BMC Medical Genetics

Study protocol

Genetic risk factors for cerebrovascular disease in children with sickle cell disease: design of a case-control association study and genomewide screen

Gaye T Adams[1], Harold Snieder[2,3], Virgil C McKie[4], Betsy Clair[1], Donald Brambilla[6], Robert J Adams[5], Ferdane Kutlar[1] and Abdullah Kutlar*[1]

Address: [1]Sickle Cell Center, Department of Medicine, Medical College of Georgia, Augusta, GA, [2]Georgia Prevention Institute, Department of Pediatrics, Medical College of Georgia, Augusta, GA, USA, [3]Twin Research and Genetic Epidemiology Unit, St. Thomas' Hospital, London, UK, [4]Department of Pediatrics, Medical College of Georgia, Augusta, GA, USA, [5]Department of Neurology, Medical College of Georgia, Augusta, GA, USA and [6]New England Research Institute, Watertown, MA, USA

Email: Gaye T Adams - gayetadams@hotmail.com; Harold Snieder - hsnieder@mcg.edu; Virgil C McKie - vmckie@mcg.edu; Betsy Clair - bclair@mcg.edu; Donald Brambilla - donb@neri.org; Robert J Adams - rjadams@mcg.edu; Ferdane Kutlar - fkutlar@mcg.edu; Abdullah Kutlar* - akutlar@mcg.edu

* Corresponding author

Published: 18 July 2003

BMC Medical Genetics 2003, **4**:6

Received: 13 March 2003
Accepted: 18 July 2003

This article is available from: http://www.biomedcentral.com/1471-2350/4/6

Abstract

Background: The phenotypic heterogeneity of sickle cell disease is likely the result of multiple genetic factors and their interaction with the sickle mutation. High transcranial doppler (TCD) velocities define a subgroup of children with sickle cell disease who are at increased risk for developing ischemic stroke. The genetic factors leading to the development of a high TCD velocity (i.e. cerebrovascular disease) and ultimately to stroke are not well characterized.

Methods: We have designed a case-control association study to elucidate the role of genetic polymorphisms as risk factors for cerebrovascular disease as measured by a high TCD velocity in children with sickle cell disease. The study will consist of two parts: a candidate gene study and a genomewide screen and will be performed in 230 cases and 400 controls. Cases will include 130 patients (TCD ≥ 200 cm/s) randomized in the Stroke Prevention Trial in Sickle Cell Anemia (STOP) study as well as 100 other patients found to have high TCD in STOP II screening. Four hundred sickle cell disease patients with a normal TCD velocity (TCD < 170 cm/s) will be controls. The candidate gene study will involve the analysis of 28 genetic polymorphisms in 20 candidate genes. The polymorphisms include mutations in coagulation factor genes (Factor V, Prothrombin, Fibrinogen, Factor VII, Factor XIII, PAI-1), platelet activation/function (GpIIb/IIIa, GpIb IX-V, Gpla/IIa), vascular reactivity (ACE), endothelial cell function (MTHFR, thrombomodulin, VCAM-1, E-Selectin, L-Selectin, P-Selectin, ICAM-1), inflammation (TNFα), lipid metabolism (Apo A1, Apo E), and cell adhesion (VCAM-1, E-Selectin, L-Selectin, P-Selectin, ICAM-1). We will perform a genomewide screen of validated single nucleotide polymorphisms (SNPs) in pooled DNA samples from 230 cases and 400 controls to study the possible association of additional polymorphisms with the high-risk phenotype. High-throughput SNP genotyping will be performed through MALDI-TOF technology using Sequenom's MassARRAY™ system.

Discussion: It is expected that this study will yield important information on genetic risk factors for the cerebrovascular disease phenotype in sickle cell disease by clarifying the role of candidate

genes in the development of high TCD. The genomewide screen for a large number of SNPs may uncover the association of novel polymorphisms with cerebrovascular disease and stroke in sickle cell disease.

Background

Sickle cell anemia (Hb SS) results from homozygosity for a A→T substitution at codon 6 of the β-globin gene (GAG→GTG) leading to a glutamic acid to valine (Glu→Val) substitution in the β globin chain of human adult hemoglobin. Despite this common genetic background, phenotypic expression of sickle cell disease is widely variable, ranging from a mild, asymptomatic course with survival into the sixth or seventh decade to a very severe course with multi-organ damage and early mortality [1]. Some of the genetic factors contributing to this phenotypic diversity (particularly those linked to the globin genes, i.e. α-thalassemia and β-globin gene cluster haplotypes) have been well recognized [2].

Stroke is a devastating complication of sickle cell disease, which occurs in 11% of the patients by 20 years of age as shown by the multi-center Cooperative Study of Sickle Cell Disease (CSSCD) [3,4]. In sickle cell patients <20 years of age, stroke is predominantly ischemic and results from the involvement of medium sized to large intracranial arteries. Ischemic stroke in the general population is considered a multi-genic disorder [5,6]. In many cases it results from multiple gene-gene and gene-environment interactions. In the case of sickle cell disease, only a few genetic factors are known to influence the stroke risk [7]. For example, α-thalassemia is the only well characterized protective genetic factor [7]. Thus, genetic factors that lead to the development of cerebrovascular disease and stroke in children with Hb SS are not well understood.

