A genome wide association study identifies a gene network of ADAMTS genes in the predisposition to pediatric stroke

Astrid Arning1*, Milan Hiersche1*, Anika Witten1, Gerhard Kurlemann2, Karin Kurnik3, Daniela Manner4, Monika Stoll1§ & Ulrike Nowak-Göttl4§

1Leibniz-Institute for Arteriosclerosis Research at the University Muenster, 2 Pediatric Neurology, University Children’s Hospital Muenster, 3 University Children’s Hospital Munich, 4 Thrombosis & Hemostasis Treatment Center, Institute of Clinical Chemistry, University Hospital Schleswig-Holstein

Running head: ADAMTS genes are involved in pediatric stroke

*AA and MH contributed equally to this study
§ MS and UNG shared the senior author

The online version of the article contain a data supplement

Addresses for correspondence:

Monika Stoll, PhD
Genetic Epidemiology of vascular disorders
Leibniz-Institute for Arteriosclerosis Research
at the Westfälische Wilhelms University of Muenster
Albert-Schweitzer-Campus 1, D3
48149 Muenster, Germany

e-mail: mstoll@uni-muenster.de
phone: +49-251-83-57206
fax: +49-251-83-56205
and

Ulrike Nowak-Göttl, MD
Thrombosis & Hemostasis Treatment Center
Institute of Clinical Chemistry
Univ. Hospital Schleswig-Holstein, Campus Kiel
Arnold-Heller-Str. 3, building 17
24105 Kiel, Germany
Email: leagottl@uksh.de
Phone +49-431-597-3225
Abstract

Pediatric stroke is a rare but highly penetrant disease with a strong genetic background. While there is an increasing number of genome-wide association studies (GWAS) for stroke in adults, such studies for stroke of pediatric onset are lacking. Here we report the results of the first GWAS on pediatric stroke using a large cohort of 270 family-based trios.

GWAS was performed using the Illumina 370 CNV SNP array and analyzed using the transmission disequilibrium test (TDT) as implemented in PLINK. An enrichment analysis was performed to identify additional true association signals among lower P-value signals, and searched for cumulatively associated genes within protein interaction data using dmGWAS. We observed clustering of association signals in four genes belonging to one family of metalloproteinases at high (ADAMTS12 $P=2.9\times10^{-6}$, ADAMTS2 $P=8.0\times10^{-6}$) and moderate (ADAMTS13 $P=9.3\times10^{-4}$, ADAMTS17 $P=8.5\times10^{-4}$) significance levels. Overrepresentation and gene-network analyzes highlight the importance of the extracellular matrix in conjunction with members of the phosphoinositide (PIP) and calcium signaling pathways in the susceptibility for pediatric stroke. Associated extracellular matrix components like ADAMTS proteins in combination with misbalanced coagulation signals as unveiled by gene network analysis suggest a major role of postnatal vascular injury with subsequent thrombus formation as the leading cause of pediatric stroke.
Introduction

Pediatric stroke (PS) is a heterogeneous disorder associated with significant morbidity and mortality. It is recognized as an important childhood disease, with an incidence of 2.6 to 6.4 per 100,000 children per year.\textsuperscript{1} Although PS is relatively rare it is devastating to those affected as half of the survivors develop cognitive or motor disabilities.\textsuperscript{2}

Risk factors for stroke are different in children and adults and classical risk factors such as smoking, arteriosclerosis or diabetes are unlikely to contribute to pediatric stroke. The most prominent risk factors for stroke in children include underlying medical conditions i.e. cardiac disorders, metabolic diseases, cerebrovascular pathologies and infectious diseases\textsuperscript{1} as well as many prothrombotic abnormalities.\textsuperscript{3,4}

Numerous association studies have established a number of genetic polymorphisms contributing to pediatric stroke risk. These candidate gene approaches identified several susceptibility genes for PS, i.e. Prothrombin, Factor-V-Leiden and GPX3\textsuperscript{3,4}

However, genetic predisposition, environmental effects and other risk factors can often not be disentangled due to the complex nature of PS etiology. Furthermore, a large proportion of missing heritability remains to be accounted for.\textsuperscript{5}

Genome-wide association studies (GWAS) offer a powerful approach to gene function discovery and are the current method of choice to dissect the genetic basis of complex diseases. Up-to-date, studies in families with a known first onset of pediatric stroke are lacking mainly due to limited sample size.

