Gene Expression Profiling of Blood for the Prediction of Ischemic Stroke

Boryana Stamova, PhD; Huichun Xu, MD, PhD; Glen Jackling, MD; Cheryl Bushnell, MD; Yingfang Tian, PhD; Bradley P. Ander, PhD; Xinhua Zhan, MD, PhD; DaZhi Liu, PhD; Renee Turner, PhD; Peter Adamczyk, MD; Jane C. Khoury, PhD; Arthur Piacioli, MD; Edward Jauch, MD; Joseph P. Broderick, MD; Frank R. Sharp, MD

Background and Purpose—A blood-based biomarker of acute ischemic stroke would be of significant value in clinical practice. This study aimed to (1) replicate in a larger cohort our previous study using gene expression profiling to predict ischemic stroke; and (2) refine prediction of ischemic stroke by including control groups relevant to ischemic stroke.

Methods—Patients with ischemic stroke (n=70, 199 samples) were compared with control subjects who were healthy (n=38), had vascular risk factors (n=52), and who had myocardial infarction (n=17). Whole blood was drawn ≤3 hours, 5 hours, and 24 hours after stroke onset and from control subjects. RNA was processed on whole genome microarrays. Genes differentially expressed in ischemic stroke were identified and analyzed for predictive ability to discriminate stroke from control subjects.

Results—The 29 probe sets previously reported predicted a new set of ischemic strokes with 93.5% sensitivity and 89.5% specificity. Sixty- and 46-probe sets differentiated control groups from 3-hour and 24-hour ischemic stroke samples, respectively. A 97-probe set correctly classified 86% of ischemic strokes (3 hour+24 hour), 84% of healthy subjects, 96% of vascular risk factor subjects, and 75% with myocardial infarction.

Conclusions—This study replicated our previously reported gene expression profile in a larger cohort and identified additional genes that discriminate ischemic stroke from relevant control groups. This multigene approach shows potential for a point-of-care test in acute ischemic stroke. (Stroke. 2010;41:2171-2177.)

Key Words: biomarkers ▪ blood ▪ gene expression profiling ▪ ischemia ▪ stroke

Stroke is a leading cause of adult death and disability.1-2 The diagnosis of ischemic stroke (IS) is made with clinical assessment in combination with brain imaging. However, the diagnosis is not always straightforward, particularly in the acute setting where an accurate, inexpensive, and rapid diagnosis is critical to optimally treat patients.

Extensive efforts have been directed toward identifying blood based biomarkers for IS. More than 58 proteins and 7 panels of proteins have been described as biomarkers of IS.3-5 RNA expression profiles in the blood have also been described in IS.6-8 We previously reported a 29-probe set expression profile predictive of IS.4 This profile required validation in a second cohort, which has been done in the current study. We also describe a 97-probe set expression profile that differentiates IS from control subjects who are healthy, have vascular risk factors, and who have myocardial infarction. These profiles represent further refinement of gene expression as a diagnostic tool in patients with acute IS, which could be used to aid in the diagnosis of stroke in the context of clinical information and evaluation.

Materials and Methods

The study had 2 objectives: (1) to demonstrate that the previously identified 29 probes distinguish IS from healthy control subjects in a new cohort; and (2) to identify additional genes that discriminate IS from vascular risk factor (Sex, Age and Variation in Vascular functionality [SAVVY]) control subjects and myocardial infarction (MI) control subjects. Whole blood was drawn from patients with IS (n=70, 199 samples) at ≤3, 5, and 24 hours (3-hour IS, 5-hour IS, 24-hour IS) as part of the Combined approach to Lysis utilizing Eptifibatide And Recombinant tissue-type plasminogen activator (CLEAR) trial9 (NCT00250991 at www.Clinical-Trials.gov). IS subjects were treated with recombinant tissue plasminogen activator with or without eptifibatide after the 3-hour blood sample was obtained. Control subjects included healthy subjects (n=38), sub-

Received April 27, 2010; final revision received June 28, 2010; accepted July 29, 2010.

From the Department of Neurology and the MIND Institute (B.S., H.X., G.J., Y.T., B.P.A., X.Z., D.L., R.T., P.A., F.R.S.), University of California at Davis, Sacramento, Calif; the Womens Health Center of Excellence (C.B.), Wake Forest University Health Sciences, Medical Center Boulevard, Winston-Salem, NC; the Division of Biostatistics and Epidemiology (J.C.K.), Cincinnati Children’s Hospital Medical Center, Cincinnati, Ohio; the Department of Neurology (A.P., J.P.B.), University of Cincinnati, Cincinnati, Ohio; and the Division of Emergency Medicine (E.J.), Medical University of South Carolina, Charleston, SC.

The online-only Data Supplement is available at http://stroke.ahajournals.org/cgi/content/full/STROKEAHA.110.588335/DC1.

© 2010 American Heart Association, Inc.

Stroke is available at http://stroke.ahajournals.org

DOI: 10.1161/STROKEAHA.110.588335
Objective 1

The predictive ability of the 29 previously identified genes was determined using the k-nearest neighbor in PAM (Prediction Analysis of Microarrays).\(^6\) IS and healthy subjects were randomly split in half stratified by group and time-point (for the IS samples) into a training set to develop the prediction algorithm and an independent test (validation) set for evaluating the accuracy of the prediction algorithm.

