Faculdade de Medicina da Universidade de Lisboa

Unidade Neurológica de Investigação Clínica

PhD Thesis

Stroke Genetics and Genomics

Tiago Krug Coelho

Host Institution: *Instituto Gulbenkian de Ciência*

Supervisor at *Instituto Gulbenkian de Ciência*: Doctor Sofia Oliveira

Supervisor at *Faculdade de Medicina da Universidade de Lisboa*: Professor José Ferro

PhD in Biomedical Sciences

Specialization in Neurosciences

2010

A ciência tem, de facto, um único objectivo: a *verdade*.
Não esgota perfeitamente a sua tarefa se não descobre a causa do todo.

Chiara Lubich

A impressão desta dissertação foi aprovada pela Comissão Coordenadora do Conselho Científico da Faculdade de Medicina de Lisboa em reunião de 28 de Setembro de 2010.

As opiniões expressas são da exclusiva responsabilidade do seu autor.

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-value < 0.025) for allelic and unadjusted genotypic tests. Performing a meta-analysis including all recently published studies and our Portugese 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 (*ROPN1*), predispose to IS in our cohort of Portuguese patients and controls. We genotyped 34 tagging SNPs in the *KALRN* and *ROPN1* 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 *ROPN1-KALRN* intergenic region (rs4499545) and two SNPs in *KALRN* (rs17286604 and rs11712619) showed significant (0.003 < p-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-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-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-

oxigenase 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-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-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 ($0.011 < \text{p-value} < 0.049$). 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 ($0.027 < \text{p-value} < 0.029$). 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.

RESUMO

Este projecto apresenta uma abordagem multifactorial para a identificação de novos genes que influenciem o risco de sofrer acidentes vasculares cerebrais (AVCs). Os AVCs são a principal causa de morte em Portugal e a terceira maior causa de morte no conjunto dos países desenvolvidos. São ainda mais incapacitantes do que letais e os distúrbios neurológicos persistentes e a incapacidade física que provocam têm um custo socioeconómico muito elevado. Além disso, o número de pessoas afectadas deverá aumentar com o actual envelhecimento da população. Os AVCs são um "ataque cerebral" que interrompe o fluxo de sangue e oxigénio para as células de determinadas regiões do cérebro e são uma doença complexa resultante de factores genéticos e ambientais. Os principais factores de risco conhecidos incluem a idade, hipertensão, hipercolesterolemia, diabetes, doenças cardiovasculares, consumo de tabaco e de álcool e história familiar. As formas mais comuns dos AVCs podem ser classificadas em dois grandes tipos clínicos: AVCs isquémicos (AVCI; mais frequentes), ou hemorrágicos (10-20% dos casos). A identificação de novos genes que aumentem a susceptibilidade para se sofrer AVCs pode ter um forte impacto na saúde pública. A componente genética da doença tem sido demonstrada em estudos feitos em gémeos, famílias e modelos animais, e foram encontradas mutações em diversos genes em formas mendelianas raras de AVCs. No entanto, poucos genes de susceptibilidade para as formas comuns de AVCs foram até hoje identificados e os estudos de associação realizados apresentam geralmente resultados contraditórios.

Neste projecto, para atingir os objectivos propostos no estudo dos AVCIs, começaram por realizar-se estudos de associação em alguns genes candidatos. Em paralelo, foi aplicada a abordagem de convergência genómica (CG) que combina estudos de ligação em todo o genoma, análises de expressão génica e estudos de associação em casos-controlos. Esta abordagem unificada, abrangente e multidisciplinar foi pela primeira vez implementada no estudo de AVCs. A disponibilidade de novas ferramentas genéticas, moleculares e estatísticas, bem como o seu sucesso no estudo de outras doenças complexas, tornam esta nova abordagem oportuna e essencial.

Uma vez que os genes que codificam a fosfodiesterase 4D (PDE4D) e a proteína activadora de araquidonato 5-lipoxigenase (ALOX5AP) têm sido nos últimos anos controversamente implicados com o risco de sofrer AVCIs, avaliou-se neste projecto a sua associação com os AVCIs na nossa amostra portuguesa. A proteína PDE4D degrada o segundo mensageiro adenosina monofosfato cíclica, uma molécula de transdução de sinal chave em diferentes tipos de células inflamatórias, vasculares endoteliais e musculares lisas. A proteína ALOX5AP está envolvida nos passos iniciais da síntese de leucotrienos e é secretada por vários tipos de células inflamatórias aglomeradas em locais lesados de vasos sanguíneos. Foram genotipados 67 polimorfismos de um só nucleótido (SNPs) na extremidade 5' do gene *PDE4D* e

24 SNPs localizados no gene *ALOX5AP* e ambas regiões flanqueadoras de 10kb, em 565 pacientes caucasianos portugueses e 518 controlos independentes. Os SNPs seleccionados são *tagging* SNPs do projecto HapMap ou SNPs que foram encontrados previamente associados. Testaram-se as associações dos seus alelos, genótipos e haplótipos com o risco sofrer AVCIs, utilizando testes padrão do qui-quadrado (χ^2) e regressões logísticas com múltiplas variáveis para ajustar as análises de associação com o risco para outros factores conhecidos, como a hipertensão, diabetes e consumo de tabaco. Nenhum dos SNPs previamente associados foram replicados na nossa amostra, e considerando o número de testes realizados, não foi encontrada nenhuma associação significativa de outras variantes destes genes na susceptibilidade para sofrer AVCs na população portuguesa. Apenas o SNP rs7442640 do gene *PDE4D* se encontra associado (p-value = 0,006) com o risco de sofrer AVCIs nos testes genotípicos ajustados para covariáveis e o SNP rs4491352 a jusante do gene *ALOX5AP* apresenta uma associação moderada (0,017 < p-value < 0,025) para os testes alélicos e genotípicos não ajustados. Realizando uma meta-análise incluindo todos os estudos publicados recentemente e as nossas amostras portuguesas e espanholas para os SNP41, SNP45, SNP56, SNP87 e SNP89 (que se encontram associados na publicação original) do gene *PDE4D*, não foram encontrados resultados de associação com o risco de sofrer AVCIs. No entanto, verificou-se que o SNP rs10507391 (ou SG13S114) do gene *ALOX5AP*, que faz parte do haplótipo HapA, está associado com o risco de sofrer AVCIs tal como descrito no seu estudo original, tanto na população ibérica, como na meta-análise realizada. Estes resultados sugerem que o gene *PDE4D* pode não ser o principal factor de risco para os AVCIs nas populações portuguesa e espanhola, contrastando com o gene *ALOX5AP* que pode conferir um risco aumentado para se sofrerem AVCIs na Península Ibérica e outras populações.

Avaliou-se também a associação do gene kalirin (*KALRN*) com os AVCIs na nossa amostra portuguesa uma vez que vários estudos recentes têm implicado algumas das suas variantes com a susceptibilidade para fenótipos cardiovasculares e metabólicos, mas nenhum estudo foi ainda realizado em AVCs. As doenças cerebrovasculares e cardiovasculares são ambas complexas, resultantes da interacção de factores genéticos e ambientais, e podem partilhar diversos genes de susceptibilidade. A proteína KALRN está envolvida, entre outras funções, na inibição da enzima óxido nítrico sintase induzida, na regulação da transdução de sinais de isquémia, e na morfogénese neuronal, plasticidade e estabilidade. O nosso objectivo foi verificar se SNPs na região do gene *KALRN*, no cromossoma 3q13, que inclui o gene roporina (*ROPN1*), predispõem para os AVCIs na nossa amostra de pacientes e controlos portugueses. Foram genotipados 34 *tagging* SNPs na região cromossómica dos genes *KALRN* e *ROPN1*, em 565 pacientes com AVCIs e 517 controlos, e realizada a imputação de genótipos para 405 marcadores no cromossoma 3. Testaram-se as associações de cada SNP individualmente e dos haplótipos que constituem com os AVCIs. Um SNP na região intergénica *ROPN1-KALRN* (rs4499545) e dois SNPs no gene *KALRN* (rs17286604 e rs11712619) apresentam associações alélicas e genotípicas (não ajustadas

e ajustadas para hipertensão, diabetes e consumo de tabaco) significativas (0,003 < p-value < 0,049) com o risco de sofrer AVCIs. Trinta e dois SNPs imputados encontram-se também associados com valores de p-value < 0,05, tendo sido validada a associação de três destes polimorfismos (rs7620580, rs6438833 e rs11712039) por genotipagem. Além disso, o SNP rs11712039 foi também associado com os AVCIs (0,001 <p-value <0,01) no estudo de associação em todo o genoma (GWAS) publicado por Ikram e seus co-autores (2009). Esses resultados sugerem que variantes no gene *KALRN* constituem factores de risco para os AVCIs, podendo ser factores de risco genéticos comuns das doenças vasculares.

Foi também testada a associação com o risco de sofrer AVCs do factor H inibidor do complemento (*CFH*), bem como de vários genes candidatos relacionados com mecanismos de neuroprotecção: eritropoietina (*EPO*), hemoxigenase-2 (*HO2*) e calicreína 1 (*KLK1*). Tem sido sugerido que a proteína CFH desempenha um papel importante na inibição do complemento em lesões ateroscleróticas. Tem sido consistentemente associada com um risco aumentado para a degeneração macular relacionada à idade (AMD) e para os enfartes do miocárdio (MI) que compartilham vários factores de risco com os AVCs. Por outro lado, as proteínas EPO, HO2 e KLK1 são neuroprotetoras, por exemplo, em casos de hipóxia cerebral e isquémia (EPO) e contra acidente vascular cerebrais induzidos (KLK1). O polimorfismo no gene *CFH* (rs1061170) que foi previamente associado com a AMD e os MI parece estar modestamente associado (388 pacientes e 461 controlos portugueses; 0,030 < p-value <0,035) com o risco de sofrer AVCIs nos testes alélicos e genotípicos não ajustados, mas nenhum *tagging* SNP neste gene foi claramente associado com AVCIs no GWAS realizado por Ikram *et al.* (2009). Apesar de só ter sido estudado um SNP neste gene, estes resultados não justificam uma análise mais aprofundada. Foram genotipados 3, 3 e 5 *tagging* SNPs localizados nas regiões codificantes e regiões flanqueadoras de 10 kb dos genes *EPO*, *HO2* e *KLK1*, respectivamente, em 565 pacientes com AVCIs e 518 controlos da nossa amostra portuguesa. Não foram encontradas associações para nenhum marcador individualmente nem para haplótipos nestes genes neuroprotetores, sugerindo que eles não constituem factores de risco genéticos para AVCIs.

Para identificar novos genes de susceptibilidade para AVCIs, aplicou-se ainda a abordagem de CG proposta. Foram realizadas análises de expressão génica em células brancas mononucleares do sangue periférico de 20 casos com AVCIs e 20 controlos equilibrados para a idade e para o género, usando *Affymetrix GeneChip Humam U133 Plus 2.0 arrays*, que representam 47.000 transcritos humanos e variantes. Foram identificadas várias vias metabólicas afectadas nos pacientes com AVCIs, tal como a via das moléculas de adesão celular. 16 dos genes diferencialmente expressos entre casos e controlos (para um *cut-off* de 1,2 *fold-change* e p-values não corrigidos < 0,05) encontram-se localizados em picos de ligação descritos em estudos de ligação do inteiro genoma em humanos. Todos os *tagging* SNPs desses genes e das suas regiões flanqueadoras de 10 kb (um total de 191 SNPs) foram genotipados em 565 casos com AVCIs e 520 controlos do nosso biobanco português. Foram realizados testes de associação de cada

SNP individualmente e dos haplótipos que constituem. Os resultados de associação sugerem que existem variantes em 6 (*HEMGN*, *GFI1B*, *TMTC4*, *TTC7B*, *SDC4* e *TUBB1*) dos 16 genes priorizados que podem constituir factores de risco para os AVCIs na população portuguesa. Vários dos SNPs associados nestes genes também fazem parte dos haplótipos que se encontraram associados. SNPs intrónicos como os rs9582406 e rs946845 do gene *TMTC4*, e o rs2284278 do gene *SDC4*, apresentam associação (0,015 < p-value < 0,050) com risco de sofrer AVCIs em todos os testes realizados. Por outro lado, o SNP intrónico rs1535321 do gene *TTC7B* encontra-se associado (p-value = 0,009) nos testes alélicos e genotípicos não ajustados, embora não apresente associação no teste ajustado para hipertensão, diabetes e para o consumo de tabaco.

Para a replicação destes resultados, os SNPs que foram encontrados associados com p-value < 0,05 em pelo menos um dos testes realizados, assim como alguns SNPs que definem os haplótipos associados, foram genotipados numa amostra espanhola. Um total de 570 casos caucasianos com AVCIs e 390 controlos foram incluídos e foram realizados testes de associação dos seus alelos, genótipos e haplótipos com o risco de sofrer AVCIs, utilizando também testes de χ^2 e análises de regressão logística de múltiplas variáveis para ajustar as análises feitas para a hipertensão, diabetes, dislipidemia e para o consumo de tabaco. As mesmas análises foram realizadas para os subtipos de AVCs aterotrombótico, cardioembolico e lacunar. Encontraram-se algumas associações significativas com o risco de sofrer AVCIs nos genes *TMTC4*, *TTC7B* e *SDC4*, no entanto, o único SNP replicado na amostra de casos-controlos espanhola foi o rs9582406 do gene *TMTC4*, significativamente associado nos testes não ajustados (p-value = 0,019). Este SNP está também associado com o risco para os subtipos aterotrombótico e lacunar de AVCs para os testes alélicos e genotípicos (0,011 < p-value < 0,049). Para os genes *TTC7B* e *SDC4*, os SNPs e os haplótipos associados na amostra espanhola não foram os mesmos que na portuguesa. O SNP rs6073708 do gene *SDC4* foi associado para todos os testes realizados (0,027 < p-value < 0,029). Nenhum dos SNPs estudados foi claramente replicado GWAS publicado recentemente por Ikram e seus colaboradores (2009).

Globalmente, os resultados obtidos sugerem que os genes *HEMGN* e *GFI1B* (que não foram replicados na amostra espanhola) podem constituir factores de risco para AVCIs na população portuguesa, sendo importante o alargamento da amostra portuguesa para validação dos resultados. Da mesma forma, o gene *TUBB1*, que não foi genotipado na amostra espanhola por problemas técnicos, pode constituir um factor de risco para AVCIs na população portuguesa, mas deve ser ainda estudado noutras amostras para validar e compreender o seu papel. Por outro lado, o gene *TMTC4* pode constituir um factor de risco para os AVCIs na população ibérica, e os genes *TTC7B* e *SDC4*, com a heterogeneidade observada nos seus resultados de associação entre diferentes amostras, podem ser factores de risco para ACVIs, sendo provável que as suas verdadeiras variantes de susceptibilidade não tenham sido estudadas

ainda. O gene *TTC7B* foi um dos mais associados para as principais doenças cardiovasculares no GWAS do projecto *Framingham Heart Study 100K* (Larson et al. 2007). Para este gene, vários SNPs e haplótipos na região intrão 5 - intrão 6 associados nas amostras portuguesa e espanhola, encontram-se modestamente associados no GWAS publicado por Ikram e seus colaboradores (2009). Existem portanto várias linhas de evidência independentes que apoiam a importância do gene *TTC7B* na susceptibilidade dos AVCs, embora futuros trabalhos sejam necessários para identificar variantes de risco exatas e para elucidar o seu potencial patogénico.

Neste projecto, dado o grande número de SNPs testados, nenhum dos resultados significativos obtidos sobrevive à correcção para múltiplos testes. No entanto, é geralmente aceite que a replicação em múltiplas amostras independentes continua a ser a melhor estratégia para os estudos de associação, mesmo para associações modestas. Se os resultados apresentados puderem ser confirmados noutras amostras independentes e um estudo mais completo de outras possíveis variações genéticas nos loci de interesse forem ser realizados, pensamos que estudos funcionais dos genes e das suas variantes de susceptibilidade, permitirão uma melhoria significativa do nosso conhecimento sobre AVCs. Pode recorrer-se à sequenciação de nova geração para identificar com precisão as verdadeiras variantes de susceptibilidade, ou variantes raras que os estudos de associação não tenham poder para detectar.

É sugerido que a identificação dos determinantes genéticos dos AVCs utilizando diferentes estratégias e populações, e a sua analise de uma forma integrada, como a realizada neste projecto, é a forma mais completa para o seu estudo, a fim de melhorar o conhecimento da doença.

KEY WORDS: Ischemic stroke, genetics, genomics, expression, association.

PALAVRAS-CHAVE: AVCs isquémicos, genética, genómica, expressão, associação.

ACKNOWLEDGEMENTS

During the development of my PhD I had the great opportunity to work with and knowledge so many people from several scientific and medical institutions with whom I share so many aspects of my scientific and personal growth.

I would like to start thanking my supervisors Dr. Sofia Oliveira and Professor José Ferro for giving me the opportunity to work with them. They always helped me to have a critic view of my work and supported me in the several steps of its evolution, motivating me for a high quality and ethical investigation. I had the opportunity to learn with them about science and about the way we should communicate it. It was challenging to understand, share and discuss their ideas and endeavours in this scientific world.

In the same way, I would like to leave a word to my thesis committee, Dr. Jörg Becker and Dr. Astrid Vicente, for the advice and availability, and to all colleagues that during my PhD years worked near me in their projects: Sara Fidalgo, Benedita Fonseca, Sara Violante, Joana Xavier, Alexandra Rosa, Madalena Martins and Patrícia Abrantes. Our discussions about our works, as well as the help and support we always gave to each other, were constant and the source of a true relation of sharing and friendship.

I am thankful to Dr. Astrid Vicente and to the members of her group team that work in collaboration with us in the study of stroke disease: Helena Manso and João Sobral. The coordinate work we developed was essential to maintain a constant auto-criticism and to be worldwide competitive. Beside our collaboration, they were always my great company in the international meetings in which I participate in our research area. The participation in these meetings let us to establish new scientific collaborations and be aware of the most recent worldwide advances.

Another very important collaboration that allowed us to replicate our positive results was the one we established with Dr. Joan Montaner and his students: Sophie Domingues-Montanari and Israel Fernandez. I am very grateful for their help and availability to reach our common goals.

I would like to acknowledge all the *Instituto Gulbenkian de Ciência* (IGC) staff. To the Genotyping Unit and to the Affymetrix Core Facility: Isabel Marques, João Costa, Júlia Lobato and Dr. Jörg Becker. Their services, help and friendship were essential to the good environment in which I always worked analysing so many results that I obtained using the different high throughput methodologies where I was able to specialize myself. It was my pleasure to be received in the IGC, to discover its corners and the excellence of the science it develops. It teached me that it is possible do be hard work and gentle at the same time. With the help of all IGC staff, it was possible to continue believing that to be a scientist is also to be conscious and responsible, with the fascinating duty of give knowledge to the world.

I am also thankful to all the *Instituto de Medicina Molecular* (IMM) staff that received me and gave me the opportunity to finish my PhD working more closely to the *Faculdade de Medicina da Universidade de Lisboa* (FMUL) where I was doing my PhD.

I am specially obliged to all the people who participated in this project, particularly the stroke patients and controls, as well as all the medical doctors whose contribution made this work possible: Dr. José Ferro, Dr. Liliana Gouveia, Dr. Miguel Viana-Baptista, Dr. Amélia Nogueira Pinto, Dr. Rita Silva, Dr. João Ramalho Fontes, Dr. Carla Ferreira, Dr. Manuel Correia, Dr. Assunção Tuna, Dr. Ricardo Taipa, Dr. Gabriela Lopes, Dr. Mário Rui Silva, Dr. João Paulo Gabriel and Dr. Ilda Matos. In the first years of my PhD, I had the great privilege of travel many times inside Portugal to collect and process samples in the laboratories of several hospitals that always make an effort to receive us in their work place. Not less frequent, the contact with potential study participants to recruit them and to explain the objectives of our study, gave me the enthusiasm to always continue. The availability of the participants to allow a better scientific knowledge of the disease were of the major importance for my perception of the importance of this project, and of their potential in the immediate lifestyle changes of the general population.

I would like to acknowledge to all the entities that supported me and the subprojects that constitute this thesis. The overall work was supported by the Marie Curie International Reintegration Grant 513760, the Marie Curie Intra-European Fellowship 024563, the grant PTDC/SAU-GMG/64426/2006 from the Portuguese *Fundação para a Ciência e a Tecnologia* (FCT), and by fellowships I received from FCT and from the Portuguese *Instituto do Emprego e Formação Profissional* (IEFP).

Finally, I would like to remember my family and friends for their constant love and care. My parents and sister were always tireless in the support and help they gave me to invest in my graduation. They are the base of the stability and security that help me to continue working each day with enthusiasm and happiness. My grandparents offer me a deep look of proud and satisfaction that invite me to be always better. In the same way, my best friends were the ones that share with me my dreams and projects. They make me optimistic and self-confident, letting me to know that we only become great when we give us completely. I am specially grateful by the unity we share among us. With their encouragement and help, the last four years in which I had developed this PhD thesis were of the major importance in my professional and personal formation. I had the opportunity of working on science in the fascinating area of the human genetics and I be able to construct a coherent route that gives me proud. I believe we build together a better world.

INDEX

ACRONYMS

AAE	age-at-examination
AAO	age-at-onset
ACE	angiotensin I-converting enzyme
AD	Alzheimer's disease
AF	atrial fibrillation
AGT	angiotensin
AGT	angiotensinogen
ALOX5AP	arachidonate 5-lipoxygenase-activating protein
AMD	age-related macular degeneration
AMP	adenosine monophosphate
ANF	atrial natriuretic factor
ANKRD9	ankyrin repeat domain 9
ANOVA	analysis of variance
API5	apoptosis inhibitor 5
ApoE	apolipoprotein E
ApoE4	apolipoprotein E4
BNF	brain natriuretic factor
C14orf64	chromosome 14 open reading frame 64
χ^2	standard qui-square test
CAD	coronary artery disease
CADASIL	cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy
cAMP	cyclic AMP
CCDC14	coiled-coil domain containing 14
CDC14B	CDC14 cell division cycle 14 homolog B
CDF	chip description file
CEU	European
CFH	complement inhibitor factor H
cGMP	cyclic GMP
CHB	Chinese
CHD	coronary heart disease
CI	confidence interval
CNS	central nervous system
CNVs	copy number variations

CO	carbon monoxide
CRP	C-reactive protein
CVD	cardiovascular disease
DGS	*Direcção Geral de Saúde*
DHS	Diabetes Heart Study
E primer	extension primer
EC	endothelial cells
ELOVL7	ELOVL family member 7, elongation of long chain fatty acids
eNOS	endothelial NOS
EPO	erythropoietin
F primer	forward primer
F13A1	coagulation factor XIII, A1 polypeptide
FAM69B	family with sequence similarity 69, member B
FDR	false discovery rate
GAPDH	glyceraldehydes-3-phosphate dehydrogenase
GC	genomic convergence
GFI1B	growth factor independent 1B
GO	gene ontology
GOLGA2	golgi autoantigen, golgin subfamily a, 2
GWAS	genome-wide association study
HEMGN	hemogen
HO2	heme-oxigenase 2
HWE	Hardy-Weinberg equilibrium
ICH	intracerebral haemorrhage
IGC	*Instituto Gulbenkian de Ciência*
iNOS	inducible nitric-oxide synthase
INSARJ	*Instituto Nacional de Saúde Ricardo Jorge*
IPA	Ingenuity Pathway Analysis
IS	ischemic stroke
ITGA2	alpha-2 integrin
JPT	Japanese
KALRN	kalirin
KLK1	kallikrein 1
LACI	lacunar infarction
LCN2	lipocalin 2
LD	linkage disequilibrium

LMNA	lamin A/C
LOD	logarithm-of-odds
LTA4	leukotriene A4
LTB4	leukotriene B4
LTB4R	leukotriene A4 receptor
LTC4	leukotriene C4
LTD4	leukotriene D4
LTE4	leukotriene E4
MAF	minor allele frequency
MDR	Multifactor Dimensionality Reduction
MELAS	mitochondrial encephalopathy, lactic acidosis, stroke-like episodes
MI	myocardial infarction
MPL	myeloproliferative leukemia virus oncogene
MS	metabolic syndrome
MTHFR	methylenetetrahydrofolate redutase gene
MYLK	myosin light chain kinase isoform
NINJ2	ninjurin2
nNOS	neuronal nitric oxide synthase
NO	nitric oxide
OR	odds ratio
oxLDL	oxidized low-density lipoproteins
PACI	partial anterior circulation infarction
PBMCs	peripheral blood mononuclear cells
PC#1	first principal component
PCA	principal component analysis
PDE4D	phosphodiesterase 4D
PKC	protein kinase C
POCI	posterior circulation infarction
PPP2R5C	protein phosphatase 2, regulatory subunit B (B56), gamma isoform
PRF	Participant Report Form
QC	quality control
QTLs	quantitative trait loci
QVSFS	Questionnaire for Verifying Stroke-Free Status
R primer	reverse primer
RAD51L1	RAD51-like 1
RIN	RNA integrity number

ROPN1	ropporin
SD	standard deviation
SDC4	syndecan 4
SELP	selectin P
SHRSP	stroke-prone spontaneously hypertensive rat
SNP	single nucleotide polymorphism
T2D	type 2 diabetes
TACI	total anterior circulation infarction
TIA	transient ischemic attack
TMTC4	transmembrane and tetratricopeptide repeat containing 4
TP53RK	TP53 regulating kinase
rTPA	recombinant tissue plasminogen activator
TPM1	tropomyosin 1-alpha
TPR	tetratricopeptide repeat
TTC7B	tetratricopeptide repeat domain 7B
TUBB1	tubulin beta 1
WHO	World Health Organization
YRI	African

1. INTRODUCTION

1.1. Stroke characterization and impact

Stroke is a "brain attack" cutting off vital blood, and consequently the nutrients and oxygen vital to the brain cells that control everything we do. It is a disease of public health importance, being even more disabling than lethal. It is the most prevalent neurological disease, forever changing the lives of many who survive. Stroke is defined according to the World Health Organization (WHO) criteria as "rapidly developing clinical signs of focal or global disturbance of cerebral function, with symptoms lasting 24 hours or longer, or leading to death with no apparent cause other than vascular origin" (WHO MONICA 1988). A transient ischemic attack (TIA) has the same definition with the difference that the acute neurological deficit lasts for less than 24 hours.

Stroke can damage most components of the neurovascular unit and lead to death of brain tissue. Beside death, the damages can include motor, sensory or cognitive dysfunctions (Kiyohara *et al.* 2003). The persistent neurological deficit and physical disability caused by stroke have also a very high socio-economic cost and impact. It is a major drain on health-care funding that will increase dramatically without more effective strategies for prevention, treatment, and rehabilitation. Even though everyone has some stroke risk, it is the most preventable neurological disease since several lifestyle risk factors such as smoking, diet, being overweight, physical inactivity, excessive drinking and use of oral contraceptives can be controlled, and several medical risk factors including a previous stroke, previous episode of TIA, hypercholesterolemia, hypertension, diabetes, sickle-cell disease, heart disease, atrial fibrillation (AF) and carotid artery disease, can also be managed. A few stroke risk factors are beyond one's control, such as being over age of 55, being a male, having diabetes, and having a family history of stroke. The race and ethnic origin are also inherent biological traits that could affect the susceptibility to suffer a stroke. Identifying genetic risk factors for stroke has the potential to cause preventive lifestyle changes in individuals at risk that will ultimately result in a decrease of the number of strokes (Feigin *et al.* 2005, Thompson *et al.* 1997, Goldstein *et al.* 2006, Marmot *et al.* 1992, Sacco 2001).

1.2. Stroke incidence

Stroke is the third cause of death in the developed world (WHO 1999, WHO 2002) and the leading cause of death in Portugal (DGS 1999, ONSA 2003). It is the number one cause of disability worldwide, forever changing the lives of many who survive. After the first stroke event, 50-70% of alive stroke patients regain functional independence, but 15-30% are permanently disabled and 20% require

institutional care at 3 months after onset (Asplund *et al.* 1998). It is estimated that, annually, 15 million people worldwide suffer a stroke. Of these, 5 million die and several millions are left permanently disabled. Stroke is uncommon in people under 40 years (WHO 2004).

The incidence of stroke is declining in many developed countries, largely as a result of better control of high blood pressure, and reduced levels of smoking. However, the absolute number of strokes continues to increase because of the ageing population, especially in rapid economic growth regions. It is estimated that stroke will account for 6.2% of the total burden of illness in 2020, remaining in the top three leading causes of death and being among the five most important causes of disability in both developing and developed countries (Menken *et al.* 2000).

Approximately one million strokes occur every year in the European Union and about 20-25% of individuals over 85 can expect to suffer a stroke. It is estimated to cost to the European Union economy over €34 billion a year.

In 1999 and for the 0-64 year old group, the mortality rate per 100,000 inhabitants attributed to stroke in Portugal (194) was much higher than the European average (100). It is more than double of the Spanish average (82) and higher than on all the other European countries (between 72 in France and 124 in Greece) (WHO 1999). Although cerebrovascular disease is the main cause of death in Portugal (DGS 1999, ONSA 2003), very few reports exist about the epidemiological situation at a national level (Médicos Sentinela 1994; DGS 2001). The mortality rates tend to be higher in the North than in the South of the country, and lower in areas more "favoured" like Lisbon (DGS 2001). Additionally, in contrast to Western Europe countries, but similar to Eastern Europe countries, Portugal (in particular the North) presents a strong contrast between urban and rural populations: the stroke mortality reaches between 254 and 298 per 100,000 in the Northeast (predominantly rural area), and only 164 per 100,000 in Porto (DGS 1999).

Besides a relatively high mortality, the incidence of stroke in central western Portugal is also high (Rodrigues *et al.* 2000). In the North of Portugal, Correia *et al.* (2004) verified that the crude annual incidence is of 305 per 100,000 inhabitants in Northeast rural areas (Mirandela and Vila Pouca de Aguiar) and of 269 per 100,000 inhabitants in Porto urban area. The age-standardized incidence of a first stroke in Porto is 173 per 100,000 inhabitants. Interestingly, the incidence per age group follows different patterns in rural and urban populations: it starts to diverge in the 65 to 74 years group and reaches the highest discrepancy for the 75 to 84 years group independently of the gender. In this last age group, the incidence of stroke in Trás-os-Montes (rural area) is the double of the incidence in Porto (2,020 and 1,090 for 100,000 inhabitants, respectively). Similar differences have also been observed in Norway (Ellekjaer *et al.* 1997), Sweden (Terent 1988, Appelros *et al.* 2002) and Bulgaria (Powles *et al.* 2002).

The pattern of the incidence of stroke in different age groups is also similar than the reported in four prospective population studies conducted in the 80s in Rochester (USA), Auckland (New Zealand), Oxford (England), and Soderhamn (Sweden) (Markus *et al.* 2003). However, while the incidence normally

increases with the age for USA, New Zealand, England, Sweden, and for Porto, it was observed by Correia *et al.* (2004)that in individuals with more than 85 years of age in Trás-os-Montes, the incidence of stroke decreases, returning to similar levels of Porto. This suggests that the Trás-os-Montes population not only has more susceptibility for stroke than the population of Porto (higher incidence between the 65 and 84 years), but it also seems that it develops stroke earlier (after 85 years, the incidence is identical).

In the USA, each year, there are about 795,000 new or recurrent stroke episodes. About 610,000 of these are first ever attacks. Data from 2006 indicate that stroke accounted for approximately 1 of every 18 deaths. From 1995 to 2005, the stroke death rate fell 29.7%, and the actual number of stroke deaths declined 13.5%. In the United States, the economic direct and indirect burden of stroke has been estimated around $68.9 billion for 2009 (Lloyd-Jones *et al.* 2009).

On the same way in Japan, stroke mortality rate has decreased significantly in the last three decades, however the incidence of stroke has remained also high in recent years, especially in elderly (Kubo *et al.* 2003). In China, each year 2.5 million people have a stroke, and more than 1 million die of stroke-related causes. Furthermore, several million patients that survived are disabled (http://www.moh.gov.cn). According to estimates from 2002, from the approximately 5 million people that died of stroke worldwide, roughly 20% occurred in South Asia. Although in the past decades have been seen a decline in the incidence of the disease in the Western countries, the burden of the disease in South Asian population (India, Pakistan, Bangladesh, and Sirilanka) has increase and is expected to rise (WHO 2004).

1.3. Clinical types and subtypes of stroke

Stroke can be broadly classified in two major clinical types: the most frequent is ischemic stroke (IS), which occurs by obstruction of blood flow through extra- or intra-cranial vessels by blood clots or by the gradual build-up of plaque and other fatty deposits, generating brain ischemia in one or more central nervous system (CNS) territories; the second type, hemorrhagic stroke, accounts for 10-20% of cases and occurs when weak spots on the wall of intracranial vessels rupture, generating an intracerebral haemorrhage (ICH), or bleeding into the subarachnoid space surrounding the brain (Mohr *et al.* 1978, Günel *et al.* 1996, Caplan 2000, Rothwell *et al.* 2004). IS can be caused by thrombosis, embolism, systemic hypoperfusion (Otero *et al.* 2007, Lee *et al.* 2007). Hemorrhagic strokes normally results from hypertension, trauma, amyloid deposition, bleeding disorders, arterio-venous malformations or aneurysm rupture (Towfighi *et al.* 2005). Hemorrhagic stroke episodes are more frequently fatal, having a 30-day mortality rate of 44% on USA (Woo *et al.* 2005a)

For the IS, the classification into etiological subtypes is usually done according to the validated TOAST diagnostic criteria (Adams *et al.* 1993) This classification denotes five subtypes of ischemic stroke:

a. Large-artery atherosclerosis - patients classified as having large-artery atherosclerosis have clinical and brain-imaging findings of cerebral cortical dysfunction and either significant (>50%) stenosis or occlusion of a major brain artery or branch cortical artery. Potential sources of cardiogenic embolism are excluded.

b. Cardioembolism - this second category includes patients with at least one cardiac source for an embolus and with potential large-artery sources of thrombosis and embolism having been eliminated.

c. Small-vessel occlusion - patients with small-artery occlusion have one of the traditional clinical lacunar syndromes and no evidence for cerebral cortical dysfunction. A potential cardiac source of embolus and stenosis > 50% in an ipsilateral extracranial artery is excluded.

d. Stroke of other determined aetiology - this fourth category includes patients with rare causes of stroke and patients with two or more potential causes of stroke.

e. Stroke of undetermined aetiology - if the causes of stroke cannot be determined despite extensive evaluation, then patients are included in this fifth category.

The differentiation of these subtypes of stroke currently depends on clinical judgment inferred from patient history, symptoms, and laboratory evidence of potential sources of thromboembolism. The clinical diagnosis of aetiology can be highly specific, but the sensitivity is modest (Ferro 2003). Also the TIA are subdivided according to this classification. TIA are strictly not defined as a stroke because the symptoms last for a short period of time, however, the same pathophysiological mechanisms are considered to be responsible for TIA and IS (Caplan 2000). Etiologic diagnosis is critical to develop an appropriate secondary prevention plan (Caplan *et al.* 2006), although, even with a thorough evaluation, the aetiology remains undetermined in 30% or more IS and TIA patients (Xu *et al.* 2008).

Alternatively, the classification of stroke into clinical subtypes may be performed using the Oxfordshire Community Stroke Project classification, which divides cerebral infarction into four categories: total anterior circulation infarction (TACI), partial anterior circulation infarction (PACI), lacunar infarction (LACI), and posterior circulation infarction (POCI) (Bamford J *et al.* 1991).