Studies conducted at the Medical College of Georgia (MCG) in the mid-1980's have shown that transcranial doppler (TCD) can identify children at high risk for stroke by detecting high flow rate in major intracranial arteries [8,9]. Children with flow velocities of 200 cm/sec or higher in middle cerebral or internal carotid arteries (normal = 140–170 cm/sec for sickle cell children) had a stroke risk of 10–15% per year, which represents a 20-fold increase over that for unselected children with sickle cell disease. These observations then led to the multicenter STOP (Stroke Prevention Trial in Sickle Cell Anemia) study in which sickle cell children age 2–16 years, from 14 centers in the U.S. and Canada, were screened by TCD [10]. One hundred thirty patients with flow velocities of >200 cm/sec were randomized to observation or to receive periodic blood transfusions to reduce % Hb S to <30. The study was stopped early by the Data and Safety Monitoring Board due to the finding of a significant reduction in the number of strokes (90% reduction, $p < .001$) in the transfusion arm (11 strokes in the observation arm vs. one in the transfusion arm) [11]. Thus, the STOP study established that transfusion was an effective means of primary stroke prevention in children with sickle cell disease at risk for stroke as determined by TCD. The recently funded STOP-II study, to be conducted in 24 centers, will investigate the duration of stroke risk and the feasibility of discontinuing transfusion after 30 months in patients whose TCD velocities have normalized after transfusion.

Data from the STOP study showed that ~10% of patients with Hb SS between the ages of 2–16 years are at risk for stroke as determined by Transcranial Doppler (TCD) velocities of 200 cm/sec or greater [11]. TCD velocity elevation can be due to two factors: reduction of arterial diameter due to stenosis and/or increased volume flow through the artery. In the case of sickle cell disease, both factors are often present. Velocity elevation corresponding to increased cerebral blood flow in all patients with anemia has been noted, with an approximate linear increase in velocity for decrease in Hb or HCT [12]. There are also data to suggest that patients with Hb SS have higher flow velocities than those with normal Hb A at comparable levels of anemia, suggesting the amount and composition of the Hb are both involved in the velocity elevation of sickle cell disease [13].

While some of the risk factors leading to the development of cerebral vasculopathy and stroke in sickle cell disease (low hematocrit, elevated white blood cell count, normal complement of α-globin genes) have been identified in previous studies (CSSCD, MCG cohort, and STOP), many remain unknown. The observation has been made that strokes in sickle cell disease are clustered in families, but no clear epidemiological confirmation of this has been published. In terms of TCD velocities, it was noted in the STOP study that sibling pairs were common among abnormals (i.e., TCD > 200 cm/sec) and there were 5 pairs among 130 randomized patients suggesting that other genetic factors may in part determine propensity for stroke (RJ Adams, unpublished observations).

Various mechanisms have been proposed to account for the "hypercoagulable state" observed in patients with sickle cell disease. These include increased platelet activation, increased thrombin generation, and more recently, elevated levels of circulating tissue factor [14]. Over the

past decade, mutations in a number of genes have been identified as the cause of thrombophilia in a large number of patients from different populations around the globe. The molecular basis of inherited thrombophilia now includes mutations in the genes for Factor V, prothrombin, methylenetetrahydrofolate reductase (MTHFR) and several others [15], in addition to those encoding proteins such as antithrombin, protein C, and protein S whose role in the homeostasis between procoagulant and anticoagulant activities in the circulation is well established. Unlike rare mutations involving antithrombin, protein C and S, which are clearly associated with a hypercoagulable phenotype, mutations we propose to study appear to have two distinct features: i) they occur with a significantly higher frequency in many populations, and ii) they seem to contribute to a thrombophilic state in the presence of additional acquired (environmental/nutritional) or inherited (interaction with other genes) factors.

It is likely that the co-inheritance of one or more of these mutations in sickle cell disease would tip the balance toward a hypercoagulable state and act as an additional risk factor for the development of cerebrovascular disease. Additionally, the effect of the thermolabile MTHFR mutation (677 C→T) on plasma homocysteine levels in heterozygotes has been shown to depend upon the availability of folate, with higher levels seen in folate deficient individuals. It is possible that increased folate requirements due to chronic hemolysis in sickle cell disease would lead to a relative deficiency state, which, in the presence of MTHFR mutation, would result in higher homocysteine levels as an additional risk factor for vasculopathy.

Thus, we propose to study polymorphisms in genes involved in a number of systems and pathways related to stroke risk [5]: genes associated with coagulation factors and thrombophilia (Factor V, Prothrombin, Fibrinogen, Factor VII, Factor XIII, PAI-1, Thrombomodulin [TM], and MTHFR), as well as polymorphisms in genes that are involved in vascular reactivity (ACE), platelet activation/function (GpIIb/IIIa, GpIb IX-V, and GpIa/IIa), endothelial cell function (MTHFR, TM, VCAM-1, E-Selectin, L-Selectin, P-Selectin, and ICAM-1), inflammation (TNFα), lipid metabolism (Apo A1 and Apo E), and cell adhesion (VCAM-1, E-Selectin, L-Selectin, P-Selectin, and ICAM-1). These candidate genes will make it possible to study not only polymorphisms associated with a hypercoagulable state, but also to study the genes involved in pathways that may lead to the development of vasculopathy (genes involved in endothelial cell function, inflammation, and adhesion, platelet activation and responsiveness). Thus, this will provide the rationale for studying potential genetic risk factors that may contribute to the pathogenesis of an "intermediate phenotype," such as vasculopathy [5] rather than a thrombotic endpoint (i.e. stroke). TCD remains the only proven indicator of cerebrovascular disease and stroke risk in Hb SS. Therefore, a large case-control association study of the role of genes in the development of cerebrovascular disease and stroke risk based upon TCD, is likely to provide more accurate information and may resolve the controversial results obtained in smaller studies [5]. Furthermore, an attempt will be made to perform a subset analysis of patients who go on to have an ischemic stroke during the study period in terms of the associations with genetic polymorphisms. As of the termination of the study in 2000, 18 of the 130 patients randomized in STOP have been adjudicated to have a stroke.