Results

Subject Demographics

Demographic information is presented in Table 1 (Objective 1) and Table 2 (Objective 2). Age was significantly different between IS and control groups (\(P<0.05\); Tables 1 and 2). Gender was significantly different (\(P<0.05\)) between IS and healthy subjects in the Tang et al study\(^6\) and the current study (Table 1) as well as between IS and vascular risk factor (SAVVY) control subjects from the current study (Table 2). Race was significantly different between IS compared with healthy and MI control subjects (Table 2). Hypertension and diabetes were not significantly different between the groups.

Replication of Tang et al IS Predictors in a Larger Cohort

Due to the different array processing protocols in the study by Tang et al,\(^6\) and the current studies, the following analyses were performed: (1) the prediction algorithm was retrained on the first random half of the new samples (training set) and the performance of the 29 probe sets evaluated in the second half (test/validation set); and (2) the samples used in the Tang et al study\(^6\) and the current study were internal gene-normalized. Overall, 92.9% sensitivity for IS and 94.7% specificity for healthy control subjects with high test set probabilities was achieved (Figure 1; Table 3). The results are similar to the ability of these predictors to classify the previously published patients\(^6\) with 88.9% sensitivity for IS and 100% specificity for healthy control subjects (Table 3). In addition, for comparison purposes to the previous study,\(^6\) Robust Mul-
tichip Averaging normalization and CV (used in the previous study [6]) on our complete set of IS and healthy samples was performed. Similar results were obtained (Supplemental Table I and Supplemental Figure II).

Refinement of Prediction of IS Against Several Different Control Groups

<table>
<thead>
<tr>
<th>Differentiation of IS Patients From Control Subjects</th>
</tr>
</thead>
</table>
| Predictive gene expression signatures were derived individually for each comparison. To discriminate the 3-hour IS group from the healthy (training set), MI (CV set, due to small sample size for MI), and SAVVY (training set) control groups, the PAM classification algorithm derived 17, 31, and 22 predictor probe sets genes, respectively. Putting these genes into PAM to predict the class of the subjects in the test groups yielded 87.9/94.7%, 93.9/88.2%, and 97/100% sensitivity/specificity for 3-hour IS compared with healthy, MI, and SAVVY control samples, respectively (Supplemental Figures III, IV, and V, respectively).

To discriminate the 24-hour IS group from the healthy (training set), MI (CV set, due to small sample size for MI), and SAVVY (training set) control groups, the PAM classification algorithm derived 20, 19, and 9 predictor probe sets genes, respectively. Putting these genes into PAM to predict the class of the subjects in the test groups yielded 90.9/94.7%, 93.9/88.2%, and 97/100% sensitivity/specificity for 24-hour IS compared with healthy, MI, and SAVVY control samples, respectively (Supplemental Figures VI, VII, and VIII, respectively).

Prediction Accuracy of 3-Hour IS Predictors on 3-Hour IS, Healthy, MI, and SAVVY Subjects

Combining the lists of the 3-hour predictors from the individual comparison analyses yielded 60 unique probe sets representing 56 annotated genes. Their prediction probability using PAM on the test set is presented in Figure 2A. The percent correctly predicted samples from PAM as well as the best performing prediction model (SVM) are presented in Table 4. Overall (normalized) accuracy was 91.2%. With

Table 2. Demographic Summary of Current Study Participants: Objective 2—Demographic Summary of IS, Healthy, SAVVY, and MI Study Participants (Current Study)

<table>
<thead>
<tr>
<th></th>
<th>IS</th>
<th>Healthy Control Subjects</th>
<th>MI Control Subjects</th>
<th>Vascular SAVVY Control Subjects</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of subjects</td>
<td>70†</td>
<td>38</td>
<td>17</td>
<td>52</td>
</tr>
<tr>
<td>Mean age, years (SD)</td>
<td>66.8±12.7</td>
<td>45.0±19.8</td>
<td>59.6±12.2</td>
<td>56.2±5.4</td>
</tr>
<tr>
<td>Gender, %</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>57.1</td>
<td>47.4</td>
<td>70.6</td>
<td>32.7</td>
</tr>
<tr>
<td>Female</td>
<td>42.9</td>
<td>52.6</td>
<td>29.4</td>
<td>67.3*</td>
</tr>
<tr>
<td>Race, %</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>8.0</td>
<td>55.3</td>
<td>47.1</td>
<td>86.5</td>
</tr>
<tr>
<td>Black</td>
<td>20.0</td>
<td>15.8</td>
<td>17.6</td>
<td>11.5</td>
</tr>
<tr>
<td>Other</td>
<td>0.0</td>
<td>28.9‡</td>
<td>35.3‡</td>
<td>2.0</td>
</tr>
<tr>
<td>National Institutes of Health Stroke Scale</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>First blood draw (3 hours)</td>
<td>14±7</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Second blood draw (5 hours)</td>
<td>11±8</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Third blood draw (24 hours)</td>
<td>10±8</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
</tbody>
</table>