1.4. Genetic component of stroke

The genetic component of stroke had been demonstrated in the last few decades in twin and family studies. Stroke shows a pattern of familial aggregation, being observed that both paternal and maternal histories of stroke were associated with an increased risk of the disease, and twin studies have

demonstrated a several fold higher risk of stroke among monozygotic as compared to dizygotic twin pairs (Flossmann *et al.* 2004, Bak *et al.* 2002, Brass *et al.* 1992, Khaw *et al.* 1986, Jousilahti *et al.* 1997, Liao *et al.* 1997, Kiely *et al.* 1993), suggesting that genetics factors are involved in the aetiology of the disease. In many of these studies, this risk persisted even when accounting for the known familiarity of individual risk factors, such as hypertension, hyperlipidemias, diabetes and smoking (Jamrozik *et al.* 1994).

An epidemiologic study of environmental and genetic risk factors estimated that two-thirds of the population-attributable risk for stroke was due to genetic factors (Jamrozik *et al.* 1994). Mutations have been also found in several genes in rare classical Mendelian forms of stroke (Joutel *et al.* 1996, Levy *et al.* 1990, Palsdottir *et al.* 1988, Vidal *et al.* 1999, Laberge-le Couteulx *et al.* 1999, Bevan and Markus 2004), corroborating the importance of primary genetic factors in the aetiology of stroke. Recently, genes causing susceptibility for the most common form of stroke have been identified and proposed (Greenberg *et al.* 1995, Gretarsdottir *et al.* 2003, Munshi *et al.* 2008), however, many genetic factors remain to be identified.

In parallel, the genetic component of stroke has been supported by animal model studies (Yamori *et al.* 1976, Rubattu *et al.* 1996). Animal models provide genetic homogeneity, not possible with a human population, to aid the search for causative genes. Correlate human individual genomic variations with stroke is extremely challenging because of the large number of variables within an individual and across different populations (Carswell *et al.* 2005).

1.4.1. Rare classical Mendelian forms of stroke

Until now, the majority of genes that have been linked to stroke, were found in monogenic diseases in rare families. Generally, the used approach is the linkage mapping of large families that display a Mendelian pattern of inheritance, followed by the resequencing of coding exons within the linkage peak to identify highly penetrant mutations. The linkage mapping allows defining the region of the genome that shows excess sharing in accordance to a pattern determined by the tested genetic-inheritance model.

These rare diseases, which include stroke as one of its phenotypes, involve primarily the cerebrovascular system and brain, such as cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy (CADASIL), hereditary ICH secondary to amyloid angiopathy (e.g. Icelandic and Dutch variants), and familial cavernous hemangiomas. In addition, other systemic diseases due to genetic mutations – such as mitochondrial encephalopathy, lactic acidosis, stroke-like episodes (MELAS) syndrome, Fabry's disease, Maeda syndrome, sickle-cell disease, homocystinuria, polycystic kidney disease, Ehler-Danlos syndrome Type IV, Marfan syndrome, and inherited coagulation disorders – have been associated with ischemic and hemorrhagic stroke in some patients (Tournier-Lasserve 2002,

Bevan and Markus 2004, Tonk *et al.* 2007). In general, the Mendelian disorders known to be associated with an increased risk of stroke include hemoglobinopathies, dyslipoproteinemias, and cardioembolic disorders (Natowicz *et al.* 1987).

According to their genetic-inheritance model, these diseases can be inherited in a classical Mendelian pattern as X-linked (e.g. Fabry's disease), autosomal recessive (e.g. Maeda syndrome, sickle-cell disease and homocystinuria), or autosomal dominant (e.g. CADASIL, polycystic kidney disease, Ehler-Danlos syndrome Type IV and Marfan syndrome) disorders. In addition, they can be also inherited as maternal disorders when they are associated with the mitochondrial DNA, such as the MELAS syndrome (Munshi *et al.* 2008).

Several genes have been identified that play roles in the pathogenesis of rare stroke syndromes, such as *NOTCH3* in CADASIL, *CST3* in the Icelandic type of hereditary ICH with amyloidosis, *APP* in the Dutch type of hereditary ICH, *KRIT1* in hereditary cavernous angiomas, *GAL* in Fabry's disease, *HBB* in sickle-cell disease, *CBS* and other genes in homocystinuria, *COL3A1* in Ehler-Danlos syndrome Type IV, and *FBN1* in Marfan syndrome (Joutel *et al.* 1996, Palsdottir *et al.* 1988, Levy *et al.* 1990, Laberge-le Couteulx *et al.* 1999, Munshi *et al.* 2008). None of the genes causing Mendelian forms of stroke has been shown to have a role in the common forms of stroke (Dong *et al.* 2003)

1.4.2. Common form of stroke

The common form of stroke, however, is a complex disease resulting from the interplay of numerous genetic and environmental factors (Rubattu *et al.* 1996, Kiely *et al.* 1993, Sharma *et al.* 1996, Liao *et al.* 1997). This polygenic genetic contribution to stroke is well established even though it has proven difficult to unequivocally identify the genes and the disease-associated alleles mediating this effect. For the overwhelming majority of patients with common stroke, stroke most likely results from the additive or multiplicative effect of this wide spectrum of pathogenic gene variants, each with a small individual contribution (Casas *et al.* 2004). The elucidation of the contributing genetic factors has been also limited by the late onset of the disease, the high short-term mortality rate, the stroke types and subtypes, the demonstrated phenotypic variability, and by the confounding presence of hypertension or other risk factors, which is commonly associated with stroke (Kannel *et al.* 1975).

Currently, aiming at better prevention and treatment of the disease, a search of the literature reveals hundreds of studies trying to identify the genetic components of stroke through different approaches, such as candidate-gene studies, whole genome linkage analysis, and, more recently, genome-wide association studies (GWASs). Several expression studies in several tissues and models have been also published.

1.5. Candidate genes

One popular method of identifying genetic risk factors for the common form of stroke, has been to conduct case-control association studies in which investigators compare frequencies of genetic variants in candidate genes among stroke cases and unrelated controls. Genes are normally chosen based on their known function. The goal of these studies is to identify specific microsatellites, single nucleotide polymorphisms (SNPs) or haplotypes (combinations of microsatellites and/or SNPs) that influence the susceptibility of developing the disease. SNPs have being increasingly accepted as powerful genetic markers for the detection of susceptibility genes through association analyses as they are more frequent and stable than microsatellites. Moreover, it has been estimated that the human genome contains more than 10 million SNPs (Kruglyak *et al.* 2001).

However, candidate gene studies relies on selecting the right candidate gene and polymorphisms, a daunting task since the human genome harbours about thirty thousand genes and the referred amount of polymorphisms. At the same time, high density chromosome-wide maps of linkage disequilibrium (LD) have been constructed with the aim of elucidating patterns of ancestral recombination and selection. These maps facilitate the identification of disease genes for complex traits in humans (Cardon *et al.* 2003). Interpreting the correlations between markers helps to interpret observed patterns of markers correlations with disease, identify efficient sets of tagging markers that provides information about nearby variants not genotyped (Goldstein *et al.* 2003, Johnson *et al.* 2001), and design more focused studies (Zondervan *et al.* 2004). The LD information helps to reduce the over- and mis-interpretation of association results (Weiss et al 2000), and can improve the cost-effectiveness of the studies. The causal genetic variants of the diseases can be within the coding exons of the genes of interest or can be nonconding variants that affect the expression and/or efficiency of splicing. The examination of only the coding variants might lead to a severely underpowered study over the candidate gene.

Hundreds of genes that may have an impact in stroke pathogenesis or intermediate phenotypes, or genes that are associated with clinical risk factors such as hypertension, diabetes, hyperlipidemia and vascular disease, have been tested in humans or in animal models but many times have generated negative or conflicting results. The big picture is difficult to interpret from the multitude of data available. In this context, Casas *et al.* (2004), performed a meta-analysis of all candidate gene association studies in IS to find truly associated genes and allow genetic risks to be quantified with more precision. Data from 120 case-control studies and a total of 51 polymorphisms in 32 genes were included, and 15 polymorphisms were analysed in detail. This study suggests that common variants in several genes, each exerting a small to moderate effect, are likely to contribute individually, together, or in combination with environment determinants to the risk of stroke, rather than a major effect on a single gene. Other meta-analyses have

been performed for some specific candidate genes.

Examples of functional candidates which have been tested for their impact in stroke pathogenesis are inflammatory genes such the ones that codify for IL1 and TNF-alpha (Lee *et al.* 2004), genes involved in oxidative stress, such as neuronal nitric oxide synthase (*nNOS*) and endothelial NOS (*eNOS*) (Berger *et al.* 2007, Casas *et al.* 2004), genes related to the coagulation system such as prothrombin, factor V Leiden genes and β-fibrinogen (Casas *et al.* 2004, Juul *et al.* 2004), and genes involved in the determination of serum homocystein levels like the methylenetetrahydrofolate redutase gene (*MTHFR*) (Keijzer *et al.* 2002, Klerk *et al.* 2002, Kelly *et al.* 2002, Casas *et al.* 2004).

Two large case-control studies on the German population involving 1901 stroke patients and 1747 controls were performed by the same group and revealed that the E298D polymorphism of the *NOS3* was significantly associated with IS independent of confounding variables. In the first case-control analysis on 63 candidate genes, five SNPs located in the *NOS3*, alpha-2 integrin (*ITGA2*), IL13, selectin P (*SELP*), and chemokine receptor 2 gene, showed a significant association with IS even though only the *NOS3* has been replicated in the second case-control and combined analysis of both studies (Berger *et al.* 2007). All of these are genes that have been largely studied for association to hemorrhagic or ischemic stroke.

Prothrombin (or factor II) has been associated with deep venous thrombosis and elevated plasma prothrombin levels through a common G to A transition at position 20210 that represents a gain-of-function mutation. This mutation causes increased cleavage site recognition, increased 3'-end processing, and increased mRNA accumulation and protein synthesis (Franco *et al.* 1999, Cattaneo *et al.*1999). Casas *et al.* (2004) found a significant association of this variant in their meta-analysis. Similarly, the Factor V Leiden (the name given to a variant of human factor V that causes a hypercoagulability disorder) was showed to be significantly associated with IS by Casas *et al.* (2004) for the variant R506Q. In the hypercoagulability disorder, the Leiden variant of factor V cannot be inactivated by activated protein C. This results in an increased thrombin generation and a hypercoagulable state, which may explain the increase risk of stroke in carriers of this mutation (Dahlback 1995). The A1691G-A transition that results in an R506Q substitution, is associated with venous thrombosis too (Juul *et al.* 2004); and the protein S and protein C deficiency are rare causes of early-onset IS.

Polymorphisms of the gene encoding for β-fibrinogen have been shown to correlate with either large-vessel stroke or carotid intima-media thickening, which is consistent with plasma levels of fibrinogen correlating with risk for future stroke (Kessler *et al.* 1997, Carter *et al.* 1997, Schmidt *et al.* 1998).

Also *MTHFR* has been extensively studied. The enzyme it encodes participates in the processing of amino acids catalyzing a reaction required in the conversion of homocysteine to methionine. Long-term differences in the serum concentration of homocysteine are associated with an increase in the risk of stroke (Wald *et al.* 2002). Keijzer *et al.* (2002) showed that hyperhomocysteinemia due to a SNP C677T

in the *MTHFR*, that encodes an amino acid substitution (A222V), was a risk factor for recurrent venous thrombosis. Additionally, Klerk *et al.* (2002) performed a meta-analysis of the risk of coronary heart disease (CHD) related to the same SNP concluding that individuals with the 677TT genotype have significantly higher risk. Kelly *et al.* (2002) performed a similar meta-analysis of the risk of IS and they conclude that the 677TT genotype has a small influence in determining the susceptibility to IS too. The meta-analysis performed by Casas *et al.* (2004) confirm this result with a significant association between IS and the genetic variant.

Also the genes apolipoprotein E (*ApoE*), angiotensin (AGT) I-converting enzyme (*ACE*), angiotensinogen (*AGT*), and paraoxonase, have been largely studied (Greenberg *et al.* 1995, Casas *et al.* 2004, Lisabeth *et al.* 2005). *ApoE* has been the candidate gene most studied in ICH because of its role in Alzheimer's disease (AD) and, in particular, its relation to sporadic cerebral amyloid angiopathy. This gene has three alleles called E2, E3, and E4, being the E3 the most common. The encoded protein is a major component of very low-density lipoproteins and is responsible for the transport of cholesterol from the blood to the liver, where it is processed (Greenberg *et al.* 1998). Woo *et al.* (2005b) showed that apolipoprotein E4 (ApoE4) is independently associated with lobar ICH and that the risk appeared to be higher in individuals older than 70 years of age. They predict that a third of all cases of lobar ICH are attributable to possession of an ApoE4 or E2 allele. More than 40 studies have also examined the role of *ApoE* in IS and the results vary greatly (Domingues-Montanari *et al.* 2008). To clarify the role of *ApoE*, meta-analyses have been performed but also with conflicting results depending on the populations they involve, such as the ones published by Casas *et al.* (2004) where no associations were found, and by Ariyaratnam *et al.* (2007) where associations were found in Chinese and Japanese populations.

The gene coding for the ACE, on the other hand, has been studied for its important role in blood pressure regulation and electrolyte balance. ACE hydrolyzes AGT I into AGT II, a potent vasopressor, and is able to inactive bradykinin, a potent vasodilator that has been suggested to stimulate vasodilator nitric oxide (NO) production. It is also involved in vascular hypertrophy and atherosclerotic processes (Kim *et al.* 2000). In 1992, Tiret *et al.* (1992) revealed an insertion/deletion polymorphism which was strongly associated with the level of circulating enzyme. The mean plasma ACE level of DD subjects was approximately twice that of II subjects, with ID subjects having intermediate levels. The insertion corresponds to an Alu repetitive sequence (Tiret *et al.* 1992, Agerholm-Larsen *et al.* 2000). Slowik *et al.* (2004) found an association between the *ACE* DD polymorphism and spontaneous ICH. For IS the results are contradictory in different populations. Two meta-analyses found a statistically significant association between IS and the *ACE* DD genotype compared with the II or ID genotypes (Casas *et al.* 2004, Ariyaratnam *et al.* 2007). On the other hand, Banerjee *et al.* (2007) did not find any significant association result. Similar conflicting results has been verified to AGT gene and consequently its association remain unclear (Domingues-Montanari *et al.* 2008). Genetic variants in regulatory regions of the *AGT*, however,

were associated with blood pressure reactivity in a study by Gu *et al.* (2005).

There are several potential explanations for the inconsistencies that have been verified, including chance findings or type I errors, small samples with low power to detect association, inconsistent clinical evaluation, cover an insufficient number of genetic variants on each studied gene, low carrier frequency of the variants in patients with stroke or both, and convenience control samples poorly matched to cases. Furthermore, many candidate-gene studies use mainly amino-acid-substituting SNPs and if important functional SNPs should arise in noncoding regions without significant LD with the site of a screened polymorphism, the association analysis may exclude the true disease susceptibility locus. Genetic and phenotypic heterogeneity and a non accurate definition or characterization of confounding variables, may also play a role in these inconsistent results. The number of individuals required for the study can also becomes insufficient when separating by the aetiology of the disease.

It is interesting that mostly nuclear genes have been intensively investigated thus far, while the role of the mitochondrial genome has been neglected (Rosa *et al.* 2008). Particular variants of the mitochondrial genome have been linked to aging (Tanaka *et al.* 1998, Ivanova *et al.* 1998), the strongest risk factor for stroke, and to several neurological and vascular disorders. We found suggestive evidence for association of the mitochondrial haplogroup H1 with IS in Portuguese patients (Rosa *et al.* 2008).

1.6. Novel candidate genes

There are also some novel interesting candidate genes that, according to their biological function, could be associated with the stroke risk and that were never studied before.

1.6.1. Kalirin (KALRN) gene

Cerebrovascular diseases such as stroke, and cardiovascular diseases (CVDs) such as coronary artery disease (CAD) and myocardial infarction (MI), are complex disorders resulting from the interplay of genetics and environment, and they share many risk factors, including age, sex, hypertension, dyslipidemia, diabetes, obesity, smoking, and physical inactivity. These atherothrombotic diseases most likely also share common pathogenic mechanisms such as inflammation and appear to have common susceptibility loci. For instance, a locus on 9p21 has been firmly associated with vascular pathologies such as heart disease, stroke, aneurysms and atherosclerosis (Matarin *et al.* 2008, Larson *et al.* 2007, McPherson *et al.* 2007, Helgadottir *et al.* 2008, Karvanen *et al.* 2009, O'Donnell *et al.* 2007, Ye *et al.* 2008).

The GENECARD study for early-onset CAD (Hauser *et al.* 2004), the Diabetes Heart Study (DHS) for CVD, type 2 diabetes (T2D) and metabolic syndrome (MS) (Bowden *et al.* 2006), and a meta-

analysis of four linkage studies for CHD (Chiodini *et al.* 2003), identified linkage peaks on chromosome 3q. Ordered subset analysis significantly increased the evidence for linkage at 3q13 (logarithm-of-odds (LOD) = 5.10, p-value = 0.008) in GENECARD families with lower-risk lipid profiles and fewer risk factors (Shah *et al.* 2006), and subsequent peak-wide association mapping led to the identification of twelve SNPs in the Ropporin (*ROPN1*, OMIM 611757) and Kalirin (OMIM 604605) genes associated with early-onset CAD (Wang *et al.* 2007). Validation in additional datasets revealed that SNP rs9289231 in the first intron of a *KALRN* alternative transcript was associated with early-onset CAD in all white data sets examined, and the risk allele of this SNP was associated with atherosclerosis burden in human aortas (Wang *et al.* 2007). Additionally, follow-up of the DHS 3q linkage peak revealed that the *KALRN* polymorphism rs4234218 is associated with T2D, MS and combined phenotype (T2D + MS + CVD + coronary calcified plaque) (Rudock *et al.* 2008). Furthermore, the recent and well-powered GWAS conducted on four European and American cohorts by Ikram *et al.* (2009) showed that several polymorphisms in Kalirin are associated with IS ($0.001 < \text{p-value} < 0.01$), even though below the genome-wide significance level.

Kalirin is an extremely large gene with 60 exons spanning over 620 kilobases, characterized by multiple promoters producing developmentally-regulated isoforms predominantly expressed in the brain (McPherson *et al.* 2004). *KALRN* encodes for a guanine nucleotide exchange factor that activates Rho proteins and is therefore a multifunctional protein involved, among others, in neuronal morphogenesis and secretory granule maturation (Ferraro *et al.* 2007, Rabiner *et al.* 2005). Additionally, Kalirin may have a neuroprotective role by inhibiting inducible nitric-oxide synthase (iNOS) activity (Ratovitski *et al.* 1999) and participates in the regulation of ischemic signal transduction (Beresewicz *et al.* 2008).

Kalirin therefore constitutes a very interesting candidate gene to test for stroke association.

1.6.2. Complement inhibitor factor H (CFH) gene

Recent studies with animal models and humans have shown that inflammation is implicated in atherothrombotic disorders like MI and some forms of stroke. *CFH* (OMIM 134370), encoding a 155 kDa serum glycoprotein (chromosome 1q32) essential in the regulation of the alternative complement pathway in fluid phase and on cellular surfaces, has been suggested to play an important role in the complement inhibition in atherosclerotic lesions. Both complement factors and complement regulatory factors have been linked to atherosclerosis.CFH is expressed in early human atherosclerotic lesions in the superficial layer of the arterial intima (Oksjoki *et al.* 2003).

A tyrosine to histidine polymorphism (rs1061170) at amino acid 402, Y402H, in exon 9 of the CFH gene, has been consistently associated with an increased risk for age-related macular degeneration (AMD; Klein *et al.* 2005, Haines *et al.* 2005, Edwards *et al.* 2005) and accounts for up to 50% of the

attributable risk, potentially through an altered inflammatory response. The Y402H polymorphism is located in a region that binds heparin and C-reactive protein (CRP), and higher CRP concentrations have been shown to indicate increased risk of cerebrovascular disease (Youssef *et al.* 2007). The substitution of a positively charged histidine for a non-charged hydrophobic tyrosine in position 402 may alter the binding properties and have functional implications (Rodriguez de Córdoba *et al.* 2004). These changes may alter CFH's ability to suppress excess complement activation, ultimately leading to complement-related damage to arterial walls and vessel injury. Kardys *et al.* (2006) suggested that the Y402H variant is a susceptibility factor for MI. Given that AMD, MI and stroke share many risk factors, they may also share some of their etiopathogenic mechanisms (e.g. inflammation) and genetic risk factors. Therefore, the Y402H polymorphism of the CFH gene is also an attractive candidate gene for stroke risk.

1.6.3 Erythropoietin (EPO), heme-oxigenase 2 (HO2), and kallikrein 1 (KLK1) genes

In the last few years, animal studies of cerebral ischemia, hypoxia and oxidative stress allowed the identification of several neuroprotective molecules, but the majority of the genes encoding for these proteins or hormones have not been tested as stroke candidate genes. These neuroprotectors include the EPO, HO2 and KLK1 genes (Table 1.1).

Table 1.1: Selected neuroprotective candidate genes.

Candidate gene	Chromosome	Function	Reference
Epo	7q22	Neuroprotective functions in brain hypoxia and ischemia	Kilic et al. (2005)
HO2	16p13	Confers neuroprotection	Namiranian et al. (2005)
KLK1	19q13	tK levels correlate with the degree of carotid stenosis and provides neuroprotection against induced stroke	Xia et al. (2006)

EPO is a cytokine that serves as the key regulator of erythropoiesis by stimulating growth, preventing apoptosis, and promoting differentiation of red blood cell precursors (Lacombe *et al.* 1999). Both EPO and its receptor are expressed in several non-hematopoietic tissues and cells, including the CNS (Marti *et al.* 1996, Bernaudin *et al.* 1999). The neuroprotective and neurotrophic role of EPO has been demonstrated in different CNS disorders, including hypoxic and ischemic injury (Brines *et al.* 2000). The presence of EPO in the brain appears to protect neurons from ischemic damage directly by inhibiting their apoptosis (Sakanaka *et al.* 1998) or indirectly by stimulating angiogenesis in the brain (Marti *et al.* 1996).

The HO system consists of two isoforms: the oxidative stress–inducible protein HO1 and the constitutive isozyme HO2. Both catalyse the oxidation of the heme ring to carbon monoxide (CO), iron and biliverdin, which is rapidly reduced to bilirubin (Doré *et al.* 2000). Under normal conditions, HO1 is

barely detectable in the brain while HO2, which activity is modulated by phosphorylation, reveals a high concentration in neural tissues (Maines 1997). HO2, has a potential neuroprotective effect (Baranano *et al.* 2001). The CO originated in the reaction may act as neurotransmitter, vasodilator, and anti-apoptotic factor, while low concentrations of bilirubin can protect neurons from oxidative stress (Scholz *et al.* 2003). It is possible to establish HO2 as an endogenous neuroprotective system in the brain, although mechanisms whereby neuroprotection is provided have not been totally clarified.

Finally, the kallikrein–kinin system, which influences the permeability of the blood-brain barrier, has been shown to protect against ischemia and is activated in stroke. Kallikreins, serine proteases present in biological fluids and tissues, mediate the release of kinins from kininogens precursors (Storini *et al.* 2006). It has been shown that kallikrein/kinin plays a key role in angiogenesis and apoptosis in response to ischemia (Emanueli *et al.* 2001). Kallikrein action during brain ischemia is known to stimulate the production of neurotransmitters resulting in vasoactive, anticoagulant, and proangiogenetic properties, thus inducing protection (Shariat-Madar *et al.* 2002, Xia *et al.* 2006). KLK1 may function independently of the kinin signalling pathway, through growth factors and several other substrates and is believed to play an essential role in a large number of processes, including blood pressure regulation and smooth muscle contraction (Emami *et al.* 2007).

1.7. Whole genome linkage screen

The whole genome linkage screen is a method that allows the identification of genetic variants linked with a disease, without relying on any knowledge of the function of such variants. The term linkage refers to finding a genetic marker within each of a series of families containing two or more affected individuals, after carrying out a whole genome scan that involves genotyping many genetic polymorphic markers situated at regular intervals throughout the whole genome. Whole-genome linkage screens identify areas of the genome that affected individuals share more often than one would expect by chance. It looks for segments of the genome that are identical by descent or which co-segregate with the disease. The degree of linkage of each marker to the disease trait is calculated and, if successful, genetic linkage can identify a large genomic region, possibly containing hundreds of genes, in which the disease gene is sought. Linkage screens typically lead to the identification of several chromosomal areas linked to the trait of interest.

Linkage analysis methods are typically carried out in one of two ways: parametric analysis, which involves tracing cosegregation and recombination phenomena between observed marker alleles and unobserved putative trait-influencing alleles among members of large pedigrees; and allele-sharing methods, the non parametric analysis, which assess the number of marker alleles shared at a particular

locus among pairs of relatives manifesting the same trait. The strength of the linkage studies is their theoretical robustness, because the methods make relatively few assumptions about either the genetic architecture of the population in which the disease is studied or about the disease risk (model-independence). The linkage-mapping strategies offer also the advantage of computational speed, and systematic identification of novel loci. However, both of the presented methods have the limitation of need a large number of families with individuals possessing the disease of interest. The low penetrance of complex diseases means that vast amounts of information are required to identify linkage. In addition, families with different environmental exposures and genetic or ethnic backgrounds can create heterogeneity and thereby increase the amount of noise obscuring the signal of a given gene effect. Another disadvantage of these studies is their relative lack of power and the low resolution of the linkage peaks identified. Linkage cannot detect genes with minimal or modest effect on the disease. Therefore, significant linkage peaks, if present, might map genes that have a more substantial effect on disease risk. A commonly taken approach has been to perform whole-genome linkage screens followed by association studies.

Linkage studies in other diseases have already shown that they can lead to the identification of significantly and consistently associated genes, such as in Crohn's disease and type 2 diabetes mellitus where the NOD2 and calpain-10 genes, respectively (Ogura *et al.* 2001, Hugot *et al.* 2001, Horikawa *et al.* 2000). These associations were confirmed in different populations (Hampe *et al.* 2001; Evans *et al.* 2001).

Two small linkage studies in Finnish and Japanese populations (85 and 104 affected sibling pairs respectively) have identified chromosomal regions linked with intracranial aneurysms (Olson *et al.* 2002, Onda *et al.* 2001). However, none of these two linkage studies showed if their linkage was specific to stroke or if it was also related with intracranial aneurysms formation or linked to some stroke risk factor(s) such as hypertension or hypercholesterolemia. The risk genes in each of these studies have not yet been identified or confirmed.

Specifically for the study of stroke, few linkage screens have been reported (Gretarsdottir *et al.* 2002, Nilsson-Ardnor *et al.* 2005 and 2007, Helgadottir *et al.* 2004). Is has proven difficult to recruit large numbers of patients with stroke who have at least one other relative alive with a history of stroke because the late onset of the disease and its high mortality. However, these studies lead to the identification of two of the most tested genes for the susceptibility of stroke, the phosphodiesterase 4D (*PDE4D*) and the arachidonate 5-lipoxygenase activating protein *APLOX5AP* (Gretarsdottir *et al.* 2003, Helgadottir *et al.* 2004).

There are also whole genome linkage screens for stroke risk published in rats. Using F2 hybrids derived from stroke-prone spontaneously hypertensive rat (SHRSP) and stroke-resistant spontaneously hypertensive rat, Rubattu and colleagues (1996) identified three major quantitative trait loci (QTLs): STR1, STR2 and STR3, with LOD scores of 7.4, 4.7 and 3.0, respectively. These QTLs accounted for

28% of the overall phenotypic variance and, in particular, the STR2 conferred a significant protective risk against stroke. STR2 maps on rat chromosome 5 and the respective linked marker localized in the atrial natriuretic factor (ANF) gene. ANF and the brain natriuretic factor (BNF), which is immediately adjacent, have a wide variety of vascular effects that suggest that they could be candidate genes for stroke. STR1 and STR3 map, respectively, to rat chromosome 1 and 4. One year later, Jeffs *et al.* (1997) confirmed the chromosome 5 linkage peak for stroke by performing a genome-wide scan in F2 hybrids derived from the SHRSP and the normotensive reference rat strain Wistar Kyoto for large infarct volumes in the SHRSP in response to a focal ischemic insult. In 2003, Kato *et al.* (2003) found new evidences for the chromosome 1 linkage with the risk of stroke using the same animal models than Rubattu. The genes behind the linkage peaks for stroke risk and severity in rats remain however to be identified.

1.7.1. PDE4D gene

In 2002, a genetic linkage analysis of Icelandic families provided the initial evidence for a susceptibility gene for stroke on chromosome 5. The genome-wide scan was performed on 476 white patients with stroke and 438 of their relatives and mapped linkage to IS, hemorrhagic stroke and TIA to a 20 cM region, named *STRK1*, on chromosome 5q12 with a LOD score of 4.4. The LOD score increased to 4.9 when 6% of patients with hemorrhagic stroke were removed, suggesting that the gene driving this linkage peak was a risk factor for IS. Additionally, it was suggested that the locus 5q12 contributes directly to IS, rather than indirectly through a known risk factor for stroke, and that there may be biological pathways independent of the known risk factors that contribute to the pathogenesis of the disease (Gretarsdottir *et al.* 2002).

Fine mapping was carried out in a case-control study of 864 affected individuals from the Icelandic population and 908 controls, using 98 microsatellite markers spanning the implicated chromosomal region. This lead to the identification of the PDE4D gene with the microsatellites *DG5S397* and AC008818-1 associated with the risk of suffering stroke. An additional set of 260 SNPs were then genotyped in the entire affected and controls cohort and common variants (SNP32, SNP41, SNP45, SNP56, SNP83, SNP87, SNP89, SNP91 and SNP100) within the PDE4D gene were found to be associated, especially for cardiogenic and carotid stroke (Gretarsdottir *et al.* 2003). It is also observed that for the combined cardiogenic and carotid subtype of stroke, the highest risk haplotype (G0, present in 9% of controls and composed by SNP45 and microsatellite AC008818-1) conferred a two-fold relative risk of suffering stroke to the 16% of the general population that carries at least one copy. A protective haplotype (AX, present in 21% of controls and composed by the same two markers) was also identified, with a reduced risk of 0.7 relative to the wild-type haplotype. The PDE4D risk haplotype has an effect that is largely independent of known risk factors. None of the associated variants are present in protein coding or

gene splicing regions, suggesting that these variants and/or variants with which they are in strong LD, could affect gene regulation (such as expression level) rather than having a direct functional effect on the protein (Gretarsdottir *et al.* 2003).

Another genome-wide linkage scan was performed later for stroke disease using 109 multicase families from northern Sweden containing multiple cases (Nilsson-Ardnor *et al.* 2005 and 2007). The authors confirmed the 5q12 linkage peak, and presented new evidence for several other loci with LOD scores > 1.2: 1p34, 5q13, 7q35, 9q22, 9q34, 13q32, 14q32, 18p11 and 20q13. The highest allele-sharing LOD scores were obtained on chromosomes 5q, 13q and 18p. Linkage calculations were performed using 3 different disease phenotypes, from all stroke cases to IS only. As presented, the authors did not identify any major susceptibility loci, but multiple minor stroke loci. In particular, they identify linkage at marker D5S424, which resides within the PDE4D gene, emphasizing the association of this gene (Nilsson-Ardnor *et al.* 2005 and 2007).

The PDE4D gene, very conserved among species, encodes a cyclic nucleotide phosphodiesterase which degrades second messenger cyclic AMP (cAMP), a key signal transduction and regulation molecule in different cell types, including inflammatory, vascular endothelial and smooth muscle cells. It is important for the regulation of the physiological responses of these cells types. The gene includes at least 22 exons and seven promoters and the resulting protein is present in at least nine isoforms differing from each other at their N-terminal regions (Gretarsdottir *et al.* 2003, Dominiczak *et al.* 2003, Dichgans 2007). Studies in animal models have shown that the elevation of cAMP reduces neointimal lesion formation and inhibits proliferation of smooth muscle cells after arterial injury (Pan *et al.* 1994). It has been also reported the involvement of PDE4D in inflammation, cell proliferation, and migration processes implicated in stroke occurrence (Ariga et al 2004, Miro *et al.* 2000, Palmer *et al.* 1998, Pan *et al.* 1994). Additionally, PDE4D seems also to regulate cyclic GMP (cGMP). The interaction of the phosphodiesterases and cyclic nucleotides is a part of complex endothelial signalling pathway that has shown to be dysfunctional in cerebrovascular disorders. The production of cGMP is linked with NO production by vascular endothelium and it is believed that the NO-cGMP pathway is dysfunctional in stroke (Birk *et al.* 2004, Stoclet *et al.* 1995). These findings and their pathogenic significance in stroke remain to be confirmed and elucidated.

These studies also provided optimism about identifying susceptibility genes even without further subtyping IS phenotypes. As mentioned before, one potential problem in stroke research is that it is a very heterogeneous disease with respect to the nature of the brain damage and to its aetiology, suggesting that subtyping of stroke could be necessary. Moreover, these studies reinforce the idea that there are genetic factors specific for stroke, and not to its risk factors. The association of *PDE4D*, however, has only been confirmed in some populations and, the risk conferred by this gene is unclear since replication studies have reported conflicting results (Table 1.2).

Table 1.2: Review of *PDE4D* association studies with IS and meta-analyses. For each publication, the number of cases and controls that were considered for the analyses of the studied polymorphisms, as well as the population where the sample are from, are indicated. Only significant results of association are indicated.

