Integrative analysis of transcriptome-wide association study and gene expression profiling identifies candidate genes associated with stroke

Jian Yang¹,²,*, Bin Yan¹,²,*, Yajuan Fan³, Lihong Yang¹, Binbin Zhao¹, Xiaoyan He³, Qingyan Ma³, Wei Wang³, Ling Bai¹, Feng Zhang⁴ and Xiancang Ma²,³

¹ Clinical Research Center, The First Affiliated Hospital of Xi’an Jiaotong University, Xi’an, China
² Center for Brain Science, The First Affiliated Hospital of Xi’an Jiaotong University, Xi’an, China
³ Department of Psychiatry, The First Affiliated Hospital of Xi’an Jiaotong University, Xi’an, China
⁴ Health Science Center, Xi’an Jiaotong University, Xi’an, China
*These authors contributed equally to this work.

ABSTRACT

Background. Stroke is a major public health burden worldwide. Although genetic variation is known to play a role in the pathogenesis of stroke, the specific pathogenic mechanisms are still unclear. Transcriptome-wide association studies (TWAS) is a powerful approach to prioritize candidate risk genes underlying complex traits. However, this approach has not been applied in stroke.

Methods. We conducted an integrative analysis of TWAS using data from the MEGASTROKE Consortium and gene expression profiling to identify candidate genes for the pathogenesis of stroke. Gene ontology (GO) enrichment analysis was also conducted to detect functional gene sets.

Results. The TWAS identified 515 transcriptome-wide significant tissue-specific genes, among which SLC25A44 (P = 5.46E−10) and LRCH1 (P = 1.54E−6) were significant by Bonferroni test for stroke. After validation with gene expression profiling, 19 unique genes were recognized. GO enrichment analysis identified eight significant GO functional gene sets, including regulation of cell shape (P = 0.0059), face morphogenesis (P = 0.0247), and positive regulation of ATPase activity (P = 0.0256).

Conclusions. Our study identified multiple stroke-associated genes and gene sets, and this analysis provided novel insights into the genetic mechanisms underlying stroke.

INTRODUCTION

Stroke is the second leading cause of death and the third leading cause of disability-adjusted life-years lost worldwide (Feigin, Norrving & Mensah, 2017; Hankey, 2017). Causal risk factors, such as hypertension, hyperlipidemia, carotid stenosis, and atrial fibrillation, account for a substantial proportion of stroke risk (Bath & Krishnan, 2014; Bellolio, Gilmore & Ganti, 2014; Benjamin et al., 2018; ENOS Trial Investigators, 2015). However,
the pathogenesis of stroke is still elusive. Family studies have indicated that genetic factors contribute substantially to the cause of stroke (Jerrard-Dunne et al., 2003). Recently, genome-wide association studies (GWAs) have identified multiple genetic variants associated with stroke (Gudbjartsson et al., 2009; Holliday et al., 2012; International Stroke Genetics Consortium et al., 2012; Kilarski et al., 2014; SiGN & ISGC, 2016). Furthermore, the MEGASTROKE Consortium conducted a large-scale GWAS meta-analysis for stroke with almost all available samples; this analysis identified a number of relevant genetic loci (Malik et al., 2018). Although GWAs have yielded many causal genetic variants, the specific mechanisms underlying this disease have still not been identified.

Transcriptome-wide association studies (TWASs) have been recently proposed as a powerful approach to prioritize candidate risk genes underlying complex traits. Briefly, this approach can be viewed as a tool for predicting trait-associated gene expression based on GWAS summary data (Gusev et al., 2016). A set of individuals for whom both gene expression and genetic variation (single nucleotide polymorphisms (SNPs)) have been measured was used to calculate the cis-genetic component of expression weights. In recent years, TWASs have been widely applied to complex diseases, such as schizophrenia, Parkinson’s disease, and prostate cancer (Gusev et al., 2018; Mancuso et al., 2018; Ratnapriya et al., 2019). Studies have shown that this method is a powerful approach to predict candidate genes by integrating gene expression panels and GWAS summary data. However, much remained to be done to support these results.