*Gender distribution significantly different (P<0.05) between IS and SAVVY control subjects.
†N=67 at 3 hours, 66 at 5 hours, 66 at 24 hours. Sixty-one subjects had all 3 time points.
‡Race significantly different (P<0.05) between IS compared with healthy and MI.
N/A indicates not applicable.
Table 3. Validation of the 29 Probe Sets From the Tang et al* *

<table>
<thead>
<tr>
<th>Class Prediction</th>
<th>Study</th>
<th>3 Hours</th>
<th>5 Hours</th>
<th>24 Hours</th>
<th>All Time Points</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS, sensitivity, %</td>
<td>Tang et al, 2006</td>
<td>73.3</td>
<td>93.3</td>
<td>100</td>
<td>88.9</td>
</tr>
<tr>
<td>Healthy, specificity, %</td>
<td>Current study</td>
<td>84.8</td>
<td>97.0</td>
<td>97.0</td>
<td>92.9</td>
</tr>
<tr>
<td></td>
<td>Tang et al, 2006</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>100</td>
</tr>
</tbody>
</table>

*These probe sets were trained on the first half (training set) of the subjects in this study (n=35 IS, n=19 healthy) and then used predict the test set probabilities on a second half of the IS subjects (test set; n=35, 99 samples) and healthy subjects (n=19, 19 samples) in the current study. In addition, the same probe sets were used to predict the test set probabilities on the original subjects in the Tang et al 2006 study.

Sensitivity indicates % correct classification of IS samples; specificity = % correct classification of healthy samples; N/A, not applicable.


Main Biological Function of Biomarkers Described

Using Ingenuity Pathway analysis software (see Supplemental Materials), the coagulation system was the only statistically overrepresented biofunction in the combined 97-probe set list of 3-hour and 24-hour IS predictors. The coagulation genes included coagulation factor V (procaccelerin, labile factor) and thrombomodulin. GO annotations and the complete list of predictors are presented in Supplemental Table III. Less stringent criteria yielded large numbers of genes with many more regulated pathways (not shown).

Discussion

Diagnosis of IS is based on clinical impression combined with brain imaging. However, in the acute setting, brain imaging is not always readily accessible, and clinical evaluation by persons experienced in stroke is not always readily available. In such patients, a blood test could be of use to diagnose IS. Several protein biomarkers have been associated with IS, but in the acute setting, these have not yet shown sufficient sensitivity nor specificity to be clinically useful.3–5

In this study, we show that gene expression profiles could be used as biomarkers of IS, replicated our previous findings, and refined the gene expression signature of IS by including more relevant control groups.

We previously reported a 29-probe set profile that distinguished IS from healthy control subjects.6 When this profile was used to predict a larger cohort of patients in this study, it distinguished IS from healthy subjects with a sensitivity of 92.9% and specificity of 94.7%. This is important in that it represents a validation of the concept that gene expression profiles can identify patients with stroke. Replication of gene expression profiles has been a challenge in the field, in large part due to false discovery associated with performing multiple comparisons. Robust biological responses and careful analyses made it possible to validate this 29-probe set profile in this study.

To obtain more biologically useful predictors of IS, we identified gene profiles that distinguish IS from patients with vascular risk factors and MI. Using the individual group comparisons, we predicted the diagnosis of IS compared with the vascular risk factor group with >95% sensitivity and specificity. Using the individual group comparisons, we differentiated patients with IS from MI with >90% sensitivity and >80% specificity. Biologically, this suggests at least some differences in the immune responses to infarction in the brain and heart.

The 3-hour time point was a focus of most comparisons because this represents the critical time when decisions are made regarding acute therapy such as thrombolysis. Thus, for the development of a point-of-care test, this time period is when gene expression profiles could be of greatest use. With the 60-probe set signature, at the 3-hour time point, we achieved correct classification rates of 85% to 94%, 92% to 96%, 88%, and 68% to 84% for IS, vascular risk factor, MI, and healthy control subjects, respectively. These are approaching clinically useful ranges.

Although RNA profiles were the focus in this study, the identified genes could be used as a guide in the evaluation of
protein biomarkers for IS. Genes for Factor 5 and thrombomodulin were both identified as differentially expressed in IS compared with control subjects. Both of these molecules have also been identified as proteins associated with IS. Many of the other genes we identified have not yet been studied but may represent potential candidates for the development of protein biomarker profiles.

The goal of this study was not to identify all differentially expressed genes between IS and control subjects, but rather identify sets of genes whose patterns of expression may be useful for stroke diagnosis. As a result, these analyses have excluded large numbers of differentially expressed genes that are biologically relevant in IS. These will be the subject of future studies. Limitations of this study include (1) lack of stroke “mimics” in the control groups; (2) lack of validation by quantitative reverse transcription–polymerase chain reaction, which would likely be used for clinical applications; (3) the confounding treatment effects in the 5-hour and 24-hour blood samples from patients with IS; (4) race was not factored in due to different distributions with zero subjects in some of the race categories; (5) age is a confounder that we tried to address by factoring it in analysis of covariance models and by selecting control groups with close age distribution.