Reference	Polymorphisms **	N cases	N controls	Population	Significant association with IS*
PDE4D					
Case-control studies:					
Gretarsdottir et al. (2003)	260 SNPs, AC, 97 MSs	864	908	Iceland	32, 45, 56, AC *cardiogenic:* 45, 89, 91, AC* *carotid:* 83*, 87, 100, AC, DG5S397* *cardiogenic + carotid:* 41*, 45*, 56*, 87*, 89*, AC*, 45-AC
Bevan et al. (2005)	2, 3, 5, 6, 13, 14, 15, 19, 20, 22, 23, 26, 30, 31, 34, 35, 37, 45, 87, AC	737	933	UK	- *large-vessel* : 19, 87 *cardioembolic:* 2, 13, 14, 15, 20, 26
Lohmussaar et al. (2005)	6, 20, 30, 34, 41, 45, 48, 49, 57, 67, 69, 73, AC	639	736	Germany	- *cardiogenic:* 73 *carotid and carotid + cardiogenic:* 57
Meschia et al. (2005)	32, 45, 56, 83, 87, AC	377	263	US (74% white)	56, 83, 56-83 *whites:* 83 *large-vessel:* 32, 83 *cardioembolic: 45* *large-vessel + cardioembolic:* 45
Saleheen et al. (2005)	32, 83, 87	200	250	Pakistan	83
Nilsson-Ardnor et al. (2005)	41, 45, rs1971940, rs716908, rs294492, AC	275	550	Sweden	rs294492, AC
van Rijin et al. (2005)	39, 45, 83	88	190	Netherlands	- *small-vessel in inbred individuals:* 39, 45
Brophy et al. (2006)	9, 26, 32, 34, 42, 45, 56, 148, 175, 199, 219, 220, 222, AC	248	560	US (women only)	9, 219, 222, 9-26-32-34-42, 148-175-199-219-220-222 *non hypertensive:* 9, 42, 219, 220, AC, 45-AC, 9-26-32-34-42, 148-175-199-219-220-222 *hypertensive:* 175, 148-175-199-219-220-222
Zee et al. (2006a)	9, 26, 32, 34, 42, 45, 56, 219, 222	259	259	US (men only)	56 *non hypertensive:* 42, 45, 56
Nakayama et al. (2006)	45, 83 + 29 SNPs in STRK1 locus, AC, 3 MSs	208	270	Japan	-* *noncardiogenic:* 83-rs153031-AC*
Woo et al. (2006)	41, 45, 56, 83, 87, 89	357	303	US (73% white)	41-56-83-87-89* *whites:* 41 *cardioembolic:* 87, 41-56-83-87-89* *cardioembolic in whites:* 41* *cardioembolic in blacks:* 83, 89 *small-vessel in whites:* 41-56-83-87-89* *large-vessel in blacks:* 41-56-83-87-89
Song et al. (2006)	41, 42, 45, 83, 89 + 18 SNPs	224	211	US (women only, 55% white)	rs918592* *whites:* 83*, rs1498606* *blacks:* 42*, 89* *atherosclerotic:* rs918592 *lacunar:* rs918592 *cardiac:* 83, rs1498606

Staton et al. (2006)	41, 45, 56, 83, 87, 89	151	164	Australia	83, 87, 89, 83-87-89
Kuhlenbäumer et al. (2006)	41, 45, 56, 83, 87, 89	1,181	1,569	Germany	- *cardioembolic:* 87
Kostulas et al. (2007)	41, 45, AC	685	751	Sweden	- *large-vessel:* 41, 41-45 *large-vessel + cardioembolic:* 41-45
Fidani et al. (2007)	45, AC	97	102	Greece	AC, 45-AC
Lövkvist et al. (2008)	34, 37, 39, 45, 87	932	396	Sweden	39, 45* *hypertensive:* 39*, 45*
Xue et al. (2009)	32, 83, 87	639	887	China	- *atherothrombotic:* 83, 32-83-87
Sun et al. (2009)	56, 83, 87 + 1 SNP	649	761	China (Han population)	- *small-artery-occlusive:* 56-rs152312 *cardiogenic + carotid:* 83
Meta-analysis:					
Staton et al. (2006)	41, 45, 56, 83, 87, 89	3,808	4,377	Meta-analysis	41*, 83*, 87*
Bevan et al. (2008a)	26, 45, 56, 83, 87, 89, AC	5,216	6,615	Meta-analysis	AC, 45-AC *whites:* AC, 45-AC *cardioembolic:* 56 *cardioembolic in whites:* 56 *small-vessel:* 89 *cardioembolic + large-vessel:* 45 *cardioembolic + large-vessel in whites:* 56 (all associations became nonsignificant after exclusion of the original study)
Lövkvist et al. (2008)	45	6,221	6,750	Meta-analysis	-

*Significant SNPs that resist adjustment for multiple testing.
**SNP ID number in original report.
AC: AC008818-1 microsatellite; MS: Other microsatellite.

Looking for the overall results, 18 case-control studies and 3 meta-analyses in European and other populations have been published as follow-up of the initial report from Gretarsdottir and colleagues (2003). 11 of these case-control studies and 1 meta-analysis claimed association of at least one genetic variant studied in the initial report with all IS (Meschia *et al.* 2005, Saleheen *et al.* 2005, Nilsson-Ardnor *et al.* 2005, Brophy *et al.* 2006, Zee *et al.* 2006a, Nakayama et al 2006, Woo *et al.* 2006, Song *et al.* 2006, Staton *et al.* 2006, Fidani *et al.* 2007, Lövkvist *et al.* 2008; Table 1.2) and 7 case-control studies and 2 meta-analyses did not (Bevan *et al.* 2005, Lohmussaar *et al.* 2005, van Rijn *et al.* 2005, Kuhlenbäumer *et al.* 2006, Kostulas *et al.* 2007, Xue *et al.* 2009, Sun *et al.* 2009, Bevan et al 2008a, Lövkvist *et al.* 2008; Table 1.2).

The Swedish study performed by Kostulas *et al.* (2007), for instance, did not find a significant association of the investigated SNPs in *PDE4D*, despite the study of Nilsson-Ardnor *et al.* (2005), in the same population, reporting positive results (Table 1.2). Additionally, Nilsson-Ardnor *et al.* (2005) report that Swedish families with IS suggestive linkage with the chromosome 5q12 locus containing the *PDE4D*. Also, among the studies from Nilsson-Ardnor *et al.* (2005) and Lövkvist *et al.* (2008) that found both

positive results, there are inconsistencies concerning the association of the SNP45 (Table 1.2).

In 2006, Staton *et al.* (2006), observed a significant association with IS of SNP41, SNP83 and SNP87 of the *PDE4D* through meta-analysis of nine case-control studies of 3,808 stroke cases and 4,377 controls. In this meta-analysis they include their own association results in the 151 IS Australian patients and 164 controls. They conclude that the differences they found among the studies in the direction of association between individual SNPs and stroke suggests that the SNPs tested are in LD with the causal allele(s) (Table 1.2). On the contrary, in a more recent and complete meta-analysis on 5,216 and 6,615 stroke patients and controls respectively, it was shown that none of the genetic variants studied was robustly and reproducibly associated with stroke, when the data from the original report was excluded (Bevan *et al.* 2008a; Table 1.2). Performing meta-analyses, however, has been difficult because there is significant heterogeneity between the studies with regard to the genetic variants analyzed, the phenotype used and the frequency of the variants among different populations.

It is verified yet that several studies that were unable to identify a positive association with IS, report trends toward association with different subtypes of stroke (Table 1.2).

1.7.2. ALOX5AP gene

Two years after the initial study of Gretarsdottir and colleagues (2003), a similar genetic linkage approach in the Icelandic population led to the identification of another stroke susceptibility gene, the ALOX5AP (or FLAP), coding for the membrane-associated arachidonate 5-lipoxygenase activating protein. The initial finding was a suggestive linkage to 13q12-13 in 296 Icelandic families with multiple affected members with MI. In an independent linkage study of males with IS or TIA, linkage to the same locus was observed, further suggesting that a cardiovascular susceptibility factor might reside at this locus. A case-control association study was then carried out using a high density of markers across the implicated region (containing 40 known genes) which led to the identification of the *ALOX5AP* as a susceptibility gene for MI. Because of the high degree of comorbidity among MI and stoke, the four-SNP haplotype HapA in the *ALOX5AP* that was found associated with MI (defined by the SNPs SG13S25, SG13S114, SG13S89 and SG13S32) was also studied for association in 702 individuals with stroke and 624 controls. It was verified that HapA doubles the risk of suffering stroke. This at-risk haplotype has higher frequency in all types of stroke, including IS, TIA and hemorrhagic stroke (Helgadottir *et al.* 2004).

ALOX5AP is a protein involved in the initial steps of leukotriene biosynthesis which is largely confined to leukocytes and can be triggered by a variety of stimuli. It was related to stroke, restenosis, MI and atherosclerosis (Helgadottir *et al.* 2004, Helgadottir *et al.* 2005, Mehrabian et al 2002, Brezinski *et al.* 1992, Spanbroek *et al.* 2003), probably through proinflammatory effects and the production of reactive

molecules such as superoxide anions (Dixon *et al.* 1990). ALOX5AP and 5-lipoxygenase together convert unesterified arachidonic acid to the leukotriene A4 (LTA4), which is further converted to LTB4 (by the action of LTA4 hydrolase), or to LTC4 (by the action of LTC4 synthase). These molecules are important proinflammatory mediators active in macrophages and leukocytes (Helgadottir *et al.* 2004). LTB4 binds to 1 of 2 receptors, LTB4R and LTB4R2, and acts as a potent attractor of neutrophils, induces recruitment of $CD8^+$ T lymphocytes, and promotes leukocyte adhesion to vascular endothelium (Goodarzi *et al.* 2003, Gimbrone *et al.* 1984). It was shown that male carriers of the HapA at-risk haplotype had significantly greater production of LTB4 in neutrophils (Helgadottir *et al.* 2004). Elevated levels of LTB4 might contribute to atherogenesis and/or plaque instability by increasing inflammation (Gulcher *et al.* 2005). LTC4, for other side, is subsequently converted to leukotriene D4 (LTD4) and leukotriene E4 (LTE4). LTC4, LTD4, and LTE4, collectively known as cysteinyl leukotrienes, bind to either cysLTR or cysLT2R and can cause altered endothelial cell permeability and vascular smooth muscle cell migration (Dahlen *et al.* 1981).

Like for *PDE4D*, the association of *ALOX5AP* has only been confirmed in some populations and replication studies have reported conflicting results (Table 1.3).

Table 1.3: Review of *ALOX5AP* association studies with IS and meta-analyses. For each publication, the number of cases and controls that were considered for the analyses of the studied polymorphisms, as well as the population where the sample are from, are indicated. Only significant results of association are indicated.

Reference	Polymorphisms **	N cases	N controls	Population	Significant association with IS*
ALOX5AP					
Case-control studies:					
Helgadottir et al. (2004)	48 SNPs	702	624	Iceland	HapA
Helgadottir et al. (2005)	25, 32, 35, 41, 89, 114, 377	450	710	Scottland	HapA
Lohmussaar et al. (2005)	4, 6, 25, 27, 32, 34, 35, 39, 41, 42, 86, 87, 89, 95, 96, 100, 114, 137, 188, 192, 372, 377	601	736	Germany	34, 86, 96, 100, 114, several haplotypes *males:* 32, 114 *females:* 89
Meschia et al. (2005)	25, 32, 89, 114, 2 MSs	377	263	US (74% white)	-
Zee et al. (2006b)	25, 30, 32, 35, 41, 42, 89, 106, 114, 377	259	259	US (men only)	-
Zhang et al. (2006)	89, 114	1,773	1,713	China	- *males:* 114
Kaushal et al. (2007)	34, 86, 89, 106, 114	357	303	US (74% white)	34-86-89-106-114* *whites:* 89*, 106*, 89-114 *cardioembolic in whites:* 89, 89-114, 34-86-89-106-114 *large-vessel in whites:* 89, 106, 89-114, 34-86-89-106-114 *small-vessel in whites:* 34-86-89-106-114

Kostulas et al. (2007)	25, 32, 89, 114	685	751	Sweden	-
Shen et al. (2007)	89, 114	1,893	1,891	China	-
Bevan et al. (2008b)	25, 32, 35, 41, 89, 377 + 12 SNPs	872	933	UK	89, rs4503649, rs3885907, rs3803278, rs4503649-rs3885907-rs3803278 *large-vessel:* 89, rs4503649, rs3803278, 89-rs4503649-rs3803278 *cardioembolic:* rs3803278 *small-vessel:* 89, rs4503649, rs3885907, rs4769060, rs4503649-rs3885907-rs4769060
Bevan et al. (2008b)	89 + 4 SNPs	601	736	Germany	- *cardioembolic:* 89
Lövkvist et al. (2008)	25 + 2 SNPs	932	396	Sweden	- *non hypertensive:* 25
Meta-analysis:					
Zintzaras et al. (2009)	25, 32, 35, 41, 42, 89, 106, 114, 377	5,194	4,566	Meta-analysis	-

*Significant SNPs that resist adjustment for multiple testing.
**SNP ID number in original report.
HapA: Haplotype A in *ALOX5AP* defined by SNPs 25, 32, 89 and 114.

Looking for the overall results, 6 case-control studies in different populations confirmed the association with IS of at least one genetic variant studied by Helgadottir *et al.* in 2004 (Helgadottir *et al.* 2005, Lohmussaar *et al.* 2005, Zhang *et al.* 2006, Kaushal *et al.* 2007, Bevan *et al.* 2008b, Lövkvist *et al.* 2008; Table 1.3), whereas 4 case-control studies and a meta-analysis did not (Meschia *et al.* 2005, Zee *et al.* 2006b, Kostulas *et al.* 2007, Shen et al 2007, Zintzaras *et al.* 2009; Table 1.3).

1.8. Genome-wide association studies

A GWAS is an approach that involves rapidly scanning markers of many samples across the complete genome, to find genetic variations associated with a particular disease. Such studies are particularly useful in finding genetic variations that contribute to common complex diseases.

GWAS with SNP markers have been extensively performed in the past few years, enabled by the decreasing genotyping costs, massively multiplexed genotyping technologies and the large-scale SNP discovery that result from the genotyping efforts of several projects and consortiums. Studies on LD patterns suggest that GWAS require the genotyping of several hundred thousands of SNPs in each individual (Evans *et al.* 2004) and the successful identification of alleles associated with disease susceptibility with modest effects require the analysis of thousands of samples (Wang *et al.* 2005). The selection of the most informative SNP panels for GWAS varies between different populations and genomic regions (Evans *et al.* 2004, Wang *et al.* 2005).

The HapMap Project (The International HapMap Project), a public database that currently contains data for more than two millions of SNP makers with verified allele frequencies in different

populations, has been, for instance, of the major importance to improve the performance of these studies. The aim of the HapMap Project is to characterize LD patterns across the genome to facilitate selection of the most informative subsets of tagging SNP (Johnson *et al.* 2001), turning possible to perform GWAS coupled with the availability of high-throughput genotyping methods. Additionally, compared with one-stage designs that genotype all samples on all markers, well-constructed two-stage association designs maintain power of the GWAS while substantially reducing genotyping requirements (Satagopan *et al.* 2004, Thomas *et al.* 2004).

In GWAS, tests of association with disease status are normally conducted one SNP at a time. Several efforts are also been made to developed computationally efficient methods to simultaneously analyse all SNPs, either in GWAS, or in fine mapping study based on re-sequencing and/or imputation, to find a set of SNPs most associated with disease risk. This efforts will avoid to ignoring the effects of all other genotyped SNPs when a single SNP is being studied, improving the performance of the single-SNP tests, since a weak effect may be more apparent when other causal effects are already accounted for. Additionally, a false signal may be weakened by inclusion in the model of a stronger signal from a true causal association (Hoggart *et al.* 2008).

In contrast to candidate gene studies, GWAS allow meticulous scan of the genome in an unbiased manner having the potential to identify totally new susceptibility genes. There are not many GWAS performed to date on stroke (Table 1.4). However, numerous GWAS have so far been published, such as on AMD, Parkinson's disease, MI, and inflammatory bowel disease, and have led to the identification of novel loci that have been independently replicated (Maraganore *et al.* 2005, Klein *et al.* 2005, Fung *et al.* 2006, Ozaki *et al.* 2002, Duerr *et al.* 2006).

In 2007, Matarin *et al.* published the first GWAS in IS. They analysed a cohort of samples of 249 IS white cases and 268 white neurologically normal controls for more than 400,000 unique SNPs. No single locus conferred a large effect on IS risk and none of the results were significant after Bonferroni's correction for multiple testing (Table 1.4). However, some of the most significant SNPs were located within or near interesting candidate genes (such as the KCNIP4 and KCNK17 genes involved in potassium transport) providing an apparent moderate-high risk. Twenty seven of the studied SNPs had an association p-value $< 1\times10^{-5}$ (Table 1.4). These results suggest that there is no single common genetic variant exerting a major risk on stroke.

In the same year, Kubo *et al.* (2007) reported another GWAS in 188 Japanese individuals with IS and 188 matched controls, using 52,608 gene-based tagging SNPs. In a second phase, they genotyped a larger cohort of 924 IS cases and 924 matched controls for the 1,098 SNPs identified associated in the first phase with p-values below 0.01. Through this analysis, they found an association of the SNP15 in *PRKCH* with lacunar infarction. Further sequencing of all exons in this gene in 48 cases and 48 controls revealed the SNP 1425G/A, which was confirmed to be associated with lacunar infarction by direct sequencing in

all cases and controls (Table 1.4). This SNP causes an amino acid substitution at position 374 from valine to isoleucine. This position is within the ATP-binding site of the protein, member of the protein kinase C (PKC) family, enhancing PKC activity. The association results in the *PRKCH* were replicated in an independent cohort of 1,137 cases and 1,875 controls selected from Biobank Japan project. Furthermore, a 14-year follow-up cohort study in Hisayama (Japan) supported the involvement of this SNP in the development of stroke. The effect was similar on the development of CHD suggesting that the studied SNP is a common genetic risk factor for these diseases. The SNP 1425G/A, however, is in almost complete LD with the SNP15 which has very small minor allele frequencies in Europeans and Africans, suggesting that the obtained results are likely to be specific to Asian populations.

Also in 2007, using SNPs from a 100K genome-wide scan, Larson *et al.* analysed 1,345 white participants of European descent from 310 families for four major CVD outcomes like the major atherosclerotic CVD (n = 142) including MI, CHD death and stroke. No association reached genome-wide significance, however significant SNPs pointed to few candidate genes of interest for any major CVD outcomes, such as ALOX5AP. Association of a 13-kb region on chromosome 9p21 was also observed for major atherosclerotic CVD and major CHD, as reported for MI and CHD (Helgadottir *et al.* 2007, McPherson *et al.* 2007). This suggests that these GWAS were able to identify true associations despite the large number of tests performed and the small number of individuals.

In 2008, Gretarsdottir *et al.* performed a GWAS with 1,661 Icelandic IS patients and 10,815 control subjects. A total of 310,881 SNPs were tested for association with IS, and the most significant signals were replicated in two large European IS samples with a total 2,224 IS cases and 2,583 controls. Additionally, the two replicated SNPs rs2200733 and rs10033464 (neighbour SNPs) in 4q25, were tested further in other two European IS samples including 2,327 IS patients and 16,760 unaffected controls. Both SNPs were strongly associated with cardioembolic stroke, and the SNP rs2200733 also associated with all IS and IS not classified as cardioembolic (Table 1.4). These two variants were previously shown to be associated with AF.

More recently, in 2009, Ikram *et al.* reported a GWAS including 19,602 white persons from the consortium the Cohorts for Heart and Aging Research in Genomic Epidemiology, in whom 1,544 incident strokes (1,164 IS) are available. Over two million autosomal SNPs were included in the analyses. They tested the most strongly associated markers with stroke in a replication sample of 2,430 black persons with 215 incident strokes (191 IS), in another cohort of 574 black persons with 85 incident stroke cases (68 IS), and in a self-reported white case-control group of 652 Dutch IS cases and 3,613 controls. The authors found two intergenic SNPs highly associated with all strokes, IS and atherothrombotic stroke on chromosome 12p13 and within 11 kb of the gene NINJ2 (ninjurin2; Table 1.4). In addition, several other SNPs within NINJ2 showed a modest association with both phenotypes (13 SNPs with all strokes, and 10 SNPs with IS). Direct genotyping showed that the SNP rs12425791 was associated with an increased risk,

yielding population attributable increase risk of 11% when all the cases were included and of 12% for the IS only. For this SNP, the findings were replicated in the larger cohort of black persons and in the Dutch sample, being nonsignificant in the underpowered analysis of the smaller black cohort. The results obtained for the other associated SNP in the discovery population (rs11833579), were not confirmed in the replication cohorts.

Ninjurin2 is one of two transmembrane proteins in the ninjurin (nerve-injury-induced protein) family. *NINJ2* encodes an adhesion molecule that interacts with matrix metalloproteinases. It is expressed in glia and shows increased expression after nerve injury (Ikram *et al.* 2009).

Table 1.4: Review of the GWAS performed for stroke. The principal characteristics of the discovery and replication cohorts, the number of genotyped SNPs, and the major findings in each GWAS are here summarized.

Study	Discovery cohort		SNPs genotyped	Replication cohort(s)		Major findings
	Origin	Size	in GWAS	Origin	Size	
Matarin et al. (2007)	USA (white only)	249 IS + 268 controls	>400,000 SNPs	-	-	No SNP associated at a genomewide level (27 SNPs with P<1E-5)
Kubo et al. (2007)	Japanese	188 IS + 188 controls	52,608 gene-based tag SNPs	Japanese	924 IS + 924 controls	SNP 1425G/A in PRKCH (chr. 14q22) associated with IS (P=5.1E-7, OR=1.40)
Gretarsdottir et al. (2008)	Iceland	1,661 IS + 10,815 controls	310,881 SNPs	Germany, Sweden, UK, and Iceland	4,551 IS + 19,343 controls	rs2200733 in chr. 4q25 associated with IS (P=2.18E-10, OR=1.26) and cardioembolic stroke (P=8.05E-9, OR=1.54)
Ikram et al. (2009)	USA + Netherlands (white only)	1,544 cases (1,164 IS) + 18,058 controls	~2.5 million (genotyped and imputed) SNPs	USA (black only) + Netherlands, (white only)	300 cases (259 IS) + 1,704 controls 652 IS + 3,613 controls	2 SNPs within 11kb of NINJ2 (chr. 12p13) associated with stroke and IS (p-value < $2x10^{-8}$, OR~1.3)

1.9. Expression studies

Several state-of-the-art technologies make the analysis of the human transcriptome (all expressed transcripts in the human genome) possible. These methods complement the traditional methods, like Northern blotting or *in situ* hybridization, used to study one gene at a time. Microarrays (or gene chips), allow the quantitative monitoring of the genetic expression patterns, measuring the mRNA levels of tens of thousands of genes in a single experiment, due to the two basic principles of the nucleic acid hybridization: the mRNA are going to bind specifically to its complementary probes and this binding occurs proportionally to the abundance of a sequence in a mixture. It provides the means to assess gene expression on a very large scale, therefore becoming a discovery tool that allows the identification of

genes and pathways linked to a wide variety of conditions. Other techniques for the measurement of genetic expression, that imply large-scale sequencing, like serial analysis of gene expression and massively parallel signature sequencing, can be more sensitive, however, no technique is as fast as a microarray and, therefore, this technique has a higher probability of being useful in the clinical practice.

In the last years, expression studies of several disorders have lead to important insights into disease aetiology, in particular, in the study of cancer.

To understand the biological cascade initiated by a stroke, several types of gene expression studies have already been published in animal models where stroke is induced (Jin *et al.* 2001, Tang *et al.* 2001, Kim *et al.* 2002, Stenzel-Poore *et al.* 2003, Roth *et al.* 2003, Dhodda *et al.* 2004), in animals that are naturally resistant or sensitive to stroke (Fornage *et al.* 2003), as well as in humans (Moore *et al.* 2005a, Moore *et al.* 2005b, Tang *et al.* 2006, Baird 2007, Xu *et al.* 2008).

For these genetic profiling studies, the selection of the used cell type is critical. In 2001, for instance, Tang and colleagues (2001) chose to examine the blood profiling to study stroke and other medical conditions in rats. Blood is frequently the selected tissue since it is the most readily accessible sample. To their surprise, genes in the blood profile were up-regulated or down-regulated depending on organ-specific injuries. In the case of the stroke, there are genes induced and other genes suppressed in white blood cells after brain ischemia and brain haemorrhage. Ischemic and hemorrhagic stroke produced different patterns of gene expression in blood that appear to be considerably specific. However, the specific factors that lead to downregulation or upregulation of the genes in each of the conditions remain unclear. The authors propose that the mechanisms by which different brain injuries affect specific gene expression in white blood cells may be related to immune surveillance.

More recently, genetic profiling studies of stroke on human peripheral blood cells has been published. The goal of those studies was to determine whether a systemic gene expression profile could be observed in the acute phase of stroke, and therefore blood of patients was collected in the first hours after the stroke event. It was observed a predominant upregulatory response in IS (Moore *et al.* 2005a, Moore *et al.* 2005b, Tang *et al.* 2006). After a stroke, there is a selective recruitment and migration of white blood cells to the ischemic focus in the brain due the consequent inflammation response. Additionally, it was observed that gene expression of the white blood cells remains relatively constant over the first week of stroke, also providing evidence that the peripheral blood gene expression signatures can be reproducibly demonstrated (Baird 2007). In the subsequent time course analysis by Moore *et al.* (2005b), blood samples were also drawn at 3 months postischemic stroke, and a longitudinal analysis was performed. The results suggested at least a partial adaptive response to the altered cerebral microenvironment, but also that some of the gene expression changes could be attributable to coexisting vascular risk conditions. Apparently, all white cell types are involved and are believed to impact significantly on tissue and clinical outcome through the exacerbation of ischemic injury, particularly after reperfusion, and conversely contributing to

tissue remodelling and repair days to weeks after stroke (Baird 2007). Finally, Xu and colleagues (2008) also suggest that it might be possible to determine the aetiology of stroke based on gene expression in peripheral blood after stroke. They define the expression profiles in blood that apparently differentiate cardioembolic from large-vessel atherosclerotic stroke in the first hours after the cerebrovascular accident. These studies performed in the acute phase or in the first months after the stroke event, however, do not address stroke risk in the first place. Rather, they examine the post-stroke phase.

Other studies were conducted in human arteries with and without atherosclerosis but these are not specific to stroke (Woodside *et al.* 2003, Archacki *et al.* 2003, Vemuganti *et al.* 2005). Vemuganti *et al.* (2005) tried to identify genetic mechanisms that promote the onset of stroke and TIA symptoms in carotid atherosclerotic patients. From all the carotid atherosclerotic plaques, commonly found in adults, only a few induce stroke symptomology and, according to the Asymptomatic Carotid Atherosclerotic Study, the size of the plaque alone is not predictive (Baker *et al.* 2000). Several expression studies using atherosclerotic plaques from humans and experimental animals indicated an increased inflammation in the plaque tissue compared with the normal blood vessel tissue (Faber *et al.* 2001, Wuttge *et al.* 2001, Hiltunen *et al.* 2002, Randi *et al.* 2003), but the molecular mechanisms that make some of them become symptomatic leading to rupture and embolization are not known. Vemuganti *et al.* (2005) compare geometrically similar carotid artery plaque samples extracted from six symptomatic stroke patients and four asymptomatic controls. Their results suggest that symptomatic plaques are molecularly and biochemically more active than the asymptomatic plaques, or that active plaque growth precipitates stroke symptoms. This study, however, had a very small sample size. No other published study in humans has tried to find gene expression changes that specifically increase the risk for a future stroke event.

A recent study on the differential expression of genes upon acute IS in human peripheral blood mononuclear cells (PBMCs) showed that the most significant expression changes between IS patients and healthy subjects were in the PDE4D gene and for inflammatory response gene cluster (Grond-Ginsbach *et al.* 2008). Also Gretarsdottir *et al.* (2003) observed that the total *PDE4D* mRNA level was significantly lower in B lymphocytes of affected individuals than in controls when they propose that this gene confers stroke risk. Kassner *et al.* (2009) report that the majority of PBMC subpopulations (cluster of differentiation CD3+, CD14+, CD19+, CD68+) show an increased expression of proinflammatory, proapoptotic or adhesion-relevant genes in patients, and significant positive correlations were observed between expression of most of these genes in PBMCs and individual plasma concentrations of oxidized low-density lipoproteins (oxLDL). The authors suggest that this elevated expression after IS may contribute to an immunodepressive syndrome, possibly due to increased plasma oxLDL levels. All these studies make promising the analysis of the PBMCs for the study of stroke.

The great number and variety of studies performed with microarrays in the last years allowed this technology to reach a maturity level such that the results produced are reliable, in contrast with what

happened a few years ago when this technology was launched. The experimental protocols, the microarrays design, the data analysis programs, have been tested and perfected with the experience acquired by many groups of investigators, creating optimal scientific and technical conditions necessary to their success.

1.10. "Genomic convergence" (GC) approach

The ultimate goal of this project is to unravel new genetic factors involved in susceptibility to the common form of stroke. We applied the novel and multifactorial GC approach (Hauser *et al.* 2003) for the first time in the field of stroke genetics. This approach combines genomic screening, expression analysis and association studies (Figure 1.1). Unlike a pure biological candidate gene analysis, this approach has the tremendous advantage of being unbiased by preconceived models of disease.

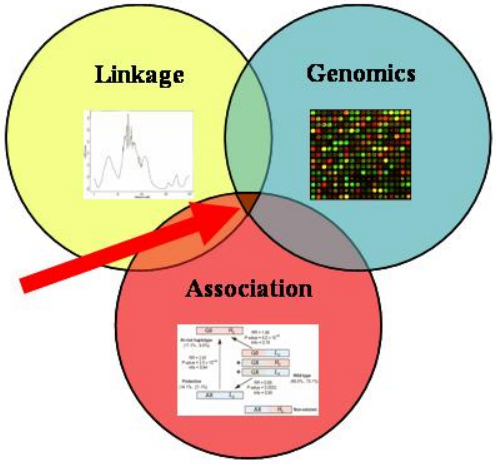


Figure 1.1: Illustration of the "genomic convergence" (GC) process. This approach combines genomic screening, expression analysis and association studies. The goal of this strategy is to identify the genes in the narrowed brown area.

Whole-genome linkage studies, the first convergence factor in this process, lead to the identification of several chromosomal areas linked to the trait of interest, each of which is typically 10-30 megabases wide and harbours hundreds of genes, far too many for a detailed follow-up analysis or an individual tissue expression analysis. Microarray expression profiling, the second convergence factor, allows us to prioritize the analysis of those genes that are significantly differentially expressed in a tissue of interest. Conversely, the usefulness of expression analyses with microarrays is often greatly hampered by the overwhelming amount of information generated, and genes with greatest levels of differential

expression or significance are not necessarily the most important to follow-up. Determining which expression changes are significant and should have valuable resources committed to their investigation can be a very difficult proposition. The use of intersecting data derived from these last two powerful resources presents the first step towards a more efficient method of focusing effort on susceptibility genes. The third and last convergence factor is association studies on the genes differentially expressed in microarray experiments and lying in the areas of linkage.

This strategy has already proved to be successful for identifying genetic risk factors in other complex diseases such as AD and Parkinson's disease (Li *et al.* 2003). It has also been applied to a number of other genetic systems. The combination of quantitative expression analysis with genetic mapping has been used to analyze the murine NOD model of type 1 diabetes (Eaves *et al.* 2002). Similar approaches have been used to identify complement factor 5 as a susceptibility locus for experimental allergic asthma (Karp *et al.* 2000), and to dissect the transcriptional regulation in budding yeast (Brem *et al.* 2002). Analysis of expression data as a quantitative trait locus has been used to phenotypically subdivide two mouse models of obesity (Schadt *et al.* 2003).

1.11. Impact of stroke genetics and genomics

There is general agreement that the field of stroke is ripe for the genomic revolution that is creating new and previously unimagined opportunities for diagnosis and treatment of neurological and other diseases. The identification of the genetic underpinnings of stroke has lagged substantially behind progress made in other polygenic neurological disorders such as AD and Parkinson's diseases. The "genetic revolution" has opened up many new directions in disease research, especially in genetically complex diseases. Genomic medicine should move the practice of medicine from a science that reacts to disease to one that prevents it. While gene expression gives us insight into the pathophysiology and outcome changes, SNP association provides us with information on aetiology, gene-gene and gene-environment interactions. Despite limited current knowledge regarding the genetics of stroke, identification of stroke genes represents the clearest path to a better understanding of the mechanisms underlying stroke. It will have an immediate and long lasting clinical and public health benefits by increasing the motivation to implement preventive lifestyle changes in individuals at risk, which will ultimately result in a decrease of the number of strokes. The identification of stroke genes, of the proteins they encode and of their functions will also provide biological and clinical information on the development, primary and secondary prevention and treatment of strokes. Eventually, this information will be used in pharmacogenetics where preventive or acute therapy is chosen based upon genetic makeup of each individual.

2. OBJECTIVES

The ultimate goal of this research is to identify new susceptibility genes for the risk of developing stroke, using for the first time the GC approach in the field of stroke genetics, and testing the association of some novel biological candidate genes. More specifically, our specific aims were:

1. To create a biobank of Portuguese Caucasian cases and unrelated healthy controls to study the genetic basis of stroke in idiopathic cases;
2. To assess the association of *PDE4D* and *ALOX5AP*, that have been in recent years controversially implicated in IS, with IS risk in our full dataset;
3. To assess the association of KALRN, CFH, EPO, HO2 and KLK1 candidate genes with the IS risk in all samples of the our full dataset;
4. To conduct the novel GC approach, that combines genomic screening, expression analysis and association studies, for the first time in the field of stroke;
   4.1 To conduct gene profiling studies on 20 patients and 20 unrelated healthy controls to identify genes whose expression pattern suggest their involvement in stroke aetiology;
   4.2 To converged the expression results with the ones from published stroke whole genome linkage screens to prioritize genes for association studies;
   4.3 To conduct association studies on prioritized genes in all samples of our full dataset to verify their association with the IS risk;
5. To confirm our positive findings in independent datasets.

We believe that novel genetic factors as well as gene-gene and gene-environment interactions identified in this project or in future complement works, will contribute to a more effective prevention of the disease through genetic screening of the population for predisposing alleles, genotypes or haplotypes.

3. SUBJECTS AND METHODS

3.1. Ethical considerations

The study was approved by the ethics committees of several participating institutions and of the *Instituto Nacional de Saúde Ricardo Jorge* (INSARJ) where works Dr. Astrid Vicente that collaborated with us in this project. Qualifying patients from participating clinicians were contacted for potential enrolment in the study. All participants were informed of the study and provided written informed consent (Appendix A.1. and A.2.) and blood samples (for DNA, RNA and/or plasma) were collected by venipuncture. Each sample collected for our *Instituto Gulbenkian de Ciência* (IGC) biobank was bar-coded upon sampling and receipt in order to insure sample integrity, and assigned a consecutive sample number that will be used in all genotyping and statistical analyses (Appendix A.3), and similar procedures were adopted in INSARJ. Dried blood cards are regularly obtained with every blood sample as a back-up source of DNA if sample identity or integrity is questioned. Participant information was gathered at the time of biological sample collection and stored in a secure BC|Gene database (Biocomputing Platforms Ltd, Finland). BC|Gene database allows an integrated, customizable relational functionality for quick and easy data management. Extensive checks of the data for valid values and internal consistency were performed. Inconsistencies were be resolved by the clinical team in collaboration with the laboratory team. Laboratory personnel who obtains or has access to genotypic data is blinded to individual personal identifiers, and vice-versa.

3.2. Study subjects of our full dataset

Our dataset is composed by samples from IGC, that were purposely collected for this project, and by samples from and INSA, that were collected in different nationwide studies.