Gene expression profiling is an approach to determine the pattern of genes expressed at the level of transcription. DNA microarrays or sequencing technologies have been applied to identify active genes for diseases. Typically, this approach provides experimental information for potential genes and may be a good complement to TWAS analysis. Moreover, the Gene Expression Omnibus (GEO) database has incorporated gene expression and hybridization array data for various diseases, providing a convenient basis for the integrative analysis of TWASs and gene expression profiling data.

Accordingly, in this study, we conducted an integrative analysis of TWASs and gene expression profiling to identify candidate genes associated with stroke. We aimed to combine the prediction approach of the TWAS with the experimental results of gene expression to provide novel insight into determination of the genes related to stroke. Gene ontology (GO) enrichment analysis was also performed to detect the pathways associated with stroke.

**MATERIALS & METHODS**

**GWAS of stroke**

The summary data for stroke were obtained from a large-scale GWAS meta-analysis conducted by the MEGASTROKE Consortium (Malik et al., 2018). All available stroke samples with published or unpublished GWAS data were combined in their essential study. We extracted the summary statistics restricted to European ancestry because the current available reference panels for TWAS were all established on European population. The final dataset was composed of 40,585 cases and 406,111 controls from 17 studies.
Genotype imputation was conducted using 10,000 Genomes Project (1000G) phase 1v3 as the reference panel. After quality control, more than 8 million SNPs and indels with minor-allele frequencies greater than or equal to 0.01 were tested in the fixed-effects meta-analysis. The meta-analysis was carried out using METAL, and only SNPs passing the MEGASTROKE filters (number of cases >50% and imputation INFO score >5) were present in the final summary dataset.

**Performing TWAS on stroke GWAS dataset**

We next performed a TWAS to identify significant expression-trait associations in stroke using FUSION software (http://gusevlab.org/projects/fusion/) (Gusev et al., 2016). Briefly, this approach took GWAS summary statistics on expression reference panels and linkage disequilibrium (LD) reference panels to identify significant expression-trait associations. The GWAS summary statistics of stroke were first converted into LD-score format, and SNPs were matched to the LD reference panel (EUR 1000 Genomes) to filter the part used for prediction. Summary-based imputation was then performed using three expression panels: RNA array data from the peripheral blood of 1,245 unrelated individuals from the Netherlands Twin Registry (NTR), RNA array data from the blood of 1,264 individuals from the Young Finns Study (YFS), and RNA-seq data from the adipose tissue of 563 individuals from the Metabolic Syndrome in Men study (METSIM) (Nuotio et al., 2014; Stancakova et al., 2012; Wright et al., 2014). We imputed the correction between predicted expression and stroke as a linear combination of GWAS z-scores (defined as $Z$) with expression weights (defined as $w$), adjusting with an SNP-correlation (LD) matrix D. Finally, the association between the cis-genetic component of expression and stroke was calculated according to the following formula as $Z_{TWAS} = \frac{w'Z}{(w'Dw)^{1/2}}$. In addition, the summarized association statistic might be inflated from by-chance quantitative trait loci (QTL) co-localization when the GWAS locus is highly significant and LD is extensive. To control for this, we further performed 5000 permutations which shuffled the expression weights and recalculated the empirical association statistic conditional on the GWAS effects at the locus. Genes with empirical permutation $P < 0.05$ were considered unlikely to be disturbed by QTLs.

**Gene expression profiling of stroke**

The gene expression profiling data were obtained from a study by Krug et al. (2012). Krug et al. performed gene expression profile analysis in peripheral blood mononuclear cells from 20 patients who experienced ischemic stroke and 20 sex- and age-matched controls. All participants were European, and ischemic stroke patients were required to have suffered only one stroke episode, at least 6 months before blood collection. Expression microarrays were generated using Affymetrix Human Genome U133 Plus 2.0 Array and extensive quality control were performed in all steps. Multi-factor analysis of variance (ANOVA) was used to identify the differentially expressed genes among cases and controls, considering group, age, sex, the interactions among them, as well as the geographic origin of the participants and the scan-date of the microarrays as factors. The genes with a fold change $>1.2$ and an uncorrected $P$ value $<0.05$ were considered differentially expressed as defined by Krug et
al. (2012). After expression profile analysis, 709 probe sets (representing 580 genes) were identified by the gene expression profile analysis.