In summary, we have designed a case-control association study to elucidate the role of genetic polymorphisms as risk factors for cerebrovascular disease, measured by a high TCD velocity, in children with Hb SS. The study will consist of two parts, a candidate gene study and a genomewide screen and will be performed in 230 cases and 400 controls. Cases will include 130 patients randomized in the STOP study (TCD ≥ 200 cm/s) as well as 100 patients found to have high TCD in STOP II screening. Four hundred sickle cell patients with a normal TCD velocity (TCD < 170 cm/s) will be used as controls. The candidate gene study will involve the analysis of 28 genetic polymorphisms in 20 candidate genes (see Table 1). References for the polymorphisms listed in Table 1 are provided in the additional file 1. The genomewide screen for a large number of SNP markers may uncover the association of novel polymorphisms with cerebrovascular disease and stroke in sickle cell disease. DNA samples will be pooled in the 230 cases and in the 400 controls and these pools will be genotyped for a large number of validated SNPs (we estimate to need at least 100,000 SNP markers) using high-throughput methods. Purpose of this paper is to give a detailed description of the design of this study.

Research Design and Methods

Subjects

African-American children who are between the ages of 2 and 16 years with Hb SS or $S\beta^0$-thalassemia from participating STOP and STOP II centers form the subjects of this study. The STOP study was conducted in 14 centers in North American (13 US and 1 Canadian). STOP II is being conducted in 28 centers (27 US and 1 Canadian). Banked DNA samples from randomized STOP patients (n = 130) stored in the STOP Core Laboratory will be utilized. An additional 100 patients with high TCD velocity and 400 control patients with normal TCD velocity will be identified through STOP-II screening. With an abnormal TCD rate of 10%, it is expected that at least 1000–1400 new patients from 24 participating Centers will need to be screened as part of the STOP-II study to identify 100 new patients with TCD velocities >200 cm/sec. DNA will be

Table 1: List of 28 polymorphisms in 20 candidate genes used in this study

	GENE	POLYMORPHISM	REFERENCE		GENE	POLYMORPHISM	REFERENCE
1	Factor V	R506Q (LeidenG1691A)	4	**15**	GpIIb/IIIa	Leu33Pro (PLA2)	16
2	Factor V	R485K (A1628G)	5	**16**	Gplb IX-V: GpIbα VNTR	(D,C,B,A):C/B	17
3	Factor V	HR2 (His1299Arg)	6	**17**	Gpla/IIa:α2 gene	nt 807 C/T	18
4	Factor V	HR3 (His1254Arg)	1	**18**	TNFα: 308	G/A	19
5	Prothrombin	20210G/A	4	**19**	Apo A1	C93T: G121A	20
6	MTHFR	C677T	7,8	**20**	Apo E	Cys112Arg	21
7	ACE	I/D	2,3	**21**	VCAM-1	G1385C	22
8	Fibrinogen – Beta chain	455 G/A	9	**22**	E-Selectin	G98T	23
9	Fibrinogen – Beta chain	148 C/T	10	**23**	E-Selectin	C1839T L554F	23
10	Factor VII	R353Q (Arg353Gln)	11	**24**	E-Selectin	A561C S128R	24
11	FactorXIII	Val34Leu	12	**25**	L-Selectin	C188G T49S	24
12	TM	Ala455Val:C/T	13	**26**	L-Selectin	T668C F206L	25
13	TM	A25Thr:G/A	14	**27**	P-Selectin	C715A	26
14	PAI-1	675 4G/5G	15	**28**	ICAM-1	G778A G214R	23

Note: MTHFR, Methylenetetrahydrofolate reductase; ACE, Angiotensin converting enzyme; TM, Thrombomodulin; PAI-1, Plasminogen activator inhibitor-1; Gp, platelet glycoprotein; TNFα, Tumor necrosis factor α; VCAM-1, vascular cell adhesion molecule 1; ICAM-1, intercellular adhesion molecule 1

extracted from the 5 ml blood collected in EDTA and shipped to the Core Laboratory, collected during TCD screening as part of STOP-II.

Cases and controls will be matched for age, sex, and weight, which are the most important covariates. The effect of any remaining differences in additional confounding variables between cases and controls will be explored by adjusting for these covariates in the analyses, for example through using conditional logistic regression models.