Five hundred sixty five unrelated patients with a clinical diagnosis of IS were recruited through Neurology and Internal Medicine Departments throughout Portugal by clinicians with special interest and training in stoke to ensure the accuracy of diagnoses. Stroke was defined by the presence of a new focal neurological deficit, with an acute onset and symptoms and signs persisting for more than 24 hours, and was confirmed (Adams *et al.* 1993) by Computed Tomography Scan in 97% of cases and/or Magnetic Resonance Imaging in 25% of patients. Patients having at least one IS episode were included in this study. All patients have a Portuguese Caucasian origin, are adult and were under the age of 75 at stroke onset. TIAs, trauma, tumours, infection, and other causes of neurological deficit were excluded. We excluded iatrogenic forms of stroke in which the episode of stroke occurs in the first 48 hours after an invasive

cerebrovascular or cardiovascular procedure or a general surgery; stroke episodes as consequence of a vasospasm after a nontraumatic subaracnoid hemorrhage; mitochondrial or monogenic forms of stroke (such as CADASIL, Fabry's disease and MELAS); and stroke episodes that occurs in a context of a bacterial endocarditis. Data collection forms were developed for the samples collected for the IGC biobank including extensive clinical information such as stroke characteristics, general clinical observation, neurological symptoms and signs, complications and interventions during hospitalization and situation at discharge (Appendix A.4.). Additional diagnostic criteria procedures such as neurological evaluation with the NIH Stroke Scale score (Lyden *et al.* 1999), electrocardiogram and extracranial Doppler, as well as evaluation procedures such as the modified Rankin Scale (van Swieten *et al.* 1988) and the Barthel index (Granger *et al.* 1979), were adopted according to internationally accepted consensus criteria. Data was also collected on relevant lifestyle aspects, including information about social environment, drugs, smoking and drinking, and on previous clinical risk factors such as hypertension, hypercholesterolemia, AF and diabetes mellitus (Grau *et al.* 2001; Murat Sumer *et al.* 2002; Lindenstrom *et al.* 1993) (Appendix A.4.) for the study of covariates and gene-environment interactions. Participants were also asked for family history (Appendix A.6.). Samples from INSARJ that were recruited in the context of a previous study aimed at characterizing stroke in patients under 65 (http://www.onsa.pt/index_17.html). For this study, a database was created following a similar protocol including information at the time of first hospitalisation and, for the patients available for follow-up, at three and twelve months after the stroke episode. The collected information includes extensive clinical data such as stroke type, neurological symptoms, complications and interventions during hospitalization; outcome measures including disability score (Rankin score) at discharge and at the subsequent evaluation points; stroke-associated risk factors present before stroke such as hypercholesterolemia, heart disease and hypertension; biochemical parameters such as lipid profile; and life-style information on consumption of alcohol and tobacco and exercise habits, in concordance with the information we collect.

Five hundred and twenty unrelated healthy individuals were included in this study as a control sample population. Since stroke is a late-onset disease, the control group was selected from a group of healthy volunteers with a higher mean age than the case group, thus minimizing the chances for mis-classification as "stroke-free". All controls have also a Portuguese Caucasian origin and are adult. Control individuals collected for the IGC biobank were verified to be free of stroke by direct interview before recruitment, applying the "Questionnaire for verifying the stroke-free status" (Meschia *et al.* 2000; Jones *et al.* 2000) (Appendix A.7.), but no brain imaging studies were performed. The interview also included questions on established clinical and life-style risk factors for stroke according to the information collected for patients (Appendix A.4.). The controls that belong to INSARJ are over 65 years old and belong to a database and biobank that includes clinical history, life style information, biochemical parameters, as well as DNA samples.

The clinical team, who provides the patient ascertainment and clinical expertise to identify, diagnose, and collect samples and information from the cases and controls for the "Stroke IGC" cohort since 2004, is currently composed of doctors working in H. Santa Maria (J. M. Ferro, L. Gouveia), H. Santo António (M. Correia, A. Tuna, G. Lopes), H. Egas Moniz (M. Viana-Baptista), H. Fernando Fonseca (A. Pinto, R. Silva), H. São Marcos (J. Fontes, C. Ferreira), H. São Bernardo (M. Rodrigues), H. São João (M. Monteiro). From INSARJ, three distinct cohorts were used in this study: the "Stroke INSA" that is a nationwide study on the epidemiology of stroke, the "APOEurope" that was designed to assess the prevalence of ApoE4 in Portugal, and the "6 regions controls" database that is composed of samples from healthy individuals. Samples from the "Stroke INSA" cohort that, as explained, were in the context of an earlier study by Dr. Marinho Falcão and his team at *Instituto Nacional de Saúde Dr. Ricardo Jorge*, and by all the clinicians that recruited study subjects from H. São João, H. Évora, H. Funchal, H. Marmeleiros, H. São Bento, H. São José, H. São Marcos, H. Garcia d'Orta, H. Faro, H. Coimbra, H. Vila Nova de Gaia, H. Aveiro, SAMS, H. Capuchos and H. Santo António.

3.3. Gene expression profiling

Gene expression profiles of PBMCs from stroke patients and controls were conducted and reported in this project in accordance with the MIAME criteria (Brazma *et al.* 2001). GEO accession number for the data is GSE22255.

3.3.1. Subjects

For this specific analysis we adapted the study manual created for the construction of our general IGC biobank to guarantee even more rigorous inclusion and exclusion criteria: IS patients were required to have suffered only one stroke episode, at least at 6 months before the blood collection, and controls did not have a family history of stroke. Participants with severe anaemia or active allergies were also excluded. Additional information about the current medication of the participants at the time of the blood collection and the current score on the modified Rankin scale and on the Barthel scale of the stroke patients was collected (Appendix A.5.). The majority of samples from the case and control groups used in these expression analysis were collected in the north of Portugal in three different regions: one urban region, Porto (H. Santo António: G. Lopes, R. Taipa), and in two rural regions from Trás-os-Montes, Vila Real (H. São Pedro: M. R. Silva, J. P. Gabriel) and Mirandela (H. Distrital de Mirandela: I. Matos), in the sequence of a prospective study of stroke conducted in 2004 (Correia *et al.* 2004). Only four samples were collected in Lisbon (H. Santa Maria: J. M. Ferro, L. Gouveia), another urban region, for logistic reasons

and to match the individuals for sex and age. Forty samples were analyzed: twenty IS cases and twenty age- and sex-matched controls.

3.3.2. Total RNA isolation

Whole blood samples were obtained by venipuncture and collected in two 8 mL BD Vacutainer CPT tubes (BD, USA). PBMCs were isolated by centrifugation of the whole blood in the CPT tubes and the cells were washed twice according to the manufacturer's protocol. The RNA from the washed cells were stabilized using RNAlater (Qiagen, Germany) within 3 hours after sample collection. Approximately 10 μg of high-quality total RNA was then extracted from PBMCs using the RNeasy Mini kit as recommended (Qiagen, Germany).

The total RNA was precipitated overnight with sodium acetate 3M and absolute ethanol at -20 ºC, the pellet was washed twice with 80% ethanol and after air drying the pellet, it was resuspended in 10 μL of DEPC-treated water to reach appropriate concentrations.

The quality and integrity of the samples was analyzed after resuspension by measuring the A_{260}/A_{280} ratio and using the RNA 6000 Nano Assay on an Agilent 2100 Bioanalyzer (Agilent, USA) (Figure 3.1). High-quality RNA is an absolute prerequisite to obtain reliable and reproducible microarray data. The A_{260}/A_{280} ratio should be between 1.8 and 2.1 for pure RNA and the RNA integrity number (RIN) given by the Bioanalyzer analysis should be close to 10. RNA concentrations were estimated by measuring the absorbance at 260 nm and samples were aliquoted and stored at -80°C. Absorvancies were measured using a NanoDrop® ND-1000 spectrophotometer. Genomic DNA were requantified using the PicoGreen reagent (Molecular Probes, USA).

3.3.3. Hybridization to human genome microarrays

3.5 μg of high-quality total RNA from each selected individual were hybridized to GeneChip® Human Genome U133 Plus 2.0 microarrays (Affymetrix, USA), which harbour more than 54,000 probe sets and 1,300,000 distinct oligonucleotides representing 47,000 human transcripts and variants, including 38,500 well-characterized genes. Samples were processed at the IGC's Affymetrix Core Facility following manufacturer's protocol (No authors listed 3, 2005-2006). Briefly, in a first step, double-stranded cDNA is generated that carries a T7 promoter at its 5´end. This promoter is then used for in vitro transcription, in which biotinylated nucleotides are incorporated into the resulting cRNA. This second step leads to an approximate 100-fold linear amplification of the starting material. After, the biotin-labelled cRNA is then fragmented into 35-200 bases fragments by metal-induced hydrolysis and hybridized to GeneChip arrays

for 16 hours. Each sample is separately hybridized to a single GeneChip® array. The array are washed and the bound biotinylated cRNA fragments are stained with a fluorescent streptavidin conjugate and fluorescent intensities for each probe cell are acquired on a GeneChip® scanner 3000 – 7G (Figure 3.2). Quality control (QC) checks of non-fragmented and fragmented cRNA on Agilent 2100 Bioanalyzer were preformed (Figures 3.3 and 3.4) before continue with the experiments.

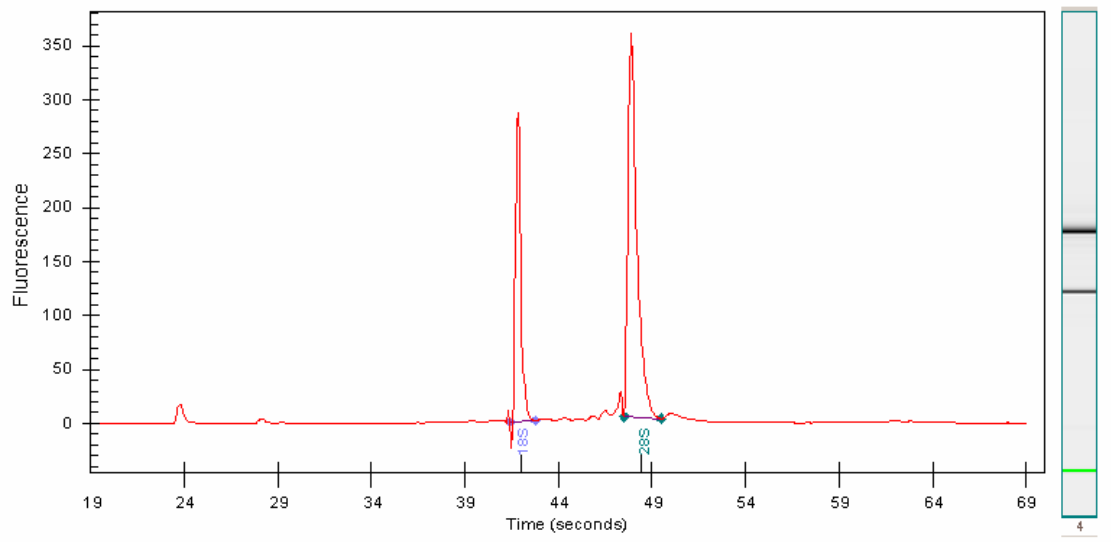


Figure 3.1: Good quality total RNA analyzed on the Agilent 2100 Bioanalyzer using the RNA 6000 Nano Assay. An electropherogram and gel-like image of mammalian total RNA are presented. RNase degradation would be detected by a shift on the RNA size distribution towards smaller fragments and a decrease in fluorescence signal of ribosomal 18S and 28S peaks. A good quality sample will typically have a ratio of 28S:18S ribosomal peaks of 2:1. Picture produced by Affymetrix (No authors listed 3, 2005-2006).

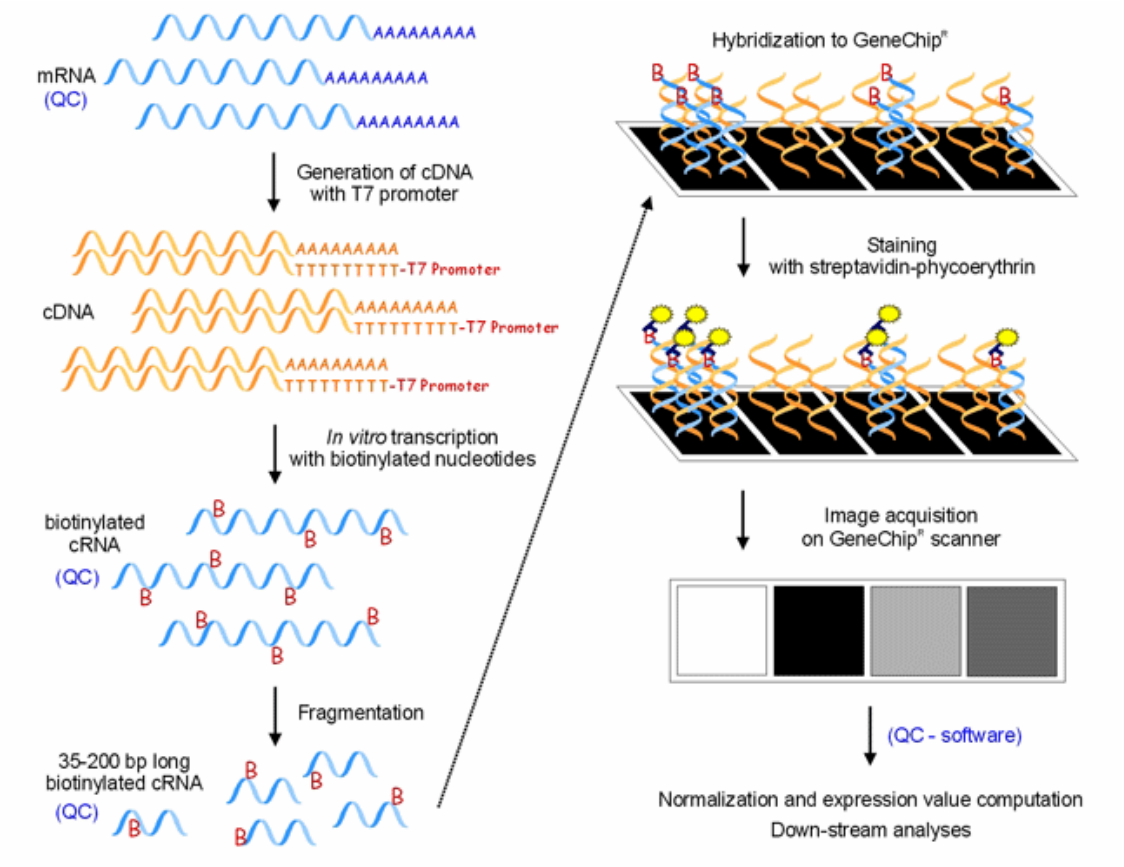


Figure 3.2: Simplified overview of the GeneChip gene expression analysis protocol. Starting from single-stranded mRNA, double-stranded cDNA is generated that is then used for in vitro transcription resulting in single-stranded, biotinylated cRNA. Fragmented biotinylated cRNA (target) is then hybridized to the immobilized single-stranded oligonucleotides (probes) on the GeneChip® and fluorescent intensities are acquired on a GeneChip® scanner. Quality checks before and during the target preparation are performed using an Agilent 2100 Bioanalyzer. Picture produced by Affymetrix.

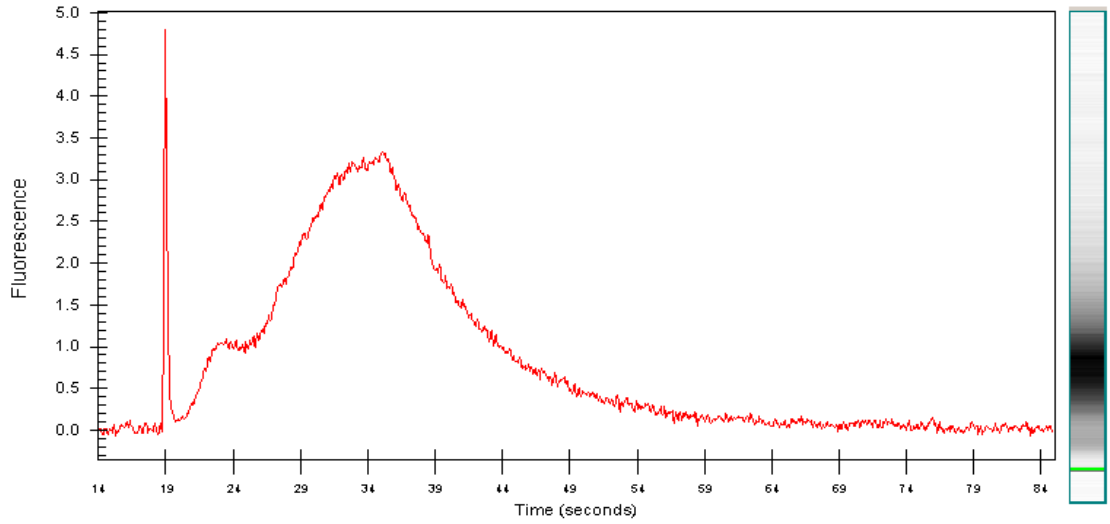


Figure 3.3: Good quality biotin-labelled cRNA analyzed on the Agilent 2100 Bioanalyzer using the RNA 6000 Nano Assay. An electropherogram and gel-like image of labelled cRNA from HeLa total RNA are presented. This electropherogram displays the nucleotide size distribution for 400 ng of labelled cRNA resulting from one round of amplification. The average size is approximately 1580 nucleotides. Picture produced by Affymetrix (No authors listed 3, 2005-2006).

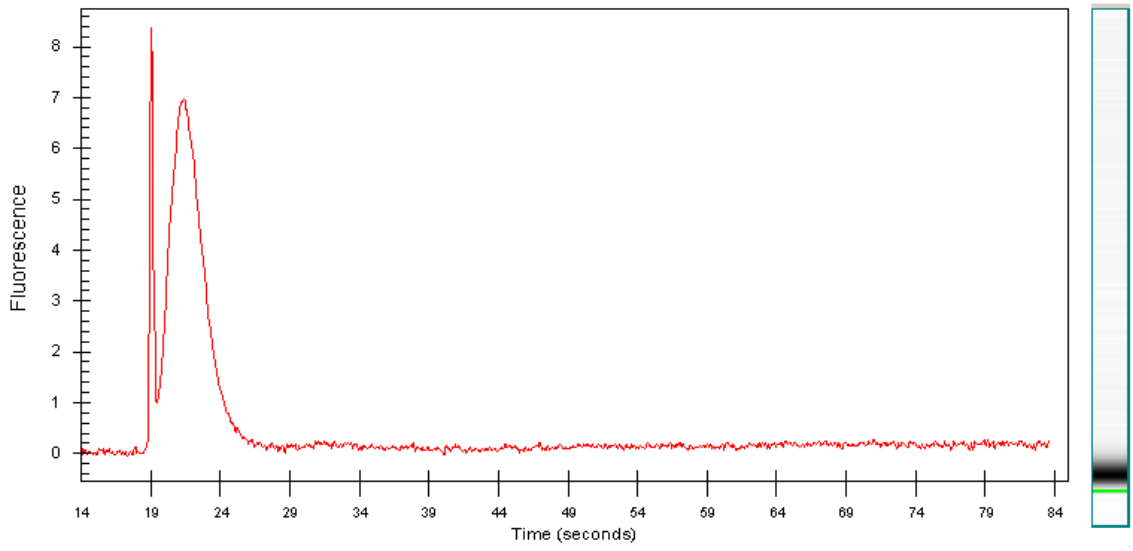


Figure 3.4: Good quality fragmented cRNA analyzed on the Agilent 2100 Bioanalyzer using the RNA 6000 Nano Assay. An electropherogram and gel-like image of fragmented labelled cRNA from HeLa total RNA are presented. This electropherogram displays the nucleotide size distribution for 150 ng of fragmented labelled cRNA resulting from one round of amplification. The average size is approximately 100 nucleotides. Picture produced by Affymetrix (No authors listed 3, 2005-2006).

3.3.4. Quality control

An extensive and detailed QC that allows monitoring all sample manipulation, hybridization, washing and staining, as well as the obtained images of the arrays, was performed before the subsequent analysis of the expression data. Such QC is possible taking in account a set of parameters and because the used arrays have probe sets from prokaryote genes that are mixed with the experimental samples at different steps of the employed procedures.

The inspection for the presence of image artefacts on the arrays, such as scratches, black spots, white spots, dark circles, or others, that can affect all the analysis including the QC procedures, was performed using the Affymetrix GCOS 1.1 software where the results are obtained. This software were also used to calculate a confidence level for each transcript to determine if it is detected (present call) or not (absent call); the background value of the arrays; and a noise value, the Q value, that assess the variation pixel by pixel of their probe cells. The values of these parameters should respect the following general guidelines:

- the percentage of present calls should always be greater than 25% and, among samples belonging to the same group, must be consistent, presenting a range of only 10% of deviations (Expression profiling - best practices for data generation and interpretation in clinical trials. 2004. *Nat Rev Genet* 5, 229-37);
- the value of the background must be typically between 20 and 100 with scanners as the used one although there are no official guidelines of how to assess this value, and the background values should be comparable between the arrays (No authors listed 2, 2004);
- the Q value, a measure of the noise of the obtained images that depends of the noise due to the quality of samples, and of the electrical noise associated with the used scanner that has a constant value, should be also comparable between arrays obtained using the same scanner (No authors listed 2, 2004).

The poly-A RNA controls are added to the studied samples of total RNA to work as exogenous positive controls and monitoring all the process of RNA manipulation until the hybridization on the arrays. These poly-A RNA controls are transcripts of *B. subtilis* genes (*lys*, *phe*, *thr* and *dap*) modified by the insertion of poly-A tails *in vitro* and they are added to the samples to achieve well determined final concentrations according to the manufacturer's protocols. These controls are then amplified and labelled together with the samples. Although they are absent in eukaryotic samples, they are presented by several probe sets on each eukaryotic GeneChip. The obtained expression results for the poly-A RNA controls,

independently from the quality of the starting RNA samples, should typically reproduce the one presented in the Figure 3.5. (No authors listed 3, 2005-2006).

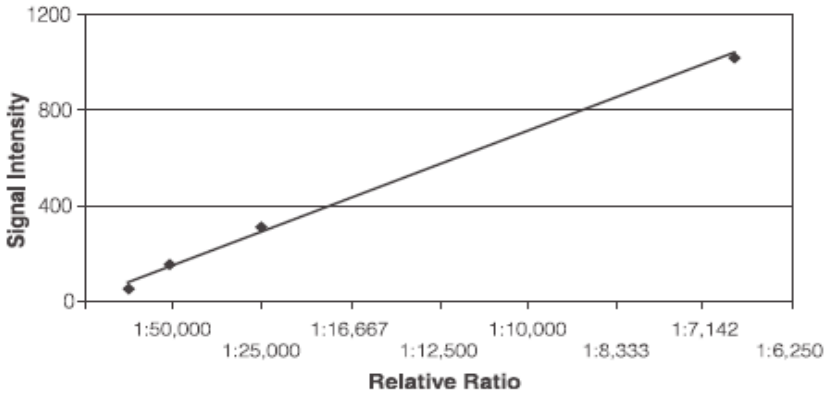


Figure 3.5: Graphical representation of the expected relation between the signal intensity of the hybridization of the Poly-A RNA controls and their initial concentration in the samples. Obtaining a linear relation as presented reflects the success of the steps of amplification and hybridization performed.

The hybridization controls are included in the hybridization cocktail. They are transcripts of the bioB, bioC, bioD and cre genes. The first three are genes from the pathway responsible for the synthesis of biotin in *E. coli* bacteria and *cre* codify for a recombinase from bacteriophage P1. These controls are biotin-labelled and they are also added to achieve well determined final concentrations. The obtained intensity values allow monitoring the efficiency of the hybridization, washing and staining procedures. The *bioB* has a concentration at the limit level of the sensitivity and should be found present in at least 50% of the prepared arrays; the other controls should always be found present showing an increasing intensity depending on their relative concentration according to the manufacturer's protocol (No authors listed 3, 2005-2006). In addition, these controls allow an indirect assessment of the quality of the used RNA samples because the obtained intensities of the controls are independent of the intensities of the sample transcripts that could be degraded or may not have been properly amplified and/or labelled with biotin. In these cases, samples with low quality may have an overall light intensity significantly different, although the intensity of the hybridization controls remains similar (No authors listed 2, 2004).

Probe sets representing actin, glyceraldehyde-3-phosphate dehydrogenase (GAPDH) and ribosomal 18S, are also used to evaluate the quality of RNA samples. The intensity values obtained for the correspondent 3' probe sets should be comparable with the values from the 5' probe sets. The ratio 3'/ 5' should not be greater than 3, otherwise it might prove the existence of total RNA or degraded cRNA, a poor transcription of the cDNA double-chain, or even an inefficient labelling of cRNA with biotin (No authors listed 2, 2004).

3.3.5. Data normalization and statistical analysis

The generated intensity array data (Affymetrix CEL files) were analysed together with their respective CDF (chip description file) file HG-U133_Plus_2.cdf that was downloaded from Affymetrix website (http://www.affymetrix.com/support/support_result.affx) on the Partek software (Partek Incorporated, USA). The background correction, normalization and summarization of the imported CEL files were performed using the robust multi-chip average algorithm. A single value was calculated that translates the relative abundance of each transcript using statistical methods applied to the fluorescence intensities. Only perfect match probe cell values (chips have pairs of probes with perfect match and mismatch single probes that form a probe set for each gene in the array) are used in this method that performs quantile normalization across all the chips in the experiment, and a $\log_2$ transformation of the expression values. If an input value is smaller than 1 the transformed value is set to 0. Partek software is available at IGC's Affymetrix Core Facility too.

In the comparative analyses, the expression level obtained for each gene is contrasted between several samples, thus allowing the identification of genes whose expression is increased or decreased and to quantify these changes. Analysis of variance (ANOVA) was used to identify the differentially expressed genes among different groups of samples such as cases and controls, taking into account known experimental (type, sex and age) and study-design co-variates (geographic origin and scan date). The genes with a fold-change greater than 1.2 were regarded as differentially expressed, and uncorrected p-values smaller or equal to the conventional level of 0.05 were considered statistically significant. To account for multiple testing, we calculated false discovery rate (FDR) using the step up, step down and q-value methods from Partek and applied Bonferroni's correction.

3.3.6. Principal component analysis (PCA) and hierarchical clustering

PCA and hierarchical clustering analysis were performed using the Partek or the dChip 2009 (http://biosun1.harvard.edu/complab/dchip/) software to visualize the relative position of each chip in a low dimensional space and the expression patterns across the samples, respectively. As these tools do not take in account the co-lateral experimental or study design batch effects, in contrast to ANOVA, prior to their use these batch effects were removed using the batch-remover tool of the Partek software. Using this statistical tool, that allows a better visualization and interpretation of the results, we can remove the effect of non-specific co-variates without changing the p-values and fold-changes for the non-removed factors of interest such as the affection status. PCA were performed with the correlation dispersion matrix and normalized eigenvector scaling, and hierarchical clustering was performed with the correlation distance

metric and centroid linkage method.

3.3.7. Gene ontology (GO) and pathway analysis

GO and pathway analysis were executed to extract the maximum of possible biological information from the data obtained. To identify significant gene ontology groups and affected pathways we used the Gene Function Enrichment tool from the dChip 2009 software with a gene p-value threshold of 0.01 and the HG-U133_Plus_2.na28.annot.csv annotation file jointly with the most recent GO structure files; the Onto-Express and the Pathway-Express tools from the Intelligent Systems and Bioinformatics Laboratory (http://vortex.cs.wayne.edu/projects.htm) using its Homo sapiens ontotools database, the KEGG pathways database, the probe set ID as input, and a hypergeomatric distribution; and the Ingenuity Pathway Analysis (IPA) 7.5 software (Ingenuity Systems, USA, http://www.ingenuity.com). IPA is built upon a very large curated and up-to-date database of genes, proteins, functions, interactions, and pathways. 81 metabolic and 202 cell signalling canonical pathways were included in the database at the time of analysis. The IPA "Core Analysis" was performed using the Ingenuity Knowledge Base as the reference set (genes only), using direct and indirect relationships for network analysis, and data from all species, tissues, cell lines and data sources. Canonical pathways analysis identified the pathways from the IPA library of canonical pathways that were most significant to the data set. Genes from the data set that met the fold-change and p-value cut-off of 1.2 and 0.05, respectively, and were associated with a canonical pathway in the Ingenuity Pathways Knowledge Base were considered for the analysis. The significance of the association between the data set and the canonical pathway was measured using the Fischer's exact test to calculate a p-value determining the probability that the association between the genes in the dataset and the canonical pathway is explained by chance alone. IPA software is available at IGC's bioinformatics unit.

3.4. Association studies

The goal of the association studies is to identify specific SNPs or combinations of SNPs that influence the risk for developing stroke. We genotyped in our case-control biobank an appropriate number of SNPs per studied gene: the *PDE4D* and *ALOX5AP* that have been controversially implicated in IS; the genes prioritized based on the convergence of published linkage results and our gene profiling data; and our selected biological candidate genes.

3.4.1. DNA extraction

Whole blood samples were obtained by venipuncture and collected in 7,5 mL EDTA S-Monovette Sarstedt tubes (Sarstedt, Germany). DNA was extracted using the Nucleo Spin Blood XL kit (Macherey-Nagel, Germany) according to the manufacturer's protocols or the phenol/chloroform method (Chomczynski *et al.* 1987). Approximately 300 μg of high-quality high molecular weight DNA was extracted for each sample and it was stored at 4ºC under chloroform (Vance *et al.* 1998). Biosafety level 3 laboratory was used to safely perform the DNA extractions from human whole blood for all IGC samples. DNA concentrations were estimated by measuring the absorbance at 260 nm using a nanodrop ND-100 spectrophotometer and their quality assessed by the A_{260}/A_{280} ratio (between 1.8 and 2.1 for pure DNA). Random samples were subject to QC measures such as gel electrophoresis and restriction digestion to check for degradation.

3.4.2. SNP selection

In general, for the selection of the analysed SNPs for each studied gene, Genotypes of 30 European (CEU) family trios were downloaded from the HapMap Release 21/phase II Jul06 (http://www.hapmap.org/), on NCBI B35 assembly, and tagging SNPs were identified in Haploview v4.0 (Barrett *et al.* 2005) based on their chromosomal location (covering all genes and 10 kb of both flanking regions), with the following options: pairwise mode; $r^2 > 0.8$; and minor allele frequency (MAF) in European population greater than 0.1. Some SNPs with MAF < 0.1 were selected when they were considered relevant to increase the number of genotyped SNPs per gene. Additionally, for some genes were also included SNPs that were associated with IS or other CVDs in previous published reports. A total of 347 SNPs were selected.

3.4.3. Genotyping

SNPs were genotyped in a 384-well format using one of the available high-throughput genotyping platforms of the IGC's genotyping facility. All genotype determinations were performed and blinded to affection status, and an extensive QC was performed using eight HapMap controls of diverse ethnic affiliation (CEU: NA07019 and NA10846; YRI (African): NA18500 and NA18505; CHB (Chinese): NA18524 and 18526; JPT (Japanese):NA18940 and NA18942; http://www.hapmap.org/), non-Mendelian inconsistency check in three large pedigrees (Figure 3.6), sample duplication within and across plate, and no-template controls. SNPs with less than 90% call rate and out of the Hardy-Weinberg equilibrium

(HWE) in the control group were excluded. We also sequenced some samples to confirm their obtained genotypes at the IGC's sequencing facility with multi-capillary gels. Genotyping data of sufficient quality were stored in a secure BC|Gene database.

345 of the SNPs were genotyped using Sequenom´s (Sequenom, USA) iPlex assays (Tables 4.2, 4.4, 4.10 and 4.16, and SNP rs1061170 for *CFH*) following manufacturer's protocol and detected in a Sequenom's MassArray K2 platform. These assays allow for multiplexing up to 40 SNPs. Briefly, after the amplification of multiplex products by PCR, a primer extension process is used to detect sequence differences at the single nucleotide level. A primer named as extension primer is designed for each studied SNP and hybridizes next to the respective SNP site being extended depending upon the template sequence. Consequently, extension products with an allele-specific difference in mass are created allowing the data analysis software to distinguish the SNP alleles. This primer extension is detected by matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. PCR and extension primer sequences (Appendix B: Table B.1) were designed in house using Sequenom's MassArray Assay Design 3.0 software according to the Cambridge reference sequence (http://wwww.mitomap.org/mitoseq.html).

TaqMan Assays-on-Demand and Assays-by-Design (Applied Biosystems, USA) were used for some SNPs that did not work with the Sequenom´s iPlex assay or when it was impossible to designed the respective extension primers. These SNPs were genotyped individually. TaqMan allelic discrimination assay from Applied Biosystems were performed according to manufacturers' instructions using the 7900HT Fast System (Applied Biosystems, USA). Briefly, one TaqMan probe for each allele is used for the allelic discrimination assay. These probes have a 5´-reporter dye and a 3´-quencher dye and a different reporter is used to the detection of each allele. When the probes are intact the 5'-reporters are quenched by the proximity of the attached quencher to the 3´ end of the probes, resulting in the suppression of the reporter fluorescence. One of the probes hybridizes to a target sequence within the PCR amplified region that contains the SNP depending of the allele it present. Then the AmpliTaq Gold enzyme cleaves the TaqMan probe with its 5´-3´ nuclease activity, resulting in increased fluorescence of the reporter as direct consequence of target amplification during PCR. The fluorescence of the reporters is measured allowing the allelic discrimination. SDS 2.3 software was used to analyze de results. The selected TaqMan assays are presented in the next two subtopics.

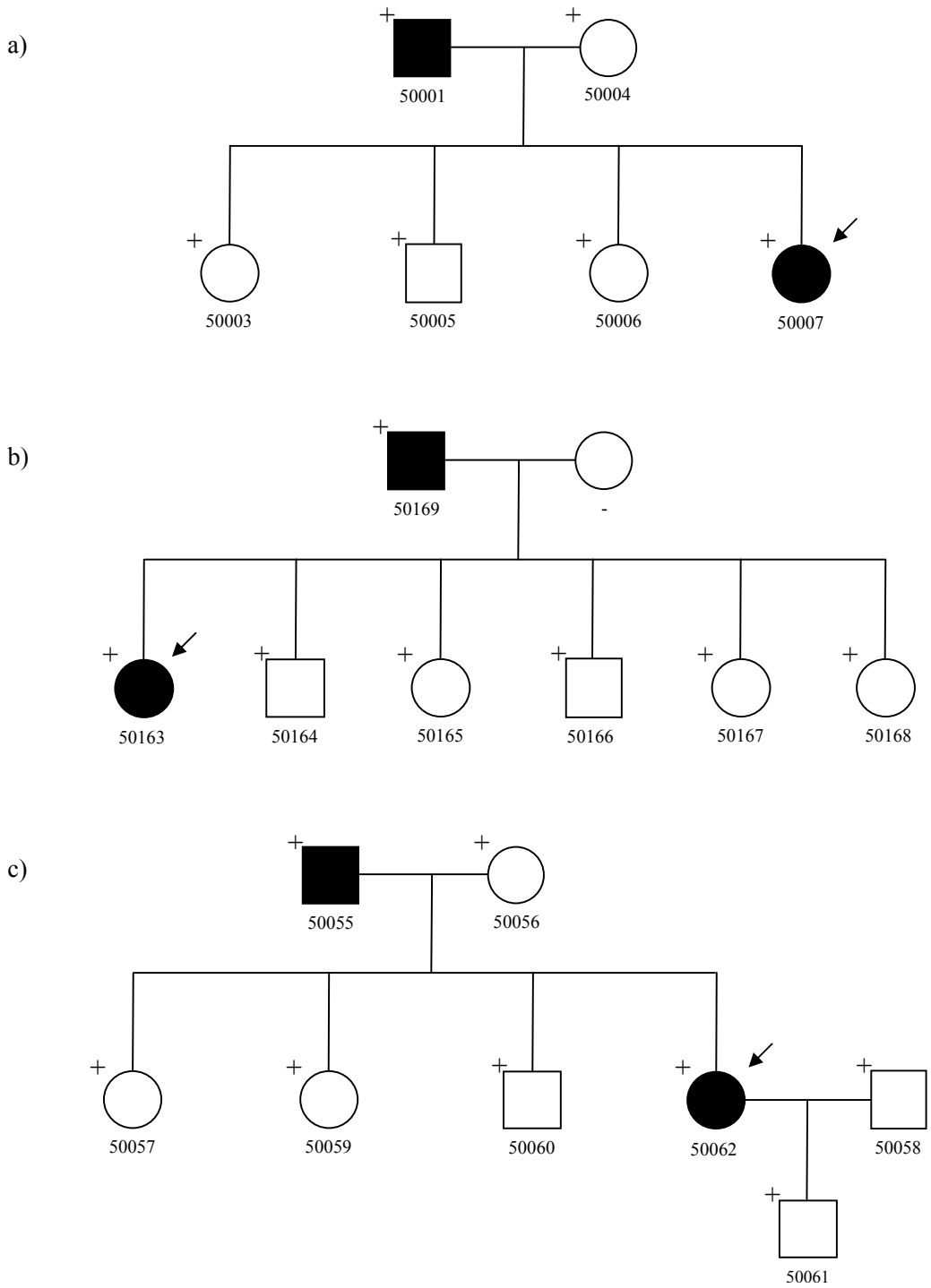


Figure 3.6: Three large pedigrees used to check non-Mendelian inconsistency in the genotyping quality control (QC). Family members marked with a "+"were collected. Males represented by squares, females by circles, and affected participants with stroke are in black. The index patients are indicated with an arrow and were also used as cases in the association studies.