**Gene Ontology enrichment analysis**

We combined TWAS results with gene expression profiling results to identify common genes for stroke. The candidate genes were then input into DAVID (https://david.ncifcrf.gov/home.jsp), a widely used tool for GO enrichment analysis. A primary enrichment \( P \) value was calculated for each GO term using DAVID. Significant GO terms were identified at \( P < 0.05 \).

**RESULTS**

The primary results for TWAS analysis were presented in Fig. 1. The distribution presented some inflation \( (\lambda = 1.13) \) of obtained versus expected results under the null hypothesis (Fig. S1). This was probably because the TWAS computed multiple models for each gene and outputted the statistics of the best performing model. In total, 11,826 tissue-specific features of gene expression were tested in the summary-based TWAS approach. After calculation, we identified 515 transcriptome-wide significant gene-stroke features \( (P_{\text{TWAS}} < 0.05 \) and \( P_{\text{permutation}} < 0.05) \) for 446 unique genes, including 58 genes that were significant in more than one panel (Table S1). The top 10 tissue-specific genes are summarized in Table 1. SLC25A44 and LRCH1 were the only two genes passing the Bonferroni correction threshold \( (P < 0.05/11,826) \) for all gene-stroke features, which implied that they were strongly associated with stroke. In order to cover more possible candidate genes, we also reported genes passing the transcriptome-wide significant threshold \( (P < 0.05) \) for integrative analysis.
Table 1  Top 10 significant tissue-specific genes identified by TWAS.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Chromosome</th>
<th>Z-score</th>
<th>( P_{TWAS} )</th>
<th>( P_{permuation} )</th>
<th>Tissue</th>
</tr>
</thead>
<tbody>
<tr>
<td>SLC25A44</td>
<td>1</td>
<td>6.2053</td>
<td>5.46E−10</td>
<td>0.00671</td>
<td>Adipose (METSIM)</td>
</tr>
<tr>
<td>LRCH1</td>
<td>13</td>
<td>−4.8058</td>
<td>1.54E−06</td>
<td>0.01153</td>
<td>Adipose (METSIM)</td>
</tr>
<tr>
<td>FLJ44606</td>
<td>5</td>
<td>4.3540</td>
<td>1.58E−05</td>
<td>0.01150</td>
<td>Peripheral blood (NTR)</td>
</tr>
<tr>
<td>PLEKHA1</td>
<td>10</td>
<td>−4.3171</td>
<td>2.39E−05</td>
<td>0.00499</td>
<td>Adipose (METSIM)</td>
</tr>
<tr>
<td>ZNF318</td>
<td>6</td>
<td>4.2252</td>
<td>7.04E−05</td>
<td>0.01974</td>
<td>Peripheral blood (NTR)</td>
</tr>
<tr>
<td>CD40</td>
<td>20</td>
<td>3.9748</td>
<td>7.67E−05</td>
<td>0.00376</td>
<td>Adipose (METSIM)</td>
</tr>
<tr>
<td>FES</td>
<td>15</td>
<td>−3.9545</td>
<td>8.06E−05</td>
<td>0.00515</td>
<td>Adipose (METSIM)</td>
</tr>
<tr>
<td>RERE</td>
<td>1</td>
<td>−3.9425</td>
<td>9.77E−05</td>
<td>0.00566</td>
<td>Peripheral blood (NTR)</td>
</tr>
<tr>
<td>SLC25A29</td>
<td>14</td>
<td>3.8962</td>
<td>1.15E−04</td>
<td>0.00100</td>
<td>Adipose (METSIM)</td>
</tr>
<tr>
<td>LARS</td>
<td>5</td>
<td>3.8558</td>
<td>1.24E−04</td>
<td>0.02459</td>
<td>Adipose (METSIM)</td>
</tr>
</tbody>
</table>

We then compared transcriptome-wide significant genes identified by TWAS with the differentially expressed genes identified by gene expression profile analysis. Finally, 19 genes within 23 tissue-specific gene-stroke features were detected (Table 2). There were more shared genes (10/19) detected in the expression panel of the YFS with expression profiling. This may be because the expression profiling was performed using tissues similar to those used in YFS. In addition, four candidate genes showed transcriptome-wide significance in more than one panel.