Therefore, we performed a family-based GWAS for pediatric stroke in 270 families from Muenster, Germany comprising affected children and their parents. We identified several single nucleotide polymorphisms (SNPs), which are associated
with PS at significance levels $P<10^{-5}$. In addition, we determined potential combined effects of single SNPs contributing to stroke risk that may be missed through conventional single marker or haplotype association. We assessed the complete GWAS dataset under a systems biology aspect incorporating protein interaction information to gain further insight into the biological networks underlying pediatric stroke.\textsuperscript{6} The latter analysis highlights the complex interplay of genes at the interface of vascular biology and coagulation in the pathogenesis of PS and supports the need of an extensive analysis of combined genetic and physiological properties of complex diseases.

**Methods**

The study was approved by the medical ethics committee of the University of Muenster (Muenster, Germany) and complies with the ethical standards established in the updated version of the 1964 Declaration of Helsinki. Written informed consent was given from both parents for all families enrolled.

**Subjects**

Patients were enrolled between 2000 and 2009 as part of a nationwide survey [ESPED registry] and centrally documented in the Muenster pediatric stroke database. Stroke subtypes were reclassified according to explicit predefined criteria based on the TOAST criteria modified for children, by substituting arteriopathy for large vessel atherosclerosis, by an independent blinded team of neuroradiologists (for details see supplementary web material).
From an initial study sample comprising 300 consecutively enrolled families, 293 trios were considered eligible for genotyping after removal of siblings and assessment of DNA quality. Of these, 270 were valid for statistical analysis after filtering for call rate (> 97%), Mendelian inheritance error rate (< 0.1%) and gender prediction match using the GenomeStudio Software (Illumina Inc. V 1.9.4) (Figure 1). Potential population stratification was assessed using a population concordance analysis implemented in PLINK V1.06 (supplementary web material).7

Genotyping

In the initial analysis, 235 trios were genotyped on Illumina Infinum 370CNV arrays. 35 additional trios were genotyped using its successor, the Infinium 660W chip, due to discontinuation of the 370CNV series. The SNP content overlap between both chip layouts is 89.7%. For statistical analysis, we used all SNPs from the 370CNV chip, accepting a slightly reduced sample size for 36.189 SNPs not present on the 660W chip. This does not pose a statistical bias because the number of observations varies for the transmission disequilibrium test (TDT) dependent on allele frequency and number of heterozygous parents. We tested the validity of integrating genotypes from these chips by running 40 samples on both chips. Allele callings match at 99.48% for common SNPs; thus we conclude that genotypes can be combined. All genotyped samples were assessed for sufficient call rate, gender estimation match, mendelian error fraction and population structure prior to statistical analysis as detailed in the supplementary text and supplementary Figure S4.

Association analysis

We performed a GWAS using a threshold of $P<1\times10^{-5}$ for suggestive genome-wide
association. *P*-values were computed using the TDT and validated by 20,000 gene-dropping permutation runs yielding empirical *P*-values based on unbiased null distributions (PLINK V1.06). This analysis was carried out for the complete dataset [n=270] and a subset excluding vascular strokes [n=201].

Systems biology analysis

In a subsequent step, we estimated the amount of true association signals by visual inspection of the QQ plot. Deviating intervals in the QQ plot are considered as indicators for population stratification or similar statistical bias. However, the observed deviation at midrange *P*-values may also be attributed to the presence of a large number of independent and weakly associated loci that reflect the genetic complexity of the disease. As family studies are immune to population stratification and individuals were ascertained consecutively, selection bias towards secondary phenotypes is unlikely in our experimental setting.

