Denes A, Brough D. NLRP3-inflammasome activating DAMPs stimulate an inflammatory response in glia in the absence of priming which contributes to brain inflammation after injury. *Front. Immunol.* 2012;3:1–11.

23. Denes A, Coutts G, Lénárt N, Cruickshank SM, Pelegrin P, Skinner J, et al. AIM2 and NLRC4 inflammasomes contribute with ASC to acute brain injury independently of NLRP3. *Proc. Natl. Acad. Sci.* 2015;112:201419090.

24. Chamorro Á, Meisel A, Planas AM, Urra X, van de Beek D, Veltkamp R. The immunology of acute stroke. Nat. Rev. Neurol. 2012;8:401–410.

25. Iadecola C, Anrather J. The immunology of stroke: from mechanisms to

translation. *Nat. Med.* 2011;17:796–808.

26. Han W, Lou Y, Tang J, Zhang Y, Chen Y, Li Y, et al. Molecular cloning and characterization of chemokine-like factor 1 (CKLF1), a novel human cytokine with unique structure and potential chemotactic activity. *Biochem. J.* 2001;357:127–135.

CHAPTER 5: Whole blood gene expression differentiates atrial fibrillation from sinus rhythm in patients with persistent atrial fibrillation

Kripa Raman, BSc,[1,2,3] Stefanie Aeschbacher, MSc,[4,5] Matthias Bossard, MD,[1, 5,6,7] Thomas Hochgruber, MD,[4,5] Andreas J. Zimmermann, MD,[4,5] Beat A. Kaufmann, MD,[6] Katrin Pumpol, MD,[5] Peter Rickenbacker, MD,[6,8] Guillaume Paré, MD, MSc[1,2,9*], David Conen, MD, MPH[4,5*]

[5] Population Health Research Institute, David Braley Cardiac, Vascular and Stroke Research Institute, 237 Barton Street East, Hamilton, ON L8L 2X2, Canada

[6] Thrombosis and Atherosclerosis Research Institute, David Braley Cardiac, Vascular and Stroke Research Institute, 237 Barton Street East, Hamilton, ON L8L 2X2, Canada

[7] Department of Medical Sciences, McMaster University, 1280 Main Street West, Hamilton ON L8S 4K1, Canada

[8] Division of Internal Medicine, Department of Medicine, University Hospital Basel, Petersgraben 4, 4031 Basel, Switzerland

[9] Cardiovascular Research Institute Basel, University Hospital Basel, Spitalstrasse 2, 4031 Basel, Switzerland

[10] Cardiology Division, Department of Medicine, University Hospital Basel, Petersgraben 4, 4031 Basel, Switzerland

[11] Division of Cardiology, Hamilton General Hospital, Hamilton Health Sciences, 237 Barton Street East, Hamilton, ON L8L 2X2, Canada

[12] Cardiology Division, Kantonsspital Bruderholz, 4101 Bruderholz, Switzerland

[13] Department of Pathology and Molecular Medicine, McMaster University, Michael G. DeGroote School of Medicine, 1280 Main Street West, Hamilton ON L8S 4K1, Canada

* G. Paré and D. Conen contributed equally to the present manuscript

5.1 FORWARD

Atrial fibrillation (AF) is the most common cardiac arrhythmia and increases stroke risk by 5-fold. Current treatment strategies to restore sinus rhythm have limited long-term success. In addition our understanding of the molecular biology underlying AF and conversion to sinus rhythm is incomplete. Recently the association between elevated levels of NT-proBNP and AF are being evaluated but additional validation is required. This study compared whole blood gene expression and circulating biomarkers in patients with AF as compared with paired sinus rhythm samples, collected after successful cardioversion. We detected elevated expression of *SLC25A20* and *PDK4* during AF. We also demonstrate that RNA biomarkers independently improve AF discrimination as compared with circulating NT-proBNP.

This manuscript has been submitted for publication in PlosOne and is currently under review. Guillaume Paré and David Conen conceptualized and designed the study. Kripa Raman created the analysis plan, conducted all analysis, qPCR-related laboratory work and wrote the manuscript. The following individuals aided with data collection and provided feedback on the manuscript: Stefanie Aeschbacher, Matthias Bossard, Thomas Hochgruber, Andreas J. Zimmermann, Beat A. Kaufmann, Katrin Pumpol, Peter Rickenbacker and David Conen.

5.2 ABSTRACT

BACKGROUND: Treatment to restore sinus rhythm among patients with atrial fibrillation (AF) has limited long-term success rates. Gene expression profiling may provide new insights into AF pathophysiology.

OBJECTIVE: To identify biomarkers and improve our understanding of AF pathophysiology by comparing whole blood gene expression before and after electrical cardioversion (ECV).

METHODS: In 46 patients with persistent AF that underwent ECV, whole blood samples were collected 1-2 hours before and 4 to 6 weeks after successful cardioversion. The paired samples were sent for microarray and plasma biomarker comparison.

RESULTS: Of 13,942 genes tested, expression of *SLC25A20* and *PDK4* had the strongest associations with AF. Post-cardioversion *SLC25A20* and *PDK4* expression decreased by 0.8 (CI 0.7-0.8, p=2.0x10^{-6}) and 0.7 (CI 0.6-0.8, p=3.0x10^{-5}) fold, respectively. Median N-terminal pro B-type natriuretic peptide (NT-proBNP) concentrations decreased from 121.6 pg/mL to 36.4 pg/mL (p=1.8 x10^{-8}) after cardioversion. AF discrimination models combining NT-proBNP and gene expression (NT-proBNP + *SLC25A20* area under the curve=0.88, NT-proBNP + *PDK4* AUC=0.86) had greater discriminative capacity as compared with NT-proBNP alone (AUC=0.84). Moreover, a model including NT-proBNP, *SLC25A20* and *PDK4* significantly improved AF discrimination as compared with other models (AUC=0.89, Net Reclassification Index >0.6, p<4.0x10^{-3}). We validated the significance of *SLC25A20* and *PDK4* expression with AF in an independent sample of 17 patients.

CONCLUSION: This study demonstrates that *SLC25A20, PDK4,* and NT-proBNP have incremental utility as biomarkers discriminating AF from sinus rhythm. Elevated *SLC25A20* and *PDK4* expression in AF indicates an important role for energy metabolism in AF.

KEYWORDS: Atrial fibrillation; Biomarker; Blood; NT-proBNP; Gene expression profiling; Electrical cardioversion

5.3 INTRODUCTION

Atrial fibrillation (AF), the most common cardiac arrhythmia, is associated with an increased risk of death, stroke[1,2] and heart failure.[2–4] Its incidence is projected to increase with the ageing of population and increased prevalence of obesity.[5–7] However, treatment strategies aiming to revert AF to sinus rhythm have limited long-term success rates and significant risks.[8,9] While it has been demonstrated that a substantial proportion of AF originates from the pulmonary veins,[10] our current understanding of the complex pathophysiology remains incomplete. Improvements in this area may not only help to develop novel treatment strategies for AF patients, but also to anticipate the rhythm stability among patients with AF, and to develop better methods to detect intermittent forms of AF.

We hypothesized that identification of novel biomarkers associated with rhythm changes among AF patients may provide insights into our pathophysiological understanding of the disease. Assessment of whole blood gene expression is an emerging and promising class of biomarkers. Gene expression levels vary rapidly in response to physiologic changes and its disease specificity may outreach conventional parameters. Recently, gene expression patterns of left atrial tissue from AF patients has been described.[11–13] In addition, the Framingham group has analyzed peripheral blood gene expression among patients with prevalent AF as compared with a large population of non-affected individiduals.[14] However, neither of these studies has assessed changes in peripheral blood gene expression within an individual patient after conversion from AF to sinus rhythm.

We therefore aimed to study AF patients pre- and post-electrical cardioversion (ECV), in order to identify AF specific whole blood RNA biomarkers potentially implicated in AF pathophysiology. We also wanted to assess their ability to discriminate AF from sinus rhythm in this setting. Novel gene expression biomarkers were validated in an independent set of participants.

5.4 METHODS

5.4.1 Study population

We prospectively enrolled consecutive patients >18 years with persistent AF, defined as a non-self-terminating episode lasting >7 days, who were scheduled for non-urgent electrical cardioversion (ECV) at two tertiary hospitals in Switzerland. We excluded patients with untreated severe valvular disease, unstable and acute heart failure, limiting active or chronic major diseases, and a history of open-heart surgery within 3 months of inclusion. Informed consent was obtained from all patients and the study was approved by the local ethics commission.

5.4.2 Study procedures

Study visits were scheduled approximately 24 hours before electrical cardioversion and after 4 ± 1 weeks of follow-up. Information on baseline characteristics, concomitant medication and co-morbidities was collected through study questionnaires both at baseline and follow-up. In addition, conventional blood pressure measurements,

standard 12-lead electrocardiogram (ECG), 24-hour Holter ECG monitoring, real time 3-dimensional echocardiography and blood sampling were obtained at both visits. At baseline, all examinations were performed 1-2 hours prior to the cardioversion procedure. ECV was performed according to local standards. After cardioversion, changes in personal medication were strongly discouraged until the follow-up visit. The second blood sampling was obtained directly after the follow-up 24-hour Holter ECG, in order to confirm stable sinus rhythm. Patients who had recurrent AF between the two scheduled visits were excluded from this study.

5.4.3 Blood sampling and biomarker measurements

Prior to ECV and at follow-up, venous blood samples were collected in EDTA tubes and PAXgene™ Blood RNA tubes (PreAnalytiX). EDTA tubes were immediately centrifuged to isolate plasma and all tubes were stored at -80°C. High-sensitivity C-reactive protein (hs-CRP), cystatin C (CYSC), and interleukin-6 (IL6) were measured on a Beckman Coulter Unicel DxC600 Synchron Clinical System (Beckman) according to the manufacturer's protocol. Myeloperoxidase (MPO) was measured using the ARCHITECT MPO immunoassay on the ARCHITECT Clinical Chemistry Analyzer (Abbott). N-terminal pro B-type natriuretic peptide (NT-proBNP) was measured on the Elecsys 2010 immunoassay analyzer (Roche).

5.4.4 RNA extraction

The PAXgene™ Blood RNA tubes were processed at the Genetic and Molecular Epidemiology Laboratory of PHRI and McMaster University, Hamilton ON. Paired

samples were processed using the same RNA extraction and amplification method. Total RNA was isolated from samples using the QIAsymphony PAXgene Blood RNA Kit (QIAGEN) or the MagMAX Stabilized Blood Tube RNA Isolation Kit (LifeTech). RNA was then quantified with RiboGreen® (LifeTech) and Nanodrop (Nanodrop).