Table 2  List of overlapping candidate genes recognized by the integrative analysis of TWAS and gene expression profiling.

<table>
<thead>
<tr>
<th>Gene</th>
<th>CHR</th>
<th>( P_{TWAS} )</th>
<th>( P_{mRNA} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>CISD1</td>
<td>10</td>
<td>0.00357</td>
<td>0.0055</td>
</tr>
<tr>
<td>RHOU</td>
<td>1</td>
<td>/</td>
<td>/</td>
</tr>
<tr>
<td>TM6SF1</td>
<td>15</td>
<td>0.0136</td>
<td>0.00586</td>
</tr>
<tr>
<td>TP53RK</td>
<td>20</td>
<td>/</td>
<td>0.00083</td>
</tr>
<tr>
<td>CLDN5</td>
<td>22</td>
<td>/</td>
<td>0.00804</td>
</tr>
<tr>
<td>NFKBIA</td>
<td>14</td>
<td>/</td>
<td>0.00957</td>
</tr>
<tr>
<td>MRVI1</td>
<td>11</td>
<td>0.00997</td>
<td>0.0359</td>
</tr>
<tr>
<td>ZNF667-AS1</td>
<td>19</td>
<td>0.01025</td>
<td>0.0019</td>
</tr>
<tr>
<td>WDSUB1</td>
<td>2</td>
<td>/</td>
<td>0.01414</td>
</tr>
<tr>
<td>WSB1</td>
<td>17</td>
<td>0.04353</td>
<td>0.01501</td>
</tr>
<tr>
<td>TPM1</td>
<td>15</td>
<td>/</td>
<td>0.01590</td>
</tr>
<tr>
<td>GOLGA8A</td>
<td>15</td>
<td>/</td>
<td>0.01704</td>
</tr>
<tr>
<td>TIPARP</td>
<td>3</td>
<td>0.02357</td>
<td>0.0095</td>
</tr>
<tr>
<td>VRK2</td>
<td>2</td>
<td>/</td>
<td>0.02953</td>
</tr>
<tr>
<td>TOR1AIP1</td>
<td>1</td>
<td>/</td>
<td>0.0339</td>
</tr>
<tr>
<td>BARD1</td>
<td>2</td>
<td>0.03580</td>
<td>0.04980</td>
</tr>
<tr>
<td>BAG5</td>
<td>14</td>
<td>/</td>
<td>0.03791</td>
</tr>
<tr>
<td>VASH1</td>
<td>14</td>
<td>0.04918</td>
<td>0.03802</td>
</tr>
<tr>
<td>PDIA6</td>
<td>2</td>
<td>/</td>
<td>0.04136</td>
</tr>
</tbody>
</table>

Yang et al. (2019), PeerJ, DOI 10.7717/peerj.7435
<table>
<thead>
<tr>
<th>Category</th>
<th>ID</th>
<th>Name</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>GOTERM_BP_DIRECT</td>
<td>GO:0008360</td>
<td>Regulation of cell shape</td>
<td>0.0059</td>
</tr>
<tr>
<td>GOTERM_BP_DIRECT</td>
<td>GO:0060325</td>
<td>Face morphogenesis</td>
<td>0.0247</td>
</tr>
<tr>
<td>GOTERM_BP_DIRECT</td>
<td>GO:0032781</td>
<td>Positive regulation of ATPase activity</td>
<td>0.0256</td>
</tr>
<tr>
<td>GOTERM_MF_DIRECT</td>
<td>GO:0004842</td>
<td>Ubiquitin-protein transferase activity</td>
<td>0.0336</td>
</tr>
<tr>
<td>GOTERM_BP_DIRECT</td>
<td>GO:0016567</td>
<td>Protein ubiquitination</td>
<td>0.0350</td>
</tr>
<tr>
<td>GOTERM_MF_DIRECT</td>
<td>GO:0008092</td>
<td>Cytoskeletal protein binding</td>
<td>0.0418</td>
</tr>
<tr>
<td>GOTERM_BP_DIRECT</td>
<td>GO:0071407</td>
<td>Cellular response to organic cyclic compound</td>
<td>0.0481</td>
</tr>
<tr>
<td>GOTERM_BP_DIRECT</td>
<td>GO:0045732</td>
<td>Positive regulation of protein catabolic process</td>
<td>0.0489</td>
</tr>
</tbody>
</table>