Therefore, we assessed the excess of observed association in an enrichment analysis for abundant stroke-related genes occurring in deviating *P*-value intervals in the QQ-plot. We assembled a set of stroke-related genes by collecting all gene identifiers that occur with the keyword ‘stroke’ in Pubmed abstracts (see supplementary web material). A second set of GWAS-derived genes was generated by setting a threshold at *P*<1×10⁻³ in our dataset to generate a set of candidate SNPs, expecting a fraction of 16% among the loci to be true positive associations according to a false discovery rate of 84% derived by Benjamini-Hochberg correction. These candidate SNPs were then assigned to one or more genes using a linkage disequilibrium (LD) mapping approach and the postgwas R package as described in the supplementary material. Overrepresentation of stroke-associated
genes within the GWAS-set was determined by Fisher’s exact test using R. To account for potential publication bias in this procedure, we performed a permutation test with 100,000 random samples matching the size of the stroke-related gene set which were drawn from the entire human genome (Ensembl database, Homo sapiens genes GRCh37.p5). The overlap with the GWAS-derived set of genes was computed and results in an approximate null distribution that is used for comparison to the observed overlap for the stroke-related set (supplementary Figure S1).

Secondly, we investigated the data regarding their relatedness within protein interaction networks to identify low-risk variants that act jointly on an intermediate phenotype (i.e. biological pathway). We used dmGWAS\textsuperscript{6}, a method that searches for densely connected modules within protein-interaction graphs, where members of such modules are constrained towards an accumulation of strong association signals. In a primary step, a representative $P$-value was assigned to each gene which is the smallest of (1) all intragenic SNPs and (2) SNPs with an LD $r^2>0.3$. Afterwards, the dense module search of dmGWAS was performed with an interaction dataset and parameter settings as described in the supplementary appendix. From the resulting set of modules, we extracted the top three for visualization. In addition, we used Gene Ontology (GO) and pathway data from REACTOME as established categorization databases in an enrichment analysis to identify primary biological functions of the resulting network. We identified enriched terms in the network of proteins using the overrepresentation analysis features of the STRING database web interface.\textsuperscript{9} Proteins containing these terms are colorized accordingly in the graphical plot.
**Results**

Genome-wide association analysis

GWAS was conducted in the complete dataset of 270 incident stroke families and a subset of nonvascular stroke (n=201). 334,581 SNPs provided valid association signals after applying missingness tests (>25% genotyped) and allele frequency filtering (>1%), resulting in a signal intensity plot as shown in Figure 2. The accompanying QQ-plot (Figure 3) confirms the presence of causative loci by showing a deviation between the numbers of SNPs at high P-values in comparison to those expected under a null model. In total, four hits exceeded the threshold of $P<1\times10^{-5}$ for suggestive genome-wide significance (rs1364044 $P=2.9\times10^{-6}$; rs2084898 $P=4.3\times10^{-6}$; rs469568 $P=8.0\times10^{-6}$; rs4663691 $P=9.8\times10^{-6}$), which are summarized in Table 1. Three SNPs are intragenic; two of them are located in genes that belong to the ADAMTS family, one is proximate to COPS8. As shown in Table 1, all loci are robust to permutation testing and range at medium to high allele frequencies (12 to 48%) in the population resulting in a large number of heterozygous parents and reliable T/U ratios. Comparative display of subphenotype $P$-values and linkage disequilibrium information is further shown in regional plots (supplementary Figure 2). Due to the smaller subphenotype datasets, $P$-values drop approximately one order of magnitude but do not show major discrepancies due to subphenotype stratification (Table 1).

Systems biology analysis

In the QQ-plot (Figure 3), we identified an interval between $10^{-3} < P < 10^{-5}$ showing an excess of observed loci beside the peak signal range. In absence of confounding...
effects, this suggests the presence of true association. To confirm that this midrange deviation is caused by association with the stroke phenotype, we performed an enrichment analysis with stroke-related genes as described in the methods section. We observed that 15.7% of moderately associated loci from our study occur in the set of stroke-related genes (1.8% vice versa), which is more than expected and reaches the threshold of significance as assessed by Fisher exact-test \((P = .019)\). Additional permutation testing confirmed the findings at \(P < .0005\) (supplementary web material).