Figure 2. PAM prediction accuracy of IS predictors in the current study. Prediction accuracy of the test set using PAM. PAM was used to perform the predictions (k-nearest neighbor, neighbors $n=10$; threshold $=0$). For A–C, the X-axis represents the patient sample number and the Y-axis represents test set probabilities. A sample is considered misclassified if its correct class predicted probability is $<0.5$. The numbers of subjects in the training set were: 3-hour IS $n=34$; 24-hour IS $n=33$; SAVVY vascular controls $n=26$; and MI $n=9$. The numbers of subjects in the test set were: 3-hour IS $n=33$; 24-hour IS $n=33$; SAVVY $n=26$; and MI $n=8$. A, Three-hour IS predictors. The 60-probe set predictors for 3-hour IS (combined from comparisons of 3-hour IS samples to healthy, MI, and SAVVY samples from the training set) were put into PAM to predict the class of the test set subject samples by calculating the probability that they were in a given class. B, Twenty-four-hour IS predictors. The 46-probe set predictors for 24-hour IS (combined from comparisons of 24-hour IS samples to healthy, MI, and SAVVY samples from the training set) were put into PAM to predict the class of the test set subject samples by calculating the probability that they were in a given class. C, Combined 3-hour and 24-hour IS predictors. The 97-probe set predictors for 3-hour IS and 24-hour IS (combined from comparisons of 3-hour IS and 24-hour IS samples to healthy, MI, and SAVVY samples from the training set) were put into PAM to predict the class of the test set subject samples by calculating the probability that they were in a given class.
to the patients with IS; and finally, (6) an analysis of variance of all of the groups combined yielded a significant number of regulated genes. However, these genes were not as predictive. This likely occurred because the PAM derivation of the training set of genes was not optimal, whereas individual group comparisons yielded more predictive genes. In the end, statistical validation was achieved by using our training set of genes to predict an independent test set of samples.

Acknowledgments

We thank the investigators of the Specialized Program for Translational Research in Acute Stroke (SPOTRIAS) Stroke Network involved in the CLEAR trial at the University of Cincinnati for supplying blood samples for analysis. We appreciate the support of the MIND Institute, the Genomics and Expression Resource at the MIND Institute, and the University of California Davis Department of Neurology.

Sources of Funding

This study was supported by National Institutes of Health/National Institute of Neurological Disorders and Stroke grants NS056302 (F.R.S.), PO21040N635110 (J.P.B.), and the American Heart Association Bugher Foundation (F.R.S.). G.J. is a fellow of the Canadian Institutes of Health Research (CIHR). H.X., B.A., and Y.T. are Bugher Fellows. This publication was also made possible by Grant Number UL1 RR024146 from the National Center for Medical Research or the National Institutes of Health.

Disclosures

J.B. is Principal Investigator of the National Institute of Neurological Disorders and Stroke (NINDS)-funded Interventional Management of Stroke (IMS) III Trial, UC SPOTRIAS Center (includes NINDS-funded CLEAR–ER and STOP–IT Clinical Trials), NINDS-funded Familial Intracranial Aneurysm (FIA) Study, and NINDS-funded T–32 Cerebrovascular Fellowship Training Program for Cerebrovascular Disease; is coinvestigator of NINDS-funded Genetic and Environmental Risk Factors for Hemorrhagic Stroke, NINDS-funded “Comparison of Hemorrhagic and Ischemic Strokes Among Blacks and Whites,” and NINDS-funded Insulin Resistance Intervention after Stroke (IRIS) Trial, Carotid Revascularization Endarterectomy Versus Stenting Trial (CREST), Carotid Occlusion Surgery Study (COSS), and SWISS Studies, Genentech Inc. (supplier of alteplase for NINDS-funded CLEAR–ER and IMS III trials). EKOS Corporation supplies catheter devices for ongoing IMS III clinical trials. Concentric Inc supplies devices for the IMS III trial. Johnson and Johnson supplies catheters for the IMS III Study. Schering Plough supplies drug for the NINDS-funded CLEAR–ER Trial. J.B. received honoraria for speaking fees from Oakstone Medical Publishing (honorarium received 3/11/10); and consulting fees for participation in stroke advisory board (by PhotoThera on 2/25/10 and by Genetech Inc. on 4/24/10); consulting fees and honoraria are placed in an educational/research stroke fund in the Department of Neurology. A.P. received a research grant from the NINDS for CLEAR–ER Study and received other research support from Genentech and Schering Plough for the study drug.