Human Biological Materials

To ensure confidentiality and protect the privacy of subjects they will not be identified by name, acrostics will be used. Authorized officials from the state and federal governments and authorized representatives of the Medical College of Georgia will have access to confidential data, which will identify the patients. Patients will not be identified in any reports or publications resulting from this study. Banked STOP DNA samples were anonymized according to a plan approved by the Institutional Review Boards of the Medical College of Georgia and the New England Research Institute. Informed consent was obtained for subjects screened as part of the STOP II study. In accordance with the National Bioethics Advisory Commission Recommendations [16] subjects and their physicians will not be notified of the test results unless all three of the following conditions are met:

1) The findings are scientifically valid and are confirmed

2) The findings have significant implications for the subject's health concerns

3) A course of action to treat these concerns is readily available

SNP Genotyping

DNA will be extracted from peripheral blood mononuclear cells and 5 μg of genomic DNA from each subject will be used for each multiplex SNP genotyping assay (the STOP DNA bank has 66–1740 μg of DNA from randomized patients). The 28 polymorphisms in 20 candidate genes and the SNP markers in the genomewide screen will be genotyped by high-throughput SNP screening using the MassARRAY™ System (Sequenom Inc., San Diego, CA). The principles of this method are detailed in references [17] and [18]. Briefly, approximately 300 bp fragments of each candidate gene, with the SNP site in the middle, will be amplified by automated PCR procedure. Multiplexing of the PCR reactions will be used as determined by the Sequenom software. One of the amplification primers includes a 5' biotin tag, which is targeted in a streptavidin-magnetic bead purification step to generate a single stranded template for the primer extension reaction. An extension primer complimentary to the template at a region directly adjacent to the SNP site is added to the single stranded template. This is followed by the MassEXTEND™ reaction during which the primer is extended across the SNP site by using a sequencing polymerase reaction. This is controlled by a mixture of dideoxy and deoxy nucleotide triphosphates, the ratio of which varies depending upon the assay design protocol obtained from the SpectroDESIGNER™ software (Sequenom). The primer extension products will differ depending upon the polymorphic base present at the SNP site. The difference in molecular weight between these products is detectable by mass spectrometry. Following the

extension reaction, the original sample template is separated and removed from the extended primers by the use of a resin. The purified samples are then transferred onto a SpectroCHIP™ (from either a 96 well or 384 well plate) with the SpectroJET™ dispenser. The CHIP is placed into the mass spectrometer, which takes advantage of the MALDI-TOF (Matrix Assisted Laser Desorption/Ionization Time-of-Flight) mass spectrometry. The mixture of the test molecules (extension products) with the organic matrix produces a crystalline dispersion. The matrix is hit with a pulse from a laser beam. The sample and matrix are vaporized and the primer extension products are expelled into the flight tube. As the primer extension products are negatively charged when an electrical field pulse is subsequently applied, they are launched down the flight tube toward the detector. The time between application of the electrical field pulse and collision of the primer extension products with the detector is known as "time-of-flight" and is a very precise measure of the product's molecular weight. The SpectroTYPER™ software gathers the time-of-flight information and applies algorithms to provide accurate, automated genotype calling. This method is fast and more accurate compared to hybridization-based methods of SNP detection [17,18]. The whole process is automated.

Data Management

Web-based data entry will be used to load the data from this study into the password-protected STOP II database at the New England Research Institutes (NERI). A data entry screen specific to this study will be developed for this purpose. Adding the data to the STOP II database will greatly facilitate linking the genotyping data to other patient characteristics (e.g. TCD status) during data analysis. The database was developed using NERI's browser-based data management system, ADEPT, and an ORACLE relational database engine. As part of standard QC, 10% of the records will be randomly selected at NERI for double data entry. Any problems with high error rates will be brought to the attention of laboratory personnel immediately. If necessary, further records entered by the same individual will also be reviewed.

Statistical analysis for the candidate gene study

As a first step in the analysis, the genotypic frequencies for each gene will be tested for departure from Hardy-Weinberg equilibrium. Statistically significant departures could indicate biased sampling or they could indicate that the study population is experiencing detectably strong natural selection at that locus.

The null hypothesis to be tested for each gene is that the genotypic frequencies in patients with normal and high TCD are the same. This will be tested using a chi-square test of association of TCD status and genotype for each gene. Heterozygotes will be pooled with homozygous mutants to form 2 × 2 tables (mutant present/absent vs. TCD status). Fisher's exact test will be used instead of chi-square tests if the frequency of a given allele is too low to justify the assumptions of the chi-square test. Rejection of the null hypothesis of no association between genotype and TCD status will lead to the conclusion that the proportion of patients with mutant alleles is different in patients with normal and high TCD. The Bonferroni correction for multiple testing will not be used because correcting for tests at 20 loci would produce an extremely small critical p-value. In the absence of such a correction, the results of this study will need to be considered with caution and replication of significant findings in an independent sample or follow-up study will be important.

A multivariate approach will then be employed to determine if TCD status varies with combinations of different genes; i.e. to determine if the probability that a patient with mutant alleles for two or more genes is a case (i.e. has high TCD) differs from the probability that is predicted from the independent effects of those genes. Loglinear models will be employed for this [19]. Suppose, for example, that the data from two genes are combined with TCD status to form a 2 × 2 × 2 table. Rejection of the null hypothesis of no three-way association in the table would lead to the conclusion that the proportion with high TCD depends upon the joint effects of the two genes rather than on their separate effects. The cell frequencies in the table would then be inspected to determine which combinations of alleles accounted for the statistically significant results; i.e. to identify the combinations that occur with elevated frequency in patients with high TCD. Bishop et al. [19] provide methods for calculating standard errors for the cell frequencies for this inspection.