5.4.5 Microarray hybridization

200ng of total RNA was amplified and biotinylated according to the manufacturer's protocol. Samples were amplified with the TotalPrep RNA Amplification Kit (LifeTech) or the Illumina TotalPrep-96 RNA Amplification Kit (LifeTech). The final biotin-labeled cRNA species were then hybridized to the Illumina HumanRef-8v4 expression BeadChips (Illumina). Each BeadChip hold 12 samples at a time so paired samples were hybridized on to the same chip. BeadChips were then washed, dried and scanned on the iScan System (Illumina) as per the manufacturer's protocol.

5.4.6 Microarray pre-processing and quality control

The Illumina HumanRef-8v4 BeadChip interrogates expression of 34,694 unique genes using 47,323 probes. The raw BeadChip sample probe profile and control probe profile were exported from GenomeStudio version 1.9.0 (Illumina). All data preprocessing and quality control was performed in R (http://r-project.org) using microarray-specific packages available through Bioconductor.[15] Four samples were deemed outliers during quality control and were excluded from further analysis, including their corresponding pairs. Data pre-processing involved background correction using the

non-genomic control probes, quantile normalization and log2 transformation.[16,17] Probes with detection P-value <0.05 in >50% of the samples were included for further statistical analysis. As a result the final pre-processed expression set consisted of expression values for 13,942 RNA probes for each of the 92 samples from 46 individuals.

5.4.7 Quantitative Real-time Polymerase Chain Reaction

Reverse transcription was performed using the QuantiTect Reverse Transcription Kit (Qiagen). *SLC25A20* expression was monitored with the Hs00386383_m1 probe (LifeTech), *PDK4* with the Hs01037712_m1 probe (LifeTech) and *ITGB5* with the Hs00174435_m1 probe (LifeTech) as per the manufacturer protocol. Each qPCR was performed in duplex with the housekeeping gene *ACTB,* measured using the Hs01060665_g1 probe (LifeTech), to normalize expression. The TaqMan qPCR was conducted on a Viia7 Real-Time System (LifeTech) and cycle threshold (CT) values were calculated automatically with default parameters. Fold change (FC) differences were calculated using the **δ**CT method.[27]

5.4.8 Statistical analysis

All statistical analyses were performed using R. Clinical demographics were grouped according to pre- or post-cardioversion status. Normally distributed variables were compared using paired Student t-tests, otherwise Wilcoxon rank sum tests were used. A two-sided p-value<0.05 was considered as statistically significant.

Microarrays (and quantitative PCR) measure relative rather than absolute gene expression, or in other words the relative increase or decrease in expression of a gene as compared with global expression (or housekeeping genes). Differential gene expression was thus reported as FC, with 95% confidence intervals (95% CI). Linear regression models were used to identify RNA transcripts that changed significantly after the cardioversion (in sinus rhythm) compared with pre-cardioversion samples (in AF). Each model tested a single gene's association with AF while adjusting for sample pairs. To correct for multiple hypothesis testing a conservative Bonferroni correction was applied, setting the significance threshold at $0.05 / 13{,}942 = 3.6 \times 10^{-6}$.

Significant genes and plasma biomarkers were also tested for association with AF risk factors, ECG, and echocardiography parameters using linear regression models. An adjusted p-value $<0.05/12=0.0042$ was considered significant. Receiver operating characteristic (ROC) curves were constructed, using pROC,[19] to determine the discriminative capacity of significant plasma and RNA biomarkers for pre-cardioversion AF. The area under the ROC curve (AUC) was determined as a measure of sensitivity and specificity. To compare models we calculated the continuous Net Reclassification Index (NRI)[20] using Hmisc.[21] We considered an NRI greater than 0.6 a strong, 0.4 an intermediate, and 0.2 a weak improvement in discriminative capacity. To verify and validate microarray expression results qPCR data was analyzed with linear regression models and adjusted for sample pairs.

5.5 RESULTS

5.5.1 Patient characteristics

Between March 2010 to April 2013, 108 consecutive patients with persistent AF were enrolled into the study. 67 patients had successful cardioversion and confirmed sinus rhythm at the follow-up visit; 50 were selected for biomarker discovery and 17 for independent validation. After microarray quality control, the biomarker discovery cohort was reduced to 46 patients. Patient demographics in the discovery cohort are presented in *Table 1*. Mean age was 65.9 ± 10.6 and 26.1% of participants were female. The post-cardioversion follow-up examination took place 35 days ± 8 days after ECV. Successful cardioversion resulted in a significant decrease in heart rate (pre 86 ± 16 vs. post 59 ± 9, $p = 6.5 \times 10^{-14}$). We also observed a decrease in E-wave velocity (pre 0.94 m/s, post 0.79 m/s, $p = 2.1 \times 10^{-4}$), an increase in E-wave deceleration time (pre 200 ms, post 255 ms, $p = 1.0 \times 10^{-4}$), and an increase in left ventricular ejection fraction (pre 45%, post 53%, $p = 8.1 \times 10^{-4}$) following cardioversion.

Table 1. Participant demographics for biomarker discovery cohort.

	Pre-cardioversion n = 46	Post-cardioversion n = 46	P-Value
Gender (n, % female)	12 (26.1)	12 (26.1)	
Age (years), mean ± SD	65.9 ± 10.6	66.0 ± 10.7	
Body mass index (kg/m^2), mean ± SD	27.8 ± 3.7	28.0 ± 3.7	0.83
Systolic BP (mmHg), mean ± SD	136.8 ± 19.7	135.3 ± 19.2	0.73
Diastolic BP (mmHg), mean ± SD	83.5 ± 20.9	77.2 ± 9.7	0.08
Holter ECG (mean ± SD)			
Heart rate (bpm)	86.0 ± 16.3	58.9 ± 9.4	$6.5x10^{-14}$
HRmax (bpm)	104.4 ± 21.7	NA	NA
HRmin (bpm)	41.8 ± 6.6	NA	NA
Echocardiography (mean ± SD)			
E wave (m/s)	0.94 ± 0.2	0.79 ± 0.2	$2.1x10^{-4}$
A wave (m/s)	NA	0.614 ± 0.3	NA
Deceleration time E wave (ms)	188.6 ± 69.3	255.2 ± 74.9	$1.0x10^{-4}$
Left ventricle ejection fraction (%)	44.9 ± 12.2	53.4 ± 9.3	$8.1x10^{-4}$

5.5.2 Association between gene expression and AF

Each of the 13,942 RNA probes was tested for association with AF while adjusting for sample pairs. The ten most significant genes from the analysis are presented in *Table 2*. *SLC25A20* expression was most significantly associated with AF, and the only gene that remained significant after adjustment for multiple hypothesis testing *(Figure 1)*. A 0.8 fold decrease in *SLC25A20* was observed in post-cardioversion samples as compared with baseline (CI 0.7-0.8, $p = 2.0 \times 10^{-6}$, *Supplementary Figure 1A*). *PDK4* was the second most significant gene and decreased by 0.7 fold post-cardioversion (CI 0.6-0.8, $p = 3.0 \times 10^{-5}$, *Supplementary Figure 2A*). *ITGB5* was the third most significant gene and increased by 1.2 fold post-cardioversion (CI 1.1-1.4, $p = 3.1 \times 10^{-5}$, *Supplementary Figure 3A*).

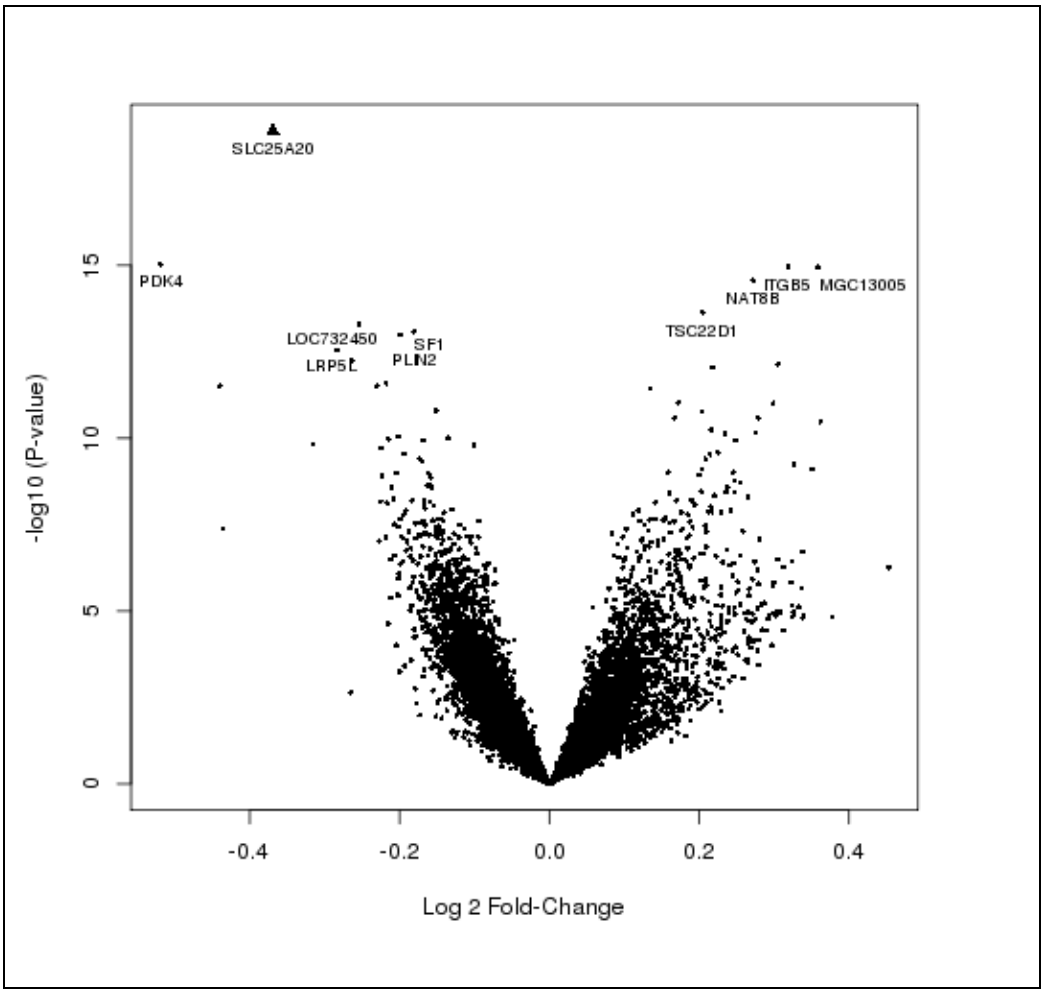
Differential expression of the top three genes, *SLC25A20, PDK4* and *ITGB5*, was verified using qPCR. A consistent decrease in *SLC25A20* expression was observed in post-cardioversion sinus rhythm samples as compared with baseline AF using qPCR (FC = 0.7, CI 0.6-0.7, $p = 1.6 \times 10^{-9}$, *Supplementary Figure 1B*). Similarly, a 0.6 fold decrease in *PDK4* expression was observed post-cardioversion (CI 0.5-0.7, $p = 4.0 \times 10^{-5}$, *Supplementary Figure 2B*). Differential expression of *ITGB5* could not be verified (*p*=0.086, *Supplementary Figure 3B*).