References


Table 4. Classification Accuracy (%) of 3-Hour and 24-Hour IS Predictors*

<table>
<thead>
<tr>
<th>Group</th>
<th>PAM</th>
<th>SVM*</th>
<th>PAM</th>
<th>SVM†</th>
<th>PAM</th>
<th>SVM‡</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>85</td>
<td>94</td>
<td>91</td>
<td>94</td>
<td>86</td>
<td>95</td>
</tr>
<tr>
<td>SAVVY</td>
<td>92</td>
<td>96</td>
<td>92</td>
<td>96</td>
<td>96</td>
<td>96</td>
</tr>
<tr>
<td>MI</td>
<td>88</td>
<td>88</td>
<td>63</td>
<td>50</td>
<td>75</td>
<td>75</td>
</tr>
<tr>
<td>Healthy</td>
<td>84</td>
<td>68</td>
<td>89</td>
<td>84</td>
<td>84</td>
<td>68</td>
</tr>
</tbody>
</table>

*Half of the subjects (training set) were used to derive the IS predictors. For the test set prediction accuracy estimate on the second half of the subjects, there were 3-hour IS (n = 33), 24-hour IS (n = 33), healthy (n = 19), SAVVY (n = 26), and MI (n = 8). The 60-probe set 3-hour IS predictors represented the sum of the 3-hour IS comparison to the 3 control groups: healthy (17), SAVVY (22), and MI (31), of which 10 were common to the 3-hour IS versus MI and 3-hour IS versus SAVVY predictors, yielding 60 probe sets. The 46-probe set 24-hour IS predictors represented the sum of the 24-hour IS comparison to the 3 control groups: Healthy (20), SAVVY (9), and MI (17). The 3-hour and 24-hour IS combined predictors represent the sum of the 3-hour IS predictors (60) and 24-hour IS predictors (46) of which 9 were common, yielding 97 probe sets. See Supplementary Materials for complete footnotes for †, ‡ and §.


Gene Expression Profiling of Blood for the Prediction of Ischemic Stroke
Boryana Stamova, Huichun Xu, Glen Jickling, Cheryl Bushnell, Yingfang Tian, Bradley P. Ander, Xinhua Zhan, DaZhi Liu, Renee Turner, Peter Adamczyk, Jane C. Khoury, Arthur Pancioli, Edward Jauch, Joseph P. Broderick and Frank R. Sharp

Stroke. 2010;41:2171-2177; originally published online August 26, 2010; doi: 10.1161/STROKEAHA.110.588335

The online version of this article, along with updated information and services, is located on the World Wide Web at:
http://stroke.ahajournals.org/content/41/10/2171

Data Supplement (unedited) at:
http://stroke.ahajournals.org/content/suppl/2010/08/30/STROKEAHA.110.588335.DC1

Permissions: Requests for permissions to reproduce figures, tables, or portions of articles originally published in Stroke can be obtained via RightsLink, a service of the Copyright Clearance Center, not the Editorial Office. Once the online version of the published article for which permission is being requested is located, click Request Permissions in the middle column of the Web page under Services. Further information about this process is available in the Permissions and Rights Question and Answer document.

Reprints: Information about reprints can be found online at:
http://www.lww.com/reprints

Subscriptions: Information about subscribing to Stroke is online at:
http://stroke.ahajournals.org/subscriptions/
Supplementary Material – Materials and Methods

Study Participants

Ischemic Stroke (IS) Patients

Participants with acute IS (n=68) were recruited from the CLEAR trial, a multicenter, randomized double blind safety study of recombinant tissue-plasminogen activator (r-tPA) and eptifibatide as previously described [1] (NCT00250991 at Clinical-Trials.gov). Blood samples were collected at ≤ 3 hours (3h IS), 5 hours (5hr IS) and 24 hours (24 IS) following ischemic stroke onset. r-tPA, with or without eptifibatide, was administered following the 3h blood draw. IS was diagnosed by a stroke neurologist with access to all clinical and diagnostic tests including neurovascular imaging data.

Control Groups

Vascular Risk Factor Subjects (SAVVY)

Subjects with at least one cardiovascular risk factor (hypertension, diabetes mellitus, hyperlipidemia, or tobacco smoking) were recruited from the SAVVY (Sex, Age and Variation in Vascular functionality) study (n=52). These subjects are referred to as vascular risk factor SAVVY Controls in the current study. Exclusion criteria were past history of cardiovascular disease (including stroke, coronary artery disease, peripheral artery disease or deep vein thrombosis), BMI > 46kg/m², history of cancer, chronic infection, autoimmune disease or blood dyscrasias.

Patients with Myocardial Infarction (MI)

Subjects with MI (n=16) were recruited from the University of California Davis Medical Center. The average time since the event was 58.0h (range 19.3-176.5). All were treated acutely with anti-platelet drugs and an anticoagulant prior to the blood draw. Angioplasty (n=8)
or CABG (n=1) were performed in some of the patients prior to the blood draw. No MI patient received r-tPA.

**Healthy Controls**

Healthy controls were recruited from the University of Cincinnati (n=15), UC Davis (n=3) and Stanford (n=20). These subjects had never been hospitalized, were on no medications, and had no known major medical, surgical or psychiatric diseases.

Baseline demographic data were compared between the previous [2] and current study as well as between current IS and control subjects using Student’s 2-tail t-test for continuous variables (age) and a \( \chi^2 \) or Fisher Exact tests for categorical variables (gender, race).