Generally, statistical tests of three-way and higher order associations are less powerful than tests of two-way associations in the same table. More specific statements regarding power are difficult without explicit statements regarding the magnitude of interaction to be detected and there is very little information available on which to base the latter. Therefore, the analysis of individual genes should be considered the primary analysis in this study. Interactions involving more than three or four variables are generally very difficult to interpret. Therefore, the analysis will be limited to tables involving no more than two or three genes. Even that restriction is unlikely to be sufficient, given the large number of tables involving either two or three genes that could be formed from the data. Therefore, multivariate modeling will begin with any genes that were statistically significantly related to TCD status when the data for each gene were analyzed separately.

Table 2: Power to detect the difference between the genotypic distributions in two populations, assuming a sample of 230 from population 1 (cases) and 400 from population 2 (controls).

P_1	P_2	P_1-P_2	POWER
0.030	0.010	0.020	0.95
	0.009	0.021	0.97
	0.008	0.022	0.99
	0.007	0.023	0.99
0.050	0.025	0.025	0.77
	0.023	0.027	0.85
	0.021	0.029	0.92
	0.020	0.030	0.94
0.100	0.057	0.043	0.78
	0.055	0.045	0.82
	0.050	0.050	0.91
0.200	0.130	0.070	0.74
	0.120	0.080	0.86
	0.110	0.090	0.94
	0.100	0.100	0.98
0.300	0.200	0.100	0.80
	0.190	0.110	0.88
	0.180	0.120	0.94
0.400	0.280	0.120	0.79
	0.270	0.130	0.84
	0.250	0.150	0.94
0.500	0.360	0.140	0.77
	0.350	0.150	0.84
	0.330	0.170	0.93

Note: P1, the proportion of homozygous mutants in population 1; P2, the proportion of homozygous mutants in population 2.

Power estimates for the candidate gene study

In the calculations below, heterogyzotes are pooled with homozygous mutants to reduce each population to two states (mutant present or absent). The power to detect a difference in genotypic frequencies between two populations was determined by Monte Carlo simulation. The simulations were based on three assumptions: (i) only two alleles are involved in the comparison for a given gene; (ii) the populations are in Hardy-Weinberg equilibrium; and (iii) genotypes will be obtained from 230 cases and 400 controls. The simulations proceeded as follows.

1. Specify the proportion of homozygous mutants in each population (p_1 for cases and p_2 for controls). The proportions of homozygous wild type and heterozygous patients in the i^{th} population are then $(1 - p_i^{1/2})^2$ and $2p_i^{1/2}(1 - p_i^{1/2})$ respectively.

2. Draw random samples of size 230 and 400 from the two multinomial populations defined above. Pool the heterozygotes and homozygous mutants to reduce each population to two states (mutant present/absent).

3. Compare the genotypic frequencies in the two samples using a chi-square test.

4. Repeat steps 2–3 5,000 times for each combination of p_1 and p_2. Estimate statistical power as the percentage of simulations in which the null hypothesis was rejected.

Results are presented in table 2. On the assumption that the goal of the study is to identify mutant alleles that occur with increased frequency in patients with high TCD, the table is limited to cases in which $p_1 > p_2$. To save space, the table is further limited to entries that just span the range of differences, $p_1 - p_2$, that can be detected with 80–90% power for each specified value of p_1. Based on these power analyses we conclude that we have good power to detect realistic differences in frequencies of alleles that increase risk of cerebrovascular disease and stroke in cases as compared to controls.

Design of DNA pooling for genomewide association screen

DNA pooling is a practical way to reduce the cost of large-scale case-control association studies, because it allows measurement of allele frequencies in groups (or pools) of individuals, thereby reducing the number of PCR reactions and genotyping assays dramatically. Primer extension is the technique that has been most commonly used in pooling studies to genotype SNPs and the results have been very good [20]. Resolution of allele frequency differ-

ences between cases and controls in pooling studies is limited, however, by so called pool-measurement and pool-formation errors. Differential amplification of different SNPs during PCR and the limited accuracy of the detection method lead to pool-measurement errors. For example, allele frequency estimates of pools using MALDI-TOF mass spectroscopy have been reported to deviate from real by approximately 3% [21]. Another source of error is caused by unequal amount of DNA being contributed by individuals that make up the pool (pool-formation errors). In principle the power of pooling studies can be improved by creating multiple pools from the same individuals (reducing pool-formation errors) and multiple measurement of allele frequencies (reducing the pool-measurement error). However, for our study we propose to follow the recommendations as recently outlined by Sham et al [20]. They state that a two-stage design, in which markers showing positive association in a pooling study are followed up by confirmatory individual genotyping, might represent the best trade-off between the cost savings of pooling and the full information provided by individual genotyping. We propose to test the full marker set of 100,000 SNPs in the genome screen using pooled assays with a relatively liberal p-value (e.g. 0.01–0.001) to allow adequate power even with information loss. Markers that show significance in the pooled assay will then be genotyped in all individuals of the original case-control sample to confirm the association.

Discussion

Identifying genes for polygenic ischemic stroke has been a difficult task and most human studies have employed a candidate gene approach [5] although at least one genomewide linkage scan in affected sibling pairs is on the way [22]. We propose to use a combination of a candidate gene association study and genomewide association scan in children with Hb SS at high (230 cases) or low (400 controls) risk for cerebrovascular disease as determined by their TCD velocity.