Table 2. Top genes associated with cardioversion.

Gene	P-value	Fold Change	Upper CI	Lower CI	Description
SLC25A20	2.0×10^{-6}	0.8	0.8	0.7	Solute carrier family 25 (carnitine/acylcarnitine translocase), member 20
PDK4	3.0×10^{-5}	0.7	0.8	0.6	Pyruvate dehydrogenase kinase, isozyme 4
ITGB5	3.1×10^{-5}	1.2	1.4	1.1	Integrin, beta 5
DDX11L2	3.2×10^{-5}	1.3	1.4	1.2	DEAD/H box helicase 11
NAT8B	4.1×10^{-5}	1.2	1.3	1.1	N-acetyltransferase 8B
TSC22D1	7.8×10^{-5}	1.2	1.2	1.1	TSC22 domain family, member 1
LOC732450	9.9×10^{-5}	0.8	0.9	0.8	
SF1	1.1×10^{-4}	0.9	0.9	0.8	Splicing factor 1
PLIN2	1.2×10^{-4}	0.9	0.9	0.8	Perilipin 2
LRP5L	1.7×10^{-4}	0.8	0.9	0.7	Low density lipoprotein receptor-related protein

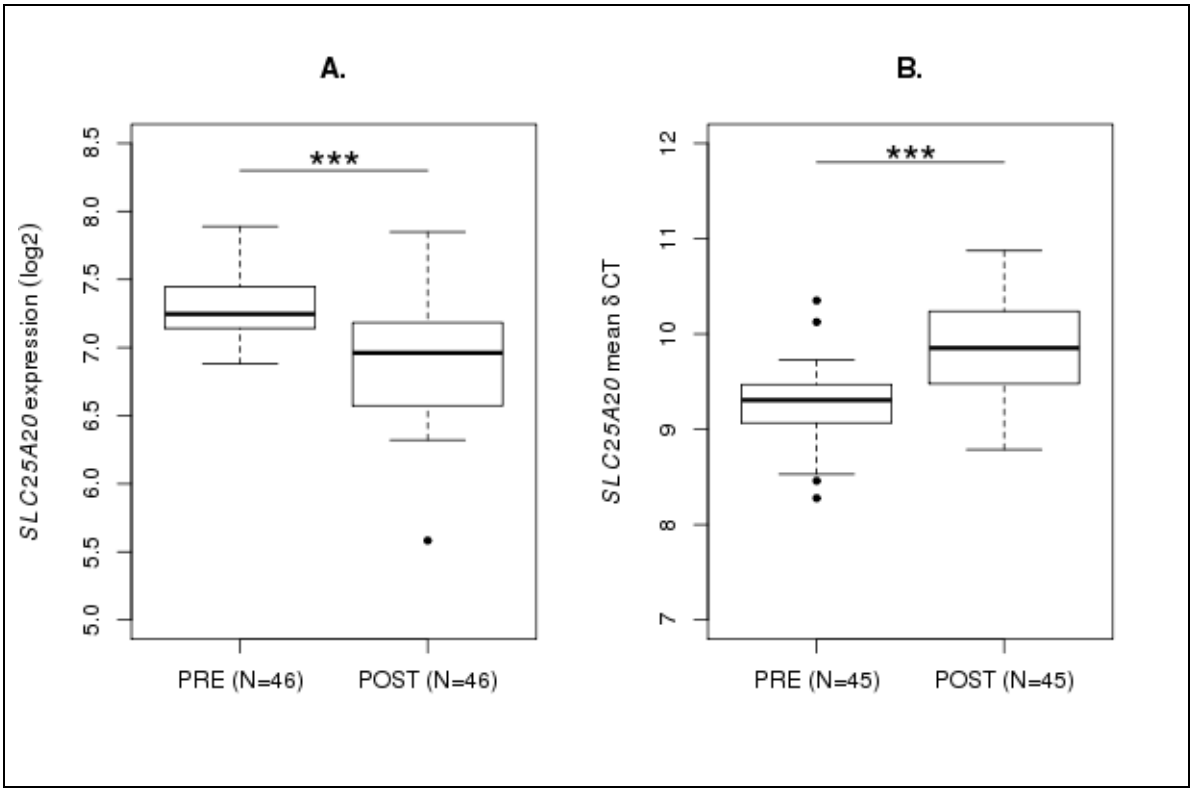
Figure 1. Volcano plot of gene expression association with cardioversion treatment.

Each point represents one of the RNA transcripts tested and the ten most significant genes have been labeled. The x-axis represents the effect of each gene, reported as log2 fold change, and a positive log2 fold change is indicative of increased expression in post-cardioversion samples. The y-axis represents the –log10(P-value). Triangle points represent genes that have significant differential expressed after Bonferroni correction (P-value $<3.6x10^{-6}$).

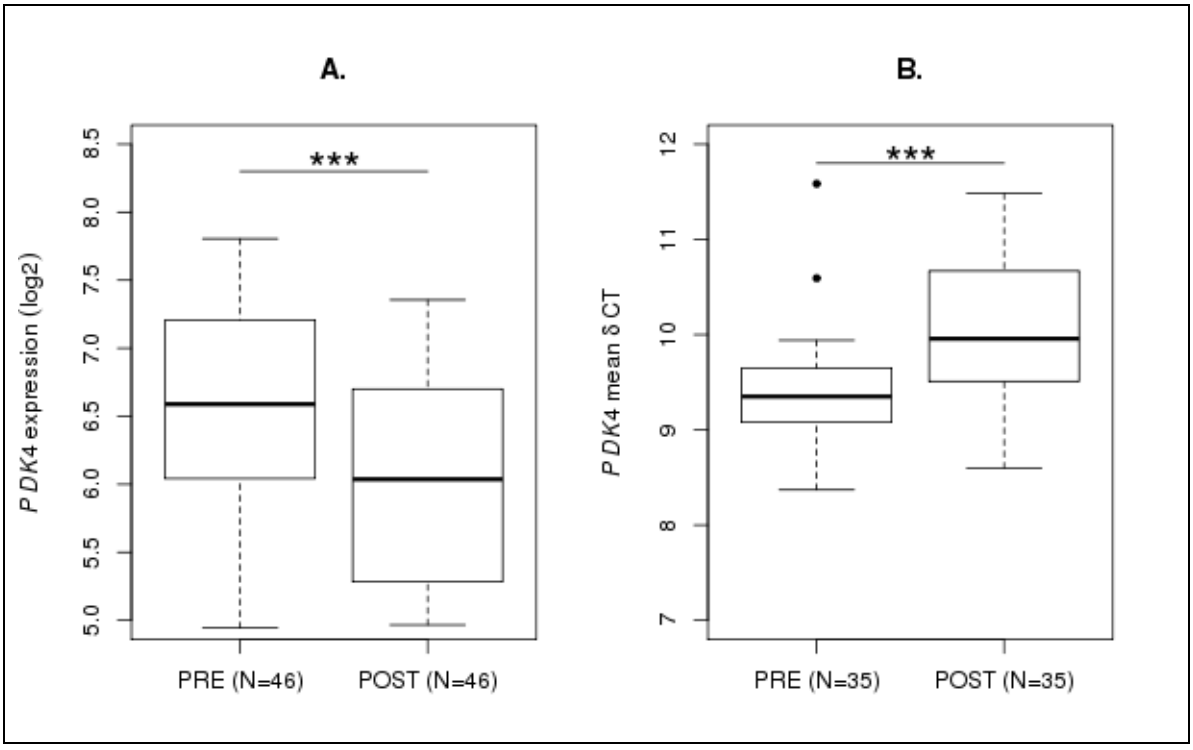


Supplementary Figure 1. Boxplots of *SLC25A20* expression pre- and post-cardioversion.

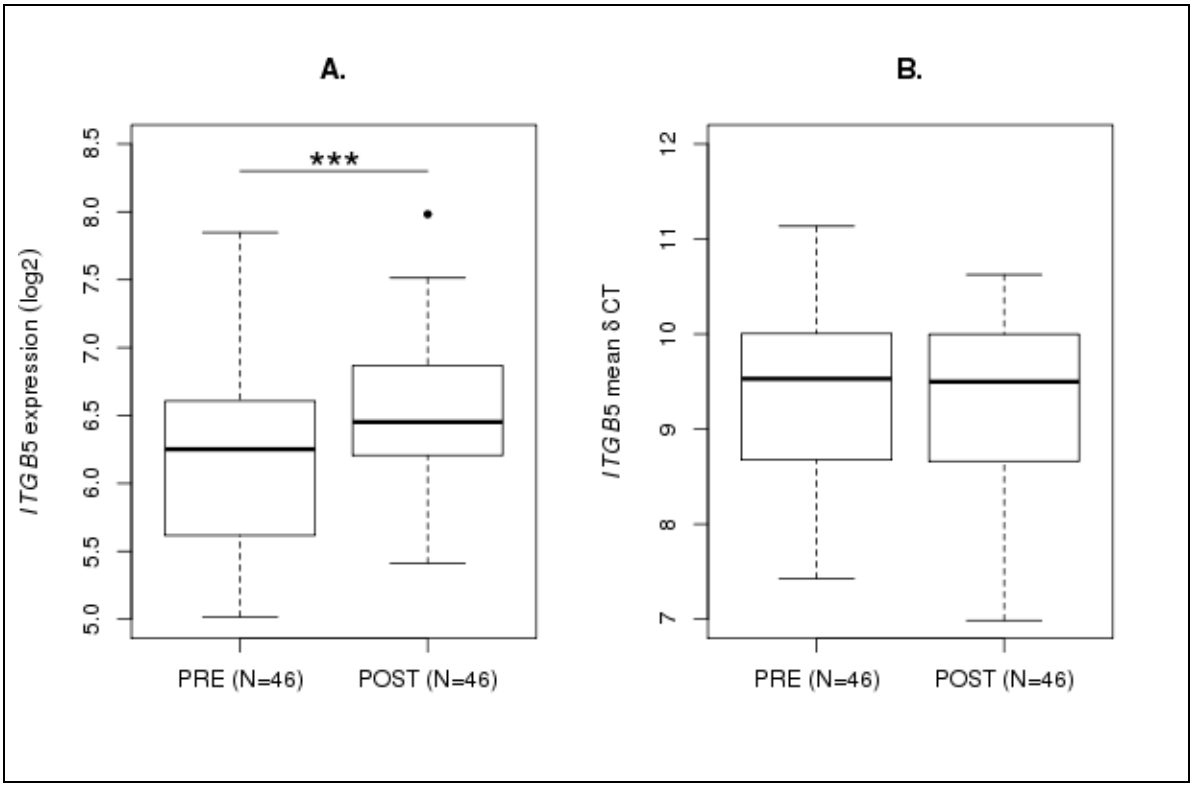
Boxes extend from the 25th to the 75th percentile, with the horizontal line representing the median. Outliers are identified as samples with an expression value 1.5 times more or less than the interquartile range. The CT (cycle threshold) is the number of PCR cycles required for the fluorescent signal to exceed background levels. Unlike microarray values, CT values are inversely proportional to the amount of target nucleic acid in a sample. A) Microarray expression of *SLC25A20* decreased following cardioversion. B) qPCR expression of *SLC25A20* also decreased following cardioversion. A symbol directly above a bar indicates a significant difference between groups; p <0.0005 (***).