**Probe-level Data Analysis**

Raw expression values of each probe from the Affymetrix U133 Plus 2.0 expression arrays were collapsed into probe set level data using Robust Multichip Averaging (RMA) normalization[3], as well as by modified internal-gene normalization (manuscript in preparation) to a subset of stably expressed internal genes [4]. This involved Median Polishing summarization step, division of each individual gene expression value by the geometric mean of the reference genes, and log2-transformation. For the analysis in Objective 1, both RMA and Internal control gene normalized values were used. For all the analysis of Objective 2, the derivation of the discriminatory genes was performed using the internal control gene normalized values. The same values were used in developing the Classifiers.

**Batch Correction**

Due to the unbalanced nature of the batches, bias is introduced when batch is used as a factor in an ANCOVA model. However, it is still desirable to account for the existing technical variation. This was accomplished by selecting genes that were common to the ANCOVA output sets with and without batch as a factor. While this technique introduced strict criteria for the
selection of discriminating genes, it was intended to improve the chance of validation of the results upon subsequent studies and to achieve greater generalization, which can be translated into IS predictive clinical test.

**Identification of Discriminatory Genes**

Analysis of each comparison (IS per time-point (3h and 24h) vs Healthy, MI and SAVVY, respectively) was performed individually. The samples were randomly split, stratified by Group, in order to perform a split-sample analysis, where the Prediction Algorithms are trained on half of the samples (Training Set), and the performance of the Classifiers is tested on the second half of the samples (Test Set). The Analysis Workflow Chart is shown in Supplementary Materials Figure 1. The feature selection for the derivation of the discriminatory genes between Healthy and IS at 3h and IS at 24h, respectively, involved finding common probe sets from four different ANCOVA analysis, referred to here as Models 1-4. All factors used in the analysis were common to all models (Group, Age, Gender) with the exception of Batch, which was only factored in Model 1 and 3. Models 1 and 2 were applied to a randomly selected one-half of the samples stratified by Group and time-point (for the IS samples) named here 1st random half, whereas Models 3 and 4 were applied to the complete data sets. Overlap of models with and without batch was performed due to the unbalanced nature of batches in an attempt to select more reliable probe sets. Overlap of complete-set and split-set models was performed to achieve greater generalization compared to the split set model which can be translated into IS predictive clinical test.

Gene lists satisfying the following criteria were developed: FDR-corrected p-value (Group) ≤0.05 and fold-change ≤-1.5 or >1.5, as well as being not-significant for the rest of the factors (uncorrected p (Age)>0.5 and uncorrected p (Gender)>0.05 and, for the models including Batch, uncorrected p (Batch) >0.05). The goal is to find genes whose expressions are not affected by significant technical (batch), gender, or age effects.
Exception to Flow Chart Analysis for IS at 24h vs Healthy was at Model 1, where the uncorrected p (Group) < 0.01 was used to generate a larger gene list. Analysis of SAVVY vs IS at 3h and IS at 24h, respectively, included only Models 2 and 4, since Batch could not be factored in, due to the complete confounding of the batches. Analysis of MI vs IS at 3h and IS at 24h, respectively, included only Models 3 and 4, since the sample size of the MI patients was very small (n=17). In this case a 10-fold cross-validation procedure was used to determine the performance of the Classification Algorithms. If the number of the probe sets at the feature selection step was large, we proceeded with excluding probe sets not annotated, annotated as chromosomal segments, annotated as hypothetical proteins, probe sets which per Affimetrix annotation may potentially detect more than one unique gene (*_x_at, *_a_at, *_s_at), and exclusion of duplicates.

Predictions/Classification

Different prediction algorithms were used. Prediction Analysis of Microarrays (PAM) uses the K-nearest neighbor as a classification engine (default k=10) as well as nearest shrunken centroid as a feature-selection method [5]. The differentially expressed genes that passed the criteria outlined above were input into PAM and the minimum numbers of genes with the optimal classification accuracy were selected. In addition, multiple other classification methods were evaluated in the analysis of the combined 3h IS predictors, 24h IS predictors and 3h plus 24h IS predictors in order to find an optimal model and to produce an unbiased estimate of prediction accuracy (analysis performed in Partek Genomics Suite, Partek Inc., St. Louis, MI, USA). We used a combination of the ANCOVA models and nearest-shrunken centroids for our feature reduction step. In addition to PAM, the classification models used in this study were K-Nearest Neighbor (K-NN) with k = 1, 3, 5, 7, and 9 number of neighbors with Euclidian Distance similarity measure; Nearest-Centroid (NC) with equal and proportional prior probabilities;
Quadratic Discriminant Analysis (QDA) with equal and proportional prior probabilities, Linear Discriminant Analysis (LDA) with equal and proportional prior probabilities, and Support Vector Machine, constituting a 121-model space. For overview of these methods, see [6, 7]. 2-level nested cross-validation (CV) was performed to generate a less biased estimate of classification success (reported as accuracy (normalized) estimate). In this approach, an “outer” cross-validation is performed in order to produce an unbiased estimate of prediction error (by holding out samples as an independent test set). To select the optimal model to be applied to the held out test sample, additional “inner” cross-validation is performed on the training data (which is the data not held out as test data by the “outer” cross-validation). Full leave-one-out cross validation (CV) was used in cases where the complete set was used to train and CV the prediction accuracy.

