Genomics Data 14 (2017) 47–52

Contents lists available at ScienceDirect

Genomics Data

journal homepage: www.elsevier.com/locate/gdata

Stroke-associated pattern of gene expression previously identified by machine-learning is diagnostically robust in an independent patient population

Grant C. O'Connell[a,b,*], Paul D. Chantler[c,d], Taura L. Barr[e]

[a] Center for Basic and Translational Stroke Research, Robert C. Byrd Health Sciences Center, West Virginia University, Morgantown, WV, United States
[b] Department of Pharmaceutical Sciences, School of Pharmacy, West Virginia University, Morgantown, WV, United States
[c] Center for Cardiovascular and Respiratory Sciences, Robert C. Byrd Health Sciences Center, West Virginia University, Morgantown, WV, United States
[d] Division of Exercise Physiology, School of Medicine, West Virginia University, Morgantown, WV, United States
[e] Valtari Bio Incorporated, Morgantown, WV, United States

ARTICLE INFO

Keywords:
GA/kNN
Genetic algorithm
Triage
Biomarker
Immunology
Pattern recognition
Cerebrovascular disease
Brain injury

ABSTRACT

Our group recently employed genome-wide transcriptional profiling in tandem with machine-learning based analysis to identify a ten-gene pattern of differential expression in peripheral blood which may have utility for detection of stroke. The objective of this study was to assess the diagnostic capacity and temporal stability of this stroke-associated transcriptional signature in an independent patient population. Publicly available whole blood microarray data generated from 23 ischemic stroke patients at 3, 5, and 24 h post-symptom onset, as well from 23 cardiovascular disease controls, were obtained via the National Center for Biotechnology Information Gene Expression Omnibus. Expression levels of the ten candidate genes (*ANTXR2*, *STK3*, *PDK4*, *CD163*, *MAL*, *GRAP*, *ID3*, *CTSZ*, *KIF1B*, and *PLXDC2*) were extracted, compared between groups, and evaluated for their discriminatory ability at each time point. We observed a largely identical pattern of differential expression between stroke patients and controls across the ten candidate genes as reported in our prior work. Furthermore, the coordinate expression levels of the ten candidate genes were able to discriminate between stroke patients and controls with levels of sensitivity and specificity upwards of 90% across all three time points. These findings confirm the diagnostic robustness of the previously identified pattern of differential expression in an independent patient population, and further suggest that it is temporally stable over the first 24 h of stroke pathology.

1. Introduction

The ability of clinicians to confidently recognize stroke during triage increases access to interventional treatments and affords patients improved odds for favorable outcome [1,2]. However, the diagnostic tools currently available to emergency medical technicians, paramedics, and hospital staff for identification of stroke have significant limitations [3,4]. Biomarker-based tests are clinically used to aid in the diagnosis of acute cardiovascular conditions such as myocardial infarction [5], however no such assay currently exists for the detection of stroke. This diagnostic limitation has resulted in a push for the identification of peripheral blood stroke biomarkers which could be rapidly measured in either the field or emergency department to guide early triage decisions [3,6].

Our group recently employed high-throughput transcriptomics in combination with a machine learning technique known as genetic algorithm/k-nearest neighbors (GA/kNN) to identify a panel of ten candidate genes whose peripheral blood expression levels were able to differentiate between 78 ischemic stroke patients and 74 control subjects with a high degree of accuracy [7]. These candidate genes include seven whose expression levels were elevated in stroke patients relative to controls (*CD163, ANTXR2, PDK4, PLXDC2, STK3, ID3, CTSZ, KIF1B*), and three whose expression levels were down regulated (*MAL, ID3, GRAP*); their coordinate pattern of differential expression was able to discriminate between groups with levels of sensitivity and specificity approaching 100%. While the levels of diagnostic performance observed in this discovery investigation were unprecedented, limitations in study design necessitate further evaluation of the candidate genes in a validation analysis before definitive conclusions can be made regarding their true diagnostic efficacy.

* Corresponding author at: West Virginia University, Robert C. Byrd Health Sciences Center, One Medical Center Drive, Morgantown, WV 26505, United States.
E-mail address: goconnell.wvu@gmail.com (G.C. O'Connell).

http://dx.doi.org/10.1016/j.gdata.2017.08.006
Received 22 March 2017; Received in revised form 1 August 2017
Available online 01 September 2017

Stroke patients and control subjects in this discovery investigation were not well matched in terms of cardiovascular disease (CVD) risk factors, leaving open the possibility that the pattern of differential expression which we observed across the ten candidate genes was driven by underlying CVD, and not by the acute event of stroke itself. Furthermore, subjects in this discovery study were almost exclusively Caucasian, and it is currently unknown whether ethnicity impacts the diagnostic efficacy the candidate genes, a possibility which deserves consideration due to the fact that there can be notable inter-ethnic differences in the pathophysiology of cardiovascular conditions [8–11]. A further limitation in of this discovery study was the fact that blood samples were only collected at a single time point, making the temporal stability of candidate gene differential expression unclear with regards to the progression of stroke pathology. While post hoc statistical analyses were used to address these potential confounds as best possible, it would be reassuring to observe similar levels of diagnostic performance across multiple time points in a more ethnically diverse subject pool which is better matched in terms of CVD risk factors.