Supplementary Figure 2. Boxplots of *PDK4* expression pre- and post-cardioversion. Boxes extend from the 25th to the 75th percentile, with the horizontal line representing the median. Outliers are identified as samples with an expression value 1.5 times more or less than the interquartile range. The CT (cycle threshold) is the number of PCR cycles required for the fluorescent signal to exceed background levels. Unlike microarray values, CT values are inversely proportional to the amount of target nucleic acid in a sample. A) Microarray expression of *PDK4* decreased following cardioversion. B) qPCR expression of *PDK4* also decreased following cardioversion. A symbol directly above a bar indicates a significant difference between groups; p <0.0005 (***).



Supplementary Figure 3. Boxplots of *ITGB5* expression pre- and post-cardioversion. Boxes extend from the 25th to the 75th percentile, with the horizontal line representing the median. Outliers are identified as samples with an expression value 1.5 times more or less than the interquartile range. The CT (cycle threshold) is the number of PCR cycles required for the fluorescent signal to exceed background levels. Unlike microarray values, CT values are inversely proportional to the amount of target nucleic acid in a sample. A) Microarray expression of *ITGB5* decreased following cardioversion. B) qPCR expression of *ITGB5* also decreased following cardioversion. A symbol directly above a bar indicates a significant difference between groups; p <0.0005 (***).



5.5.3 *SLC25A20* and *PDK4* expression are not associated with clinical variables

Restricting the analysis to pre-cardioversion AF samples, we tested *SLC25A20* and *PDK4* for association with AF risk factors, such as age, gender, BMI, systolic and diastolic blood pressure; Holter ECG and echocardiography parameters. We observed no association between pre-cardioversion expression of either gene and measured variables (all p>0.05=NS). We also restricted the analysis to samples collected post-ECV. After correction for multiple hypotheses testing, we observed a modest association between post-cardioversion *SLC25A20* expression and elevated diastolic blood pressure (p = 0.0029). *PDK4* had no association with the measured variables.

5.5.4 Association between plasma biomarkers and AF

NT-proBNP levels were significantly decreased in post-cardioversion samples, as compared with baseline (median 121.6 vs. 36.4 pg/mL, $p = 1.8x10^{-8}$). Circulating levels of hs-CRP, CYSC, IL6, and MPO did not change between pre and post-cardioversion samples *(Table 3)*. Limiting the analysis to either pre-cardioversion AF samples or post-cardioversion sinus rhythm samples, we observed no association between NT-proBNP concentrations and AF risk factors, Holter ECG or echocardiography parameters (all *p*=NS).

Table 3. Plasma biomarker concentrations in participants pre- and post-cardioversion.

	Pre-cardioversion n = 46	Post-cardioversion n = 46	P-Value
hs-CRP (mg/L)	1.65 (2.7 - 5.9)	2.215 (0.5 - 5.9)	0.31
CYSC (mg/L)	0.93 (0.7 - 1.1)	0.925 (0.7 - 1.2)	0.97
IL6 (pg/mL)	2.24 (2.4 - 4.8)	2.77 (1.8 - 4.9)	0.28
MPO (pmol/L)	1114.8 (202.1 - 2100.3)	1113.4 (176.8 - 2560.6)	0.38
NT-proBNP (pg/mL)	121.6 (42.9 - 270.5)	36.4 (2.4 - 119)	$1.8x10^{-8}$

Data are medians (interquartile range)

5.5.5 Discriminative capacity of NT-proBNP, *SLC25A20* and *PDK4* for AF

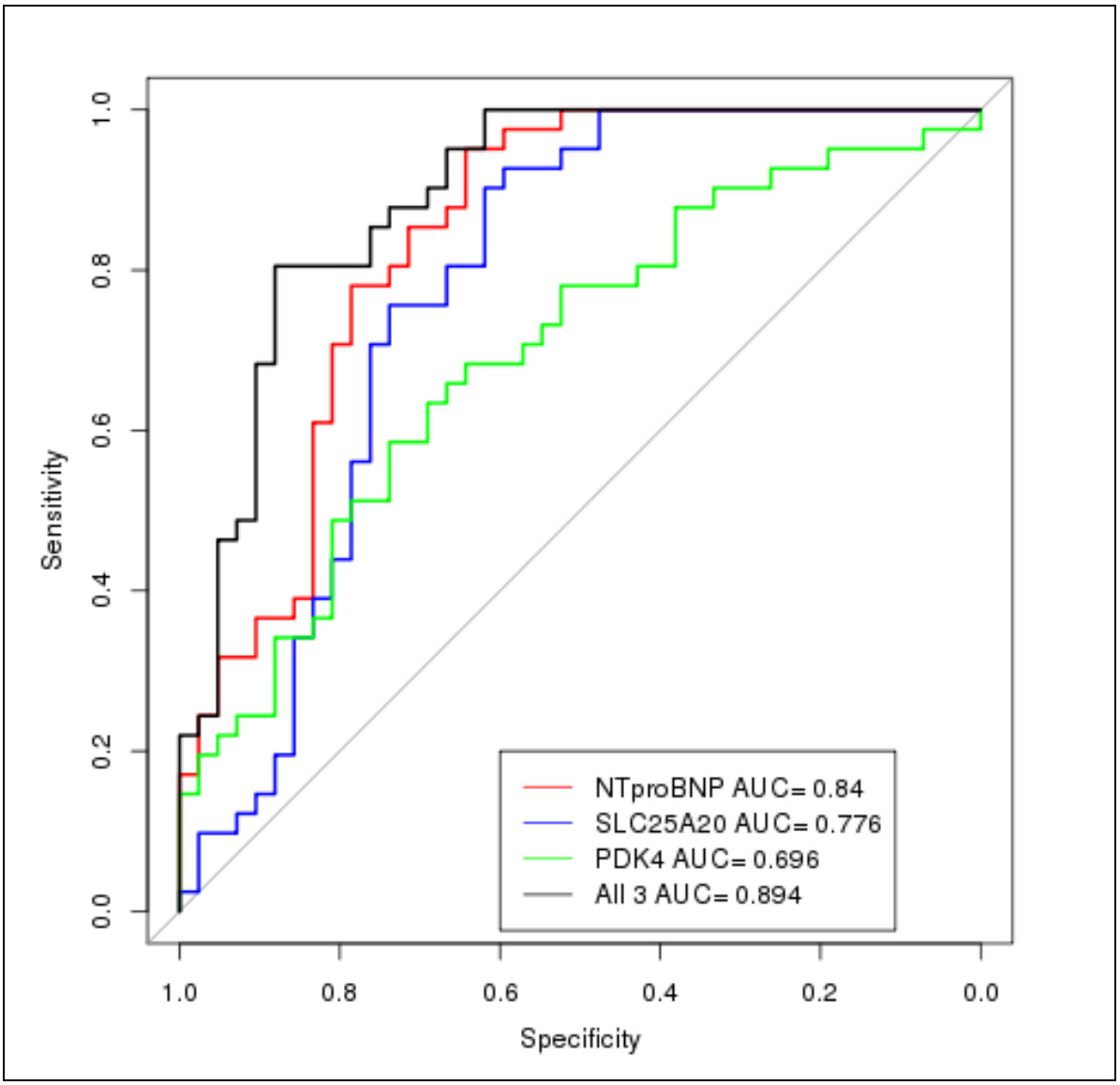
A multivariable logistic regression model for AF including NT-proBNP, *SLC25A20* and *PDK4,* indicated that all three biomarkers were independently associated with AF *(Supplementary Table 1)*. To determine the discriminative capacity of each biomarker we constructed receiver operator characteristics (ROC) curves. In single variable models, AUC of the ROC curves discriminating between pre-cardioversion and post-cardioversion samples was 0.84 (CI 0.75-0.93) for NT-proBNP, 0.78 (CI 0.67-0.88) for *SLC25A20* and 0.70 (CI 0.58-0.81) for *PDK4* (*Supplementary Figure 4*). A two variable model including NT-proBNP and expression of either gene strongly improved discrimination as compared with NT-proBNP alone (NT-proBNP + *SLC25A20* AUC = 0.88 (CI 0.81-0.96), NRI = 0.85, $p = 1.6x10^{-5}$; NT-proBNP + *PDK4* AUC = 0.86 (CI 0.78-0.94), NRI = 0.65, $p = 1.7x10^{-3}$). Moreover a model including all three biomarkers had the greatest discriminative capacity (AUC = 0.89, CI 0.82-0.96). The combination of NT-proBNP, *SLC25A20* and *PDK4* further improved discrimination as compared with other models (all vs NT-proBNP NRI = 0.79, $p = 7.8x10^{-5}$; all vs *SLC25A20* NRI = 0.89, $p = 5.6x10^{-6}$; all vs *PDK4* NRI = 1.2, $p = 2.5x10^{-11}$; all vs NT-proBNP+*SLC25A20* NRI = 0.60, $p = 4.0x10^{-3}$; all vs NT-proBNP+*PDK4* NRI = 0.70, $p = 5.8x10^{-4}$).

Supplementary Table 1. Results of multiple regression between biomarkers and rhythm status.

A logistic regression model was constructed for rhythm status. The model included only NT-proBNP, *SLC25A20* and *PDK4*.

Biomarker	Beta	P-value
SLC25A20	2.17	0.016
PDK4	0.95	0.024
NT-proBNP	0.018	0.0010

Supplementary Figure 4. Receiver-operating characteristic curves for the discrimination of pre-cardioversion AF from post-cardioversion sinus rhythm.