Association studies of candidate genes for complex diseases have been criticized because of non-replication of results [23] and studies of genetic risk factors in ischemic stroke have been no exception [5,22]. Rish [24] has argued that the most likely reason for the high false-positive rate is the low prior probability in most candidate gene studies that the examined polymorphisms are causally related to the disease outcome. All 28 polymorphisms in 20 candidate genes that we propose to study have been shown to be associated with a hypercoagulable state or have been shown to be involved in pathways leading to the development of vasculopathy and, therefore, have a high probability to be related to the TCD velocity phenotype.

One of the causes for conflicting results in candidate gene studies and in particular for spurious associations is believed to be population stratification or admixture [25]. Population stratification refers to the presence of subgroups, e.g. ethnic groups, in the sample and can potentially cause a spurious association between the locus and trait. A spurious association due to population stratification can only occur when two conditions hold: (i) the population strata differ with respect to the phenotype, and (ii) the population strata differ in allele frequencies. Although population stratification is frequently used as an explanation for non-replicable associations in the literature, there are few actual examples to support this assumption [24] and experts in the field now agree that the problem has probably been overstated [25]. For example, Wacholder et al. [26] argue that population stratification of an extent large enough to distort results is unlikely to occur in many realistic situations. Ardlie et al. [27] evaluated 4 moderately sized case-control studies for the presence of population structure and concluded that carefully matched case-control samples in cosmopolitan US and European populations are unlikely to contain levels of population stratification that would result in significantly inflated numbers of false positive associations. We therefore believe that hidden stratification is unlikely to be a problem in our case-control study of African-American children with sickle cell disease at high or low risk for cerebrovascular disease based on their TCD values. However, methods are being developed by which unlinked genetic markers can be used to detect stratification and even correct for it when it is present [28,29]. In the unlikely event of spurious association results, we will be able to use these methods to detect and adjust for the effect of the population stratification present. Furthermore, the study of other genetic markers, namely the distribution of β-globin haplotypes and the frequency of deletional α-thalassemia among African-American patients with sickle cell disease from different centers in the United States do not show significant differences arguing against significant population admixture in this group [4,11,30,31].

Our genomewide association scan aimed at identifying novel polymorphisms associated with stroke risk will necessarily have a somewhat exploratory character. The reason is simply that this approach has only recently become possible with the advent of high throughput SNP genotyping and we are not currently aware of any published studies that have used genomewide association for identification of susceptibility loci of complex traits. Two issues are worth discussing at this point: (i) the optimization of selection criteria for the case and control pools and (ii) the total number of SNP markers needed to provide adequate coverage of the whole genome.

Our definition of cases and controls in this study is based on absolute cut-off values for TCD velocity. For quantitative traits like TCD velocity, loss of efficiency in a pooling design is due to loss of information from within-pool phenotypic differences. In the absence of experimental errors, this information loss can be minimized by optimizing the criteria for selection of individuals from the extreme tails of the distribution for the two pools, which turns out to be the upper and lower 27% of this distribution [20]. Interestingly, this optimal pooling fraction is largely independent of marker frequency and mode of inheritance of the trait. Exploration of alternative selection criteria for cases and controls in our genomewide association study will be dependent on the combined TCD distribution of the 1000 to1400 new patients that will be screened as part of the STOP-II study and the 130 patients (TCD ≥ 200 cm/s) from the original STOP study that are already available.

The number of SNPs that are required for whole-genome linkage disequilibrium (LD) mapping has been a hotly debated issue in the last few years. Initial simulation estimates based on monotonic population expansion suggested that useful LD was unlikely to extend beyond 3 kb, implying a number of 500,000 SNPs necessary for a whole-genome scan [32]. Another study looking at real data found that LD extended over much larger distances and arrived at an estimate as small as 30,000 SNPs, or 1 SNP per 100 kb [33]. Recent evidence suggests that the genome consists of blocks of high LD (haplotype blocks) separated by recombination hot spots [34]. Within each block, a limited number of common haplotypes (three to five) typically capture about 90% of all chromosomes, which means that a reduced number of SNPs would be needed to characterize these haplotypes [35]. Thus, the recently initiated construction of a haplotype map of the human genome may facilitate selection of SNP markers for genomewide association studies like ours. In the meantime, replication of significant associations detected in the candidate gene part of the study with our genomewide approach should provide us with practical estimates for the likely density of SNP markers needed for successful LD mapping.

The practical and clinical implications of potential findings of this study remain unclear. The association of one or more polymorphisms in the candidate genes with a high TCD (high risk) phenotype will enable the investigators to streamline the screening and follow-up programs towards those with high risk as determined by genetic testing. The incorporation of any preventive or therapeutic measures into this program will depend not only on the findings of association of high TCD phenotype with certain SNPs but also on the results of the ongoing STOP II study, which is studying the optimal duration of transfusion in these high risk children.

In conclusion, we believe that our dual approach of a candidate gene and whole-genome association study will yield important information on genetic risk factors for cerebrovascular disease in sickle cell disease. Not only will this generate important knowledge for improving treatment and prevention options of patients with Hb SS, polymorphisms showing a significant association will also be strong candidates for stroke risk in the general population. Such associations can efficiently be confirmed in banked DNA of large cohorts of stroke cases and controls [22].