3.4.3.1. *PDE4D and ALOX5AP genes*

We genotyped using TaqMan assays the SNPs rs12188950 and rs12153798 in the *PDE4D*, and SNPs rs17216473, rs10507391 and rs4769874 in the *ALOX5AP* that failed our QC requirements using the iPlex assays. For the SNPs rs10507391 and rs4769874 we used the C__27330284_10 and C__27929140_10 assays-on-demand, respectively, and for the others we ordered the assays-by-design using the forward strand sequence with 100 bp flanking each side of the SNP. All of them have VIC as reporter 1 dye, FAM as reporter 2 dye, and NFQ (non-fluorescent quencher) as quencher of the two reporters.

3.4.3.2. *KALRN gene*

In the *ROPN1-KALRN* region, the SNPs rs12634530, rs12637456 and rs13075202 that failed at least one QC, and the SNP rs1444754 that could not be genotyped using the iPlex assay, were genotyped using the TaqMan allelic discrimination assays. Namely we used respectively the C__11236577_10, C__11236574_10, C___1720000_10 and C___9532531_10 assays-on-demand. As in the obtained results using the iPlex assay, the SNP rs13075202 was out of HWE in the control dataset (p-value < 0.05). As it was genotyped consistently in the two assays (iPlex and assay-on-demand) and passed all other quality checks it was considered as a well genotyped SNP.

3.4.4. Statistical analyses

An unpaired Student's *t* test and a χ^2 test were used to compare quantitative and qualitative clinical and demographic data, respectively, between cases and controls to determine which are the covariates that should be take in account to adjust the results of association.

3.4.4.1. *Genotyped SNPs*

For all genotyped SNPs deviations of genotype distribution from HWE were assessed at each marker in the case and control samples separately using the SNPassoc v.1.4-9 package (González *et al.* 2007) implemented in the R freeware (http://cran.r-project.org/). Since deviations from HWE in affected individuals could indicate that the marker is in LD with the disease locus (Nielsen *et al.* 1998), markers with significant deviations from HWE in the affected sample but not in unaffected sample were noted as possible associated loci and further investigated. Excepting these, only markers in HWE will be considered since HWE is assumed in haplotype reconstruction. LD plots were performed with Haploview v4.0 (Barrett 2005). Using this software, all possible pairwise correlation coefficients r^2 in each gene were

calculated, and LD plots constructed with the r^2 colour scheme ($r^2 = 0$: white; $0 < r^2 < 1$: shades of grey; $r^2 = 1$: black). We computed also Lewontin's D' for each pair of SNPs since it captures different information about the nature of the LD between alleles.

We performed single-marker and haplotype association tests. Association between the risk of IS and specific classes of alleles/genotypes/haplotypes will be analyzed for significance using a standard qui-square test (χ^2). Haplotypes were estimated in Haploview v4.0 using the confidence intervals algorithm. Multivariate logistic regression (log-additive model) with backward elimination of risk factors was performed using SNPassoc to adjust the association analyses with risk for confounding co-variates, namely hypertension, diabetes and ever smoking. Additive, dominant and recessive multivariate logistic regression models were performed for the associated SNPs. The interaction *i* among covariates in regression models was not strong ($-0.2<i<0.2$). Logistic regressions were performed using the SNPassoc v.1.4-9 package. Results were considered statistically significant below the conventional level of 0.05. Odds ratios (ORs) and their associated 95% confidence intervals (CIs) were calculated to assess the relative disease risk conferred by a particular associated allele/genotype. ORs were corrected for covariates in adjusted regression models. Since some of the markers are in LD and the haplotype comparisons are not independent, we did not perform corrections for multiple testing and uncorrected p-values are reported. Sample with more than 50% of missing genotypes were excluded from the analysis.

3.4.4.2. *Imputed SNPs*

For a detailed study of the KALRN gene, ungenotyped SNPs in chromosome 3 were imputed with PLINK v1.04 (Purcell *et al.* 2007) using HapMap data (Release 22, 154783 SNPs in chromosome 3 with MAF greater than 0.01 and genotyping rate greater than 0.95 in the 60 CEU founders). For every imputed SNP, PLINK provides an information content metric INFO, ranging from 0 to 1 (although it can be greater than 1 occasionally). A higher INFO value generally means a better SNP imputation. All imputed SNPs with MAF in controls smaller than 0.05 and with INFO smaller than 0.5 were excluded. For SNPs that have been genotyped, PLINK calculates the concordance rate among observed and imputed genotypes.

3.4.5. Validation of the imputed results

To confirm the imputed results, three imputed SNPs (rs7620580, rs6438833 and rs11712039) significantly associated with IS were fully genotyped in our case-control biobank and tested for association as described previously. These SNPs were genotyped using Sequenom´s (Sequenom, USA) iPlex assays following manufacturer's protocol and detected in a Sequenom's MassArray K2 platform, as explained above. PCR and extension primer sequences were designed in house using Sequenom's MassArray Assay Design 3.0 software (Appendix B: Table B.2).

3.4.6. Replication of the results

Testing of the GC genes was carried out in two stages. On a first stage, a large number of SNPs per gene were tested in all our Portuguese case-control biobank. On a second stage, SNPs with positive association at a low-stringency significance (p-value < 0.05) and some SNPs that constitute positively associated haplotypes were genotyped in the Spanish population. During this project, we established collaborations with Dr. Joan Montaner and colleagues, from the Neurovascular Research Laboratory and Neurovascular Unit of the Universitat Autònoma de Barcelona, that genotyped our selected SNPs in their samples.

The Spanish stroke cases and controls were ascertained and collected as described by Montaner *et al.* (2006). Briefly, five hundred seventy participants with an acute IS involving the vascular territory of the middle cerebral artery were enrolled by the specialized clinicians in H. Vall d'Hebron, H. Santa Creu i Sant Pau, H. Basurto and H. Universitario Dr Josep Trueta since 2003. IS was confirmed by Computed Tomography Scan and/or Magnetic Resonance Imaging. To identify potential mechanisms of cerebral infarction, a set of diagnostic tests was performed that included electrocardiogram, chest radiography, carotid and transcranial doppler, complete blood count and leukocyte differential and blood biochemistry in all patients; some patients also underwent special coagulation tests, transthoracic echocardiography and Holter monitoring. With this information and the neuroimaging data, stroke etiologic subgroups were determined following the TOAST criteria. Three hundred ninety Spanish control participants were recruited in the same hospitals since 2007 and classified free of stroke by direct interview before recruitment. All controls were free of neurological or vascular related diseases (including stroke and familiar history of stroke) and elder than 65 years. Details on socio-economic and demographic characteristics were obtained from all participants by questionnaires, together with information on smoking, dyslipidemia, hypertension, diabetes mellitus and current medication use. Informed written consent was obtained from all subjects, who were all of European Caucasian ancestry, and the local Ethics Committee approved the study.

Genomic DNA was extracted for each subject from peripheral blood stored in EDTA tubes by standard methods. Twenty SNPs of the six GC prioritized genes (*HEMGN*: rs10760017, *GFI1B*: rs633153, *TMTC4*: rs9582406 and rs946845, *TTC7B*: rs2343, rs12147413, rs11629065, rs942738, rs12893100, rs1742100, rs1742098, rs1535321, rs13379124 and rs7154098, *SDC4*: rs6104115, rs6073708, rs4599, rs2251252 and rs2284278, *TUBB1*: rs151348) were selected and investigated. The presented polymorphisms were genotyped in the Spanish samples using the Sequenom's iPlex assays (Sequenom, USA) at the Spanish National Genotyping Centre.

Statistical analyses were performed as described before for the genotyped SNPs in the Portuguese population.

4. RESULTS

4.1. Study subjects

The principal demographic and clinical characteristics and the frequency of the risk factors of the participants that contribute to our full dataset are shown in Table 4.1. From our biobank, the "Stroke IGC" cohort, 130 cases and 82 controls were included, and from "Stroke INSA", "APOEurope", and "6 regions controls" cohorts, that belongs to INSARJ, 435 cases, 71 controls and 367 controls were selected, respectively. All patients were more than 21 and less than 75 years old and all controls were more than 30 and less than 80 years old. As expected, for the known nongenetic risk factors of stroke, male to female ratio, and the frequencies of hypertension, diabetes, ever smoking and ever drinking, were significantly higher in IS patients than in controls. The age-at-examination (AAE) is deliberately significantly different among patients and controls (the control group being older than the case group) to minimize mis-classification biases. In our full dataset, the sex, ever smoking and ever drinking seem to be correlated with correlation factors near 0.5 (data not shown). Consequently, only the hypertension, diabetes and ever smoking were included in the analyses adjusted for covariates.

Table 4.1: Portuguese sample characterization. Principal demographic, clinical and lifestyle characteristics of the IS case-control study sample.

Characteristic	Cases	Controls	p-value[a]
N	565	520	
Sex (n, % male)	361 (63.9 %)	238 (45.8 %)	$< 10^{-4}$
Age-at-examination (mean ± SD, years)	52.4 ± 9.3	62.9 ± 6.9	$< 10^{-4}$
Age-at-onset (mean ± SD, years)	51.7 ± 9.5	-	-
Risk factors (n/N[b], %)			
Hypertension (> 140-85 mmHg)	289/505 (57.2 %)	192/511 (37.6 %)	$< 10^{-4}$
Hypercholesterolemia (> 200 mg/dL)	328/526 (62.4 %)	326/518 (62.9 %)	0.847
Hypertriglycemia (> 200 mg/dL)	42/229 (18.3 %)	68/438 (15.5 %)	0.352
Diabetes	95/538 (17.7 %)	58/499 (11.6 %)	0.006
Ever smoking	272/556 (48.9 %)	146/510 (28.6 %)	$< 10^{-4}$
Ever drinking	326/558 (58.4 %)	218/503 (43.3 %)	$< 10^{-4}$

[a]p-value of an unpaired Student's t test or a χ^2 test for a quantitative and qualitative data, respectively

[b]Number of individuals for whom the data was available

The TOAST stroke subtype classification system has been systematically collected only in the cases (130 cases) of the "Stroke IGC" cohort. Since these only represent less than a quarter of the total number of the cases, stratified analysis of the results by stroke subtype cannot be performed. The "Stroke IGC" cohort includes 20 large artery atherosclerosis, 14 cardioembolisms, 28 small-vessel occlusions, 6 IS of other determined aetiology, 25 IS of other undetermined aetiology with incomplete evaluation, 34 IS of

other undetermined aetiology with negative evaluation, 1 IS of other undetermined aetiology with two or more causes identified, and 2 cases with missing data.

4.2. *PDE4D* and *ALOX5AP* association studies

Since *PDE4D* and *ALOX5AP* have been in recent years controversially implicated in the risk of IS, we assessed their association with IS in our Portuguese cohort. We genotyped in our complete dataset 67 SNPs on a 740kb region in the 5'end of *PDE4D* most strongly implicated in IS (chr5: 59140000 to 59900546 bp on NCBI B35 according to the original paper; Gretarsdottir *et al.* 2003), and 24 SNPs in *ALOX5AP* and 10kb of both flanking regions (chr13: 30197000 to 30247000 bp on NCBI B35). These SNPs have either been previously found associated with the risk of stroke (Table 1.2 and 1.3) or constitute haplotype tagging SNPs from HapMap project (Table 4.2). Ten additional SNPs of the *PDE4D* and three additional SNPs of the *ALOX5AP* in almost complete LD with ten and three of the selected ones, respectively, were also genotyped to substitute SNPs with bad genotyping results according to our QC requirements or to verify the reproducibility of the association results in our Portuguese sample of SNPs described in high LD in the HapMap project (Table 4.2). For the subsequent analysis of the genotyping results, we included 565 affected individuals and 518 controls, as well as all SNPs that passed all our QC requirements (Table 4.2).

Table 4.2: Characterization of the investigated SNPs in the PED4D and ALOX5AP genes. For each SNP it is indicated its chromosomic position, the possible alleles, the minor allele frequency (MAF) in the CEU population, and the genotyping platform that was used. The SNP name in the original report is indicated when available. SNPs that did not pass our quality controls requirements are in grey.

Gene	Chr.	SNP ID	Position (NCBI build 35)	Allele 1:2	MAF	Genotyping assay	SNP name in original report/ Comments
PDE4D	5q12	rs2963821	59.144.658	A:C	0,500	I	
		rs2963820	59.144.771	A:G	0,250	I	
		rs2938784	59.147.106	C:T	0,367	I	
		rs2938787	59.149.660	T:C	0,258	I	to replace rs2963820
		rs11951422	59.176.944	C:T	0,246	I	
		rs159608	59.209.860	A:G	0,350	I	
		rs168883	59.214.508	T:C	0,467	I	
		rs159616	59.227.369	T:C	0,383	I	
		rs7710463	59.237.228	T:C	0,133	I	
		rs6879326	59.244.059	C:T	0,492	I	
		rs1435077	59.264.373	T:C	0,342	I	
		rs1529842	59.265.936	C:T	0,175	I	
		rs10066510	59.270.315	C:A	0,108	I	
		rs6889660	59.308.563	C:T	0,424	I	

rs17315957	59,314,726	C:T	0.288	I	to replace rs12518928
rs12518928	59,328,752	G:A	0.292	I	
rs1533019	59,343,570	T:C	0.175	I	
rs16890455	59,367,123	C:T	0.192	I	
rs16890459	59,371,867	T:C	0.192	I	to replace rs16890455
rs13179619	59,401,178	C:T	0.458	I	
rs7732249	59,402,026	T:C	0.109	I	
rs2136203	59,418,081	T:C	0.367	I	
rs1396476	59,432,399	T:G	0.172	I	SNP 89
rs2910831	59,497,502	T:C	0.425	I	
rs2910829	59,505,656	A:G	0.483	I	SNP 87
rs966221	59,538,277	G:A	0.408	I	SNP 83
rs2898269	59,541,250	A:T	0.117	I	
rs2409741	59,543,730	C:T	0.241	I	
rs11745887	59,545,702	C:G	0.133	I	
rs2962964	59,553,798	G:C	0.224	I	to replace rs2409741
rs6889641	59,554,825	G:A	0.138	I	to replace rs11745887
rs7442640	59,556,202	A:G	0.203	I	
rs4604150	59,578,681	T:C	0.383	I	
rs11739760	59,589,499	T:G	0.212	I	
rs12658881	59,594,902	C:T	0.202	I	
rs12523473	59,601,876	G:A	0.500	I	
rs12515974	59,605,280	T:C	0.283	I	
rs10471476	59,605,397	A:C	0.408	I	
rs10471477	59,605,488	G:C	0.325	I	
rs4700371	59,623,477	C:T	0.358	I	to replace rs12522161
rs6449467	59,663,890	A:G	0.225	I	
rs7733705	59,671,110	C:T	0.175	I	
rs12522161	59,680,542	T:C	0.358	I	
rs4580739	59,691,437	G:A	0.305	I	to replace rs1423246
rs10514895	59,694,833	G:C	0.100	I	
rs1423246	59,697,213	A:G	0.300	I	
rs12189147	59,721,330	T:C	0.275	I	
rs6887545	59,742,557	G:A	0.179	I	
rs37707	59,750,596	A:T	0.306	I	
rs13166292	59,752,307	T:C	0.330	I	
rs37702	59,753,932	G:A	0.368	I	
rs37684	59,769,122	T:G	0.350	I	
rs11745887	59,545,702	C:G	0.133	I	to replace rs6889641
rs2962964	59,553,798	G:C	0.224	I	to replace rs2409741
rs6889641	59,554,825	G:A	0.138	I	
rs7442640	59,556,202	A:G	0.203	I	
rs4604150	59,578,681	T:C	0.383	I	
rs11739760	59,589,499	T:G	0.212	I	
rs12658881	59,594,902	C:T	0.202	I	
rs12523473	59,601,876	G:A	0.500	I	
rs12515974	59,605,280	T:C	0.283	I	
rs10471476	59,605,397	A:C	0.408	I	
rs10471477	59,605,488	G:C	0.325	I	
rs4700371	59,623,477	C:T	0.358	I	to replace rs12522161

		rs456009	59,832,491	G:A	0.295	I	SNP 32
		rs35387	59,838,426	G:C	0.424	I	
		rs35386	59,838,540	C:T	0.425	I	to replace rs35387
		rs40512	59,841,224	T:C	0.425	I	
		rs35385	59,842,938	T:A	0.442	I	
		rs35384	59,856,627	G:A	0.275	I	
		rs35383	59,856,990	C:T	0.275	I	to replace rs35384
		rs35382	59,857,300	G:A	0.383	I	
		rs7727343	59,858,955	T:C	0.175	I	
		rs10051720	59,873,233	G:A	0.316	I	
		rs27565	59,873,348	T:C	0.433	I	
		rs10939851	59,877,118	A:G	0.133	I	
		rs152341	59,881,720	T:A	0.450	I	
ALOX5AP	13q12	rs17222814	30.197.553	G:A	0,116	I	SG13S25 (HapA)
		rs4769870	30.201.511	C:T	0,100	I	
		rs17216473	30.201.965	G:A	0,124	I and T	SG13S377 (HapB)
		rs4076128	30.203.143	A:G	0,250	I	
		rs4769055	30.207.830	C:A	0,267	I	
		rs9579645	30.208.506	A:C	0,089	I	to replace rs4769870, MAF<0.1
		rs10507391	30.210.096	T:A	0,325	I and T	SG13S114 (HapA/B)
		rs12429692	30.210.178	A:T	0,290	I	
		rs3885907	30.212.455	A:C	0,475	I	
		rs17612031	30.216.068	T:G	0,103	I	
		rs9671182	30.219.138	G:C	0,425	I	
		rs9671124	30.222.253	C:T	0,394	I	
		rs4769874	30,224,441	G:A	0.033	I and T	SG13S89 (HapA), MAF<0.1
		rs9579648	30,225,032	G:C	0.183	I	
		rs17074975	30,229,016	A:T	0.034	I	to replace rs4769874, MAF<0.1
		rs9551963	30,230,547	C:A	0.455	I	SG13S32 (HapA)
		rs9551964	30,231,517	A:T	0.276	I	
		rs9315050	30,234,045	A:G	0.050	I	SG13S41 (HapB), MAF<0.1
		rs3935644	30,235,640	G:A	0.258	I	
		rs17222842	30,238,117	G:A	0.093	I	SG13S35 (HapB), MAF<0.1
		rs4445746	30,239,435	G:A	0.217	I	
		rs9578200	30,241,864	C:T	0.186	I	
		rs9579653	30,242,712	T:C	0.158	I	
		rs4491352	30,244,983	A:C	0.375	I	
		rs4769062	30,245,221	G:A	0.142	I	
		rs4769063	30,245,274	C:T	0.121	I	to replace rs4769062
		rs4238139	30,245,589	A:G	0.217	I	

MAF: Minor Alele Frequency, I: iPlex, T: TaqMan

In *PDE4D*, the SNP rs17315957 was found modestly associated with IS risk (0.034 < p-value < 0.041) in the allelic test as well as in the unadjusted and adjusted (log-additive model) genotypic tests that were performed (Figure 4.1; Appendix C: Table C.1). On the other hand, the SNP rs7442640 shows only a good association (p-value = 0.006) with IS risk when genotypic tests were adjusted for covariates (Figure 4.1; Appendix C: Table C.1). This association remains rather significant using the dominant genetic model

(p-value = 0.007; Table 4.3). Conversely, in *ALOX5AP*, we found only a modest evidence of association with IS risk (0.017< p-value < 0.025) for the downstream SNP rs4491352 for allelic tests and for unadjusted (log-additive model) genotypic tests (Figure 4.2; Appendix C: Table C.1). This SNP also belongs to the associated haplotype with IS risk (p-value = 0.013), the haplotype CTAGCA defined by block 5 (rs9578200-rs9579653-rs4491352-rs4769062-rs4769063-rs4238139), that we found in *ALOX5AP* (Figure 4.2; Appendix C: Table C.2). Curiously, the rs4491352 minor allele (less frequent allele), that seems to confer risk of IS (OR [95% CI] = 1.23 [1.08 - 1.41]; Table 4.3), is the C allele, and the allele that is present in the associated haplotype, that is more frequent in controls than in cases (0.423 vs. 0.370), is the A allele (Appendix C: Table C.2).

It is interesting to note that the associated SNP rs17315957 in the *PDE4D* was one of the additional SNPs that were genotyped to replace the initially selected haplotype tagging SNP rs12518928 (Table 4.2) that was not associated with the risk of IS in our analysis (Appendix C: Table C.1). According to the HapMap project these two SNPs are in almost complete LD with $r^2 = 0.98$, and their MAFs are 0.288 and 0.292 (Table 4.2), respectively, for the European population. In our control group their calculated MAFs were 0.230 and 0.223 (Appendix C: Table C.1). Despite these calculated values are slightly smaller than the expected ones, they continue to be very similar among them, suggesting that the marginal associations verified for the SNP rs17315957 could probably disappear with the enlargement of our biobank. The SNPs rs2938787 and rs789395 that were added to replace the rs2963820 and rs371424, respectively, in the *PDE4D*, present also very similar calculated MAFs in the control group among them (0.214, 0.313, 0.233 and 0.313 for the SNPs in the same order; Appendix C: Table C.1) and with the presented ones in the HapMap project (0.258, 0.331, 0.250 and 0.330 for the SNPs in the same order; Table 4.2). These added SNPs had also very similar p-values in the performed tests (Appendix C: Table C.1) comparing with the SNPs they were used to replace. The same was verified for the SNP rs4769063 (calculated MAF = 0.133, HapMap MAF = 0.121) in the *ALOX5AP* that was selected to replace the rs4769062 (calculated MAF = 0.135, HapMap MAF = 0.142; Table 4.2; Appendix C: Table C.1). These evidences support our obtained results and validate that it is not wrong to use the HapMap project to estimate the frequencies and to selected haplotype tagging SNP for the Portuguese population.

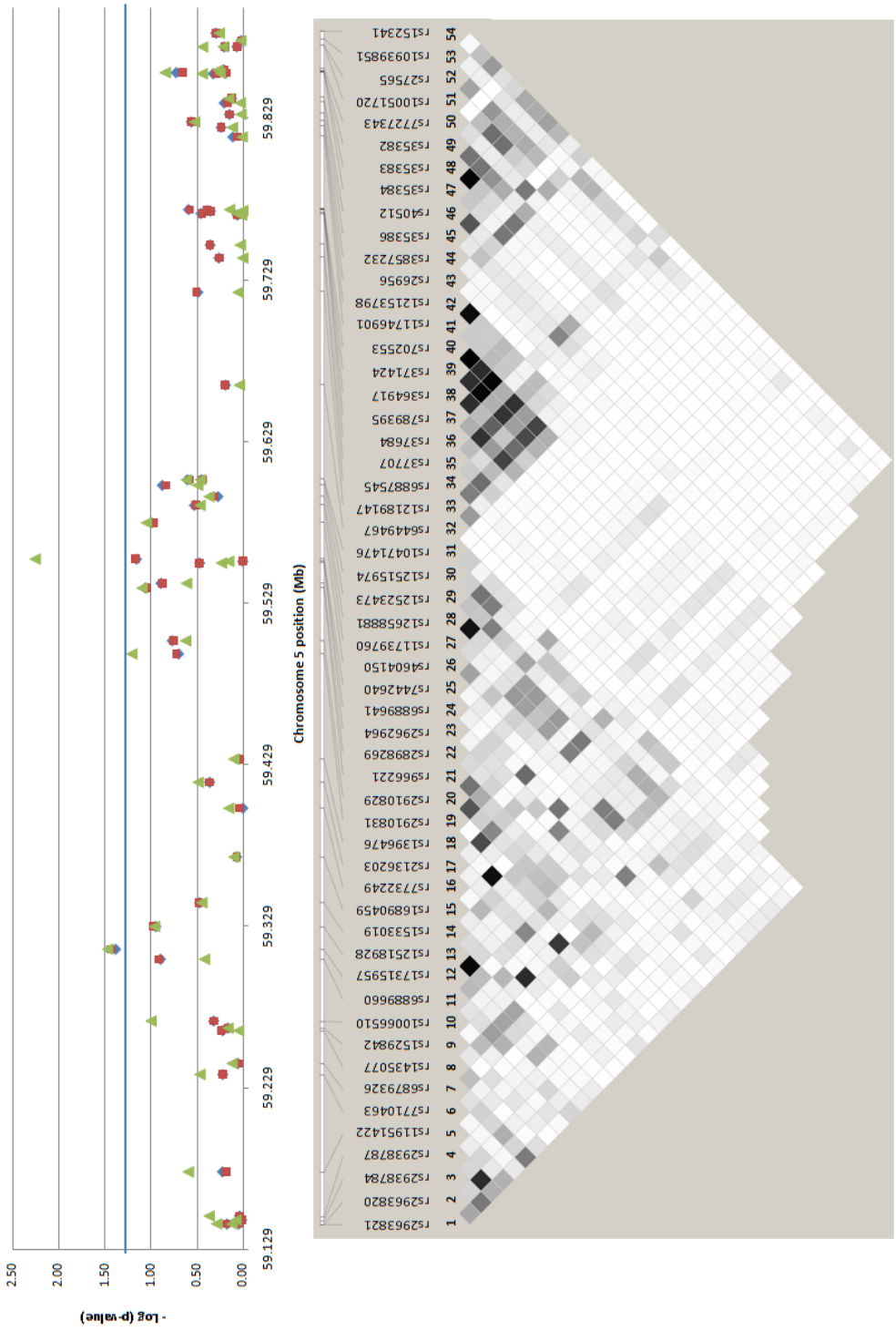


Figure 4.1: Association results and pairwise linkage disequilibrium (LD) for genotyped polymorphisms in *PDE4D*. Allelic (blue diamonds) as well as crude (red squares) and adjusted (green triangles) genotypic (log-additive model) association results are shown. Results were considered significant over the blue line (p-value = 0.05). In the LD plot, the white-to-black gradient shading within each diamond represents the magnitude of LD using the pairwise statistic r^2.

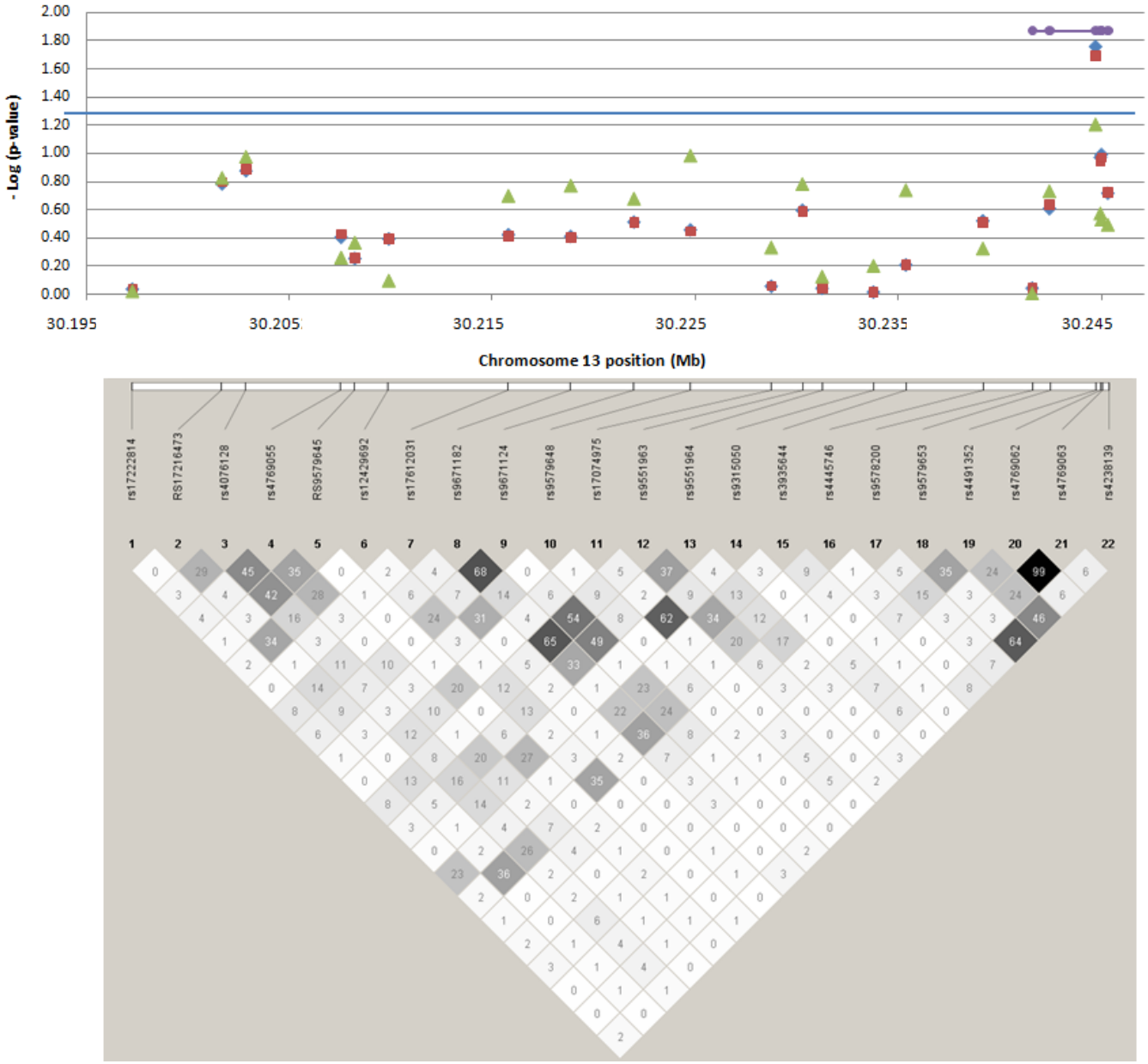


Figure 4.2: Association results and pairwise linkage disequilibrium (LD) for genotyped polymorphisms in *ALOX5AP*. See legend in Figure 4.1. Associated haplotypes are presented in purple joined circles. The value of the pairwise statistic r^2 is shown inside the diamonds of the LD plot.

None of the significant association results survive to Bonferroni's correction for multiple testing, and none of the genotyped SNPs previously found to be associated with the risk of stroke (Table 1.2 and 1.3; Table 4.2) or SNPs selected to replace them (the SNP rs17074975 in ALOX5AP) were significantly associated with the IS risk for any test performed in our Portuguese biobank.

Table 4.3: Detailed association results for associated SNPs in *PDE4D* and *ALOX5AP*. The allelic and genotype frequencies in cases and controls as well as the association results for the allelic and genotype association tests are shown here. Unadjusted (without co-variates) and adjusted (for hypertension, diabetes and ever smoking) genotype association testing was performed using different models (dominant, recessive and log-additive). Significant p-values are bolded and the respective odds ratio and 95% confidence intervals are indicated.

Gene	SNP ID/ Model	Unadjusted						Adjusted					
		Cases		Controls		OR [95% CI]	p-value	Cases		Controls		OR [95% CI]	p-value
		N	%	N	%			N	%	N	%		
PDE4D	**rs17315957**												
	Alleles												
	C	823	73.2	795	77.0	1.00	**0.041**						
	T	301	26.8	237	23.0	1.23 [1.03-1.45]							
	Genotypes												
	Dominant												
	C/C	297	52.8	303	58.7	1.00	0.052	259	53.1	285	59.6	1.00	0.068
	C/T-T/T	265	47.2	213	41.3	1.27 [1.00-1.62]		229	46.9	193	40.4	1.28 [0.98-1.67]	
	Recessive												
	C/C-C/T	526	93.6	492	95.3	1.00	0.208	456	93.4	456	95.4	1.00	0.105
	T/T	36	6.4	24	4.7	1.40 [0.83-2.39]		32	6.6	22	4.6	1.6 [0.90-2.84]	
	log-Additive												
	0,1,2					1.24 [1.01-1.51]	**0.037**					1.27 [1.02-1.58]	**0.034**
	rs7442640												
	Alleles												
	A	906	81.5	858	84.4	1.00	0.069						
	G	206	18.5	158	15.6	1.23 [1.00-1.52]							
	Genotypes												
	Dominant												
	A/A	368	66.2	363	71.5	1.00	0.064	315	65.2	342	72.6	1.00	**0.007**
	A/G-G/G	188	33.8	145	28.5	1.28 [0.99-1.66]		168	34.8	129	27.4	1.49 [1.11-1.99]	
	Recessive												
	A/A-A/G	538	96.8	495	97.4	1.00	0.510	465	96.3	461	97.9	1.00	0.184
	G/G	18	3.2	13	2.6	1.27 [0.62-2.63]		18	3.7	10	2.1	1.73 [0.76-3.91]	
	log-Additive												
	0,1,2					1.24 [0.98-1.55]	0.068					1.43 [1.11-1.84]	**0.006**
ALOX5AP	**rs4491352**												
	Alleles												
	A	626	56.2	625	61.3	1.00	**0.017**						
	C	488	43.8	395	38.7	1.23 [1.08-1.41]							
	Genotypes												
	Dominant												
	A/A	186	33.4	195	38.2	1.00	0.099	165	34.1	180	38.1	1.00	0.107
	C/A-C/C	371	66.6	315	61.8	1.23 [0.96-1.59]		319	65.9	292	61.9	1.26 [0.95-1.66]	
	Recessive												
	A/A-C/A	440	79.0	430	84.3	1.00	**0.025**	388	80.2	396	83.9	1.00	0.149
	C/C	117	21.0	80	15.7	1.43 [1.04-1.96]		96	19.8	76	16.1	1.29 [0.91-1.83]	
	log-Additive												
	0,1,2					1.22 [1.03-1.45]	**0.020**					1.19 [0.99-1.44]	0.062

OR: Odds ratio; CI: Confidence interval

4.3. Association studies in biological candidate genes

4.3.1. KALRN gene

Since several recent studies have implicated *KALRN* variants with susceptibility to cardiovascular and metabolic phenotypes which may share susceptibility genes with stroke, we also assessed the association of SNPs in the *ROPN1-KALRN* region with IS in our Portuguese population (Krug *et al.* 2010). We first tested the association with IS of the twelve SNPs originally found associated with CAD in the CATHGEN initial dataset (SNPs 2, 3, 4, 7, 8, 9, 12, 14, 20, 22, 31, and 37 in Wang *et al.* 2007; Table 4.4), which includes the SNP rs4234218 (SNP 31 in Wang *et al.* 2007) associated with cardiovascular risk, T2D and MS (Rudock *et al.* 2008). These polymorphisms are located in *ROPN1*, *ROPN1-KALRN* intergenic region, and in the 5' region of the *KALRN* (Table 4.4). In our dataset, all of these polymorphisms are also haplotype tagging SNPs, with the exception of the SNPs rs7613868 and rs12637456 (SNPs 7 and 9, respectively, in Wang *et al.* 2007) which are in almost complete LD ($r^2 = 0.97$, Figure 4.3). We included in this analysis the 565 affected individuals and 517 controls selected from our biobank.