5.5.6 Replication of *SLC25A20* and *PDK4* in the validation cohort

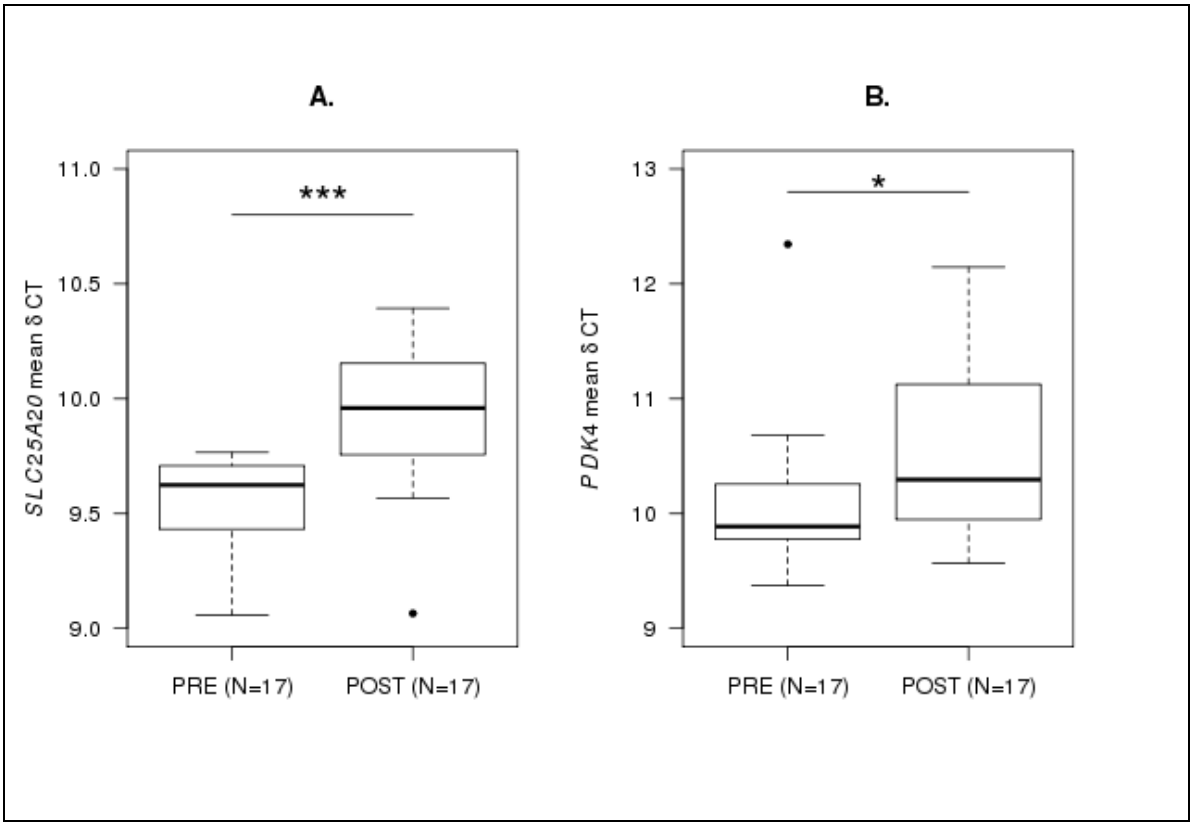
We validated the decrease in *SLC25A20* and *PDK4* expression post-cardioversion in an independent sample consisting of 17 individuals. Patient demographics are shown in *Supplementary Table 2*. The successfully cardioverted patients had significantly decreased heart rate (pre = 83.8 bpm, post = 56.5 bpm, $p = 2.9 \times 10^{-6}$) and decreased diastolic blood pressure (pre = 83.8. post = 76.0, p = 0.015), similar to the main sample. qPCR validated the initial microarray findings. We observed a 0.8 fold decrease in *SLC25A20* (CI 0.7-0.9, $p = 2.0 \times 10^{-4}$, *Supplementary Figure 5*) and 0.7 fold decrease in *PDK4* (CI 0.5-1.0, p = 0.05, *Supplementary Figure 5B*) in post-cardioversion sinus rhythm samples as compared with baseline AF. Restricting the analysis to pre-cardioversion AF samples or post-cardioversion sinus rhythm, we tested both *SLC25A20* and *PDK4* for association with AF risk factors and observed no association (*p*=NS for all comparisons).

Supplementary Table 2. Demographics for the independent validation cohort samples.

	Pre-cardioversion n = 17	Post-cardioversion n = 17	P-value
Gender (n, % female)	6 (35.3)	6 (35.3)	
Age, mean ± SD	69.1 ± 8.9	64.7 ± 19.2	
Body mass index (kg/m^2), mean ± SD	25.3 ± 4.3	25.5 ± 4.4	
Systolic BP, mean ± SD	139.9 ± 18.1	131.0 ± 11.1	0.12
Diastolic BP, mean ± SD	85.9 ± 12.4	76.0 ± 7.3	0.015
Heart rate, mean ± SD	83.8 ± 13.2	56.5 ± 12.4	$2.9x10^{-6}$

Supplementary Figure 5. qPCR gene expression in the independent validation cohort.

Boxes extend from the 25th to the 75th percentile, with the horizontal line representing the median. Outliers are identified as samples with an expression value 1.5 times more or less than the interquartile range. The CT (cycle threshold) is the number of PCR cycles required for the fluorescent signal to exceed background levels. Unlike microarray values, CT values are inversely proportional to the amount of target nucleic acid in a sample. A) qPCR expression of *SLC25A20* pre- and post- cardioversion in the independent validation cohort. B) qPCR expression of *PDK4*.A symbol directly above a bar indicates a significant difference between groups; $p < 0.0005$ (***), $p < 0.05$(*).



5.6 DISCUSSION

The present study evaluated peripheral blood gene expression and plasma protein biomarkers associated with AF rhythm by comparing paired patient samples pre- and post-ECV. We identified novel associations between whole blood gene expression of *SCL25A20* and *PDK4* with AF. Expression of both genes was elevated in AF as compared with post-ECV sinus rhythm. Adding either RNA marker to a model with NT-proBNP strongly improved AF discrimination. A model including *SLC25A20, PDK4* and NT-proBNP had the greatest ability to discriminate between AF and sinus rhythm. The association between both *SLC25A20* and *PDK4* with rhythm status was confirmed in an independent validation cohort.

Using transcriptome-wide expression profiling we identified an association between AF and *SLC25A20. SLC25A20* encodes the carnitine-acylcarnitine translocase (CACT), which transports fatty acids into the inner mitochondrial membrane for β-oxidation.[22,23] CACT deficiency, an autosomal recessive disorder, results in fatal cardiomyopathy early in life.[24] We observed decreased expression of *SLC25A20* following successful cardioversion. Hence suggesting that fatty acid metabolism is associated with AF, and up-regulated during AF episodes. In conjunction, decreased expression of *PDK4* was identified following successful cardioversion. Since PDK4 inhibits the pyruvate dehydrogenase complex[25], the study results indicate that glucose metabolism may be inhibited during AF episodes.

Taken together, the elevated levels of *SLC25A20* and *PDK4* pre-cardioversion are suggestive of an adaptive response to the increased metabolic demand during AF

episodes, which are characterized by high atrial rates and a consecutive high metabolic demand. As such *SLC25A20* and *PDK4* expression may be associated with AF burden and may have utility for the diagnosis of paroxsysmal AF. In this context, recent studies showing that weight reduction was associated with reduced AF burden may also point towards the importance of energy metabolism in the occurrence of AF episodes.[26,27]

Recently a gene expression study observed that atrial tissue expression of *SLC25A20* was significantly decreased in patients that had no history of AF as compared with patients that had AF.[11] In addition, the researchers reported a decrease in *SLC25A20* and *PDK4* in patients currently in sinus rhythm that had a history of AF, as compared with patients currently in AF. Our study confirms the potential importance of these markers in the pathophysiology of AF, and shows that these changes can be observed not only across different patients, but also in an individual patient, if a sustained change in rhythm occurs. Considering a potential clinical applicability of these markers, it is of crucial importance that our changes were detected in peripheral blood samples, given that atrial biopsies are not feasible in clinical practice.

The Framingham whole blood expression study[14] did not report an association between AF and *SLC25A20* or *PDK4.* These potential differences are not surprising since our study evaluated expression changes occurring during AF episodes as compared to sinus rhythm within the same individual, while the Framingham study assessed differential expression between individuals with and without AF.

There are limitations of our study, which need to be taken into account. First, we included only patients with persistent AF, and therefore generalizability to other AF

populations remains uncertain. Second, all participants were of European origin thus the generalizability to other ethnicities remains uncertain. Third, fasting may have impacted gene expression. Pre-cardioversion samples were mostly collected after several hours of fasting, whereas fasting was not specified prior to post-cardioversion sampling. Studies have shown that free fatty acid concentrations increase with long term fasting.[27,28] Expression profiling of PBMCs after 24-hours of fasting has revealed increased expression of genes involved in fatty acid metabolism, including *SLC25A20* and *PDK4*.[28] However, studies have not described the time-course of expression changes with respect to the duration of fasting. As such the impact of shorter fasting episodes on gene expression, as in our study, has yet to be published. We have evaluated expression of *SLC25A20* and *PDK4* in a control population for up to 12 hours of fasting and observed no association between expression and hours since last meal (data not shown). The significance of *SLC25A20* and *PDK4* in AF is further supported by the atrial tissue study that observed elevated expression of both genes in AF patients as compared with patients in sinus rhythm that had a history of AF.[11] Since surgery is required to collect atrial tissue, all individuals were fasting prior to sampling. Therefore, the results indicate that *SLC25A20* and *PDK4* are truly associated with AF. Finally, our study populations were relatively small which may have hindered the detection of subtle expression differences in other genes.

In conclusion, the results of this study demonstrate that expression of *SLC25A20* and *PDK4* are independently associated with rhythm status among patients with persistent AF. These findings indicate that alterations in metabolic pathways are associated with the

prevalent cardiac rhythm in an individual AF patient, providing not only novel pathophysiological insights but also new potential intervention targets that can be tested in future studies. In addition, our study demonstrates that NT-proBNP, *SLC25A20* and *PDK4* have incremental utility as biomarkers discriminating AF from sinus rhythm. Future studies should explore whether these markers may be helpful for predicting AF recurrence in clinical practice.