List of Abbreviations

ACE angiotensin converting enzyme

CSSCD Cooperative Study of Sickle Cell Disease

Gp platelet glycoprotein

Hb SS sickle cell anemia

ICAM-1 intercellular adhesion molecule 1

LD linkage disequilibrium

MALDI-TOF Matrix Assisted Laser Desorption/Ionization Time-of-Flight

MCG Medical College of Georgia

MTHFR Methylenetetrahydrofolate reductase

PAI-1 plasminogen activator inhibitor-1

SNP single nucleotide polymorphism

STOP study Stroke Prevention Trial in Sickle Cell Anemia study

TCD transcranial doppler

TM thrombomodulin

TNFα tumor necrosis factor α

VCAM-1 vascular cell adhesion molecule 1

Competing Interests

None declared.

Author's Contributions

GTA participated in primer design and in the drafting of the manuscript.

HS participated in the design of the study and in the drafting of the manuscript.

VCM participated in the design of the study and by consenting and collecting patient samples.

BC participated in the design of the study and in its coordination.

DB participated in the design of the study and will perform all statistical analysis of the data.

RJA participated in the design of the study and serves as the PI of the parent STOP and STOP II studies.

FK participated in the design of study methods.

AK conceived of the study, participated in its design and coordination, and finalized the manuscript.

Additional material

Additional File 1

The 28 polymorphisms that will be studied are referenced in an attached document: Design Paper.Reference list for polymorphisms.doc.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2350-4-6-S1.doc]

Acknowledgements

This study is supported by NHLBI grant HL67682-01

References

1. Steinberg M and Embury S: **Natural history: Overview** In: *Sickle Cell Disease: Basic Principles and Practice* Edited by: *Embury SH, Hebbel RP, Mohandas N, Steinberg M. New York, Raven*; 1994:349-352.
2. Powars DR, Chan L and Schroeder WA: **B^s-Gene-cluster haplotypes in sickle cell anemia: clinical implications** *Am J Pediatr Hematol/Oncol* 1990, **12(3):**367-374.
3. Powars ER, Wilson B, Imbus C, Pegelow C and Allen J: **The natural history of stroke in sickle cell disease** *Am J Med* 1978, **65:**461-471.
4. Ohene-Frempong K, Weiner SJ, Sleeper LA, Miller ST, Embury S, Moohr JW, Wethers DL, Pegelow CH and Gill FM: **Cerebrovascular accidents in sickle cell disease: Rates and risk factors** *Blood* 1998, **91(1):**288-294.
5. Hassan A and Markus HS: **Genetics and Ischaemic Stroke** *Brain* 2000, **123:**1784-1812.
6. Lane DA and Grant PJ: **Role of Hemostatic Gene Polymorphisms in Venous and Arterial Thrombotic Disease** *Blood* 2000, **95:**1517-1532.
7. Adams RJ, Kutlar A, McKie VC, Carl EM, Nichols FT, Liu JC, McKie K and Clary A: **Alpha Thalassemia and Stroke Risk in Sickle Cell Anemia** *Am J Hemat* 1994, **45:**279-282.
8. Adams RJ, McKie V, Nichols FT, Carl E, Zhang DL, McKie K, Figueroa R, Litaker M, Thompson W and Hess DC: **The Use of Transcranial Ultrasonography to Predict Stroke in Sickle Cell Disease** *N Engl J Med* 1992, **326(9):**605-610.
9. Adams RJ, McKie VC, Carl EM, Nichols FT, Perry R, Brock K, McKie K, Figueroa R, Litaker M, Weiner S and Brambilla DJ: **Long Term Stroke Risk in Children with Sickle Cell Disease Screened with Transcranial Doppler** *Ann Neurol* 1997, **42(5):**699-704.
10. Adams RJ, McKie VC, Brambilla DJ, Carl EM, Gallagher D, Nichols FT, Roach S, Abboud M, Berman B, Driscoll C, Files B, Hsu L, Hurlet A, Miller S, Olivieri N, Pegelow C, Scher C, Vichinsky E, Wang W, Woods G, Kutlar A, Wright E, Hagner S, Tighe F, Lewin J, Cure J, Zimmerman RA and Waclawiw M: **Stroke Prevention Trial in Sickle Cell Anemia ("STOP"): Study Design** *Controlled Clinical Trials* 1997, **19:**110-129.
11. Adams RJ, Brambilla DJ, McKie VC, Hsu L, Files B, Vichinsky E, Pegelow C, Abboud M, Woods G, Olivieri N, Driscoll C, Miller S, Wang W, Hurlet A, Scher C, Berman B, Carl EM, Nichols FT, Roach S, Kutlar A, Wright E, Zimmerman RA, Gallagher D, Waclawiw M and Bonds D: **Transfusion Prevents First Stroke in Children with Sickle Cell Disease: The "STOP" Study** *N Engl J Med* 1998, **339(1):**5-11.
12. Brass L, Pavlakis S, DeVivo D, Piomelli S and Mohr J: **Transcranial Doppler Measurements of the Middle Cerebral Artery. Effect of Hematocrit** *Stroke* 1988, **19:**1466-1469.
13. Hurlet-Jensen AM, Prohovnik I, Pavlakis SG and Piomelli S: **Effects of Total Hemoglobin and Hemoglobin S Concentration on Cerebral Blood Flow During Transfusion Therapy to Prevent Stroke in Sickle Cell Disease** *Stroke* 1994, **25(8):**1688-1692.
14. Solovey A, Gui L, Key NS and Hebbel RP: **Tissue Factor Expression By Endothelial Cells in Sickle Cell Anemia** *J Clin Invest* 1998, **101(9):**1899-1904.
15. Nowak-Gottl U, Strater R, Heinecke A, Junker R, Koch HG, Schuierer G and von Edckardstein A: **Lipoprotein (a) and Genetic Polymorphisms of Clotting Factor V, Prothrombin, and Methylenetetrahydrofolate Reductase are Risk Factors of Spontaneous Ischemic Stroke in Childhood** *Blood* 1999, **94(11):**3678-3682.
16. National Bioethics Advisory Commission: **Research Involving Human Biological Materials: Ethical Issues and Policy Guidance** *Rockville, Md, National Bioethics Advisory Commission* 1999, **1:**.
17. Buetow KH, Edmonson M, MacDonald R, Clifford R, Yip P, Kelley J, Little DP, Strausberg R, Koester H, Cantor CR and Braun A: **High-throughput Development and Characterization of a Genomewide Collection of Gene-Based Single Nucleotide Polymorphism Markers by Chip-Based Matrix-Assisted Laser Desorption/Ionization Time-of-Flight Mass Spectrometry** *PNAS* 2001, **98(2):**581-584.
18. Heaton MP, Grosse WM, Kappes SM, Keele JW, Chitko-McKown CG, Cundiff LV, Braun A, Little DP and Laegreid WW: **Estimation of DNA Sequence Diversity in Bovine Cytokine Genes** *Mammalian Genome* 2001, **12:**32-37.
19. Bishop YMM, Fienberg SE and Holland PW: **Discrete multivariate analysis: Theory and Practice** *Cambridge MA: MIT Press* 1975.
20. Sham P, Bader JS, Craig I, O'Donovan M and Owen M: **DNA Pooling: A Tool for Large-scale Association Studies** *Nature Rev Genet* 2002, **3:**862-871.
21. Werner M, Sych M, Herbon N, Illig T, König IR and Wjst M: **Large-Scale Determination of SNP Allele Frequencies in DNA Pools Using MALDI-TOF Mass Spectrometry** *Hum Mutat* 2002, **20:**57-64.
22. Meschia JF, Brown RD, Brott TG, Chukwudelunzu FE, Hardy J and Rich SS: **The Siblings Ischemic Stroke Study (SWISS) Protocol** *BMC Med Genet* 2002, **3:**1.
23. Tabor HT, Rish NJ and Myers RM: **Candidate-gene Approaches for Studying Complex Genetic traits: Practical Considerations** *Nature Rev Genet* 2002, **3:**1-7.
24. Rish N: **Searching for Genetic Determinants in the New Millennium** *Nature* 2000, **405:**847-856.
25. Cardon LR and Bell JI: **Association Study Designs for Complex Disease** *Nature Rev Genet* 2001, **2:**91-99.
26. Wacholder S, Rothman N and Caporaso N: **Population Stratification in Epidemiologic Studies of Common Genetic Variants and Cancer: Quantification and Bias** *J Natl Cancer Inst* 2000, **92:**1151-1158.