For Table 4 in the Results section, the following parameters were used: †Accuracy (normalized) estimate of 121-Model Space=91.2% (80.3/88). Best Model: SVM (shrink=yes, cost=101, nu=0.5, tol=0.001, kern rbf deg = 3, radial basis function (gamma) = 0.01, coef=0.0). Kappa =0.83. ‡Accuracy (normalized) estimate of 121-Model Space=87.9% (76.4/87). Best Model: SVM (shrink=yes, cost=101, nu=0.5, tol=0.001, kern rbf deg = 3, radial basis function (gamma) = 0.0001, coef=0.0). Kappa=0.83. §Accuracy (normalized) estimate of 121-Model Space=91.2% (110/121). Best Model: SVM (shrink=yes, cost=701, nu=0.5, tol=0.001, kern rbf deg = 3, radial basis function (gamma) = 0.00001, coef=0.0). ¶Correct classification at 3h=76%, at 24h=97%. ‖Correct classification at 3h=94%, at 24h=97%.
Gene Enrichment Analysis of Discriminatory Genes to Identify Biological Themes in the Combined 3h and 24h IS Predictors

Ingenuity Pathway Analysis (IPA 8.0, Ingenuity® Systems) was used for identifying over-represented biological functions in the combined 97 probe set list of 3h and 24h predictors. A Fisher’s exact test (p<0.1) was used to determine whether there was over representation of the 97 probe sets/genes in any given biological function. Gene ontology of the stroke predictors was extracted from Affymetrix NetAffix website (https://www.affymetrix.com/user/login.jsp?toURL=/analysis/netaffx/index.affx).
SUPPLEMENTARY MATERIALS

Supplementary Figure 1. Diagram of the analysis work flow for the identification of IS predictors.

Supplementary Figure 2. PAM prediction accuracy of IS and healthy using the 29 probe set predictors of IS from Tang et al, 2006. The internal gene normalized expression values of all IS (n=70, 199 samples) and healthy (n=38) for the 29 IS predictors from Tang et al, 2006 were used as input in PAM. K-NN (number of neighbors n=10) threshold =0 (including all 29 predictors, and a 10-fold cross-validation was used to estimate prediction accuracy. X-axis represents sample number and the Y-axis represents cross-validated probability of diagnosis. A sample is considered misclassified if the predicted class does not match the known class with a probability greater than 0.5.

Supplementary Figure 3. PAM 3h vs. Healthy test set + test set confusion matrix

Supplementary Figure 4. PAM 3h vs. MI CV + CV confusion matrix

Supplementary Figure 5. PAM 3h vs. SAVVY test set + test set confusion matrix

Supplementary Figure 6. PAM 24h vs. healthy test set + test set confusion matrix

Supplementary Figure 7. PAM 24h vs. MI CV + CV confusion matrix

Supplementary Figure 8. PAM 24h vs. SAVVY test set + test set confusion matrix

Supplementary Figure 9. PAM on Combined 3h, 24h and 3+24h IS predictors. CV Probabilities.

Figure 9A. 3h IS predictors. Combined 60-probe set predictors from combined analysis on 3h IS vs all controls (healthy, MI and SAVVY) were input in PAM.

Figure 9B. 24h IS predictors. Combined 46-probe set predictors from combined analysis on 24h IS vs all controls (healthy, MI and SAVVY) were input in PAM.

Figure 9C. Combined 3h and 24h IS predictors. Combined 97-probe set predictors from combined analysis on 3h IS and 24h IS vs all controls (healthy, MI and SAVVY) were input in PAM.
Supplementary Table 1. Validation of the 29 probe sets from the Tang et al, 2006 study [2]. Cross-validated Probabilities. Trained and cross-validated on current study samples (IS: n = 70, 199 samples) and Healthy (n=38, 38 samples).

<table>
<thead>
<tr>
<th>Normalization Method</th>
<th>Class Prediction</th>
<th>Study</th>
<th>3h</th>
<th>5h</th>
<th>24h</th>
<th>All Time Points</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMA</td>
<td>IS, Sensitivity, %</td>
<td>Tang et al, 2006</td>
<td>66.7</td>
<td>86.7</td>
<td>100</td>
<td>84.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Current Study</td>
<td>86.6</td>
<td>98.5</td>
<td>89.4</td>
<td>91.5</td>
</tr>
<tr>
<td>Internal Genes</td>
<td>Healthy, Specificity, %</td>
<td>Tang et al, 2006</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Current Study</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>84.2</td>
</tr>
<tr>
<td>IS, Sensitivity, %</td>
<td></td>
<td>Tang et al, 2006</td>
<td>73.3</td>
<td>93.3</td>
<td>100</td>
<td>88.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Current Study</td>
<td>86.6</td>
<td>98.5</td>
<td>95.5</td>
<td>93.5</td>
</tr>
<tr>
<td>Healthy, Specificity, %</td>
<td></td>
<td>Tang et al, 2006</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Current Study</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>89.5</td>
</tr>
</tbody>
</table>