Table 4.4: Characterization of the investigated SNPs in the *ROPN1-KALRN* region. For each SNP it is indicated its chromosomic position, the possible alleles, the minor allele frequency (MAF) in the CEU population, and the genotyping platform that was used. The SNP name in the original report (Wang *et al.* 2007) is indicated for the twelve SNPs that were first genotyped. SNPs that did not pass our quality controls requirements are in grey.

Gene	Chr.	SNP ID	Position (NCBI build 35)	Allele 1:2	MAF	Genotyping assay	SNP name in Wang *et al.*
ROPN1-KALRN	3q13	rs7633408	125.166.534	C:G	0,133	I	
		rs12695434	125.168.847	G:A	0,225	I	
ROPN1		rs6810298	125.177.665	A:G	0,275	I	SNP 2
ROPN1		rs6774735	125.177.888	G:T	0,117	I	
ROPN1		rs17376453	125.179.139	G:C	0,142	I	SNP 3
ROPN1		rs2280422	125.179.620	G:A	0,325	I	
		rs4499545*	125.188.532	G:A	0,133	I	SNP 4
		rs2332719	125.195.656	A:G	0,208	I	
		rs7613868	125.224.719	C:T	0,217	I	SNP 7
		rs12634530	125.227.160	C:T	0,125	I and T	SNP 8
		rs12637456	125.227.353	T:A	0,217	I and T	SNP 9
		rs1317671	125.245.276	T:C	0,217	I	
		rs9289231	125.256.768	T:G	0,092	I	SNP 12

KALRN	rs13075202	125.304.977	A:G	0,217	I and T**	SNP 14
KALRN	rs4678085	125.314.057	G:A	0,183	I	
KALRN	rs4678086	125.319.942	G:C	0,144	I	
KALRN	rs1950091	125.383.007	G:A	0,150	I	
KALRN	rs1444760	125.402.970	G:A	0,283	I	
KALRN	rs1444768	125.406.612	A:G	0,342	I	SNP 20
KALRN	rs1444766	125.407.961	A:G	0,259	I	
KALRN	rs1444754	125.420.907	T:C	0,433	T	SNP 22
KALRN	rs1373609	125.432.219	T:C	0,475	I	
KALRN	rs17286604	125.434.907	C:T	0,458	I	
KALRN	rs2141664	125.437.808	A:G	0,133	I	
KALRN	rs1158012	125.440.085	G:A	0,417	I	
KALRN	rs4608634	125.443.748	G:C	0,383	I	
KALRN	rs4234218	125.443.900	G:C	0,450	I	SNP 31
KALRN	rs4608635	125,444,186	G:C	0.297	I	
KALRN	rs6781700	125,450,642	T:C	0.100	I	
KALRN	rs1444763	125,457,660	A:G	0.283	I	
KALRN	rs17221479	125,460,168	G:A	0.142	I	
KALRN	rs11929003	125,462,011	A:G	0.425	I	SNP 37
KALRN	rs9880957	125,462,341	A:G	0.425	I	
KALRN	rs13064819	125,468,805	G:C	0.219	I	
KALRN	rs11712619	125,502,492	C:T	0.417	I	
KALRN	rs17377867	125,504,693	G:A	0.108	I	
KALRN	rs9838361	125,517,439	G:T	0.283	I	
KALRN	rs6784664	125,524,089	C:A	0.367	I	
KALRN	rs3821525	125,537,714	A:G	0.317	I	
KALRN	rs11708466	125,544,276	T:C	0.258	I	

*rs4499545 is the current name for SNP rs7434266, MAF: Minor Alele Frequency, I: iPlex, T: TaqMan

**Out of HWE in the control dataset (p-value<0.05), but it was genotyped consistently in two assays

The SNP rs9289231, which has been associated with early-onset CAD and atherosclerosis burden in human aortas, and the rs4234218 were not associated with IS in any test performed (Figure 4.3, Appendix C: Table C.3). However, we found that a cluster of SNPs in low LD (pairwise $r^2 \leq 0.56$) in the *ROPN1-KALRN* intergenic region (rs4499545, rs7613868 / rs12637456, and rs12634530) was modestly associated with IS ($0.024 < \text{p-value} < 0.043$) in allelic and unadjusted genotypic (log-additive model) tests (Figure 4.3, Appendix C: Table C.3). The association of SNP rs4499545 remained significant (log-additive model p-value = 0.028) after adjustment for covariates (Figure 4.3, Appendix C: Table C.3). As expected the results obtained for the SNPs rs7613868 and rs12637456 were very similar supporting the quality of the used procedures and of the performed analysis.

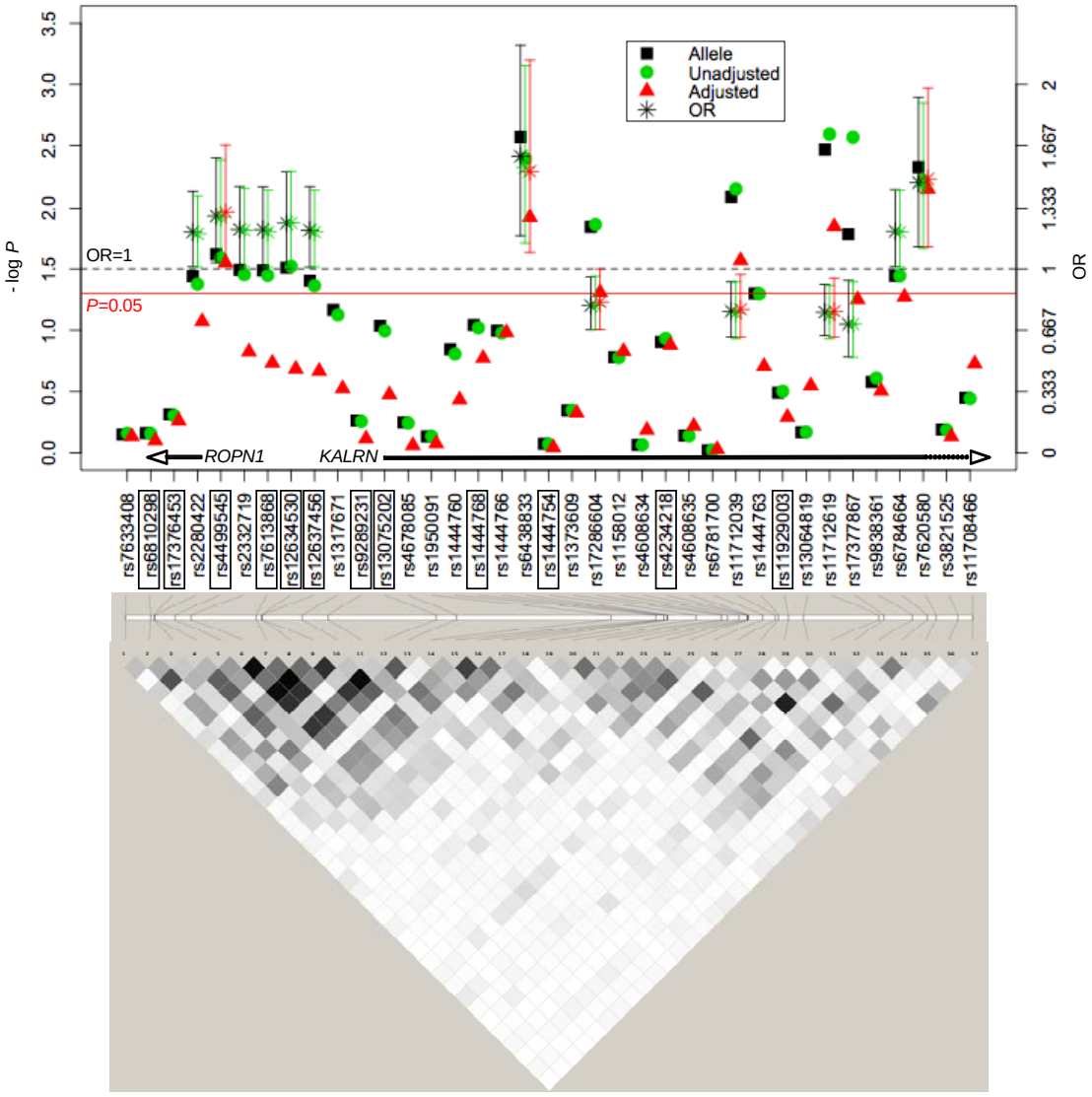


Figure 4.3: Association results and pairwise linkage disequilibrium (LD) for genotyped polymorphisms in the *ROPN1-KALRN* intergenic region. Allelic (black squares) as well as crude (green discs) and adjusted (red triangles) genotypic (log-additive model) association results are shown. Odds ratios (ORs) and 95% confidence intervals (CIs) are depicted with stars and bars for significantly associated (p-value < 0.05) polymorphisms. In the LD plot, the white-to-black gradient shading within each diamond represents the magnitude of LD using the pairwise LD statistic r^2 (from lower to higher values, respectively). The positions of the SNPs relative to the *ROPN1* and *KALRN* genes are indicated. Boxes surrounding SNP names indicate the twelve polymorphisms that we first genotyped.

Given the evidence for association of genetic variants in the *ROPN1-KALRN* region with IS, although with a different variant than previously reported to be associated with cardiovascular risk, we

decided to further investigate this region by analyzing additional SNPs. We genotyped twenty eight additional haplotype tagging SNPs in the *ROPN1* region, *ROPN1-KALRN* intergenic region and 5' region of *KALRN* most strongly implicated in IS (chr3: 125166534 to 125544276 bp on NCBI build 35). Since the *KALRN* gene is extremely large, we chose to concentrate our efforts on the *KALRN* region with the reported associations (Table 4.4).

Two newly analyzed SNPs in the *ROPN1-KALRN* intergenic region LD block (rs2280422 and rs2332719) and two SNPs in *KALRN* (rs17377867 and rs6784664) were associated ($0.003 < \text{p-value} < 0.042$) in allelic and unadjusted (log-additive model) tests, but these associations became marginal ($0.05 < \text{p-value} < 0.15$) when adjusted for co-variates (Figure 4.3, Appendix C: Table C.3). Inclusion of any combination of two of these co-variates had the same overall effect, but it is not always the same polymorphisms that become marginally significant, suggesting that this may be mostly a power issue and not a biologically-relevant phenomenon. On the other hand, two new SNPs in *KALRN* (rs17286604 and rs11712619) demonstrated association ($0.003 < \text{p-value} < 0.049$) in all tests performed (Figure 4.3, Appendix C: Table C.3).

Table 4.5 presents more detailed association results for the ten genotyped SNPs demonstrating an association with IS in the *ROPN1-KALRN* intergenic region. Similarly to the rs4234218 association with cardiovascular risk, T2D and MS (Rudock *et al.* 2008), the most significant results for the three SNPs associated in all tests performed (rs4499545, rs17286604 and rs11712619) were under the dominant genetic model ($0.003 < \text{p-value} < 0.030$; Table 4.5). These polymorphisms are in very low LD (all pairwise $r^2 < 0.34$; Figure 4.3). SNP rs4499545 remains significantly associated when the other two markers are used as co-variates, while the association subsides for rs11712619 when adjusted for rs17286604. These data suggest at least two independent lines of evidence for association of KALRN with IS. The higher association for the dominant genetic model is also verified for the SNPs rs2332719, rs7613868 and rs12634530 that were associated with IS in unadjusted genotypic tests ($0.023 < \text{p-value} < 0.049$) but not in the adjusted tests for covariates (Table 4.5). Conversely, for the SNP rs17377867 the higher association (p-value = 0.003) is obtained for the recessive genetic model and, in this model, persists in the adjusted genotypic test (p-value = 0.007; Table 4.5).

The majority of the ten associated SNPs are also part of the associated haplotypes obtained in the *ROPN1-KALRN* region (Appendix C: Table C.4). The haplotypes GG and AA defined by block 2 (rs2280422-rs4499545) are modestly associated with IS (GG, p-value = 0.017; AA, p-value = 0.032), being the combination that contains the minor alleles of the two SNPs (AA, both with OR > 1; Figure 4.3 and Table 4.5) most frequent in the cases (0.211 vs. 0.174), and the other, most frequent in the controls (0.579 vs. 0.528). The haplotype ACCTTTA defined by block 3 (rs2332719-rs7613868-rs12634530-rs12637456-rs1317671-rs9289231-rs13075202), on its side, is the unique associated haplotype in this block (p-value = 0.023) and the unique that has the sequence ACCT for the first four SNPs that are

associated in our results. The alleles in this sequence are the complementary alleles of the minor alleles of these SNPs that seem to confer risk for IS (OR > 1; Figure 4.3 and Table 4.5). Concordantly, the referred haplotype is more frequent in controls than in cases (0.659 vs. 0.611). The haplotypes CT and CC defined by block 6 (rs1373609-rs17286604), and GCGGA and GTAGC defined by block 9 (rs13064819-rs11712619-rs17377867-rs9838361-rs6784664) are also associated (CT, p-value = 0.016; CC, p-value = 0.042; GCGGA, p-value = 0.006; GTAGC, p-value = 0.029). For these haplotypes, the associated SNPs rs17286604, rs11712619 and rs17377867 have minor alleles that seem to confer protection against IS (OR < 1; Figure 4.3 and Table 4.5) and consequently the haplotypes that harbour that alleles are most frequent in controls than in cases (e.g. 0.364 vs. 0.315 for haplotype CT defined by block 6), and the ones that harbour the major alleles are most frequent in cases than in controls (e.g. 0.152 vs. 0.122 for haplotype CC defined by block 6). The SNP rs6784664, for other side, seems to confer risk for IS (OR [95% CI] = 1.20 [1.01 - 1.43]; Figure 4.3 and Table 4.5) and consequently the associated haplotype GCGGA defined by block 9 that is most frequent in cases than in controls (0.135 vs. 0.097) harbours its minor allele (A), and the haplotype GTAGC defined by block 9 that is most frequent in controls than in cases (0.144 vs. 0.077) harbours the major allele (C). Finally, haplotypes TA and TG defined by block 8 (rs6781700-rs1444763) were obtained associated without containing any of the associated SNPs to IS risk (p-value = 0.030 and 0.014, respectively).

Table 4.5: Detailed association results for associated SNPs in *KALRN*. See legend in table 4.3.

Gene	SNP ID/ Model	Unadjusted						Adjusted					
		Cases		Controls		OR [95% CI]	p-value	Cases		Controls		OR [95% CI]	p-value
		N	%	N	%			N	%	N	%		
KALRN	**rs2280422**												
	Alleles												
	G	599	53.7	598	58.2	1.00	**0.036**						
	A	517	46.3	430	41.8	1.20 [1.01-1.42]							
	Genotypes												
	Dominant												
	G/G	169	30.3	182	35.4	1.00	0.074	149	30.8	168	35.3	1.00	0.168
	G/A-A/A	389	69.7	332	64.6	1.26 [0.98-1.63]		335	69.2	308	64.7	1.22 [0.92-1.62]	
	Recessive												
	G/G-G/A	430	77.1	416	80.9	1.00	0.120	374	77.3	387	81.3	1.00	0.141
	A/A	128	22.9	98	19.1	1.26 [0.94-1.70]		110	22.7	89	18.7	1.28 [0.92-1.77]	
	log-Additive												
	0,1,2					1.19 [1.01-1.40]	**0.042**					1.18 [0.98-1.41]	0.085

rs4499545												
Alleles												
G	846	78.6	819	82.6	1.00	**0.024**						
A	230	21.4	173	17.4	1.29 [1.03-1.60]							
Genotypes												
Dominant												
G/G	331	61.5	344	69.4	1.00	**0.008**	293	62.6	320	69.7	1.00	**0.016**
A/G-A/A	207	38.5	152	30.6	1.42 [1.09-1.83]		175	37.4	139	30.3	1.42 [1.07-1.89]	
Recessive												
G/G-A/G	515	95.7	475	95.8	1.00	0.974	447	95.5	442	96.3	1.00	0.615
A/A	23	4.3	21	4.2	1.01 [0.55-1.85]		21	4.5	17	3.7	1.19 [0.60-2.35]	
log-Additive												
0,1,2					1.28 [1.03-1.59]	**0.026**					1.31 [1.03-1.67]	**0.028**
rs2332719												
Alleles												
A	707	62.7	691	67.1	1.00	**0.032**						
G	421	37.3	339	32.9	1.21 [1.02-1.45]							
Genotypes												
Dominant												
A/A	223	39.5	239	46.4	1.00	**0.023**	194	39.6	220	46.1	1.00	0.165
G/A-G/G	341	60.5	276	53.6	1.32 [1.04-1.69]		296	60.4	257	53.9	1.21 [0.93-1.58]	
Recessive												
A/A-G/A	484	85.8	452	87.8	1.00	0.344	423	86.3	420	88.1	1.00	0.374
G/G	80	14.2	63	12.2	1.19 [0.83-1.69]		67	13.7	57	11.9	1.20 [0.81-1.78]	
log-Additive												
0,1,2					1.21 [1.01-1.44]	**0.035**					1.15 [0.95-1.40]	0.150
rs7613868												
Alleles												
C	708	62.7	692	67.1	1.00	**0.033**						
T	422	37.3	340	32.9	1.21 [1.02-1.45]							
Genotypes												
Dominant												
C/C	226	40.0	237	45.9	1.00	**0.049**	197	40.1	219	45.8	1.00	0.286
T/C-T/T	339	60.0	279	54.1	1.27 [1.00-1.62]		294	59.9	259	54.2	1.16 [0.89-1.51]	
Recessive												
C/C-T/C	482	85.3	455	88.2	1.00	0.165	423	86.2	423	88.5	1.00	0.258
T/T	83	14.7	61	11.8	1.28 [0.90-1.83]		68	13.8	55	11.5	1.26 [0.84-1.87]	
log-Additive												
0,1,2					1.20 [1.01-1.43]	**0.036**					1.14 [0.94-1.38]	0.185
rs12634530												
Alleles												
C	819	74.5	797	78.4	1.00	**0.031**						
T	281	25.5	219	21.6	1.25 [1.02-1.53]							
Genotypes												
Dominant												
C/C	304	55.3	312	61.4	1.00	**0.043**	266	55.5	283	60.1	1.00	0.363
C/T-T/T	246	44.7	196	38.6	1.29 [1.01-1.65]		213	44.5	188	39.9	1.13 [0.87-1.48]	
Recessive												
C/C-C/T	515	93.6	485	95.5	1.00	0.188	447	93.3	450	95.5	1.00	0.177
T/T	35	6.4	23	4.5	1.43 [0.83-2.46]		32	6.7	21	4.5	1.50 [0.83-2.7]	
log-Additive												
0,1,2					1.25 [1.02-1.53]	**0.030**					1.15 [0.92-1.44]	0.207

rs12637456												
Alleles												
T	665	62.5	671	66.8	1.00	**0.040**						
A	399	37.5	333	33.2	1.21 [1.01-1.45]							
Genotypes												
Dominant												
T/T	212	39.8	229	45.6	1.00	0.061	185	40.0	211	45.5	1.00	0.353
A/T-A/A	320	60.2	273	54.4	1.27 [0.99-1.62]		278	60.0	253	54.5	1.14 [0.87-1.49]	
Recessive												
T/T-A/T	453	85.2	442	88.0	1.00	0.171	396	85.5	410	88.4	1.00	0.247
A/A	79	14.8	60	12.0	1.28 [0.90-1.84]		67	14.5	54	11.6	1.27 [0.85-1.89]	
log-Additive												
0,1,2					1.20 [1.01-1.43]	**0.043**					1.13 [0.93-1.38]	0.215
rs17286604												
Alleles												
C	774	68.6	658	63.6	1.00	**0.014**						
T	354	31.4	376	36.4	0.80 [0.67-0.96]							
Genotypes												
Dominant												
C/C	270	47.9	201	38.9	1.00	**0.003**	231	47.0	184	38.4	1.00	**0.010**
T/C-T/T	294	52.1	316	61.1	0.69 [0.54-0.88]		260	53.0	295	61.6	0.70 [0.54-0.92]	
Recessive												
C/C-T/C	504	89.4	457	88.4	1.00	0.613	436	88.8	423	88.3	1.00	0.898
T/T	60	10.6	60	11.6	0.91 [0.62-1.33]		55	11.2	56	11.7	0.97 [0.65-1.47]	
log-Additive												
0,1,2					0.80 [0.67-0.96]	**0.014**					0.82 [0.67-1.00]	**0.049**
rs11712619												
Alleles												
C	775	69.3	646	63.3	1.00	**0.003**						
T	343	30.7	374	36.7	0.76 [0.64-0.92]							
Genotypes												
Dominant												
C/C	263	47.0	196	38.4	1.00	**0.004**	229	47.1	186	39.3	1.00	**0.030**
C/T-T/T	296	53.0	314	61.6	0.70 [0.55-0.90]		257	52.9	287	60.7	0.74 [0.57-0.97]	
Recessive												
C/C-C/T	512	91.6	450	88.2	1.00	0.068	445	91.6	420	88.8	1.00	0.087
T/T	47	8.4	60	11.8	0.69 [0.46-1.03]		41	8.4	53	11.2	0.68 [0.43-1.06]	
log-Additive												
0,1,2					0.75 [0.62-0.91]	**0.003**					0.77 [0.63-0.95]	**0.014**
rs17377867												
Alleles												
G	1022	92.1	912	89.1	1.00	**0.017**						
A	88	7.9	112	10.9	0.70 [0.52-0.94]							
Genotypes												
Dominant												
G/G	467	84.1	408	79.7	1.00	0.058	404	83.8	378	79.6	1.00	0.126
G/A-A/A	88	15.9	104	20.3	0.74 [0.54-1.01]		78	16.2	97	20.4	0.76 [0.54-1.08]	
Recessive												
G/G-G/A	555	100.0	504	98.4	1.00	**0.003**	482	100.0	469	98.7	1.00	**0.007**
A/A	0	0.0	8	1.6	0.00 [0.00-/]		0	0.0	6	1.3	0.00 [0.00-/]	
log-Additive												
0,1,2					0.75 [0.62-0.91]	**0.003**					0.73 [0.52-1.01]	0.057

rs6784664												
Alleles												
C	506	48.7	536	53.4	1.00	**0.036**						
A	534	51.3	468	46.6	1.20 [1.01-1.43]							
Genotypes												
Dominant												
C/C	122	23.5	144	28.7	1.00	0.057	106	23.6	133	28.6	1.00	0.055
C/A-A/A	398	76.5	358	71.3	1.31 [0.99-1.74]		344	76.4	332	71.4	1.35 [0.99-1.84]	
Recessive												
C/C-C/A	384	73.8	392	78.1	1.00	0.113	334	74.2	359	77.2	1.00	0.203
A/A	136	26.2	110	21.9	1.26 [0.95-1.68]		116	25.8	106	22.8	1.23 [0.89-1.69]	
log-Additive												
0,1,2					1.20 [1.01-1.43]	**0.036**					1.21 [1.00-1.47]	0.050

OR: Odds ratio; CI: Confidence interval

Even though we investigated 34 haplotype tagging SNPs in the region of the 5' end of *KALRN*, a large portion of its natural genetic variation was not assessed because it is a very large gene with a high degree of genetic diversity. We therefore performed genotype imputation for SNPs in chromosome 3 using data from HapMap as well as the genotypes observed at the 34 fully genotyped polymorphisms. We obtained imputed genotypes meeting minimum quality standards (MAF in controls ≥ 0.05 and SNP INFO ≥ 0.5) for 405 SNPs in and around *ROPN1* and *KALRN* (Figure 4.4).

As expected, the imputed SNPs with higher information content metric (SNP INFO > 0.8) are located in the region of genotyped SNPs. As an additional QC, the genotyped SNPs are dropped one at a time and imputed using the other observed genotypes. A concordance rate between observed and imputed genotypes can be calculated for these markers, and ninety percent of the genotyped SNPs had a concordance rate > 85% between imputed and observed genotypes (Figure 4.4). Thirty two of the imputed polymorphisms have an allelic association with IS risk at a more stringent p-value threshold of 0.01 (Figure 4.4; Appendix C: Table C.5). Among the top five SNPs (p-value < 0.001), the first one (rs6790975) has a modest SNP INFO metric (0.598) and therefore its association has to be taken with caution, but the remaining four markers (rs4678111, rs7620580, rs9820396, and rs1444770) have a very high SNP INFO metric (0.93; Appendix C: Table C.5) and are in complete LD (Figure 4.4). The LD among the top 32 imputed associated SNPs (Figure 4.4) also suggests the existence of several independent clusters of association with IS in the 5' region of the KALRN gene, such as the rs6438833-rs1373612, rs6779809-rs11712039-rs11719349, and rs4678111-rs7620580-rs9820396-rs1444770 clusters.

To validate the imputed results, we additionally genotyped in our dataset three SNPs (rs7620580, rs6438833 and rs11712039) representing the above-mentioned clusters of association, and tested their association with IS (Table 4.6).

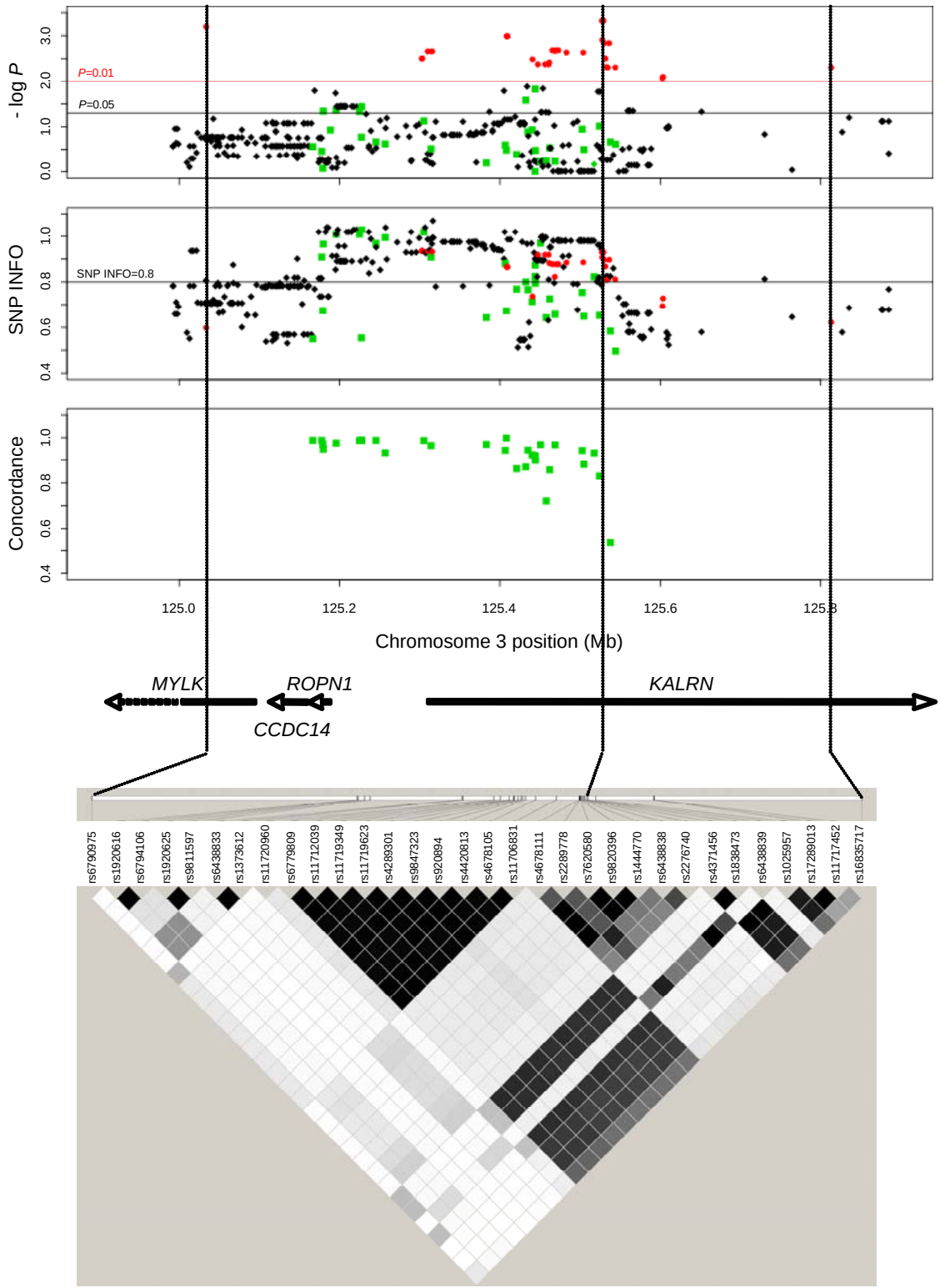


Figure 4.4: Association results of imputed SNPs. The top plot displays the negative logarithm of the p-value for the allelic association test of all 405 imputed SNPs in chromosome 3. The second plot shows the SNP information content metric SNP INFO, and the third plot displays the concordance rate of observed

and imputed genotypes (for genotyped SNPs only). In all plots, the 32 imputed SNPs with p-value < 0.01 are represented with red spheres, and the 34 SNPs that have been genotyped are represented with green squares. Known genes in this region are indicated (*MYLK*: myosin light chain kinase isoform, *CCDC14*: coiled-coil domain containing 14). The LD plot at the bottom shows all the r^2 pairwise correlation coefficients among the 32 imputed SNPs with p-value < 0.01.

Table 4.6: Characteristics of the three imputed SNPs of the KALRN gene to be validated. For each SNP it is indicated its chromosomic position, the possible alleles and the minor allele frequency (MAF) in the CEU population. All SNPs were genotyped using the iPlex assay.

Gene	Chr.	SNP ID	Position (NCBI build 35)	Allele 1:2	MAF	Comment
KALRN	3q13	rs6438833	125,408,285	T:A	0.050	MAF<0.1
		rs11712039	125,456,276	C:T	0.420	
		rs7620580	125,527,993	A:G	0.090	MAF<0.1

MAF: Minor Alele Frequency

The average concordance rate among observed and imputed genotypes was 92%. All three polymorphisms were associated with IS for any of the association tests performed (0.003 < p-value < 0.027; Figure 4.3; Table 4.7; Appendix C: Table C.6). The SNPs rs7620580 and rs6438833 maintain significant results under the dominant genetic model in unadjusted and adjusted tests (0.005 < p-value < 0.018), and the SNP rs11712039 under the recessive genetic model (0.026 < p-value < 0.045; Table 4.7). These association results validate the imputing procedures highlighting the biological importance of the imputed associated SNPs.

Table 4.7: Detailed association results for the three validated imputed SNPs in *KALRN*. See legend in table 4.3.

Gene	SNP ID/ Model	Unadjusted						Adjusted					
		Cases		Controls		OR [95% CI]	p-value	Cases		Controls		OR [95% CI]	p-value
		N	%	N	%			N	%	N	%		
KALRN	**rs7620580**												
	Alleles												
	A	697	83.8	850	88.4	1.00	**0.005**						
	G	135	16.2	112	11.6	1.47 [1.14-1.89]							
	Genotypes												
	Dominant												
	A/A	293	70.4	378	78.6	1.00	**0.005**	257	70.2	344	77.7	1.00	**0.005**
	G/A-G/G	123	29.6	103	21.4	1.54 [1.14-2.09]		109	29.8	99	22.3	1.60 [1.15-2.22]	
	Recessive												
	A/A-G/A	404	97.1	472	98.1	1.00	0.317	355	97.0	434	98.0	1.00	0.355
	G/G	12	2.9	9	1.9	1.56 [0.65-3.73]		11	3.0	9	2.0	1.54 [0.61-3.87]	
	log-Additive												
	0,1,2					1.45 [1.11-1.90]	**0.006**	366	45.2	443	54.8	1.49 [1.12-1.98]	**0.007**

rs6438833												
Alleles												
T	738	87.9	873	92.1	1.00	**0.003**						
A	102	12.1	75	7.9	1.61 [1.19-2.17]							
Genotypes												
Dominant												
T/T	329	78.3	404	85.2	1.00	**0.007**	290	78.2	372	85.3	1.00	**0.018**
T/A-A/A	91	21.7	70	14.8	1.60 [1.13-2.25]		81	21.8	64	14.7	1.57 [1.08-2.28]	
Recessive												
T/T-T/A	409	97.4	469	98.9	1.00	0.076	362	97.6	432	99.1	1.00	0.136
A/A	11	2.6	5	1.1	2.52 [0.87-7.32]		9	2.4	4	0.9	2.41 [0.72-8.04]	
log-Additive												
0,1,2					1.55 [1.14-2.10]	**0.004**	371	46.0	436	54.0	1.53 [1.09-2.13]	**0.012**
rs11712039												
Alleles												
C	586	69.1	607	63.2	1.00	**0.009**						
T	262	30.9	353	36.8	0.77 [0.66-0.90]							
Genotypes												
Dominant												
C/C	196	46.2	187	39.0	1.00	**0.027**	175	46.7	175	39.6	1.00	0.093
T/C-T/T	228	53.8	293	61.0	0.74 [0.57-0.97]		200	53.3	267	60.4	0.78 [0.59-1.04]	
Recessive												
C/C-T/C	390	92.0	420	87.5	1.00	**0.026**	345	92.0	389	88.0	1.00	**0.045**
T/T	34	8.0	60	12.5	0.61 [0.39-0.95]		30	8.0	53	12.0	0.61 [0.38-1.00]	
log-Additive												
0,1,2					0.76 [0.62-0.93]	**0.007**	375	45.9	442	54.1	0.78 [0.63-0.97]	**0.027**

OR: Odds ratio; CI: Confidence interval

Among the SNPs tested and associated with IS in this study, SNP rs11712039 has been associated (0.001 < p-value < 0.01) with IS in the GWAS performed by Ikram *et al.* (2009). The rs17286604 and rs11712619, which are associated with IS in our Portuguese sample, were marginally associated in the GWAS (0.01 < p-value < 0.10) and SNP rs4499545 was not tested.

Even though these association results would not survive to the Bonferroni's correction for multiple testing, the obtained results and the presented multiple independent evidences of association suggest that variants in *KALRN* may be important risk factors to IS and a risk factor for vascular diseases.

4.3.2. CFH gene

Since CFH gene was previously found to be associated with an increased risk of AMD and MI which share several risk factors with stroke, we tested its association with IS. This study was performed on 388 affected individuals and 461 controls selected from our biobank (Table 4.8). This sample has approximately the same characteristics than our complete biobank (Table 4.1).

Table 4.8: Sample characterization. Principal demographic, clinical and lifestyle characteristics of the IS case-control sample used for the study of the CFH gene.