Subsequently, we assessed the entire GWAS dataset for dysregulated molecular networks arising from a cumulative effect of loci with moderate P-values. The results of the corresponding dmGWAS analysis are displayed in Figure 4. We identified three different domains of biological function in this network. The first comprises genes of transcriptional control \((\text{CNOT2, NCOR1, TRIM29})\) and cell division \((\text{CEP70})\) without known functional implication in stroke. More interestingly, \text{ADAMTS2} and \text{CHRM3} appear as members of the extracellular matrix and endothelium in a submodule that is tightly coupled to members of the phosphoinositide (PIP) and calcium signaling pathways \((\text{CALM1, PIK3CB})\). Together with \text{ANGPT1}, a carrier of several fibrinogen domains, these are essential mediators of platelet activation. Profilin-2 \((\text{PFN2})\) occupies another central role in this interface between coagulation signaling and extracellular matrix, though appearing in an isolated position in the network. It is an actin polymerizing protein found in the plasma membrane of platelets that depends on the activity of PIP.\(^{10}\) An assessment of enriched GO biological processes in the network confirms the central role of blood coagulation with a \(P\)-value of \(9.5 \times 10^{-5}\) but slightly misses significance after correction for multiple testing using the false discovery rate \((P = .055)\). Further classes of GO-terms are metabolic processes \((P = 2.2 \times 10^{-6}, \text{corrected } P = .01)\) and transcription \((P = 8.6 \times 10^{-5}, \text{corrected}\)
as listed in supplementary Table S1. *DMD* as a prominent cardiovascular risk gene is not addressed here because it spans a large genomic region and resides on the X-chromosome, and is thus likely a false positive association, which requires careful interpretation.

Overlap with GWAS signals previously identified for adult stroke

To investigate whether some genetic components overlap between phenotype windows, we searched for loci that were previously reported to be associated with adult stroke in the literature. We extracted 710 SNP loci as putative candidates and mapped them to our dataset using a proxy SNP approach (supplementary web material). However, none of these loci reached significance after correction for multiple testing.

**Discussion**

Here we present results of the first GWAS on pediatric stroke in 270 nuclear families. To our knowledge, this cohort represents the largest study sample currently available worldwide. We identified several novel susceptibility genes that directly relate to coagulation and vascular biology. The most important finding is the identification of four members of the *ADAMTS* gene family at confident association levels. While SNPs residing in *ADAMTS2* and *12* are strongly associated ($P = 8.0 \times 10^{-6}$, $P = 2.9 \times 10^{-5}$), SNPs in *ADAMTS13* and *17* showed moderate association ($P = 9.3 \times 10^{-4}$, $P = 8.5 \times 10^{-4}$) with pediatric stroke. *ADAMTS13*, also known as von-Willebrand (VWF) cleaving protease, is already established as a crucial factor in
blood coagulation and stroke\textsuperscript{11}, reduces platelet adhesion and aggregation\textsuperscript{12} and down regulates thrombus formation and inflammation.\textsuperscript{13} Reduction or absence of ADAMTS13 activity results in impaired degradation of ultra-large VWF multimers and leads to excessive VWF-induced platelet aggregation on the endothelium. Mutations in the $ADAMTS13$ gene have been shown to cause familial thrombotic thrombocytopenic Purpura (TTP) which is associated with brain ischemia and neurologic symptoms due to microthrombi in the microvasculature.\textsuperscript{14} These observations are further supported by an experimental study, where $ADAMTS13$ gene deletion aggravated ischemic brain damage in a reperfusion mouse model.\textsuperscript{15}

In the clinical setting, $ADAMTS13$ has been proposed as a target for treatment of stroke due to its negative regulation of thrombosis and inflammation.\textsuperscript{16} In addition, acquired TTP has been described in the presence of IgG, IgA or IgM antibodies inhibiting ADAMTS13 activity. In these cases the inhibitory antibody effect could be overridden by administration of recombinant ADAMTS13 concentrate.\textsuperscript{17} In an experimental study, Zhao and colleagues reported that infusion of high dose ADAMTS13 concentrates administered immediately before reperfusion injury significantly reduces infarct volume in mice and improves functional outcome without evidence of hemorrhage.\textsuperscript{16} Taken together, these studies render ADAMTS13 an attractive candidate for clinical treatment of stroke and a potential diagnostic kit that incorporates genetic variation as predictors for stroke risk.