27. Ardlie KG, Lunetta KL and Seielstad M: **Testing for Population Subdivision and Association in Four Case-Control Studies** *Am J Hum Genet* 2002, **71:**304-311.
28. Pritchard JK and Rosenberg NA: **Use of Unlinked Genetic Markers to Detect Population Stratification in Association Studies** *Am J Hum Genet* 1999, **65:**220-228.
29. Satten GA, Flanders D and Yang Q: **Accounting for Unmeasured Population Substructure in Case-Control Studies of Genetic Association Using a Novel Latent Class Model** *Am J Hum Genet* 2001, **68:**466-477.
30. Hattori Y, Kutlar F, Kutlar A, McKie VC and Huisman THJ: **Haplotypes of β^s Chromosomes Among Patients with Sickle Cell Anemia from Georgia** *Hemoglobin* 1986, **10(6):**623-642.
31. Schroeder WA, Powars DR, Kay LM, Chan LS, Huynh V, Shelton JB and Shelton JR: **β-cluster Haplotypes, α-gene Status, and Hematological Data from SS, SC, and S-β-Thalassemia Patients in Southern California** *Hemoglobin* 1989, **13(4):**325-353.
32. Kruglyak L: **Prospects for Whole-Genome Linkage Disequilibrium Mapping of Common Disease Genes** *Nature Genet* 1999, **22:**139-144.
33. Collins A, Lonjou C and Morton NE: **Genetic Epidemiology of Single-Nucleotide Polymorphisms** *PNAS* 1999, **96:**15173-15177.
34. Goldstein DB: **Islands of Linkage Disequilibrium** *Nature Genet* 2001, **29:**109-111.
35. Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A and Faggart M *et al.*: **The Structure of Haplotype Blocks in the Human Genome** *Science* 2002, **296:**2225-2229.

Pre-publication history

The pre-publication history for this paper can be accessed here:

http://www.biomedcentral.com/1471-2350/4/6/prepub