5.7 REFERENCES

1. Wolf PA, Abbott RD, Kannel WB: Atrial fibrillation as an independent risk factor for stroke: the Framingham Study. Stroke 1991; 22:983–988.
2. Conen D, Chae CU, Glynn RJ, Tedrow UB, Everett BM, Buring JE, Albert CM: Risk of death and cardiovascular events in initially healthy women with new-onset atrial fibrillation. JAMA 2011; 305:2080–2087.
3. Stewart S, Hart CL, Hole DJ, McMurray JJ V: A population-based study of the long-term risks associated with atrial fibrillation: 20-Year follow-up of the Renfrew/Paisley study. Am J Med 2002; 113:359–364.
4. Benjamin EJ, Wolf PA, D'Agostino RB, Silbershatz H, Kannel WB, Levy D: Impact of Atrial Fibrillation on the Risk of Death: The Framingham Heart Study. Circulation 1998; 98:946–952.
5. Go AS, Hylek EM, Phillips KA, Chang Y, Henault LE, Selby J V, Singer DE: Prevalence of diagnosed atrial fibrillation in adults: national implications for rhythm management and stroke prevention: the AnTicoagulation and Risk Factors in Atrial Fibrillation (ATRIA) Study. JAMA 2001; 285:2370–2375.
6. Chugh SS, Havmoeller R, Narayanan K, et al.: Worldwide epidemiology of atrial fibrillation: A global burden of disease 2010 study. Circulation 2014; 129:837–847.
7. Miyasaka Y, Barnes ME, Gersh BJ, Cha SS, Bailey KR, Abhayaratna WP, Seward JB, Tsang TSM: Secular trends in incidence of atrial fibrillation in Olmsted County, Minnesota, 1980 to 2000, and implications on the projections for future prevalence. Circulation 2006; 114:119–125.
8. Van Gelder IC, Hagens VE, Bosker HA, Kingma JH, Kamp O, Kingma T, Said SA, Darmanata JI, Timmermans AJM, Tijssen JGP, Crijns HJGM, Group RC versus EC for PAFS: A comparison of rate control and rhythm control in patients with recurrent persistent atrial fibrillation. N Engl J Med 2002; 347:1834–1840.
9. Cappato R, Calkins H, Chen S-A, Davies W, Iesaka Y, Kalman J, Kim Y-H, Klein G, Natale A, Packer D, Skanes A: Prevalence and causes of fatal outcome in catheter ablation of atrial fibrillation. J Am Coll Cardiol American College of Cardiology Foundation, 2009; 53:1798–1803.
10. Haïssaguerre M, Jaïs P, Shah DC, Takahashi a, Hocini M, Quiniou G, Garrigue S, Le Mouroux a, Le Métayer P, Clémenty J: Spontaneous initiation of atrial fibrillation by ectopic beats originating in the pulmonary veins. N Engl J Med

1998; 339:659–666.

11. Deshmukh A, Barnard J, Sun H, et al.: Left Atrial Transcriptional Changes Associated with Atrial Fibrillation Susceptibility and Persistence. Circ Arrhythm Electrophysiol 2014; 8:CIRCEP.114.001632 – .
12. Barth AS, Merk S, Arnoldi E, et al.: Reprogramming of the human atrial transcriptome in permanent atrial fibrillation: expression of a ventricular-like genomic signature. Circ Res 2005; 96:1022–1029.
13. Ohki R, Yamamoto K, Ueno S, Mano H, Misawa Y, Fuse K, Ikeda U, Shimada K: Gene expression profiling of human atrial myocardium with atrial fibrillation by DNA microarray analysis. 2005; 102:233–238.
14. Lin H, Yin X, Lunetta KL, et al.: Whole blood gene expression and atrial fibrillation: the framingham heart study. PLoS One 2014; 9:e96794.
15. Gentleman RC, Carey VJ, Bates DM, et al.: Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 2004; 5:R80.
16. Shi W, Oshlack A, Smyth GK: Optimizing the noise versus bias trade-off for Illumina whole genome expression BeadChips. Nucleic Acids Res 2010; 38:e204.
17. Schmid R, Baum P, Ittrich C, Fundel-Clemens K, Huber W, Brors B, Eils R, Weith A, Mennerich D, Quast K: Comparison of normalization methods for Illumina BeadChip HumanHT-12 v3. BMC Genomics BioMed Central, 2010; 11:349.
18. Livak KJ, Schmittgen TD: Analysis of relative gene expression data using real-time quantitative PCR and. Methods 2001; 25:402–408.
19. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, Müller M: pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics BioMed Central Ltd, 2011; 12:77.
20. Pencina MJ, D'Agostino RB, Vasan RS: Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. Stat Med 2008; 27:157–172.
21. Harrel FE: Hmisc: A package of miscellaneous R functions.
22. Ramsay RR, Tubbs PK: The mechanism of fatty acid uptake by heart mitochondria: an acylcarnitine-carnitine exchange. FEBS Lett 1975; 54:21–25.
23. Pande S V: A mitochondrial carnitine acylcarnitine translocase system. Proc Natl Acad Sci U S A 1975; 72:883–887.
24. Iacobazzi V, Invernizzi F, Baratta S, et al.: Molecular and functional analysis of

SLC25A20 mutations causing carnitine-acylcarnitine translocase deficiency. Hum Mutat 2004; 24:312–320.

25. Patel MS, Korotchkina LG: Regulation of the pyruvate dehydrogenase complex. Biochem Soc Trans 2006; 34:217–222.

26. Pathak RK, Middeldorp ME, Meredith M, Mehta AB, Mahajan R, Wong CX, Twomey D, Elliott AD, Kalman JM, Abhayaratna WP, Lau DH, Sanders P: Long-Term Effect of Goal-Directed Weight Management in an Atrial Fibrillation Cohort. J Am Coll Cardiol 2015; 65:2159–2169.

27. Abed HS, Wittert G a, Leong DP, et al.: Effect of weight reduction and cardiometabolic risk factor management on symptom burden and severity in patients with atrial fibrillation: a randomized clinical trial. Jama 2013; 310:2050–2060.

28. Dole VP: A relation between non-esterified fatty acids in plasma and the metabolism of glucose. J Clin Invest 1956; 35:150–154.

29. Bouwens M, Afman L a, Müller M: Fasting induces changes in peripheral blood mononuclear cell gene expression profiles related to increases in fatty acid beta-oxidation: functional role of peroxisome proliferator activated receptor alpha in human peripheral blood mononuclear cells. Am J Clin Nutr 2007; 86:1515–1523.

CHAPTER 6: GENERAL DISCUSSION

6.1 GENERL OVERVIEW

Biomarkers of stroke have the potential to facilitate timely-diagnosis, estimate patient prognosis and to identify patients at higher risk for stroke. Protein biomarkers have had limited robustness (Whiteley *et al.*, 2009) and are not currently used in the clinical setting. With the advent of high throughput genomic technology, studies have sought to identify novel RNA biomarkers of stroke. However many of these studies have been limited by small sample sizes and have lacked microarray verification and validation (Moore *et al.*, 2005; Tang *et al.*, 2006; Barr *et al.*, 2010). In contrast, this thesis has identified novel whole blood RNA biomarkers by: utilizing a large discovery cohort, assessing global gene networks or applying a unique study design. Many of the microarray findings have also undergone verification with qPCR and validation in small independent cohorts. Cumulatively, the thesis results demonstrate that RNA biomarkers have clinical value since they improve discrimination, as well as lending further insight into the pathogenesis of stroke and stroke risk factors. This section briefly summarizes the main research findings, the clinical implications, biologic significance and current challenges translating RNA biomarkers to the clinic.

6.2 CHAPTER 3 SUMMARY

In the univariate analysis of peripheral blood RNA expression in INTERSTROKE participants, elevated expression of *MCEMP1* was observed in stroke cases as compared

with controls. The association remained significant even after adjustment for available risk factors. *MCEMP1* expression differed between primary stroke subtypes. Furthermore, *MCEMP1* expression was highest within 24-hours of symptom onset and also had prognostic discriminative capacity. One-month disability and mortality discrimination models comprised of *MCEMP1* expression, baseline modified Rankin score (mRS), and primary stroke type performed significantly better than models without *MCEMP1* expression. The novel association between stroke and *MCEMP1* expression was verified and validated in an independent cohort.

6.3 CHAPTER 4 SUMMARY

Global gene co-expression network analysis was performed on ischemic stroke cases and controls from the INTERSTROKE study. *NLRC4*, *CKLF* and *HS.546375* were highly interconnected genes within three gene networks and were each independently associated with stroke. A model including all three genes improved ischemic stroke discrimination as compared with single gene models or clinical variables. Similarly, one-month disability discrimination was also improved using the three-gene model as compared with single gene models. Disability discrimination could be further improved with the addition of *MCEMP1* and baseline mRS to the three gene model.

6.4 CHAPTER 5 SUMMARY

Whole blood RNA and plasma biomarkers were evaluated from persistent atrial fibrillation (AF) patients before and after successful cardioversion. Decreased expression of *SLC25A20* and *PDK4* were detected during sinus rhythm. Similarly, circulating NT-proBNP concentrations decreased after successful cardioversion. The three markers, *SLC25A20*, *PDK4* and NT-proBNP, were independently associated with AF. Furthermore, an AF discrimination model consisting of all three biomarkers had better performance than single biomarker models. The novel association between AF and expression of *SLC25A20* and *PDK4* was verified and validated in an independent cohort.

6.5 CLINICAL IMPLICATIONS

The thesis results indicate that RNA biomarkers can provide additional diagnostic and prognostic information to what is currently available. Specifically, RNA biomarkers were able to discriminate: stroke cases from controls, ischemic stroke from hemorrhage stroke, patients with one-month post-stroke disability from those without, patients with one-month mortality from survivors and atrial fibrillation from sinus rhythm. The association between RNA biomarkers and both diagnostic and/or prognostic outcomes remained significant even after adjustment for clinical variables and available risk factors. Moreover discrimination models that included gene expression had better performance than models without gene expression. For instance, the model consisting of *MCEMP1* expression, baseline mRS and stroke type improved stroke prognosis discrimination. Similarly, *SLC25A20*, *PDK4* and NT-proBNP had incremental utility for AF

discrimination. Thus the results indicate that RNA biomarkers have independent clinical utility.

6.6 BIOLOGIC SIGNIFICANCE

Non-invasive RNA biomarkers can provide further insights into the pathophysiology of stroke and AF. Literature indicates that inflammation plays an important role in the pathogenesis of stroke (Jin *et al.*, 2010; Iadecola and Anrather, 2011), but the specific molecular pathways and cells involved still require elucidation. Similarly, AF is an established risk factor of stroke however our understanding of the molecular biology underlying AF remains incomplete. Chapter 3 and 4 of the thesis suggest that the mast cells and the NLRC4 inflammasome have an important inflammatory role in stroke, while Chapter 5 implicates energy metabolism genes with AF. Cumulatively the results identify putative therapeutic target and provide new avenues for research.

6.6.1 MAST CELLS IN STROKE

Chapter 3 identified and evaluated the association between *MCEMP1* expression and stroke. *MCEMP1* encodes mast cell expressed membrane protein 1, a newly identified protein for which function has yet to be determined. However, mast cells are known to be both sensor and effector cells of the innate immune system. Upon activation, mast cells secrete granules. Mast cell granules contain a plethora of pre-activated

molecules with vasoactive, pro-inflammatory, anticoagulant and proteolytic activity (Wernersson and Pejler, 2014). Recent studies indicate that brain resident mast cells act as early regulator of blood-brain barrier (BBB) permeability and neutrophil infiltration (Strbian, Karjalainen-Lindsberg, *et al.*, 2006; McKittrick *et al.*, 2015). Specifically gelatinase, released from mast cell granules, may play a role in mediating BBB disruption during stroke (Mattila *et al.*, 2011). Since mast cell degranulation can trigger a multitude of pathways, mast cell stabilization may be a new pharmacologic target for stroke.