GO enrichment analysis was conducted for the 19 candidate genes identified by the integrative analysis. We detected eight significant GO functional gene sets for stroke, including six biological process terms and two molecular function terms (Table 3). The most significant GO term for biological process was regulation of cell shape ($P = 0.0059$) while the most significant GO term for molecular function was the ubiquitin-protein transferase activity ($P = 0.0336$). Identification of these gene sets contributed to a better understanding of the pathogenesis of stroke.

**DISCUSSION**

In this study, we conducted an integrative analysis of TWAS and gene expression profiling data to identify candidate genes for stroke. In total, 19 causal genes were detected, and eight significant GO functional gene sets were identified. To the best of our knowledge, this was the first TWAS conducted on stroke, and our approach provided novel insights into understanding the genetic mechanism of stroke.

Our TWAS identified two genes, i.e., *SLC25A44* and *LRCH1*, that were significant for stroke after Bonferroni correction. *SLC25A44* is a known genetic risk factor for intracerebral hemorrhage (ICH), which is a common cause of stroke (Carpenter et al., 2016; Chen, Chang & Chen, 2018; Falcone et al., 2014). This gene was first mentioned as a risk gene for stroke in a GWAS conducted by the International Stroke Genetics Consortium (Woo et al., 2014). *SLC25A44* is a gene on locus 1q22 and encodes a mitochondrial carrier protein, suggesting that stroke may be associated with mitochondrial dysfunction. Moreover, a recent study found that *SLC25A44* was upregulated in obese patients throughout the arterial network based on rat models. Because the signal for *SLC25A44* was detected in adipose tissue, we suspected that this gene may have important roles in lipid metabolism. However, the biological mechanisms of *SLC25A44* remain unclear. *LRCH1* was also recognized as a risk gene for stroke in the GWAS included in our study. *LRCH1* was recently reported to be associated with atrioventricular nodal delay. Additionally, the regulatory factor for *LRCH1* was found to be active in multiple tissues, including the left ventricle, atherosclerotic aorta, atherosclerotic-lesion-free arteries, and blood. This suggested that *LRCH1* was strongly associated with cardiac mechanisms of stroke, such as large-artery atherosclerotic stroke, cardioembolic stroke, and small-vessel stroke. Recently, Xu et al. found that LRCH1
restrained T cell migration through binding partner to sequester DOCK8 from Cdc42 (Xu et al., 2017). T cell response after stroke is increasingly recognized in previous studies (Gill & Veltkamp, 2016). Evidence showed that antibody-mediated depletion of CD4(+), CD8(+) and gamma delta T cells could reduce infarct volume and improve functional outcome (Liesz et al., 2011; Mracsko et al., 2014; Shichita et al., 2009). Our study also showed that depressed expression of LRCH1 was associated with stroke (Table 1). Thus, LRCH1 might act a role in interfering T cell migration in stroke.

Of the 19 candidate genes identified by integrative analysis, CISD1, which encodes CDGSH iron sulfur domain 1, has been extensively studied. The main function of CISD1 is modulation of mitochondrial iron uptake and respiratory capacity (Yuan et al., 2016). Loss of CISD1 results in oxidative injury in the mitochondria, thereby affecting lipid and glucose metabolism. Furthermore, a recent clinical study showed that CISD1 mRNA and protein levels were significantly decreased in subcutaneous and visceral adipose tissues of obese patients (Moreno-Navarrete et al., 2016). Moreover, CISD1 is highly associated with adipose tissue dysfunction via regulation of mitochondrial dysfunction (Comas et al., 2018). Interestingly, SLC25A44 also plays a role in mitochondrial function and is associated with lipid metabolism. Thus, we speculate that SLC25A44 may have a pathophysiological mechanism similar to that of CISD1. Taken together, these findings suggested that stroke may be closely related to mitochondrial abnormalities.