Several other genes from the $ADAMTS$ family, such as $ADAMTS7$, have already been connected to endothelial differentiation and matrix physiology.\textsuperscript{18} Recent meta-analyzes revealed an association of $ADAMTS7$ and 8 with pulse pressure\textsuperscript{19} and coronary artery disease\textsuperscript{20} at genome-wide significance. Notably, $ADAMTS7$ is a direct phylogenetic neighbor of $ADAMTS12$ (supplementary Figure S3), which is the
gene with the most stringent association in our study. The large amount of sequence identity between the two genes is a strong indicator of redundant functionality. ADAMTS18 has been shown to be protective against stroke in mice as it is capable of dissolving platelet aggregates. Thus, a general role of the ADAMTS gene family in extracellular matrix biology and coagulation is conceivable based on our data and existing knowledge. Nevertheless, for ADAMTS2 and ADAMTS12, the susceptibility genes with most significant association signals, there is no information on biological function available yet. ADAMTS2 is known to be involved in procollagen processing, inhibition of angiogenesis and tumor growth. ADAMTS12 shows similar functionality, being involved in the degradation of cartilage oligomeric matrix protein and showing anti-angiogenic and tumor suppressive properties as well as a modest association with heart failure. However, the mechanisms through which the genetic variants in ADAMTS genes lead to pediatric stroke remain unknown.

We observed further loci at suggestive genome-wide significance that do not map unambiguously to a distinct gene. One variant tags a LD block containing the genes TRIM29, OAF and POU2F3 and the other variant resides in an intergenic region proximal to COPS8. The TRIM29 block is a cassette of genes involved in regulation of transcription. COPS8 (as part of the protein complex COP9) is a regulator of cyclooxegenase-2 and might therefore influence coagulation. It has been argued that variants identified in GWAS explain only a small proportion of the genetic variance underlying most complex traits.

Particularly for a phenotype such as stroke, the physiological heterogeneity can manifest as locus heterogeneity, where a larger number of loci with small effects affects relevant pathways in a cumulative fashion. The severity of pediatric diseases with an early onset hints towards the presence of individual rare mutations scattered over a larger number of causal genes, resulting in poor association strength. Therefore, we applied dmGWAS to identify densely connected modules in the human
protein interactome that conjoins loci of strong to moderate significance. This network provides first insights into the interplay of putative causative genes on a proteome level and represents a first attempt to understand gene variants in the context of whole cells and cellular communication. Accordingly, we found interacting components of extracellular matrix biology (\textit{ADAMTS2}, \textit{CHRM3}, \textit{PFN2}) and coagulation signaling (\textit{CALM1}, \textit{PIK3CB}, \textit{PFN2}) that suggest a complex relation between multiple causative genes with low to medium individual effects. Significance of these genes is further substantiated by the literature. For \textit{PIK3CB}, crucial function in platelet aggregation and adhesion to fibrinogen has been shown with targeted alleles in mice.\textsuperscript{27} Available antithrombotic therapy targeting \textit{PI3K}\textsuperscript{28} could be established in the future as a treatment conditional on genetic testing, given successful replication of our results. For \textit{ANGPT1}, lethal vascular defects have been observed in mouse knockout models, and a special role in injured vasculature has been proposed.\textsuperscript{29} Interestingly, there is an additional peak at \( P=2.0\times10^{-4} \) with the next gene, \textit{ANGPT2}, residing 380kb apart, which was not included in the dmGWAS analysis due to our restrictive SNP to gene mapping. \textit{CHRM3}, also known as muscarinic receptor 3, plays a major role in dilation of coronary arteries upon stimulation by acetylcholine in mice.\textsuperscript{30} Finally, \textit{CALM1} has been shown to be significant in an association study for adult stroke\textsuperscript{31} and is known to be an essential trigger of platelet activation. The abundance of potential risk factors supports a model of heterogeneity as outlined above. To our knowledge there is no comparable cohort of pediatric stroke with suitable DNA samples available, mainly due to the low incidence of the disease and restrictions in several countries with such cohorts to conduct genetic association studies in children.
In lack of such an independent study sample, we were not able to validate the association by replication analysis at this point in time. And we are aware of the fact that the lack of replication limits our ability to generalize our findings to other cohorts.

In summary, this GWAS provides important insights into the genetic underpinnings of stroke in children. We identified a gene family with several members exhibiting considerably strong association signals in the dataset. The observation of two genes from the same family within the peak signal range is extraordinary and a novel observation in GWAS and suggests a crucial role for the class of ADAMTS metalloproteinases in pediatric stroke pathogenesis. Secondly, we have found a cluster of genetic variants that act on common physiological mechanisms leading to hyper-reactive platelet activation and altered vascular function, and are thus conceivable to cause prothrombotic susceptibility in a cumulative fashion.

