

ABSTRACT

This project presents a comprehensive approach to the identification of new genes that influence the risk for developing stroke. Stroke is the leading cause of death in Portugal and the third leading cause of death in the developed world. It is even more disabling than lethal, and the persistent neurological impairment and physical disability caused by stroke have a very high socioeconomic cost. Moreover, the number of affected individuals is expected to increase with the current aging of the population. Stroke is a “brain attack” cutting off vital blood and oxygen to the brain cells and it is a complex disease resulting from environmental and genetic factors. Major known risk factors include family history, age, hypertension, hypercholesterolemia, diabetes, cardiovascular disease, smoking and alcohol consumption. The common forms of stroke can be classified in two major clinical types: ischemic stroke (IS; most frequent) or hemorrhagic (10-20% of cases) stroke. Identification of genes increasing susceptibility to stroke could have far-reaching public health impact. The genetic component of this disease has been demonstrated in twin, family and animal model studies, and mutations have been found in several genes in rare classical Mendelian forms of stroke. However, very few susceptibility genes for the common forms of stroke have been identified and association studies have mostly reported conflicting results.

In this project, to accomplish our goals in the study of IS, we first performed some candidate gene association analysis. Concomitantly, we applied the genomic convergence (GC) approach combining whole genome linkage screens, expression analysis, and case-control association studies. This unified, comprehensive, and multidisciplinary approach has not yet been implemented in other studies of stroke but the availability of new genetic, molecular and statistical tools, as well as their success in the study of other complex diseases, made such an approach both timely and essential.

Since phosphodiesterase 4D (*PDE4D*) and arachidonate 5-lipoxygenase-activating protein (*ALOX5AP*) genes have been in recent years controversially implicated in the risk of IS, we assessed their association with IS in a Portuguese cohort. *PDE4D* degrades second messenger cyclic adenosine monophosphate, a key signal transduction molecule in different cell types, including inflammatory, vascular endothelial and smooth muscle cells. *ALOX5AP* is involved in the initial steps of leukotriene synthesis and is secreted by various types of inflammatory cells clustering at the injured sites in blood vessels. We genotyped 67 single nucleotide polymorphisms (SNPs) in the 5'end of *PDE4D* and 24 SNPs in *ALOX5AP* and both 10kb flanking regions on 565 Caucasian Portuguese patients and 518 unrelated controls. These SNPs are either tagging SNPs from HapMap or SNPs previously found associated. We tested their allelic, genotype and haplotype associations with IS risk, using standard qui-square tests (χ^2) and multivariate logistic regression to adjust the analyses of association with risk for confounding factors, namely hypertension, diabetes and ever smoking. None of the previously associated SNPs were found associated with IS risk in our cohort, and considering the number of tests performed, we found no major involvement of other variants in these genes in stroke susceptibility in

the Portuguese population. Only the SNP rs7442640 in *PDE4D* shows an association (p-value = 0.006) with IS risk when genotypic tests were adjusted for covariates, and SNP rs4491352 downstream of the *ALOX5AP* shows a modest evidence of association with IS risk ($0.017 < p\text{-value} < 0.025$) for allelic and unadjusted genotypic tests. Performing a meta-analysis including all recently published studies and our Portuguese and Spanish samples for SNP41, SNP45, SNP56, SNP87, and SNP89 (found associated in the original report) in *PDE4D*, no significant association results with IS risk were found. However, we found that SNP rs10507391 (or SG13S114) in *ALOX5AP*, which is part of the HapA haplotype, was associated with IS risk as described in the original study both in the Iberian population and in the meta-analysis performed. These results suggest that *PDE4D* may not constitute a major risk factor for IS in the Portuguese or Spanish populations, contrasting with *ALOX5AP* which may confer an increased risk of IS in the Iberian and other populations.