CLDN5 was another gene identified in this study and showed the strongest differential expression among all 19 candidate genes. CLDN5 encodes an integral membrane protein that is a critical component of endothelial tight junctions (Jang et al., 2011). Previous studies have demonstrated that CLDN5 plays important roles in the blood–brain barrier (Jiao et al., 2015). Additionally, an animal experiment showed that CLDN5-knockout mice show increased blood–brain barrier permeability (Nitta et al., 2003). In contrast, inhibiting CLDN5 expression contributes to reduced brain edema and hemorrhagic transformation (Ishiguro et al., 2010). Increased blood–brain barrier permeability is a common pathological symptom after ICH, which could explain the differential expression of CLDN5 in patients with stroke.

We also conducted GO enrichment analysis to determine the functions of the identified genes. We observed five GO gene sets for biological process, and the most significant GO term in this category was regulation of cell shape. According to a previous report, 92% of abnormal erythrocytes are present in smears of patients who have experienced thromboembolic ischemic stroke (Pretorius & Lipinski, 2013). Altered red blood cells form close interactions with abnormal fibrin fibers, typically resulting in a diseased clot. Moreover, protein modifications with ubiquitin are common responses for ischemia-reperfusion injury, accounting for the GO term of protein ubiquitination (Hochrainer, 2018). The regulation of ATPase activity may reflect the regulatory roles of mitochondria.

There were several limitations to our study. First, gene expression is a highly complex process, varying temporally and spatially in humans. The expression profiling in our study was performed only with peripheral blood mononuclear cells. Further work is needed to discover more candidate genes active in other tissues. Second, we extracted the summary statistics restricted to European ancestry as current available expression
reference panels were all generated from European population. More samples from other ethnic groups should be collected to provide available reference panels for the entire multiethnic population. Third, the expression profiles analysis was performed with a relaxed criterion for determining differential expressed genes. Further work should be conducted with a larger samples and stricter criterion. Fourth, our study provided a novel method to determine candidate genes for stroke. However, functional studies are needed to determine the biological mechanisms underlying the roles of these genes in stroke.

CONCLUSIONS
In summary, we conducted an integrative analysis of TWAS and gene expression profiling to determine the candidate genes for stroke. From this analysis, we identified 19 unique genes and eight corresponding functional gene sets for stroke. The annotation of some genes implied that mitochondria dysfunction may contribute to the cause of stroke. Some candidate genes also showed signals with regulation of cell shape and protein ubiquitination. Our study provided novel insights into the genetic mechanisms underlying stroke.

ACKNOWLEDGEMENTS
We would like to thank the High-Performance Computing Cluster of the First Affiliated Hospital of Xi’an Jiaotong University for data computing.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding
This work was supported by the Clinical Research Award of the First Affiliated Hospital of Xi’an Jiaotong University, China (No. XJTU1AF-CRF-2017-025). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures
The following grant information was disclosed by the authors:
Clinical Research Award of the First Affiliated Hospital of Xi’an Jiaotong University, China: XJTU1AF-CRF-2017-025.

Competing Interests
The authors declare no competing financial interests.

Author Contributions
• Jian Yang and Bin Yan conceived and designed the experiments, analyzed the data, approved the final draft.
• Yajuan Fan and Lihong Yang analyzed the data, prepared figures and/or tables, approved the final draft.
• Binbin Zhao contributed reagents/materials/analysis tools, prepared figures and/or tables, approved the final draft.
• Xiaoyan He and Qingyan Ma contributed reagents/materials/analysis tools, approved the final draft.
• Wei Wang, Ling Bai and Feng Zhang authored or reviewed drafts of the paper, approved the final draft.
• Xiancang Ma conceived and designed the experiments, authored or reviewed drafts of the paper, approved the final draft.

Data Availability
The following information was supplied regarding data availability:
The GWAS summary results of stroke are available from the GWAS Catalog (https://www.ebi.ac.uk/gwas/), Publication ID 29531354.
Tissue specific expression reference panels are publicly available on FUSION website (http://gusevlab.org/projects/fusion/).
The data on gene expression profiling is available as a supplementary file named “10.1038_jcbfm.2012.24x2.doc” from this publication: https://journals.sagepub.com/doi/suppl/10.1038/jcbfm.2012.24.

Supplemental Information
Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.7435#supplemental-information.

REFERENCES