Mast cell stabilizers block calcium channels essential for mast cell degranulation. When granules are not secreted, enzymes such as gelatinase are not released. As a result, mast cell stabilizer may reduce BBB dysregulation and minimize other inflammatory responses. Thus these stabilizers may be used to reduce progressive tissue damage following stroke. Initial proof of concept has been demonstrated in animal models (Strbian, Tatlisumak, *et al.*, 2006; Strbian, Karjalainen-Lindsberg, *et al.*, 2006; Jin *et al.*, 2009). Experimental stroke was conducted on rats deficient in mast cells, rats given mast cell blocking agents and wild-type rats. Study results indicate that rats without mast cells and rats with inactive mast cells had significantly better neurologic outcomes as compared with wild-type rats. Further research will be required to unravel the molecular mechanisms underlying mast cell activation during stroke. Currently, mast cell stabilizing drugs are used to prevent allergic reactions (Finn and Walsh, 2013). In future, clinical trials may be conducted to evaluate the utility of mast-cell blocking agents for stroke management.

6.6.2 INFLAMMATORY MECHANISMS OF STROKE

Inflammation is a protective response to injury, but the inflammatory response triggered by stroke also contributes to the progression of neurologic injury. During stroke, molecules such as damage associated molecular patterns (DAMPs) are released from stressed and necrotic cells. DAMPs are recognized by intracellular immune receptors. One family of immune receptors is the nucleotide-binding domain and leucine-rich repeat containing receptor (NLR) family (Kanneganti *et al.*, 2007). Several members of the NLR family, including NLRP3 and NLRC4, can assemble into multi-molecular complexes known as inflammasomes. Inflammasomes control activation of proteolytic caspase-1 (Rathinam *et al.*, 2012). Caspase-1 regulates the maturation of pro-inflammatory cytokines (Martinon and Tschopp, 2007) and pyropotosis (Bergsbaken *et al.*, 2009; Miao *et al.*, 2011).

NLRP3 is the best-characterized inflammasome and previous studies report that NLRP3 mediates sterile inflammatory responses (Cassel and Sutterwala, 2010; Yang-Wei Fann *et al.*, 2013; Yang *et al.*, 2014), such as inflammation during ischemic brain injury. However, in Chapter 4, elevated expression of *NLRC4* was independently associated with ischemic stroke. In support of these findings a recent gene knockout study also reported an association between *NLRC4* and stroke (Denes *et al.*, 2015). The study observed that *NLRC4*$^{-/-}$ mice had reduced neurologic injury as compared with *NLRP3*$^{-/-}$ or wild type mice. Although the NLRC4 inflammasome has been regarded as a sensor of pathogenic bacteria via NAIP co-receptors (Franchi *et al.*, 2012; von Moltke *et al.*, 2013), NLRC4 may also have an active role in sterile inflammation. NLRC4 may be a sensor of DAMPs

released following brain inflammation. NLRC4 can also be activated by phosphorylation (Qu *et al.*, 2012). Therefore an alternative method of NLRC4 activation during stroke may be through post-translational modification. These results give insight into the regulation of NLRC4 and may trigger new research into the role of NLRC4 during stroke. With additional research, NLRC4 may prove as a useful therapeutic target to regulate post-stroke inflammation.

6.6.3 CARDIAC METABOLISM AND STROKE

AF is characterized by irregular atrial contractions that cause blood to pool in the ventricles thus increasing the likelihood of clot formation and cardioembolic stroke. Chapter 5 demonstrated that expression of *SLC25A20* and *PDK4* were elevated during AF. *SLC25A20* encodes carnitine-acylcarnitine translocase (CACT), which is responsible for fatty acid transport into the inner mitochondrial membrane for β-oxidation (Pande, 1975; Ramsay and Tubbs, 1975). PDK4 inhibits the pyruvate dehydrogenase complex and thus acetyl-CoA formation for glycolysis (Patel and Korotchkina, 2006). Elevated expression of both genes indicates increased fatty acid oxidation and decreased glucose metabolism in response to the increased metabolic demands of the heart during AF episodes. *SLC25A20* and *PDK4* expression may have utility as AF biomarkers. However additional studies are required to determine the specificity of *SLC25A20* and *PDK4* as markers of cardiac stress. Although RNA biomarkers are not necessary for detection of permanent or persistent AF, they may have utility for discrimination of cardioembolic stroke or reclassification of cryptogenic stroke cases.

Stroke prognosis, recurrence risk and secondary prevention differ based on stroke etiology. However 22 to 35% of ischemic stroke cases are deemed cryptogenic, also known as stroke with undetermined cause (Petty *et al.*, 2000; Grau *et al.*, 2001; Kolominsky-Rabas *et al.*, 2001; Bang *et al.*, 2003). Randomized clinical trials (RCTs) assessing secondary prevention have not been conducted for cryptogenic stroke patients and as result there are currently no standardized treatment guidelines (European Stroke Organisation (ESO) Executive Committee, 2008; Furie *et al.*, 2011). Consequently, it is not surprising that cryptogenic stroke patients have the second highest rate of stroke recurrence (Petty *et al.*, 2000; Grau *et al.*, 2001; Kolominsky-Rabas *et al.*, 2001; Bang *et al.*, 2003).

Recent literature suggests that the underlying cause of stroke in a subgroup of cryptogenic patients, may be low-risk cardioembolic sources such as: paroxysmal AF, mitral annular calcification, aortic valve stenosis, patent foramen ovale (PFO), and atrial septal defects (Hart *et al.*, 2014). These low-risk cardioembolic sources may increase metabolic demands on the heart. Since elevated *SLC25A20* and *PDK4* may be the result of increased metabolic demands, these RNA biomarkers may facilitate reclassification of cryptogenic stroke cases and in turn improve patient management. Patients with cryptogenic stroke and low-risk cardioembolic sources may benefit from anticoagulant therapy similar to cardioembolic stroke patients. RCTs will be required to evaluate the utility of anticoagulation therapy to prevent stroke in cryptogenic patients, but RNA biomarkers could facilitate identification of eligible patients.

6.7 CHALLENGES ASSOCIATED WITH IMPLEMENTATION OF ROUTINE RNA BIOMARKER TESTING

Although the study results indicate that RNA biomarkers can improve diagnosis and prognosis of disease, there are a few challenges to overcome in order to implement routine RNA testing.

6.7.1 BIOMARKER SPECIFICITY

Foremost, well-validated and high-quality evidence supporting clinical utility and specificity is required. The studies using INTERSTROKE data are the first, large-scale gene expression studies for stroke, while the AF study is the second whole blood expression study for AF. Each of the studies was conducted for the purpose of biomarker discovery. Large sample sizes with independent validation are crucial to the identification of general biomarkers for heterogeneous disorders. However additional research is required to characterize the clinical specificity of these biomarkers. Elevated expression of *MCEMP1, NLRC4, CKLF* and *HS.546375* was detected in stroke cases as compared with controls, but elevated expression may also be expected in other inflammatory or neurologic conditions. Determining stroke specificity is crucial since the results may be used to support the prescription of life saving or life threatening tPA therapy. Therefore future validation studies including patients with stroke mimics (migraine and seizure), inflammatory disorders (encephalitis), and neurologic conditions (dementia and depression), will be required to determine the specificity of the identified RNA biomarkers for stroke.

Similarly, an association between *SLC25A20* and *PDK4* with AF was identified, but baseline expression must be evaluated in healthy individuals. In addition elevated expression must be assessed in patients with structural cardiac abnormalities that may also have increased metabolic demands (PFO and atrial septal defects). Unlike stroke, all patients with cardiac abnormalities may benefit from the prescription of novel anticoagulants. However, identifying specific biomarkers of AF may be useful to identify patients with intermittent AF episodes, predict AF recurrence or to quickly determine success after cardioversion.

6.7.2 ADDED VALUE VS COST

Studies are also required to determine the added diagnostic and prognostic value of new biomarkers as compared to commercially available tests. Currently there are no diagnostic tests for stroke and scores to determine stroke prognosis are currently underused. Therefore a simple test for *MCEMP1* may facilitate diagnosis and improve resource allocation regardless of its comparative predictive value. Statistically the studies have shown that biomarker panels are superior to single RNA biomarkers, but the added clinical value of biomarker panels has yet to be determined. Net reclassification index (NRI) was used to compare the discriminative capacity of different models. NRI determines the change in true and false positive assignment between two tests. Given a null hypothesis that a new score does not change the true or false positive rates, a statistically significant NRI indicates that the null hypothesis can be rejected. Therefore statistical significance refers to the likelihood of a chance finding that cannot be

replicated, but does not take into account the magnitude of the difference. NRI provides some information regarding magnitude of improvement; an NRI>0.6 was considered strong improvement, NRI 0.4 considered moderate and NRI 0.2 considered weak. However the magnitude of improvement to confer clinical improvement has not been established. Due to the added cost and complexity of RNA biomarker panels, additional studies are required to determine the clinical value of the modest improvement in discrimination.

AF diagnosis is most commonly made using standard 12-lead ECG, which is effective for the identification of permanent and persistent AF. Conversely, standard 12-lead ECG is poor at identifying paroxysmal AF, intermittent episodes of AF. This limitation is important since the ASSERT study reports that paroxysmal AF is associated with increased risk of stroke (Healey *et al.*, 2012). It has been suggested that long-term monitoring techniques may improve paroxysmal AF detection, but these methods are moderately invasive and cumbersome to patients. NT-proBNP, *SLC25A20* and *PDK4* may be independent biomarkers of AF, but the cost associated with RNA biomarker testing for all possible AF candidates would be exorbitant as compared with the cost of ECG. However elevated expression of the RNA biomarkers may be expected in paroxysmal AF patients since AF increases the metabolic demands of the heart. Thus although biomarkers would not be cost effective for the diagnosis of persistent AF, future studies may assess the utility of RNA biomarkers for detection of paroxysmal AF.

6.7.3 IMPACT ON PATIENT MANAGEMENT

Clinical trials are required to determine whether RNA biomarkers can improve patient management. The studies identify RNA biomarkers and determine their discriminative capacity. However, there is no clinical consensus as to the discriminative capacity necessary for a biomarker to be of value or improve stroke or AF management. Therefore following biomarker discovery, further research is required to ascertain the diagnostic accuracy required for clinical utility. In such studies it is crucial to have a large control population with similar comorbidities, regulate the timing of the test, and to report changes in medical treatments - such as prescription of tPA. In addition, biomarker validation studies should be designed to evaluate the biomarker test that will ultimately be used in the clinic.