Characteristic	Cases	Controls	p-value[a]
N	388	461	
Sex (n, % male)	227 (58.5 %)	221 (47.9 %)	$< 10^{-4}$
Age-at-examination (mean ± SD, years)	51.1 ± 9.9	63.4 ± 6.9	$< 10^{-4}$
Age-at-onset (mean ± SD, years)	50.2 ± 10.0	-	-
Risk factors (n/N[b], %)			
Hypertension (> 140-85 mmHg)	197/351 (56.1 %)	172/455 (37.8 %)	$< 10^{-4}$
Hypercholesterolemia (> 200 mg/dL)	229/366 (62.6 %)	283/459 (61.7 %)	0.788
Hypertriglycemia (> 200 mg/dL)	25/144 (17.3 %)	58/386 (15.0 %)	0.511
Diabetes	65/374 (17.4 %)	49/441 (11.1 %)	0.010
Ever smoking	169/383 (44.1 %)	134/452 (29.6 %)	$< 10^{-4}$
Ever drinking	225/382 (58.9 %)	196/446 (43.9 %)	$< 10^{-4}$

[a]p-value of an unpaired Student's t test or a χ^2 test for a quantitative and qualitative data, respectively

[b]Number of individuals for whom the data was available

For this gene with 95.5 kb, we only genotyped the SNP rs1061170 previously found to be associated with an increased risk for AMD and MI (Klein *et al.* 2005, Haines *et al.* 2005, Edwards *et al.* 2005, Kardys *et al.* 2006). This SNP is localized in the position 194,925,860 of chromosome 1, and it was genotyped using an iPlex assay. It is a T:C biallelic SNP, and its MAF is 0.35 (Thakkinstian *et al.* 2006).

In our dataset, this marker was in HWE in both case and control groups (p-value = 0.158 and 0.159, respectively) and had a MAF similar to that of other populations (0.340). We found a modest allelic and unadjusted genotypic association (p-value = 0.030 and log-additive model p-value = 0.035) of this polymorphism with IS (Table 4.9). There was no adjusted genotypic association with the risk of IS using the log-additive genetic model. However, a marginal association (p-value = 0.045) is obtained using the recessive genetic model (Table 4.9).

Table 4.9: Detailed association results for associated SNPs in *CFH*. See legend in table 4.3.

Gene	SNP ID/	Unadjusted						Adjusted					
	Model	Cases		Controls		OR [95% CI]	p-value	Cases		Controls		OR [95% CI]	p-value
		N	%	N	%			N	%	N	%		
CFH	**rs1061170**												
	Alleles												
	T	533	68.7	587	63.7	1.00	**0.030**						
	C	243	31.3	335	36.3	0.80 [0.68-0.94]							
	Genotypes												
	Dominant												
	T/T	189	48.7	194	42.1		0.053	164	47.5	181	42.4		0.206
	T/C-C/C	199	51.3	267	57.9	0.77 [0.58-1.00]		181	52.5	246	57.6	0.83 [0.62-1.11]	
	Recessive												
	T/T-T/C	344	88.7	393	85.2		0.142	306	88.7	362	84.8		**0.045**
	C/C	44	11.3	68	14.8	0.74 [0.49-1.11]		39	11.3	65	15.2	0.64 [0.41-1.00]	
	log-Additive												
	0,1,2					0.81 [0.67-0.99]	**0.035**					0.82 [0.66-1.01]	0.060

OR: Odds ratio; CI: Confidence interval

These results suggest that the studied polymorphism in CFH can affect also the risk of IS on the Portuguese population. It is not associated with IS in the GWAS performed by Ikram *et al.* (2009).

4.3.3. EPO, HO2, KLK1 genes

We also selected three neuroprotective genes to test for association with IS risk. For *EPO* (2.9 kb), *HO2* (33.9 kb) and *KLK1* (4.6 kb) we genotyped 3, 3 and 5 tagging SNPs respectively, both in the coding and 10 kb flanking regions (on NCBI B35). Two additional SNPs in almost complete LD in the HapMap project with two of the selected tagging SNPs were also genotyped to compare the LD patterns in our cohort to those of the Caucasian HapMap sample (Table 4.10). The association studies were performed in the 565 patients with IS and 518 unrelated controls from our Caucasian dataset (Table 4.1).

Table 4.10: Characterization of the investigated SNPs in the EPO, HO2 and KLK1 genes. For each SNP it is indicated its chromosomic position, the possible alleles and the minor allele frequency (MAF) in the CEU population. All SNPs were genotyped using the iPlex assay. The SNPs that did not pass our quality controls requirements are shown in grey.

Gene	Chr.	SNP ID	Position (NCBI build 35)	Allele 1:2	MAF	Comments
EPO	7q22	rs1617640	99,961,949	A:C	0.392	
		rs564449	99,965,789	G:T	0.125	
		rs4729607	99,971,395	T:A	0.250	
HO2	16p13	rs3761680	4,464,061	A:C	0.383	
		rs2160567	4,475,558	T:C	0.383	to replace rs3761680
		rs7702	4,500,930	C:G	0.364	
		rs7665	4,502,513	G:A	0.339	
KLK1	19q13	rs266116	56,010,373	T:A	0.492	
		rs266117	56,011,329	A:T	0.280	
		rs2739454	56,012,392	G:A	0.407	
		rs2659058	56,017,918	T:C	0.358	
		rs3212820	56,018,170	C:A	0.173	
		rs3212810	56,020,548	C:T	0.231	to replace rs3212820

MAF: Minor Alele Frequency

We did not find any association with IS risk for any SNP or haplotype in these genes (Figures 4.5 to 4.7; Appendix C: Tables C.7 and C.8), suggesting that they are not involved in stroke pathogenesis. None of these SNP are associated with IS in the GWAS performed by Ikram *et al.* (2009) too.

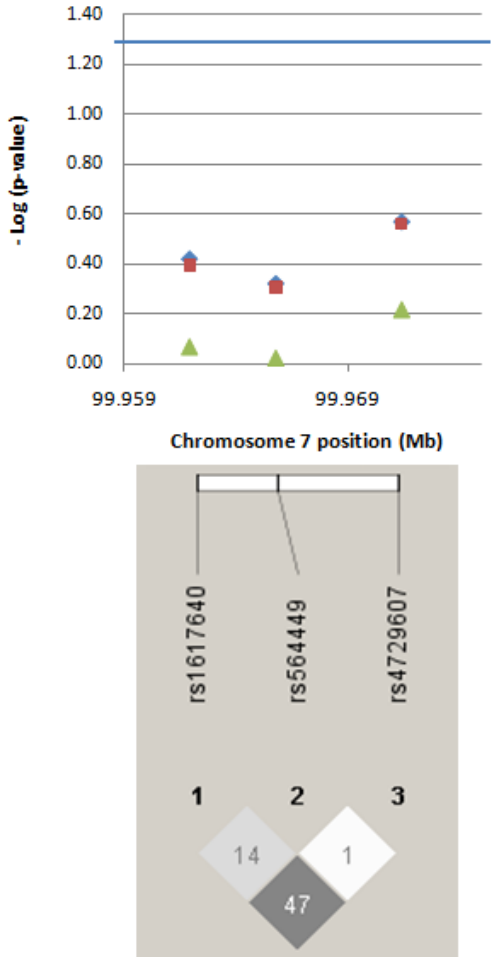


Figure 4.5: Association results and pairwise linkage disequilibrium (LD) for genotyped polymorphisms in *EPO*. See legend in Figure 4.1. The value of the pairwise statistic r^2 is shown inside the diamonds of the LD plot.

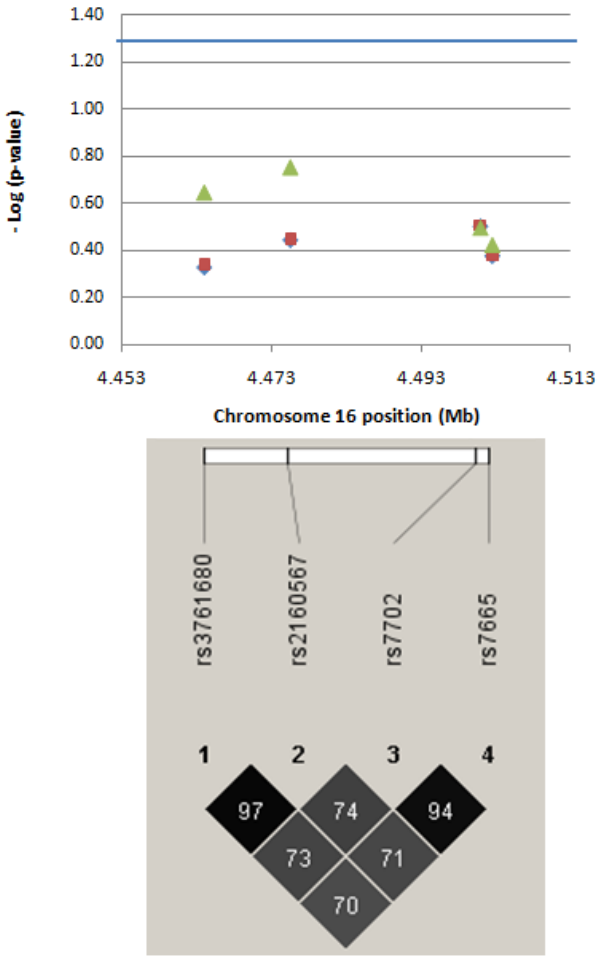


Figure 4.6: Association results and pairwise linkage disequilibrium (LD) for genotyped polymorphisms in *HO2*. See legend in Figure 4.1. The value of the pairwise statistic r^2 is shown inside the diamonds of the LD plot.

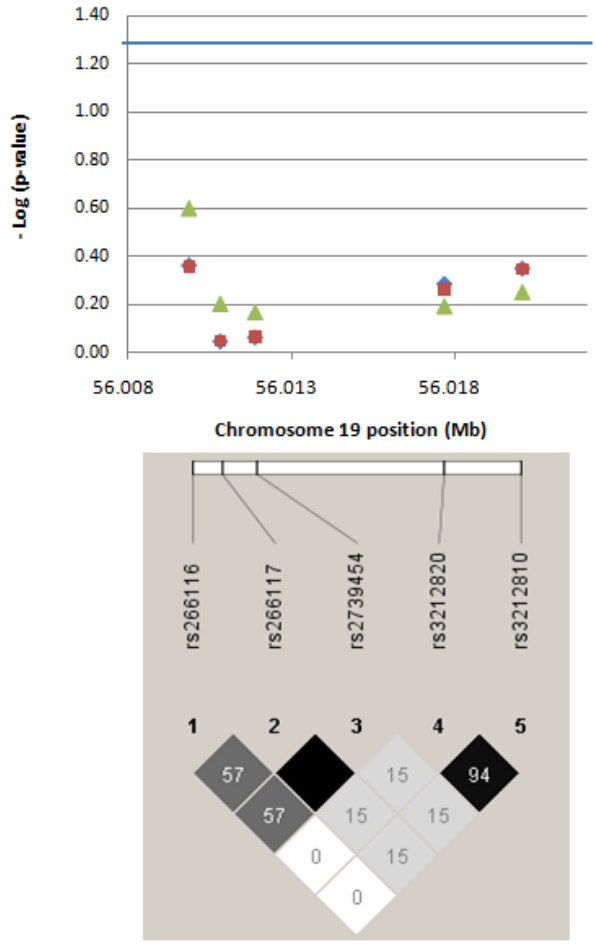


Figure 4.7: Association results and pairwise linkage disequilibrium (LD) for genotyped polymorphisms in *KLK1*. See legend in Figure 4.1. The value of the pairwise statistic r^2 is shown inside the diamonds of the LD plot.

4.4. Prioritizing genes by CG approach

4.4.1. Gene profiling studies

To investigate the gene expression differences between IS cases and controls in the non-acute phase of stroke, we compared and contrasted the genetic profiles in PBMCs from twenty IS cases (from whom the samples were collected at least six months after the first and only stroke event) and twenty controls. The principal demographic, clinical and risk factor information of the expression profiling study participants are shown Table 4.11. The controls were sex- and age-matched with cases. The older

participants were between 65 and 74 years old, and younger participants were between 45 and 54 years old, at the time of the blood collection and at the time of the IS with only one exception (sample 60162; Table 4.11). For this group of participants, from all the presented clinical and lifestyle characteristics, only the frequency of diabetes was significantly higher in IS patients than in controls (p-value = 0.035). There are four patients with diabetes that are controlled by medication and no controls with this disease. The great majority of hypertensive participants are medicated too. Due to our rigorous inclusion and exclusion criteria, and for logistic reasons, it was not possible to perfectly match the samples according their urban or rural origin to analyse if there are any genetic factors that predispose differently the stroke patients from different origins for the ischemic attack. Samples were prepared and hybridized to the Affymetrix GeneChip Human U133 Plus 2.0 microarrays at the IGC's Affymetrix core facility.

The RNA samples of the selected participants met all quality parameters (described in the methods section) and had RIN numbers higher than 6.8 (92,5% have RIN ≥ 7,0 and 50% have RIN ≥ 8,0; data not shown).

Table 4.11: Characterization of the samples used for the expression studies on Affymetrix GeneChip Human U133 Plus 2.0 microarrays. The principal demographic, clinical and lifestyle characteristics of the study subjects are presented. Cases and controls were matched for sex and age. The mean and the standard deviation (SD) of the age-at-examination (AAE) and age-at-onset (AAO) are indicated when applicable. Additionally, the geographic origin of the participants and the hybridization group of the samples in the Affymetrix microarrays were presented. Vila-Real and Mirandela are rural regions, while Porto and Lisbon are urban regions.

Status	Sex	N	ID	AAE	AAO	Clinical / lifestyle characteristics	Origin	Hyb[a]
Cases	Female	1	60092	71	65		Porto	F
		2	60164	65	65	Hypertension	Vila-Real	F
		3	60165	73	70		Vila-Real	B
		4	60169	70	69	Hypertension, hypercholesterolemia	Vila-Real	B
		5	60337	65	65	Hypertension, hypercholesterolemia	Porto	F
		6	60089	53	45	Hypertension, hypercholesterolemia, smoker, drinker	Porto	B
		7	60160	54	49	Hypertension, diabetes, hypercholesterolemia, drinker	Vila-Real	A
		8	60162	47	43	Hypercholesterolemia	Vila-Real	F
		9	60170	54	50	Hypertension	Vila-Real	A
		10	60340	49	48	Diabetes, hypercholesterolemia, smoker, drinker	Porto	F
	Male	11	60014	69	65	Hypertension, diabetes, hypercholesterolemia, drinker	Mirandela	B
		12	60008	74	72	Hypertension, drinker	Mirandela	C
		13	60012	74	73	Hypertension, hypercholesterolemia, drinker	Mirandela	C
		14	60020	68	66	Drinker	Mirandela	C
		15	60090	71	65	Hypertension, smoker, drinker	Porto	C
		16	60010	54	53	Diabetes, hypercholesterolemia, drinker	Mirandela	C
		17	60018	52	48	Hypercholesterolemia, smoker, drinker	Mirandela	C
		18	60022	46	45	Hypertension, drinker	Mirandela	B
		19	60024	49	47	Smoker, drinker	Mirandela	B
		20	60167	46	45	Hypertension, drinker	Vila-Real	A
Mean ± SD				60.2 ± 10.6	57.4 ± 10.8			

Controls	Female	21	60013	68	-		Mirandela	A
		22	60015	68	-	Hypertension	Mirandela	A
		23	60017	73	-	Hypertension	Mirandela	B
		24	60021	65	-	Hypertension, hypercholesterolemia	Mirandela	C
		25	60088	73	-	Hypertension, hypercholesterolemia	Porto	C
		26	60019	45	-		Mirandela	A
		27	60023	48	-	Drinker	Mirandela	A
		28	60025	47	-	Hypertension	Mirandela	A
		29	60342	47	-	Hypertension, smoker	Porto	D
		30	60413	48	-	Hypercholesterolemia, drinker	Lisboa	F
	Male	31	60159	69	-	Smoker, drinker	Vila-Real	C
		32	60166	72	-	Drinker	Vila-Real	C
		33	60338	66	-	Hypertension, hypercholesterolemia, smoker, drinker	Porto	F
		34	60411	68	-	Hypercholesterolemia, drinker	Lisboa	F
		35	60421	69	-		Lisboa	F
		36	60168	45	-		Vila-Real	F
		37	60401	52	-	Hypertension, hypercholesterolemia, drinker	Vila-Real	E
		38	60402	53	-	Smoker	Vila-Real	E
		39	60412	48	-	Smoker	Lisboa	F
		40	60422	50	-	Hypercholesterolemia, smoker, drinker	Lisboa	F
Mean ± SD				58.7 ± 11.0				

ID: sample identification number, Hyb: hybridization group

[a]Chips in different hybridization group were hybridized in different scan dates

4.4.1.1. *GeneChip quality controls*

Total RNA from each individual was hybridized to an Affymetrix GeneChip® Human Genome U133 Plus 2.0 microarray. We got very good results with all the QC requisites during the amplification, fragmentation, hybridization, washing and staining steps. The obtained electropherograms with the Agilent 2100 Bioanalyzer showed good size distributions of the prepared biotin-labelled cRNA and of the fragmented cRNA for each sample (data not shown); none of the arrays presented any significant image artefacts (data not shown); and the average and the standard deviation of the percent present calls and of the background of the prepared arrays were respectively 44.8 ± 1.6 % and 45.3 ± 5.7 of intensity (Table 4.12) in agreement with the generally followed guidelines. The Q value and the intensities of the QC prokaryote probe sets obtained for each array are presented in the Table 4.12. We can clearly see that there are a great concordance between the raw intensity values for samples prepared on the same group independently of the QC we analyse. Comparing the values among the different groups, there is an evident increase of the intensities to the following group as consequence of the use of a new lot of the staining dye and of the upgrade of the Affymetrix fluidics station from the IGC' facility. However, this increase is gradual and it is solved with the normalization of the raw results. Looking with more detail for the Table 4.12 we can underline the expected linear relation of the signal intensity of the poly-A RNA controls similar to the presented in Figure 3.5; the high intensity values of the 3' Cre probes; and the values of the ratio 3'/ 5' for GAPDH, actin and ribosomal 18S not greater than 3, with an exception for the actin ratio 3'/ 5' of the sample 60165 that is slightly greater. This exception, however, is not worrying since the value is only slightly greater and all the other QC requisites are respected for the sample. The hybridization

controls that are not presented respected also the quality requirements.

All expression results were then normalized and $\log_2$ transformed before their downstream analyses.

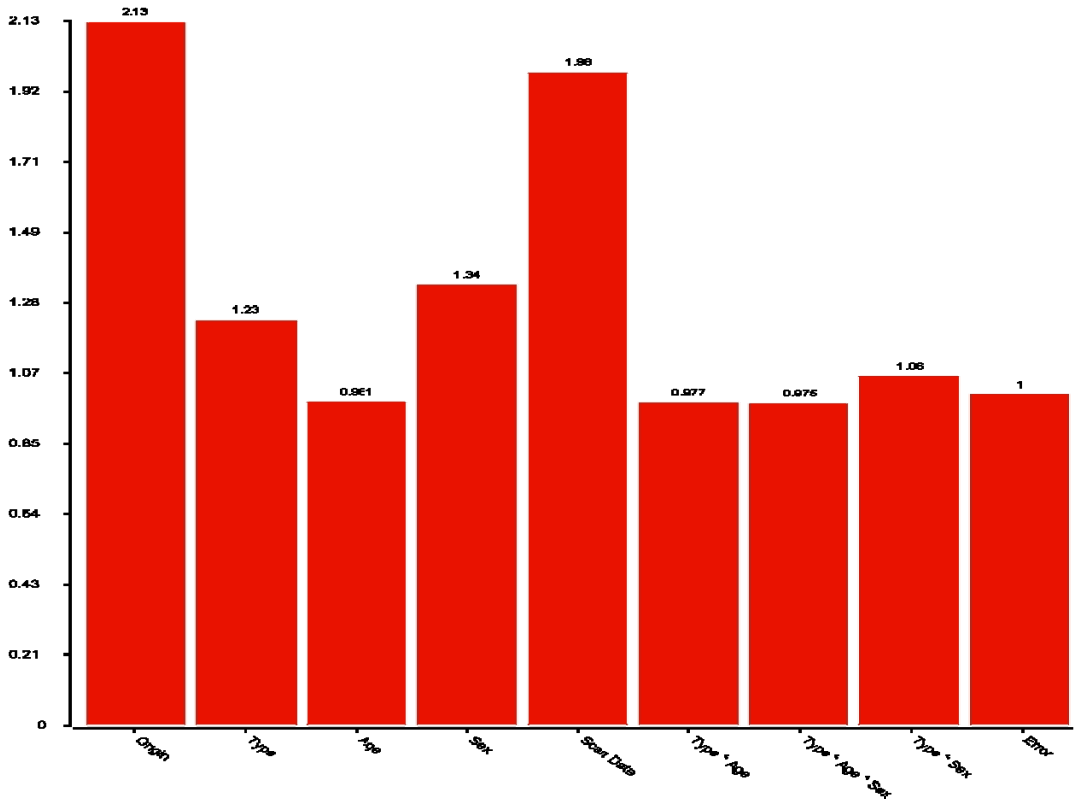
Table 4.12: Obtained quality control parameters for the prepared GeneChips. Samples are listed by groups in which the Affymetrix microarrays were prepared. The intensity values, the percentages or the ratios are presented for each of the quality control parameters (described in the methods section).

Sample	Q value	Background	P calls	GAPDH 3' - 5' Ratio	Actin 3' - 5' Ratio	18SRNA 3' - 5' Ratio	3' Cre	Poly-A controls 3' DAP	3' Thr	3' Phe	3' Lys
Group A											
60013	1.35	40.1	42.7%	1.41	1.30	0.72	5415.2	496.3	140.4	84.7	26.3
60015	1.32	41.6	42.2%	1.21	1.22	0.42	5046.8	504.9	143.0	92.3	27.1
60019	1.35	42.1	43.0%	1.75	1.92	0.25	6056.6	623.9	200.1	108.5	48.9
60023	1.36	39.2	43.3%	1.39	1.49	0.50	6472.2	559.5	151.2	113.0	41.0
60025	1.37	41.2	44.8%	1.32	1.41	0.59	5652.1	600.0	161.6	130.0	43.9
60170	1.25	37.9	43.8%	1.68	2.15	0.56	5917.9	481.6	81.5	99.9	36.3
60160	1.38	40.3	44.7%	1.34	1.72	0.77	7823.0	610.1	141.5	119.7	39.7
60167	1.37	41.1	42.3%	1.18	2.01	0.56	7162.0	674.3	125.3	92.1	33.4
Group B											
60014	1.29	39.3	43.2%	1.19	1.49	0.50	8037.0	897.2	269.8	169.6	75.5
60017	1.30	39.1	43.9%	1.39	2.40	0.92	7178.3	1158.8	363.7	195.0	81.8
60022	1.17	36.5	42.7%	1.65	2.34	1.07	6585.6	699.4	210.9	146.0	84.2
60024	1.30	40.5	43.0%	1.46	1.74	0.72	6705.7	690.7	216.7	137.8	73.7
60089	1.32	39.6	44.3%	1.31	2.10	1.25	7416.5	819.4	231.6	157.9	70.2
60169	1.34	40.8	43.8%	1.21	1.47	0.70	7443.8	736.9	230.6	151.9	67.5
60165	1.34	40.8	43.3%	1.02	3.13	0.91	7690.1	1291.7	347.7	246.0	98.3
Group C											
60008	1.57	44.9	43.8%	1.08	1.60	1.02	6711.3	1298.9	391.6	255.2	102.0
60010	2.07	62.0	42.6%	1.07	1.81	1.27	7972.9	1473.7	461.4	300.1	114.5
60012	1.53	45.9	45.4%	1.18	2.00	0.98	7804.6	1468.9	363.9	244.5	98.0
60018	1.50	46.1	44.4%	1.18	2.06	1.00	8305.7	1376.2	394.5	292.4	115.3
60020	1.55	46.4	45.3%	1.25	2.32	1.00	8186.5	1209.6	339.1	215.0	105.1
60021	1.46	43.6	46.2%	1.17	1.77	1.14	7938.3	1475.9	386.4	268.6	111.4
60088	1.55	46.4	46.0%	1.12	1.59	1.08	6743.0	1339.4	382.8	250.8	106.9
60090	1.45	44.1	45.2%	1.13	1.87	1.19	7094.0	1021.9	306.9	193.1	77.8
60159	1.39	41.8	44.8%	1.11	1.81	1.28	6127.1	923.4	267.2	214.6	64.7
60166	1.43	39.9	43.6%	1.07	2.20	1.20	7368.1	1178.0	362.0	198.1	83.4
Group D											
60342	1.62	43.9	46.5%	1.03	1.19	0.72	9548.7	1378.0	363.2	220.4	106.3
Group E											
60401	1.69	44.7	48.3%	1.04	1.30	0.73	11598.3	1651.4	456.2	262.9	137.4
60402	1.87	48.9	47.9%	1.23	1.29	0.54	10763.0	1557.4	475.2	281.0	144.5
Group F											
60092	1.80	45.8	46.0%	1.23	1.25	1.11	10869.3	1042.8	293.0	211.7	79.5
60337	1.84	49.4	47.0%	1.07	1.19	0.66	12005.5	1600.9	416.2	258.8	121.0
60340	1.89	49.4	44.6%	1.06	1.13	0.81	13047.0	1316.7	321.6	155.1	114.6
60338	1.91	50.6	47.0%	0.99	1.18	0.95	11090.2	1276.4	361.4	231.8	91.5
60168	1.96	50.8	46.6%	1.15	1.24	0.63	11556.2	1527.0	378.4	251.6	113.7
60164	2.01	53.9	45.1%	1.08	1.35	0.93	10982.8	958.9	241.5	153.6	64.0
60162	1.96	52.4	44.6%	1.10	1.37	1.25	12498.7	1376.7	312.7	197.5	99.9
60413	1.98	52.8	47.1%	1.09	1.35	0.97	11696.5	1277.7	354.3	218.2	78.4
60411	2.00	51.9	45.9%	1.05	1.24	1.66	11753.2	1120.5	285.3	235.6	77.7
60421	1.93	51.6	46.6%	1.02	1.36	1.43	12033.6	1362.1	326.1	224.6	104.4
60412	1.89	50.0	46.0%	1.04	1.31	1.47	12199.9	1106.4	357.8	200.8	88.7
60422	2.06	55.2	45.1%	1.10	1.36	1.47	11615.2	1207.7	296.0	178.7	79.7

4.4.1.2. *Batch effects removal*

ANOVA was used to identify the differentially expressed genes among cases and controls, taking into account our known experimental and study-design co-variates. ANOVA *per se* does not need a batch remover; it removes batch effects by simply including them in the calculations. However, for the visualization of the true biological effects (rather than the co-lateral experimental and study-design effects) in tools such as the PCA and hierarchical clustering, we removed from the normalized expression results the non-specific co-variates of our data following batch-remove procedures implemented in the Partek software.

When we preformed the ANOVA of the normalized expression results including as factors the type, sex, age, and the combinations among them (type*age, type*age*sex, type*sex), as well as the geographic origin of the participants and scan-date of the microarrays (Figure 4.8), we observed that, for the average of all probe sets, the geographic origin and scan-date have a larger effect on the expression variation among the samples than the biological factor of interest (type).



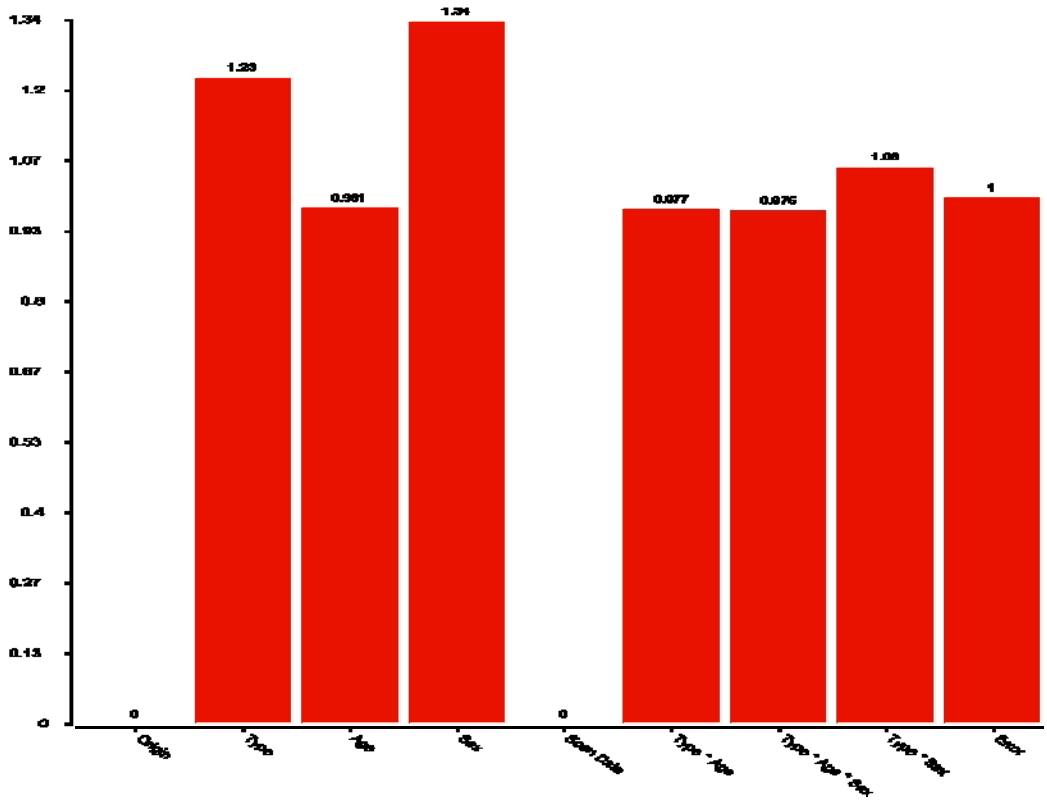


Figure 4.8: Analysis of variance. Average of the F ratio of the different ANOVA factors: origin, type, age, sex, scan date, and combinations among type, sex and age, before (at top) and after (at bottom) the batch-remove procedures of the origin and scan-date effects. Results were obtained using the Partek software.

As expected, the 3D PCA plot of the samples obtained using all probes sets without batch removal, shows a clear geographic origin effect on sample distribution, with samples from Mirandela in a separate cluster (Figure 4.9). This effect may result from lifestyle differences among participants from different regions that are not important for the determination of the genetic factors of the IS, or could result from differences in experimental procedures at the collecting centres (e.g. the use of different centrifuges to the treatment of the bold samples to get the PBMCs from which the total RNA was extracted). On the other hand, samples seem to be slightly separated in the first principal component (PC#1) depending if their microarrays were prepared before or after the upgrade of the Affymetrix fluidics station at the IGC's facility (Figure 4.10).

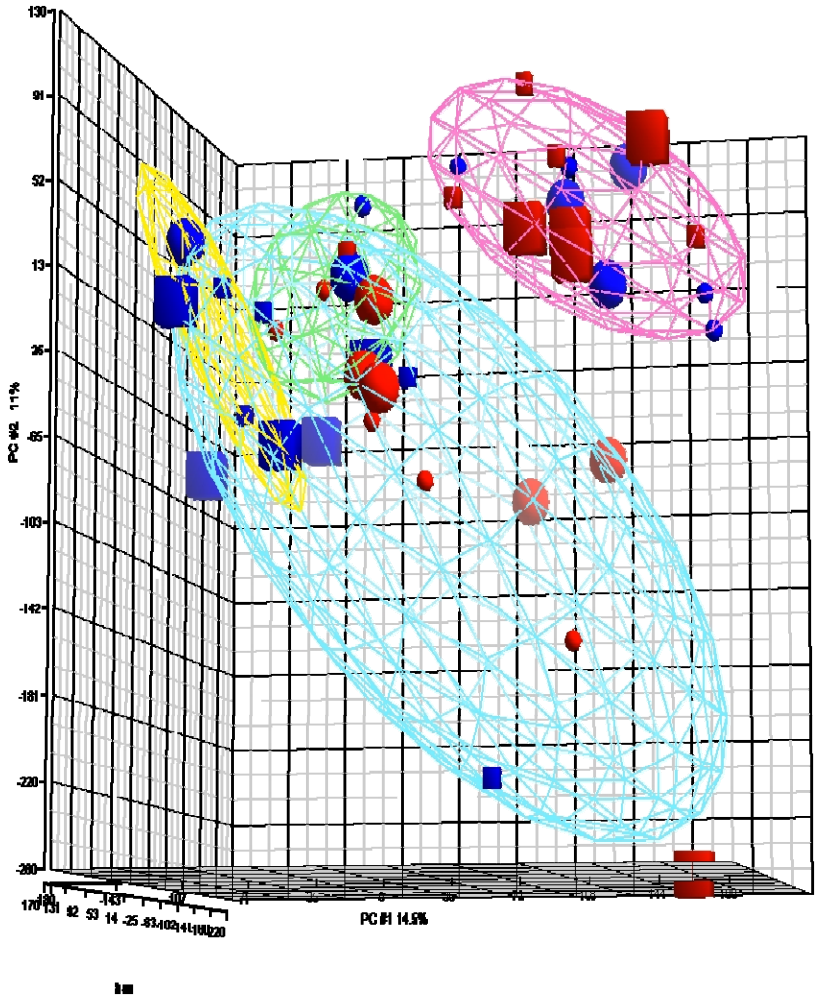


Figure 4.9: Illustration of the geographic origin effect. 3D PCA plot of the samples using all the probes from Affymetrix GeneChip Human U133 Plus 2.0 microarrays. Ellipsoids identify different Portuguese regions where samples were collected: Porto (green), Vila-Real (blue), Mirandela (pink) and Lisbon (yellow). Samples from Mirandela are clearly separated for the remaining ones. IS cases are represented by red symbols and controls by blue symbols, males by cubes, females by spheres, and the younger participants with bigger symbols than the older ones. Figure created with the Partek software.

These co-lateral batch effects could mask the genes differentially expressed between IS cases and controls and consequently, they must be removed for a correct the visualization of the relevant biological differences. After the effects of the referred non-specific co-variates were removed, we obtained all samples balanced in the first principal component of the PCA (Figure 4.11). The removed factors must still be included in posterior ANOVAs to account for the degrees of freedom used in the batch removal process (Figure 4.8).

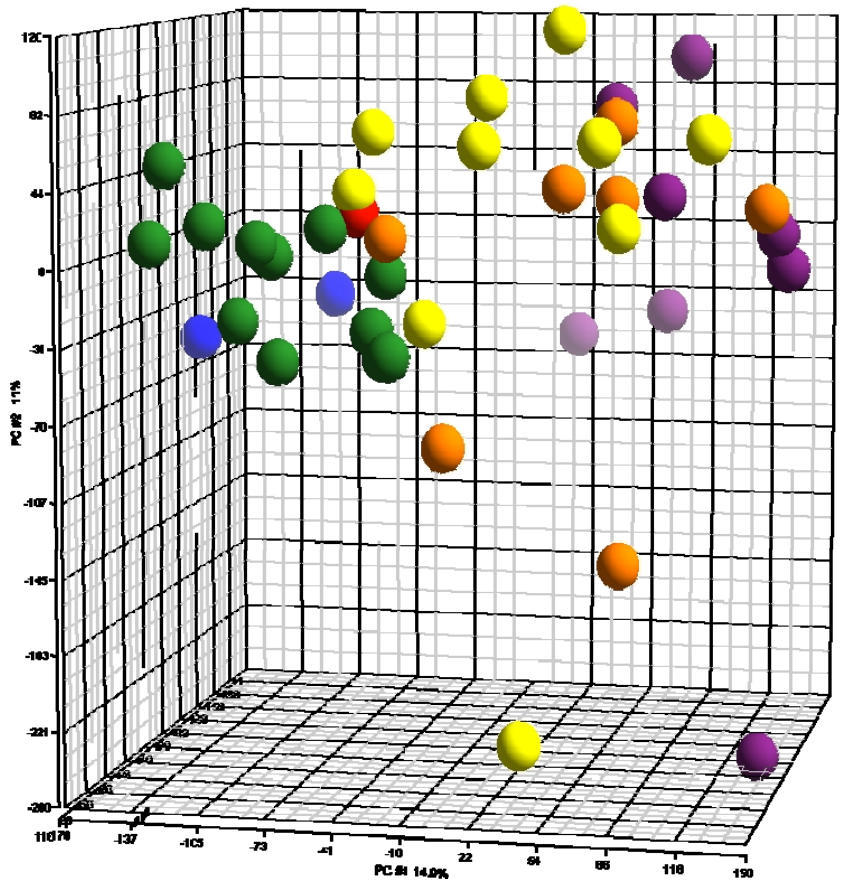


Figure 4.10: Illustration of the scan date effect. 3D PCA plot of the samples using all the probes from Affymetrix GeneChip Human U133 Plus 2.0 microarrays. Samples are coloured by scan date: group A (purple), group B (orange), group C (yellow), group D (red), group E (blue), group F (green). Samples have a bias in the first principal component (PC#1) depending on whether their microarrays were prepared before or after the upgrade of the Affymetrix fluidics station at the IGC's facility. Figure created with the Partek software.