In conjunction, these results give rise to the hypothesis that a disposition towards a reduced integrity of the endothelial wall as e.g. conferred by ADAMTS defects together with hypersensitive platelet activation and/or vasoconstriction is a major cause of stroke in early childhood. While our cohort is comparable to similar studies in white children\textsuperscript{1,3}, the generalizability of our findings hinges upon future studies using larger study samples with joint analyzes of susceptibility variants. Such studies would also enable the inclusion of intermediate phenotypes and rare genetic variants (identified through next generation sequencing efforts currently underway) to unravel genetic susceptibility and underlying biological pathways.
Acknowledgments
The authors thank all participants from the ESPED (survey on rare paediatric diseases in Germany) registry.
The study was supported by grants from the Förderverein “Schlaganfall und Thrombosen im Kindesalter e.V.” and Interdisziplinäres Zentrum für Klinische Forschung (IZKF: CRA01-09), University of Münster.

Authorship Contributions
M Stoll and U Nowak-Göttl were responsible for the study concept and design. U Nowak-Göttl, A Arning and M Hiersche acquired data and M Hiersche did the statistical analysis. G Kurlemann, K Kurnik, D Manner and U Nowak-Göttl were responsible for patient recruitment. A Arning, M Hiersche, U Nowak-Göttl and M Stoll drafted the report. All genome-wide genotypings were performed on the Core-facility of the LIFA by A Witten. All authors contributed to the interpretation of data and critically revised the report for important intellectual content.

Conflicts of interest disclosures
None declared
References


Table 1: Summary of SNPs with suggestive genome-wide significance ($P<10^{-5}$).

<table>
<thead>
<tr>
<th>SNP</th>
<th>chr</th>
<th>Position, build 36</th>
<th>MI / MA allele</th>
<th>MIF</th>
<th>TDT (all)</th>
<th>TDT (nonvascular subset)</th>
<th>Next gene (distance in bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>T/U ratio</td>
<td>P-value</td>
<td>P-value perm</td>
</tr>
<tr>
<td>rs1364044</td>
<td>5</td>
<td>33915549 A / G</td>
<td>0.296</td>
<td></td>
<td>2.91E-06</td>
<td>5.00E-05</td>
<td>1.91</td>
</tr>
<tr>
<td>rs2084898</td>
<td>11</td>
<td>119531958 A / G</td>
<td>0.126</td>
<td></td>
<td>4.30E-06</td>
<td>5.00E-05</td>
<td>2.37</td>
</tr>
<tr>
<td>rs469568</td>
<td>5</td>
<td>178596014 T / G</td>
<td>0.483</td>
<td></td>
<td>8.00E-06</td>
<td>5.00E-05</td>
<td>1.77</td>
</tr>
<tr>
<td>rs4663691</td>
<td>2</td>
<td>237565291 G / A</td>
<td>0.238</td>
<td></td>
<td>9.79E-06</td>
<td>5.00E-05</td>
<td>1.9</td>
</tr>
</tbody>
</table>

Figure 1: Outline of the study cohort. Numbers relate to cases or trios, respectively. Additionally, baseline phenotype statistics are given for all groups that were analyzed.
Figure 2: Manhattan plot showing the distribution of P-values over all chromosomes. Loci are colored in correspondence to P-value thresholds (red: $P < 10^{-5}$, blue $P < 5 \times 10^{-4}$), annotation shows the closest gene in each direction (up/downstream, intragenic).
Figure 3: Quantile plot comparing observed distribution of $P$-values from our study with the expected distribution of $P$-values under a null model.
Figure 4: Analysis of associated loci accumulated in protein interaction network modules as computed by dmGWAS. Proteins that belong to a Reactome-pathway containing the string ‘platelet’ have been marked black and grey for ‘endothel’ and ‘extracellular’ accordingly.
A genome wide association study identifies a gene network of ADAMTS genes in the predisposition to pediatric stroke

Astrid Arning, Milan Hiersche, Anika Witten, Gerhard Kurlemann, Karin Kurnik, Daniela Manner, Monika Stoll and Ulrike Nowak-Götti