We also assessed the association of the kalirin gene (*KALRN*) with IS in our Portuguese cohort since several recent studies have implicated its variants with susceptibility to cardiovascular and metabolic phenotypes, but no studies have yet been performed in stroke. Cerebrovascular and cardiovascular diseases are both complex disorders resulting from the interplay of genetic and environmental factors, and may share several susceptibility genes. *KALRN* is involved, among others, in the inhibition of inducible nitric oxide synthase, in the regulation of ischemic signal transduction, and in neuronal morphogenesis, plasticity and stability. Our goal was to determine whether SNPs in the *KALRN* region on 3q13, which includes the ropporin gene (*ROPNI*), predispose to IS in our cohort of Portuguese patients and controls. We genotyped 34 tagging SNPs in the *KALRN* and *ROPNI* chromosomal region on 565 IS patients and 517 unrelated controls from our Portuguese case-control sample, and performed genotype imputation for 405 markers on chromosome 3. We tested the single-marker and haplotype association of these SNPs with IS as explained above. One SNP in the *ROPNI-KALRN* intergenic region (rs4499545) and two SNPs in *KALRN* (rs17286604 and rs11712619) showed significant ($0.003 < p\text{-value} < 0.049$) allelic and genotypic (unadjusted and adjusted for hypertension, diabetes and ever smoking) association with IS risk. Thirty-two imputed SNPs also showed an association at $p\text{-value} < 0.05$, and actual genotyping of three of these polymorphisms (rs7620580, rs6438833 and rs11712039) validated their association. Furthermore, rs11712039 was associated with IS ($0.001 < p\text{-value} < 0.01$) in the genome-wide association study (GWAS) published by Ikram and co-authors (2009). These studies suggest that variants in the *KALRN* constitute risk factors for IS, and that *KALRN* may be a common genetic risk factor for vascular diseases.

Additionally, we tested the association with the IS risk of the complement inhibitor factor H gene (*CFH*), as well as of several candidate genes related to neuroprotection: erythropoietin (*EPO*), heme-oxygenase 2 (*HO2*), and kallikrein 1 (*KLK1*) genes. *CFH* has been suggested to play an important role in the complement inhibition in atherosclerotic lesions and has been consistently associated with an increased risk for age-related macular degeneration (AMD) and myocardial infarction (MI) which share several risk factors with stroke. On the other hand, *EPO*, *HO2* and *KLK1* are neuroprotectors, for instance, in brain hypoxia and ischemia (*EPO*) and against induced stroke

(*KLK1*). The polymorphism in *CFH* (rs1061170) previously associated with AMD and MI seems to be modestly associated (388 Portuguese patients and 461 controls; $0.030 < p\text{-value} < 0.035$) with the IS risk in allelic and unadjusted genotypic tests, but no haplotype tagging SNP in this gene was clearly associated with IS in the GWAS performed by Ikram *et al.* (2009). Although we only study one polymorphism in the gene, these results did not justify a more in depth analysis. We genotyped 3, 3 and 5 tagging SNPs in the coding and 10 kb flanking regions of the *EPO*, *HO2* and *KLK1*, respectively, on 565 IS patients and 518 controls from our Portuguese sample. No single-marker and haplotype associations were found for the studied SNPs in these neuroprotector genes, suggesting that they do not constitute genetic risk factors of IS.

To identify novel susceptibility genes for IS, we applied the proposed GC approach. We performed gene expression analysis in peripheral blood mononuclear cells of 20 IS cases and 20 age- and sex-matched controls using Affymetrix GeneChip Human U133 Plus 2.0 arrays, which represent 47,000 human transcripts and variants. We identified several affected biological pathways in stroke patients, such as the cell adhesion molecules pathway. 16 out of the differentially expressed genes among cases and controls (1.2 fold-change cut-off and uncorrected p-value < 0.05) map to linkage peaks reported in published human whole-genome linkage studies. All tagging SNPs from these prioritized genes and from their 10 kb flanking regions (a total of 191 SNPs) were genotyped in 565 IS cases and 520 controls from our Portuguese biobank. Single-marker and haplotype association tests were performed. Association results suggest that variants in 6 (*HEMGN*, *GFI1B*, *TMTC4*, *TTC7B*, *SDC4* and *TUBB1*) out of the 16 prioritized genes may constitute risk factors for IS in the Portuguese population. Several of the associated SNPs in these genes are also part of associated haplotypes. SNPs like the intronic rs9582406 and rs946845 polymorphisms in *TMTC4*, and the intronic rs2284278 in *SDC4*, were associated ($0.015 < p\text{-value} < 0.050$) with IS risk in all tests performed. On the other hand, the intronic SNP rs1535321 in *TTC7B* showed an association (p-value = 0.009) in allelic and unadjusted genotypic tests, even though no association in the adjusted test for hypertension, diabetes and ever smoking was verified.