Sensitivity = % correct classification of IS samples

Specificity = % correct classification of healthy samples
Supplementary Table 2. Classification Accuracy (% correct classification) of 3h and 24h Ischemic Stroke (IS) Predictors. Sample sizes used for Cross-Validation were n=67 at 3h IS, n=66 at 24h IS, n=52 for SAVVY, n=17 for MI. Sample sizes used for split-sample prediction performance estimate on the test set were n=33 at 3h IS, n=33 at 24h IS, n=26 for SAVVY, n=8 for MI. The 60–probe set 3h IS predictors represented the sum of the 3h IS comparison to the three control groups: Healthy (17 probe sets), SAVVY controls (22 probe sets) and MI (31 probe sets). The 46–probe set 24h IS predictors represented the sum of the 24h IS comparison to the three control groups: Healthy (20 probe sets), SAVVY controls (9 probe sets) and MI (17 probe sets). The 3h and 24h IS Combined predictors represent the sum of the 3h IS predictors (60 probe sets) and 24h IS predictors (n=46) of which 9 were common, thus yielding 97 probe sets.

*Accuracy (normalized) estimate of 121-Model Space=86.4% (150/174). Best Model: SVM

†Accuracy (normalized) estimate of 121-Model Space=89.2% (154/173). Best Model: SVM

‡Accuracy (normalized) estimate of 121-Model Space=88.2% (212/240). Best Model: SVM

Correct classification at 3h=87%, at 24h=96%

Supplementary Table 3. Annotations for the combined 3h and 24h IS predictors – in a separate Excel file.
Supplementary Material References


Supplementary Materials Figure 1
Analysis Workflow
Supplementary Materials Figure 2
29-probe set from Tang et al, 2006 study on our IS and Healthy Subjects. Cross-Validated Probability.
Supplementary Materials Figure 3
Ischemic Stroke at 3h versus Healthy

Test set Prediction Confusion Matrix (Threshold=0)

<table>
<thead>
<tr>
<th>True\Predicted</th>
<th>Healthy</th>
<th>IS_3h</th>
<th>Correct Classification, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Healthy</td>
<td>18</td>
<td>1</td>
<td>94.73</td>
</tr>
<tr>
<td>IS_3h</td>
<td>4</td>
<td>29</td>
<td>87.9</td>
</tr>
</tbody>
</table>
Supplementary Materials Figure 4
Ischemic Stroke at 3h versus Myocardial Infarction

Cross-validated Probabilities (Threshold=3.2349)

CV Confusion Matrix (Threshold=3.23495)

<table>
<thead>
<tr>
<th></th>
<th>True\Predicted</th>
<th>IS_3h</th>
<th>MI</th>
<th>Correct Classification, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS_3h</td>
<td>66</td>
<td>1</td>
<td></td>
<td>98.5</td>
</tr>
<tr>
<td>MI</td>
<td>3</td>
<td>14</td>
<td></td>
<td>82.4</td>
</tr>
</tbody>
</table>
Supplementary Materials Figure 5
Ischemic Stroke at 3h versus SAVVY

Test Probabilities (Threshold=4.948)

Test set Prediction Confusion Matrix (Threshold=4.948)

<table>
<thead>
<tr>
<th>True\Predicted</th>
<th>IS_3h</th>
<th>SAVVY</th>
<th>Correct Classification, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS_3h</td>
<td>33</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>SAVVY</td>
<td>1</td>
<td>25</td>
<td>96.2</td>
</tr>
</tbody>
</table>
Supplementary Materials Figure 6
Ischemic Stroke at 24h versus Healthy

Test set Prediction Confusion Matrix (Threshold=0)

<table>
<thead>
<tr>
<th>True\Predicted</th>
<th>Healthy</th>
<th>IS_24h</th>
<th>Correct Classification, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Healthy</td>
<td>18</td>
<td>1</td>
<td>94.7</td>
</tr>
<tr>
<td>IS_24h</td>
<td>3</td>
<td>30</td>
<td>90.9</td>
</tr>
</tbody>
</table>
Supplementary Materials Figure 7
Ischemic Stroke at 24h versus Myocardial Infarction

Cross-validated Probabilities (Threshold=3)

<table>
<thead>
<tr>
<th>IS_24h</th>
<th>MI</th>
</tr>
</thead>
<tbody>
<tr>
<td>62</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>15</td>
</tr>
</tbody>
</table>

Correct Classification, %
- IS_24h: 93.9%
- MI: 88.2%
Supplementary Materials Figure 8
Ischemic Stroke at 24h versus SAVVY

Test Probabilities (Threshold=6.1803)

Test set Prediction Confusion Matrix (Threshold=6.1803)

<table>
<thead>
<tr>
<th>True\Predicted</th>
<th>IS_24h</th>
<th>SAVVY</th>
<th>Correct Classification, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS_24h</td>
<td>32</td>
<td>1</td>
<td>97</td>
</tr>
<tr>
<td>SAVVY</td>
<td>0</td>
<td>26</td>
<td>100</td>
</tr>
</tbody>
</table>
Supplementary Materials Figure 9. Cross Validated Probabilities of
A. 3h IS predictors.
B. 24 IS predictors.
C. Combined 3h and 24h IS predictors.