6.7.4 POINT-OF-CARE TESTING

The clinical utility of RNA biomarkers is currently limited by the availability of point-of-care RNA testing. Transcriptome-wide microarray expression profiling is an excellent method of biomarker discovery, but the: associated costs, sample preparation time, requirement of multiple pieces of specialized equipment, and need for experienced personnel to interpret the data, prevents this technology from being used for routine diagnostic testing. In Vitro Diagnostic Multivariate Assays (IVDMIAs) are emerging classes of tests that combine multiple markers using an interpretation function to produce a diagnostic, prognostic and/ or predictive value for patients. Unlike microarrays used for biomarker discovery that assess thousands of targets, IDVMIAs probe for a few specific

targets. An example of such a test would be the 21-gene Onctotype DX test (Genomic Health) that is under evaluation for guiding breast cancer treatment (Sparano *et al.*, 2015). Initial results of the clinical trial indicate that the 21-gene IDVMIA provides clinically useful information, however the tests are still costly and labour intensive due to the RNA isolation and cDNA synthesis steps.

Diagnostic tests that bypass RNA isolation and sample preparation would have the greatest clinical utility and are currently being developed by DxTerity Diagnostics (DxTerity, n.d.) and Luminex (Luminex, n.d.). The DxTerity DxDirect assays utilize chemical ligation dependent probe amplification chemistry (CLPA) to conduct direct-from blood testing of up to 40 RNA transcripts in a single tube. Quantification is relative to one or more housekeeping genes in the assay, similar to qPCR. The Luminex QuantiGenePlex assay also facilitates direct-from blood testing by combining branched DNA signal amplification technology and multi-analyte profiling xMAP beads. Both methods measure RNA at the sample source, which eliminates the variation introduced by traditional RNA extraction and amplification techniques. Currently, DxDirect and QuantiGenePlex are used only for research purpose. However further validation and proof of clinical utility may one day lead to routine use of these technologies and RNA biomarkers.

6.8 CONCLUSION

Peripheral blood RNA is emerging as a new class of biomarkers. This thesis has 1) identified novel RNA biomarkers, 2) determined that RNA biomarkers improve

discrimination of stroke, AF and stroke prognosis, and 3) has provided insights into the pathogenesis of stroke and AF. Despite successfully identifying RNA biomarkers there are several challenges to overcome in order to translate these initial results to clinical practice. However with advancements in technology and further validation studies, RNA biomarkers have the potential to transform stroke management.

6.10 REFERENCES

Bang, O.Y., Lee, P.H., Joo, S.Y., Lee, J.S., Joo, I.S., and Huh, K. (2003) Frequency and mechanisms of stroke recurrence after cryptogenic stroke. *Ann Neurol* **54**: 227–34.

Barr, T.L., Conley, Y., Ding, J., Dillman, a, Warach, S., Singleton, a, and Matarin, M. (2010) Genomic biomarkers and cellular pathways of ischemic stroke by RNA gene expression profiling. *Neurology* **75**: 1009–14.

Bergsbaken, T., Fink, S.L., and Cookson, B.T. (2009) Pyroptosis: host cell death and inflammation. *Nat Rev Microbiol* **7**: 99–109.

Cassel, S.L., and Sutterwala, F.S. (2010) Sterile inflammatory responses mediated by the NLRP3 inflammasome. *Eur J Immunol* **40**: 607–11.

Denes, A., Coutts, G., Lénárt, N., Cruickshank, S.M., Pelegrin, P., Skinner, J., *et al.* (2015) AIM2 and NLRC4 inflammasomes contribute with ASC to acute brain injury independently of NLRP3. *Proc Natl Acad Sci* **112**: 201419090.

DxTerity http://www.dxterity.com/.

European Stroke Organisation (ESO) Executive Committee (2008) Guidelines for management of ischaemic stroke and transient ischaemic attack 2008. *Cerebrovasc Dis* **25**: 457–507.

Finn, D.F., and Walsh, J.J. (2013) Twenty-first century mast cell stabilizers. *Br J Pharmacol* **170**: 23–37.

Franchi, L., Muñoz-Planillo, R., and Núñez, G. (2012) Sensing and reacting to microbes through the inflammasomes. *Nat Immunol* **13**: 325–332.

Furie, K.L., Kasner, S.E., Adams, R.J., Albers, G.W., Bush, R.L., Fagan, S.C., *et al.* (2011) Guidelines for the prevention of stroke in patients with stroke or transient ischemic attack: a guideline for healthcare professionals from the american heart association/american stroke association. *Stroke* **42**: 227–76.

Grau, a. J., Weimar, C., Buggle, F., Heinrich, a., Goertler, M., Neumaier, S., *et al.* (2001) Risk Factors, Outcome, and Treatment in Subtypes of Ischemic Stroke: The German Stroke Data Bank. *Stroke* **32**: 2559–2566.

Hart, R.G., Diener, H.-C., Coutts, S.B., Easton, J.D., Granger, C.B., O'Donnell, M.J., *et al.* (2014) Embolic strokes of undetermined source: the case for a new clinical construct. *Lancet Neurol* **13**: 429–38.

Healey, J.S., Connolly, S.J., Gold, M.R., Israel, C.W., Gelder, I.C. Van, Capucci, A., *et al.* (2012) Subclinical Atrial Fibrillation and the Risk of Stroke. *N Engl J Med* **366**: 120–129.

Iadecola, C., and Anrather, J. (2011) The immunology of stroke: from mechanisms to translation. *Nat Med* **17**: 796–808.

Jin, R., Yang, G., and Li, G. (2010) Inflammatory mechanisms in ischemic stroke: role of inflammatory cells. *J Leukoc Biol* **87**: 779–89.

Jin, Y., Silverman, a. J., and Vannucci, S.J. (2009) Mast Cells Are Early Responders After Hypoxia-Ischemia in Immature Rat Brain. *Stroke* **40**: 3107–3112.

Kanneganti, T.-D., Lamkanfi, M., and Núñez, G. (2007) Intracellular NOD-like Receptors in Host Defense and Disease. *Immunity* **27**: 549–559.

Kolominsky-Rabas, P.L., Weber, M., Gefeller, O., Neundoerfer, B., and Heuschmann, P.U. (2001) Epidemiology of Ischemic Stroke Subtypes According to TOAST Criteria: Incidence, Recurrence, and Long-Term Survival in Ischemic Stroke Subtypes: A Population-Based Study. *Stroke* **32**: 2735–2740.

Luminex https://www.luminexcorp.com/.

Martinon, F., and Tschopp, J. (2007) Inflammatory caspases and inflammasomes: master switches of inflammation. *Cell Death Differ* **14**: 10–22.

Mattila, O.S., Strbian, D., Saksi, J., Pikkarainen, T.O., Rantanen, V., Tatlisumak, T., and Lindsberg, P.J. (2011) Cerebral mast cells mediate blood-brain barrier disruption in acute experimental ischemic stroke through perivascular gelatinase activation. *Stroke* **42**: 3600–5.

McKittrick, C.M., Lawrence, C.E., and Carswell, H.V.O. (2015) Mast cells promote blood brain barrier breakdown and neutrophil infiltration in a mouse model of focal cerebral ischemia. *J Cereb Blood Flow Metab* **35**: 638–647.

Miao, E. a, Rajan, J. V, and Aderem, A. (2011) Caspase-1-induced pyroptotic cell death. *Immunol Rev* **243**: 206–14.

Moltke, J. von, Ayres, J.S., Kofoed, E.M., Chavarría-Smith, J., and Vance, R.E. (2013) *Recognition of bacteria by inflammasomes.* .

Moore, D.F., Li, H., Jeffries, N., Wright, V., Cooper, R. a, Elkahloun, A., *et al.* (2005) Using peripheral blood mononuclear cells to determine a gene expression profile of acute ischemic stroke: a pilot investigation. *Circulation* **111**: 212–21.

Pande, S. V (1975) A mitochondrial carnitine acylcarnitine translocase system. *Proc Natl Acad Sci U S A* **72**: 883–887.

Patel, M.S., and Korotchkina, L.G. (2006) Regulation of the pyruvate dehydrogenase complex. *Biochem Soc Trans* **34**: 217–222.

Petty, G.W., Brown, R.D., Whisnant, J.P., Sicks, J.D., O'Fallon, W.M., and Wiebers, D.O. (2000) Ischemic Stroke Subtypes : A Population-Based Study of Functional Outcome, Survival, and Recurrence. *Stroke* **31**: 1062–1068.

Qu, Y., Misaghi, S., Izrael-Tomasevic, A., Newton, K., Gilmour, L.L., Lamkanfi, M., *et al.* (2012) Phosphorylation of NLRC4 is critical for inflammasome activation. *Nature* **490**: 539–542.

Ramsay, R.R., and Tubbs, P.K. (1975) The mechanism of fatty acid uptake by heart mitochondria: an acylcarnitine-carnitine exchange. *FEBS Lett* **54**: 21–25.

Rathinam, V. a K., Vanaja, S.K., and Fitzgerald, K. a (2012) Regulation of inflammasome signaling. *Nat Immunol* **13**: 333–332.

Sparano, J.A., Gray, R.J., Makower, D.F., Pritchard, K.I., Albain, K.S., Hayes, D.F., *et al.* (2015) Prospective Validation of a 21-Gene Expression Assay in Breast Cancer. *N Engl J Med* **9**: 960–6.

Strbian, D., Karjalainen-Lindsberg, M.-L., Tatlisumak, T., and Lindsberg, P.J. (2006) Cerebral mast cells regulate early ischemic brain swelling and neutrophil accumulation. *J Cereb Blood Flow Metab* **26**: 605–12.

Strbian, D., Tatlisumak, T., Ramadan, U.A., and Lindsberg, P.J. (2006) Mast cell blocking reduces brain edema and hematoma volume and improves outcome after experimental intracerebral hemorrhage. *J Cereb Blood Flow & Metab* 795–802.

Tang, Y., Xu, H., Du, X., Lit, L., Walker, W., Lu, A., *et al.* (2006) Gene expression in blood changes rapidly in neutrophils and monocytes after ischemic stroke in humans: a microarray study. *J Cereb Blood Flow Metab* **26**: 1089–102.

Wernersson, S., and Pejler, G. (2014) Mast cell secretory granules: armed for battle. *Nat Rev Immunol* **14**: 478–494.

Whiteley, W., Chong, W.L., Sengupta, A., and Sandercock, P. (2009) Blood markers for the prognosis of ischemic stroke: a systematic review. *Stroke* **40**: e380–9.

Yang, F., Wang, Z., Wei, X., Han, H., Meng, X., Zhang, Y., *et al.* (2014) NLRP3 deficiency ameliorates neurovascular damage in experimental ischemic stroke. *J Cereb Blood Flow Metab* **34**: 660–667.

Yang-Wei Fann, D., Lee, S.-Y., Manzanero, S., Tang, S.-C., Gelderblom, M., Chunduri, P., *et al.* (2013) Intravenous immunoglobulin suppresses NLRP1 and NLRP3 inflammasome-mediated neuronal death in ischemic stroke. *Cell Death Dis* **4**: e790.