To follow-up these results, SNPs with single low-stringency significance association (p-value < 0.05) in at least one of the tests performed, and some SNPs that define associated haplotypes, were then genotyped in a Spanish dataset. A total of 570 Caucasian IS cases and 390 controls were included, and allelic, genotype and haplotype association tests with IS risk were conducted using also χ^2 tests and multivariate logistic regression to adjust the analyses for hypertension, diabetes, dyslipidemic status and cigarette smoking. The same analyses were performed for the atherothrombotic, cardioembolic and lacunar forms of stroke. We found some significant associations with IS risk in *TMTC4*, *TTC7B* and *SDC4*, however, the only replicated SNP was rs9582406 in *TMTC4* which was associated with IS risk in unadjusted tests (p-value = 0.019) in the Spanish case-control dataset. This SNP is also associated with the risk of atherothrombotic and lacunar forms of stroke for allelic and genotypic tests ($0.011 < p\text{-value} < 0.049$). For *TTC7B* and *SDC4*, the single SNPs

and haplotypes associated in the Spanish sample were not the same as in the Portuguese cohort. The SNP rs6073708 in *SDC4* was associated in the Spanish dataset for all tests performed ($0.027 < p\text{-value} < 0.029$). However, none of the studied SNPs were clearly replicated in the recent well-powered GWAS reported by Ikram and colleagues (2009).

The overall results suggest that *HEMGN* and *GFI1B* (that were not replicated in the Spanish dataset) may constitute risk factors for IS in the Portuguese population, being important the enlargement of the Portuguese sample to validate the positive results. Similarly, *TUBB1*, which could not be genotyped in the Spanish cohort for technical problems, may constitute a risk factor for IS in the Portuguese population but should be further studied in other datasets to validate and understand its role in IS. *TMTC4* may constitute a risk factor for IS in the Iberian population, and *TTC7B* and *SDC4*, with the observed heterogeneity of their significant association results among cohorts, may be novel risk factors for IS, being likely that their true susceptibility variants have not been studied yet. *TTC7B* was one of the top hits for major cardiovascular diseases in the Framingham Heart Study 100K project GWAS (Larson *et al.* 2007). For this gene, several SNPs and haplotypes in the intron 5 – intron 6 region associated in the Portuguese and Spanish datasets individually and combined, were modestly associated in the GWAS reported by Ikram and colleagues (2009). Multiple independent lines of evidence therefore support the role of *TTC7B* in stroke susceptibility, but further work is warranted to pinpoint the exact risk variant and to elucidate its pathogenic potential.

In this project, given the very large number of SNPs tested, none of the significant findings would survive to multiple testing correction. However, it is generally accepted that replication in multiple independent datasets remains the gold-standard of association studies, even for modest associations. If our findings could be confirmed in other independent datasets and a most complete study of other possible genetic variants in the loci of interest could be performed, we think that functional studies of the genes and of their causative variants will allow a significant improvement of our knowledge on stroke disease. Deep sequencing may have to be used to precisely identify the true susceptibility genetic variants, or rare variants that the association studies have no power to detect.

We suggest that identifying the genetic determinants of stroke using different strategies and populations and analysing them in an integrate view as performed in this project, is a most complete form to study the stroke in order to improve our knowledge of the disease.