Stamova et al. recently used microarray to examine gender differences in the response of the peripheral immune system to stroke [12]. This investigation produced a publicly available data set which includes genome-wide whole blood expression data generated from 23 cardioembolic ischemic stroke patients at three replicate time points post-symptom onset (3, 5, and 24 h), as well as from 23 neurologically asymptomatic control subjects; this patient population was ethnically diverse and groups were well matched in terms of risk factors for CVD. In the study reported here, we assessed the diagnostic robustness of the ten previously identified candidate genes in the aforementioned publicly available data set.

2. Methods

2.1. Data procurement

Raw whole blood microarray data (Affymetrix Human Genome U133 Plus 2.0 Array) were downloaded as .CEL files from the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) via accession number GSE58294 (Supplementary File 1). Patient clinical and demographic characteristics were aggregated from the gender-wise information reported by Stamova et al. [12].

2.2. Microarray analysis

Analysis of microarray data was performed using the 'affy' package for R (R project for statistical computing) [13,14]. Raw perfect match probe intensities were background corrected, quantile normalized (Fig. 1), and summarized at the set level via robust multi-array averaging using the rma() function [15]. Probe set level data associated with the ten candidate genes were then extracted for differential expression analysis; in the case of candidate genes with more than one associated probe set, data were further summarized at the gene level via simple averaging. Gene level summarized expression levels were then compared between stroke patients and controls across all three post-onset time points.

2.3. Diagnostic evaluation

The diagnostic robustness of candidate gene expression levels was tested in terms of their ability to discriminate between stroke patients and controls using k-nearest neighbors (kNN) at each time point post-symptom onset. Classification was performed using standardized expression values, five nearest neighbors, and majority rule via the knn.cv() function of the 'class' package for R [16]. Same-set leave one out cross-validation was performed, and the resultant prediction probabilities were used to generate receiver operator characteristic (ROC) curves using the roc() function of the 'pROC' package for R [17]. Areas under curves were then compared between time points via the roc.test() function according the non-parametric method described by DeLong et al. [18].

2.4. Statistics

All statistics were performed using R 3.3. Fisher's exact test was used for comparison of dichotomous variables. *t*-Test or one-way ANOVA was used for comparisons of continuous variables where appropriate. The null hypothesis was rejected when $p < 0.05$. In the case of multiple comparisons, *p*-values were false discovery rate adjusted using the Benjamini-Hochberg procedure [19].

3. Results

3.1. Clinical and demographic characteristics

Stroke patients were significantly older than control patients, but well matched in terms of gender and ethnicity. In terms of cardiovascular disease risk factors, groups were well matched with regards to rates of hypertension and diabetes, however control subjects displayed a significantly higher prevalence of dyslipidemia relative to stroke patients. All stroke patients received thrombolytic intervention via recombinant tissue plasminogen activator (rtPA) following 3 h blood collection, but prior to 5 h blood collection (Table 1).

3.2. Microarray data processing

Distributions of perfect match probe intensities were visually similar following normalization, providing indication that normalized expression data were suitable for inter-sample comparison (Fig. 1). Probe sets extracted for differential expression analysis are listed in Table 2.

3.3. Candidate gene differential expression

Six of the seven candidate genes which we had previously reported as being elevated in stroke in our prior investigation displayed similar up-regulation in stroke patients relative to controls (Fig. 2A, B, D, E, F, J), however one exhibited no significant differences in expression levels at any time point post-symptom onset (Fig. 2H). In terms of the candidate genes which we had previously reported as being down regulated in stroke, all three displayed significantly lower expression levels in stroke patients relative to controls (Fig. 2C, G, I). Collectively, these observations largely confirmed the pattern of candidate gene differential expression reported in our prior investigation.

3.4. Temporal profile of candidate differential expression

Most candidate genes displayed some degree of differential expression by 3 h post-symptom onset, and the magnitude of the overall response appeared to increase over time. Several candidate genes appeared to achieve maximal differential expression at 5 h post-onset and then plateau, while a few displayed steady increases in the degree of differential expression through 24 h (Fig. 3), providing evidence that the expression levels of the candidate genes are likely directly responsive to acute stroke pathology.

3.5. Candidate gene diagnostic performance

In terms of diagnostic ability, the coordinate expression levels of the ten candidate genes were able to discriminate between stroke patients and controls using kNN with levels of sensitivity and specificity upwards of 90% at all three time points post-symptom onset (Fig. 4A, B, C). While the overall diagnostic capacity of the ten candidate genes appeared slightly more robust at five and 24 h, no statistically significant differences in area under ROC curve were observed between