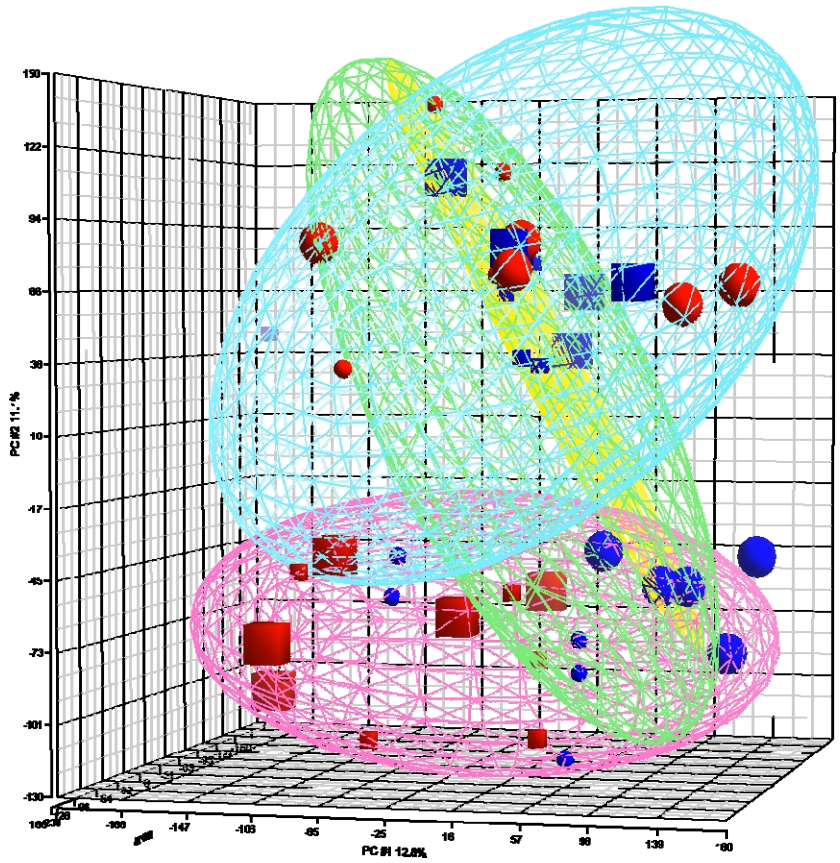


Figure 4.11: Illustration of the batch-remover tool effect. 3D PCA plot of the samples using all the probes from Affymetrix GeneChip Human U133 Plus 2.0 microarrays after the batch-remove procedures of the geographic origin and scan-date effects. Ellipsoids identify different Portuguese regions were samples were collected: Porto (green), Vila-Real (blue), Mirandela (pink) and Lisbon (yellow). Samples are all mixed in the first principal component (PC#1). IS cases are represented by red symbols and controls by blue symbols, males by cubes, females by spheres, and the younger participants with bigger symbols than the older ones. Figure created with the Partek software.

4.4.1.3. *Differentially expressed genes*

Performing an ANOVA of the normalized expression results including as factors the type, sex, age, the combinations among them, the geographic origin of the participants and the scan-date of the microarrays (Figure 4.8), we found 1,675 probe sets differentially expressed among IS cases and controls with a threshold of 1.2 fold-change (data not shown). From all these probe sets, 709 of them, representing 580 genes (Appendix D: Table D.1), were considered statistically significant with an uncorrected p-value < 0.05 (331 probe sets representing 287 genes were down-regulated in cases vs. controls). All of these 709 probe sets remain differentially expressed after using the q-value method (Storey 2002) to determine the FDR for a stringent significance of q-value < 0.05. The q-value of an individual test is the minimum FDR at which the test may be called significant. It was used as the false discovery rate method as it has a higher apparent power when compared to other standard methods (Qian *et al.* 2005). None of the probe sets resist to the step up or step down method implemented in Partek or the Bonferroni's correction. However FDR is the most appropriate method to apply to microarrays multiple comparisons problem (Allison *et al.* 2006).

The 3D PCA plots obtained using the referred 709 probe sets and for lists of probe sets with different cut-offs of the uncorrected p-value are shown in Figure 4.12. All of them revealed sets of genes with an expression that allows a clear separation between cases and controls, mainly on the first principal component. For an uncorrected p-value < 0.05, the hierarchical clustering diagram also supports this observation showing an almost perfect distinction among cases and controls according to their expression patterns (Figure 4.13). Only the samples 60020 and 60022 that correspond to a male case and a male case, respectively, are clustered as if they are most similar to the controls than to the other cases.

The 580 genes represented by the referred 709 probe sets localize to all human chromosomes and the genes with top fold-change are shown in Table 4.13. Neither the *PDE4D* or the *ALOX5AP* were found significantly differentially expressed among our IS cases and controls.

Face to these promising results, we opted to use the indicated less stringent list of 709 probe sets (with an uncorrected p-value < 0.05) in the following analysis and to apply the GC approach we proposed to do. We also verified that differentially expressed genes among age-matched cases and controls (older cases vs. controls, and younger cases vs. controls), generally belong to this list of 709 probes (data not shown). We did not find genes differentially expressed among older cases and younger cases, or among older controls and younger controls (data nor shown), which supports the joint analysis of individuals belonging to different age classes.

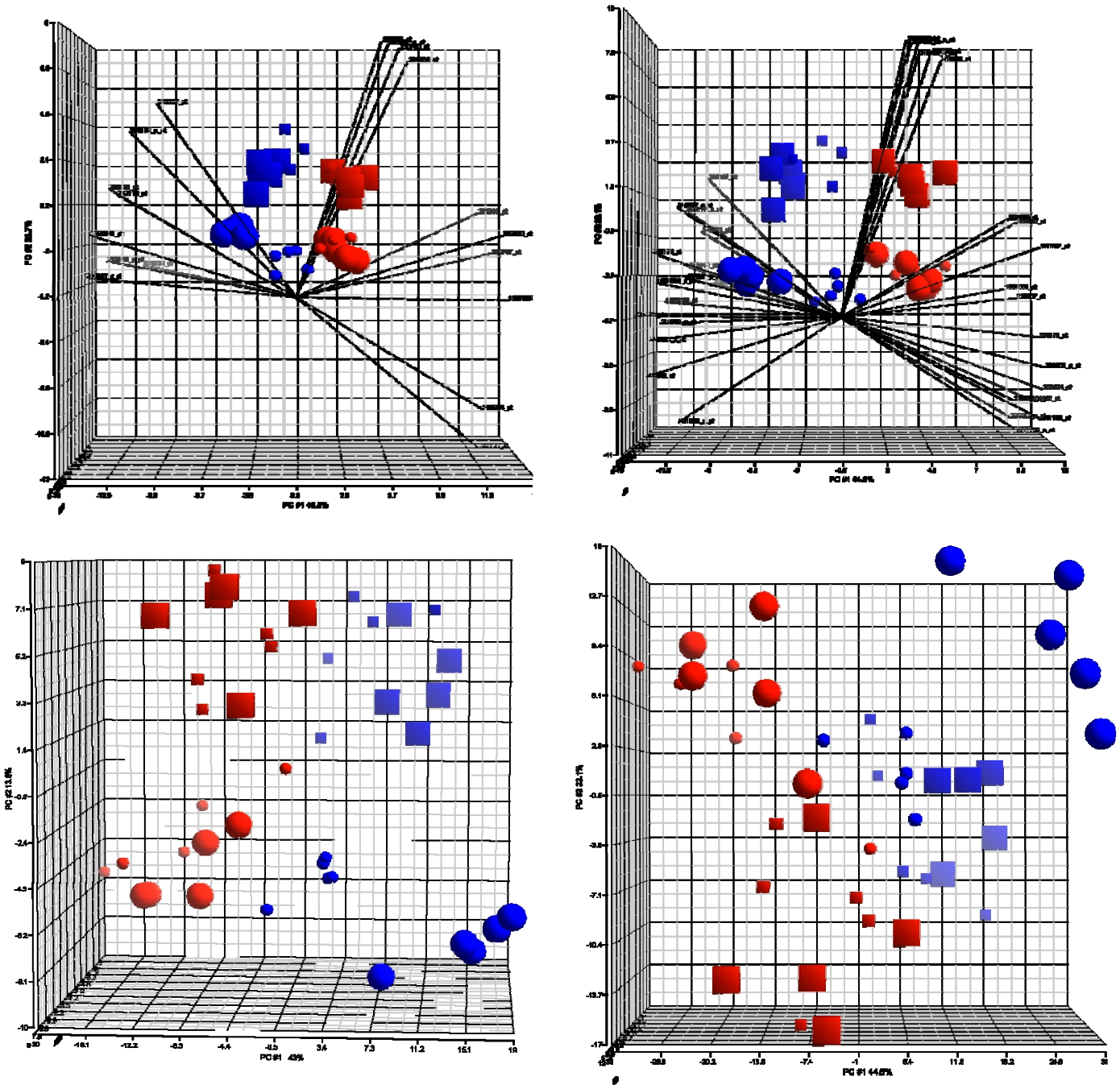


Figure 4.12: 3D PCA plot of the analyzed samples. These plots were created using lists of probe sets differentially expressed among IS cases and controls with a threshold of 1.2 fold-change and different cut-offs of the uncorrected p-value: p<0.0005, 18 probes (at top left corner); p<0.001, 35 probes (at top right corner); p<0.01, 230 probes (at down left corner); p<0.05, 709 probes (at down right corner). IS cases (red symbols) are separated from controls (blue symbols) in the first principal component (PC#1). Males are represented by cubes, females by spheres and the younger participants with bigger symbols than the older ones. Figures created with the Partek software.

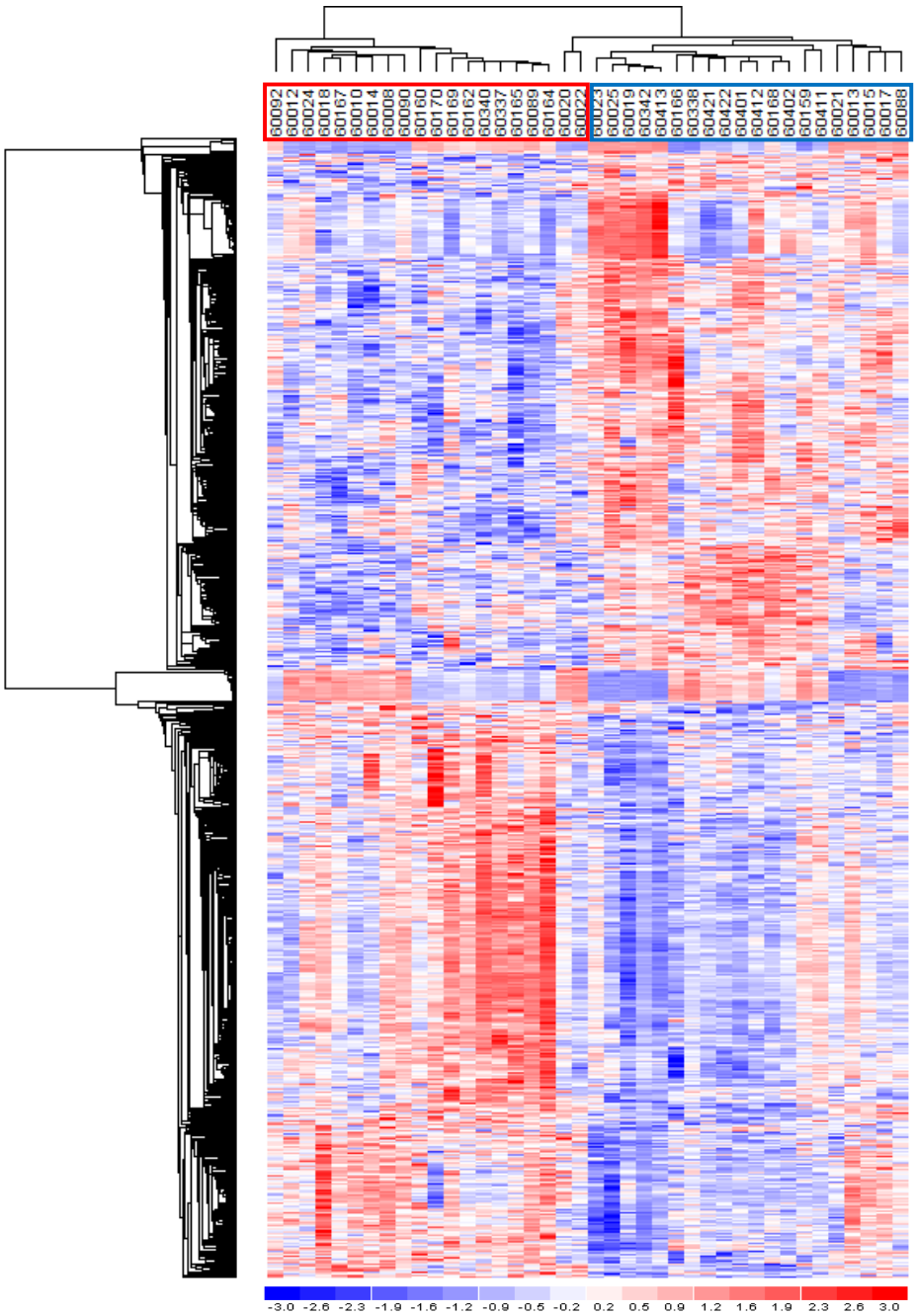


Figure 4.13: Hierarchical clustering of the analyzed samples. This plot was constructed using the 709 probe sets differentially expressed among IS cases and controls with a threshold of 1.2 fold-change and an uncorrected p-value < 0.05. Each column represents an individual and each row a probe set. The higher expression levels are dark red and lower levels are dark blue. Control samples are inside the blue square and IS patients inside the red square. Figure created with the dChip 2009 software.

Table 4.13: Top differentially expressed genes among IS cases and controls. 30 top genes out to the 580 genes differentially expressed among IS cases and controls with a threshold of 1.2 fold-change and an uncorrected p-value <0.05. Results were obtained using the Partek software.

Probeset ID	Chr.	Gene symbol	Gene name	p-value	Fold-change (cases/controls)
224588_at	X	XIST	X (inactive)-specific transcript	0.005	-2.69
211430_s_at	14	IGH@/IGHG1/IGHG2/IGHG3/IGHM	Immunoglobulin heavy locus/heavy constant gamma/heavy constant mu	0.003	-2.53
206641_at	16	TNFRSF17	Tumor necrosis factor receptor superfamily, member 17	0.002	-2.42
217022_s_at	14	IGHA1/IGHA2	Immunoglobulin heavy constant alpha	0.011	-2.38
242961_x_at	9	DDX58	DEAD (Asp-Glu-Ala-Asp) box polypeptide 58	0.012	-1.78
211644_x_at	2	IGKC	Immunoglobulin kappa constant	0.016	-1.76
218340_s_at	4	UBE1L2	Ubiquitin-activating enzyme E1-like 2	0.010	-1.75
212592_at	4	IGJ	Immunoglobulin J polypeptide, linker protein for immunoglobulin alpha and mu polypeptides	0.008	-1.73
234764_x_at	22	IGL@	Immunoglobulin lambda locus	0.010	-1.72
214777_at	2	---	Immunoglobulin Kappa light chain V gene segment	0.012	-1.72
215118_s_at	14	IGHG1	Immunoglobulin heavy constant gamma 1 (G1m marker)	0.048	-1.67
1556209_at	12	CLEC2B	C-type lectin domain family 2, member B	0.008	-1.63
221651_x_at	2	IGKC/IGKV1-5	Immunoglobulin kappa constant/immunoglobulin kappa variable 1-5	0.007	-1.62
212843_at	11	NCAM1	Neural cell adhesion molecule 1	0.044	-1.60
211343_s_at	10	COL13A1	Collagen, type XIII, alpha 1	0.015	-1.60
204627_s_at	17	ITGB3	Integrin, beta 3 (platelet glycoprotein IIIa, antigen CD61)	0,034	1,91
241881_at	1	OR2W3	Olfactory receptor, family 2, subfamily W, member 3	0,044	1,95
240103_at	8	---	Full-length cDNA clone CS0DI080YO16 of Placenta Cot 25-normalized	0,026	2,00
206494_s_at	17	ITGA2B	Integrin, alpha 2b (platelet glycoprotein IIb of IIb/IIIa complex, antigen CD41)	0,018	2,03
204409_s_at	Y	EIF1AY	Eukaryotic translation initiation factor 1A, Y-linked	0,003	2,12
230760_at	Y	ZFY	Zinc finger protein, Y-linked	0,000	2,19
207113_s_at	6	TNF	Tumor necrosis factor (TNF superfamily, member 2)	0,033	2,24
212077_at	7	CALD1	Caldesmon 1	0,003	2,25
201058_s_at	20	MYL9	Myosin, light polypeptide 9, regulatory	0,031	2,26
205476_at	2	CCL20	Chemokine (C-C motif) ligand 20	0,045	2,32
214974_x_at	4	CXCL5	Chemokine (C-X-C motif) ligand 5	0,003	2,55
221491_x_at	6	HLA-DRB1/HLA-DRB3/HLA-DRB4	Major histocompatibility complex, class II, DR beta	0,029	2,63
201909_at	Y	RPS4Y1	Ribosomal protein S4, Y-linked 1	0,001	3,09
205000_at	Y	DDX3Y	DEAD (Asp-Glu-Ala-Asp) box polypeptide 3, Y-linked	0,000	3,12
241133_at	7	PRSS1	Protease, serine, 1 (trypsin 1)	0,000	3,31

Chr.: Chromosome

4.4.1.4. *Classes of genes and pathways over-represented among the differentially expressed genes*

Gene ontology and pathway analysis were then executed to extract the maximum of possible biological information from the data obtained. Using the Bio functions tool from IPA software to cluster the genes according to their molecular and cellular functions and ranking the groups of genes according their p-value, we found a significant over-representation (3.8×10^{-12} < p-value < 1.7×10^{-3}) of groups of genes related cell death (142 genes), cellular growth and proliferation (141 genes), cellular movement (78

genes), cell-to-cell signalling and interaction (82 genes) and cellular function and maintenance (34 genes) among the 580 differentially expressed genes (Appendix D: Table D.2). This suggests an active role of the PBMCs (from where the total RNA were extracted) of the IS patients in the complex immune and homeostatic responses to the vascular injuries that cause the ischemic attacks.

These results are concordant with those obtained using the Gene Function Enrichment tool from dChip 2009 software, which also cluster genes according to their cellular function. Groups related with the antigen binding (13 genes), immune and inflammatory responses (38 and 16 genes respectively), platelet alpha granule membrane (5 genes), response to virus (11 genes), oxidoreductase activity (8 genes), and response to DNA damage stimulus (17 genes), were significantly over-represented (1.0×10^{-9} < p-value < 8.8×10^{-4}) among the differentially expressed genes (Appendix D: Table D.3). The platelet alpha granule membrane group, for instance, includes the SELP gene, as well as other platelet receptors.

When analysing the differentially expressed genes by their physiological system development and function (in Bio functions tool from IPA software), we found a significant over-representation (4.2×10^{-8} < p-value < 1.7×10^{-3}) of genes related with haematological system development and function (96 genes), tissue morphology and development (64 and 81 genes, respectively) and organismal survival and development (64 and 66 genes, respectively) (Appendix D: Table D.2). 263, 76 and 59 of the obtained differentially expressed genes were associated with genetic disorders, haematological diseases, and organismal injury and abnormalities, respectively (Appendix D: Table D.4). The ischemia, thrombosis, injury of tissue, injury of organ, and damage of tissue groups of genes are expected to be involved in the pathogenic process of IS.

Using the Pathway-Express tool of the free available Intelligent Systems and Bioinformatics Laboratory, we identified 24 pathways (Figure 4.14 and Table 4.14) significantly different among IS cases and controls. None of them has more than 12% of its involved genes differentially expressed. Antigen processing and presentation pathway has only down-regulated genes, while gap junction, complement and coagulation cascades, and Parkinson's disease related pathways have only up-regulated genes (Figure 4.14). Using the IPA software 34 canonical pathways were obtained with significant p-values, with a maximum of 21% of its involved genes differentially expressed (Appendix D: Table D.5). Several pathways, like the most significant cell adhesion molecules pathway (Figure 4.14 and Table 4.14), appear to be related with the inflammatory process of the stroke disease.

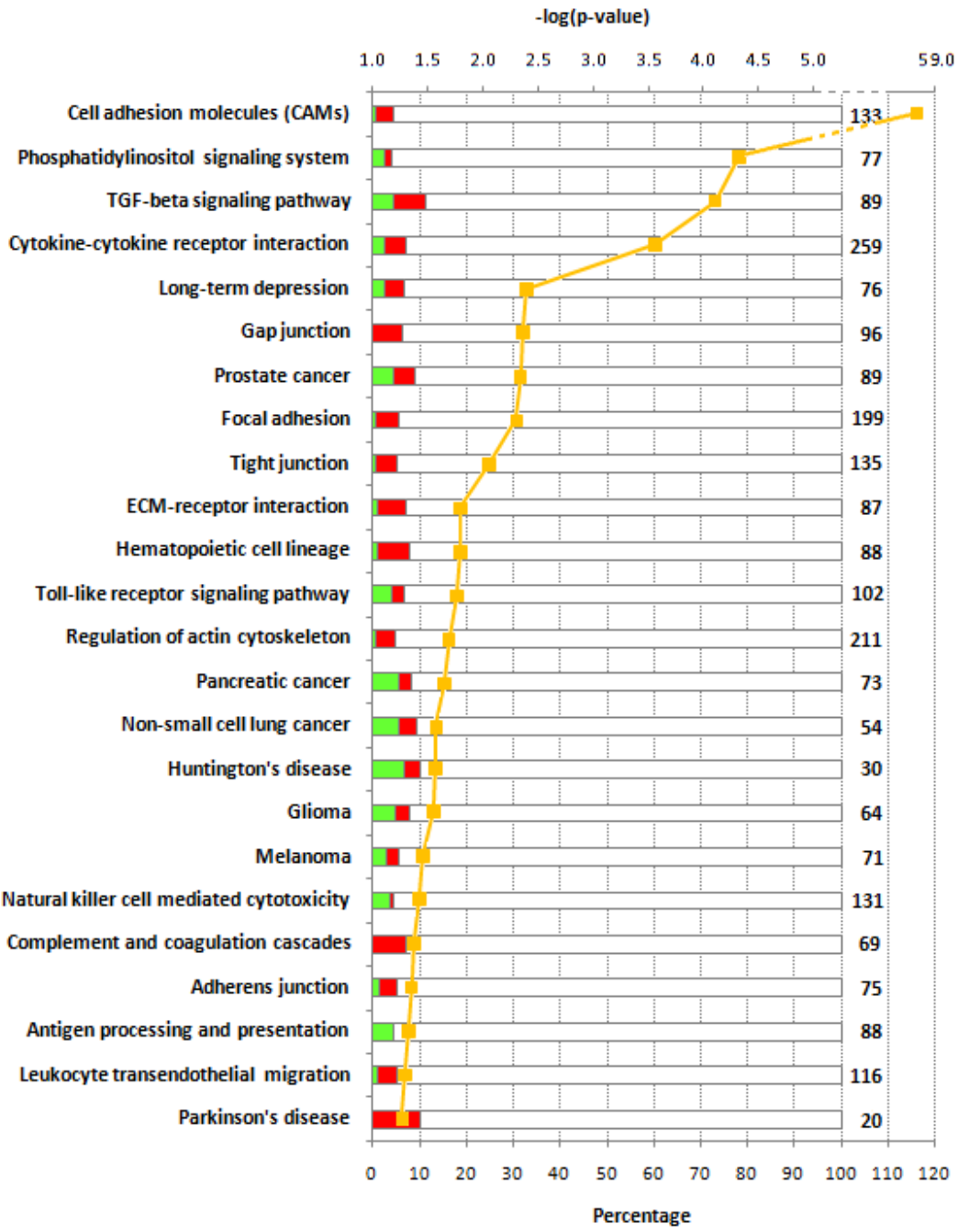


Figure 4.14: Pathways significantly over-represented by the differentially expressed genes. In each bar, the percentage of down-regulated genes in the corresponding pathway is shown in green and the percentage of up-regulated genes in red. The percentage of unchanged genes is in white, and the total number of genes of each pathway described by the software dataset is indicated in the right side of each bar. In orange squares are the negative logarithms of the p-value of the pathways. Results were obtained using the Pathway-Express tool of the Intelligent Systems and Bioinformatics Laboratory.

Table 4.14: Pathways significantly over-represented by the differentially expressed genes. The genes that are differentially expressed in each pathway are indicated. Results were obtained using the Pathway-Express tool of the Intelligent Systems and Bioinformatics Laboratory.

Pathway	p-value	Molecules
Cell adhesion molecules (CAMs)	$2.42x10^{-59}$	CLDN5, ESAM, JAM3, NCAM1, SDC4, SELP
Phosphatidylinositol signaling system	$4.22x10^{-5}$	DGKG, PIK3CG, SYNJ1
TGF-beta signaling pathway	$6.95x10^{-5}$	ACVR1B, ACVR2A, BMP6, EP300, LEFTY1, LTBP1, PPP2R1B, SMAD5, SP1, TNF
Cytokine-cytokine receptor interaction	$2.47x10^{-4}$	ACVR1B, ACVR2A, CCL20, CCL7, CXCL2, CXCL5, EGF, IFNAR2, IFNGR1, IL10, LEPR, MPL, PF4, TNF, TNFRSF17, TNFRSF18, TNFSF4, TNFSF8
Long-term depression	$3.61x10^{-3}$	GNAZ, GUCY1A3, GUCY1B3, PPP2R1B, RYR1
Gap junction	$3.82x10^{-3}$	EGF, GUCY1A3, GUCY1B3, PDGFA, PRKY, TUBB1
Prostate cancer	$4.09x10^{-3}$	CREB1, EGF, EP300, NFKBIA, PDGFA, PIK3CG, TGFA, TP53
Focal adhesion	$4.43x10^{-3}$	ACTN1, COL6A3, EGF, ITGA2B, ITGB3, ITGB5, MYL9, MYLK, PARVB, PDGFA, PIK3CG
Tight junction	$7.84x10^{-3}$	ACTN1, CLDN5, CSDA, CTTN, JAM3, MYL9, PPP2R1B
ECM-receptor interaction	$1.42x10^{-2}$	COL6A3, GP9, HMMR, ITGA2B, ITGB3, ITGB5, SDC4
Hematopoietic cell lineage	$1.44x10^{-2}$	CD38, CD55, GP9, GYPA, ITGA2B, ITGB3, TNF
Toll-like receptor signaling pathway	$1.55x10^{-2}$	IFNAR2, IRF7, MAP2K3, NFKBIA, PIK3CG, TLR3, TNF
Regulation of actin cytoskeleton	$1.82x10^{-2}$	ACTN1, ARHGEF12, EGF, ITGA2B, ITGB3, ITGB5, MYL9, MYLK, PDGFA, PIK3CG
Pancreatic cancer	$1.99x10^{-2}$	ACVR1B, EGF, PIK3CG, PLD1, TGFA, TP53
Non-small cell lung cancer	$2.37x10^{-2}$	EGF, FOXO3, PIK3CG, TGFA, TP53
Huntington's disease	$2.40x10^{-2}$	EP300, HIP1, TP53
Glioma	$2.53x10^{-2}$	EGF, PDGFA, PIK3CG, TGFA, TP53
Melanoma	$3.12x10^{-2}$	EGF, PDGFA, PIK3CG, TP53
Natural killer cell mediated cytotoxicity	$3.37x10^{-2}$	IFNAR2, IFNGR1, MICA, PIK3CG, SYK, TNF
Complement and coagulation cascades	$3.80x10^{-2}$	CD55, F13A1, F3, PROS1, TFPI
Adherens junction	$3.97x10^{-2}$	ACTN1, ACVR1B, EP300, WASF3
Antigen processing and presentation	$4.24x10^{-2}$	CREB1, CTSL1, HSPA4, PSME3
Leukocyte transendothelial migration	$4.58x10^{-2}$	ACTN1, CLDN5, ESAM, JAM3, MYL9, PIK3CG
Parkinson's disease	$4.83x10^{-2}$	SEPT5, UBB

4.4.2. Convergence with whole genome linkage screens

According to our proposed GC approach, we prioritized candidate genes for association studies by converging our expression results with the linkage peaks reported in published human whole-genome linkage screens for stroke. Previous linkage studies found evidence for linkage to chromosome 5q12 (Gretarsdottir *et al.* 2002), and to chromosomes 1p34, 5q13, 7q35, 9q22, 9q34, 13q32, 14q32, 18p11 and 20q13 (Nilsson-Ardnor *et al.* 2007). We found 16 differentially expressed genes which map to these linkage peaks (Table 4.15). All of these 16 genes are represented in the Affymetrix microarrays by probe sets identified with the "_at" suffix. These probe sets are specifically designed to recognize unique sequences of each gene and their suffix indicates that the probe recognizes the antisense strand of the gene of interest. Differentially expressed genes mapping on the referred linkage peaks but exclusively

represented by less specific probe sets with other suffixes (e.g. "s_at", "x_at") were not included. Only one gene, the *ELOVL7*, emerged as differentially expressed on the chromosome 5q12. This is the most strongly implicated IS linkage peak in the published studies. This gene maps relatively close to the 5' end region of the *PDE4D*.

Table 4.15: Prioritized genes. Differentially expressed genes (1.2 fold-change cut-off and uncorrected p-value < 0.05) mapping on reported whole genome linkage peaks for stroke.

Linkage peak (reference)	Probeset ID	Gene symbol	Gene name	p-value	Fold-change (cases/controls)
1p34 (Nilsson-Ardnor *et al.*)	207550_at	***MPL***	Myeloproliferative leukemia virus oncogene	0.029	1.71
5q12 (Gretarsdottir *et al.*)	227180_at	***ELOVL7***	ELOVL family member 7, elongation of long chain fatty acids (yeast)	0.022	1.78
9q22 (Nilson-Ardnor *et al.*)	221556_at	***CDC14B***	CDC14 cell division cycle 14 homolog B (S. cerevisiae)	0.010	1.47
	223669_at	***HEMGN***	Hemogen	0.042	1.64
9q34	212531_at	***LCN2***	Lipocalin 2 (oncogene 24p3)	0.038	1.79
(Nilsson-Ardnor *et al.*)	204384_at	***GOLGA2***	Golgi autoantigen, golgin subfamily a, 2	0.006	-1.24
	237403_at	***GFI1B***	Growth factor independent 1B (potential regulator of CDKN1A, translocated in CML)	0.010	1.46
	229002_at	***FAM69B***	Family with sequence similarity 69, member B	0.004	1.24
13q32 (Nilsson-Ardnor *et al.*)	225666_at	***TMTC4***	Transmembrane and tetratricopeptide repeat containing 4	0.001	-1.23
14q32	226152_at	***TTC7B***	Tetratricopeptide repeat domain 7B	0.020	1.54
(Nilsson-Ardnor *et al.*)	1559097_at	***C14orf64***	Chromosome 14 open reading frame 64	0.046	1.24
	237181_at	***PPP2R5C***	Protein phosphatase 2, regulatory subunit B (B56), gamma isoform	0.014	-1.27
	230972_at	***ANKRD9***	Ankyrin repeat domain 9	0.018	1.31
20q13	202071_at	***SDC4***	Syndecan 4 (amphiglycan, ryudocan)	0.010	1.55
(Nilsson-Ardnor *et al.*)	225402_at	***TP53RK***	TP53 regulating kinase	0.020	-1.21
	230690_at	***TUBB1***	Tubulin, beta 1	0.024	1.79

4.4.3. Association studies in GC genes

The goal of the association studies in the prioritized genes is to identify novel polymorphisms or combinations of polymorphisms that influence the risk of developing stroke.

For each of the 16 genes, we genotyped all known haplotype tagging SNPs. These markers were selected based on their chromosomal location, covering all genes and 10 kb of both flanking regions (on NCBI B35), and based on their MAF in European population as explained in the methods section. Some SNPs with MAF < 0.1 were selected when they were considered relevant to increase the number of genotyped SNPs per gene. A total of 191 tagging SNPs were selected (Table 4.16).

Table 4.16: Characterization of the investigated SNPs in the GC genes. See legend in table 4.10.

Gene	Chr.	SNP ID	Position (NCBI build 35)	Allele 1:2	MAF	Comment
MPL	1p34	rs710252	43,575,809	G:A	0.425	
		rs12731981	43,576,927	G:A	0.043	MAF<0.1
ELOVL7	5q12	rs10440618	60,074,790	T:C	0.422	
		rs1563905	60,078,081	G:A	0.120	
		rs6449493	60,082,253	C:A	0.233	
		rs702564	60,083,731	A:T	0.092	MAF<0.1
		rs7715147	60,088,977	C:A	0.225	
		rs16878412	60,102,601	A:G	0.136	
		rs898622	60,115,938	G:A	0.125	
		rs1870682	60,117,053	T:C	0.100	
		rs6872863	60,123,905	G:A	0.442	
		rs7730827	60,144,681	G:C	0.150	
		rs13159372	60,152,290	C:T	0.083	MAF<0.1
		rs17331746	60,152,370	T:C	0.183	
		rs1807017	60,163,481	A:G	0.458	
		rs6869332	60,165,118	G:A	0.117	
		rs17332108	60,167,641	T:C	0.292	
		rs7731760	60,172,925	T:A	0.145	
CDC14B	9q22	rs6477496	98,296,233	C:T	0.175	
		rs7852498	98,320,242	A:G	0.392	
		rs1411885	98,364,198	T:C	0.433	
		rs10978403	98,427,818	G:A	0.158	
		rs10121511	98,428,445	T:A	0.417	
HEMGN	9q22	rs10760017	99,727,744	G:C	0.333	
		rs4743146	99,727,888	A:T	0.092	MAF<0.1
		rs1059003	99,729,514	T:C	0.158	
		rs10984462	99,747,076	T:G	0.233	
LCN2	9q34	rs6478823	129,946,669	G:A	0.300	
		rs878400	129,947,865	T:C	0.367	
		rs10760543	129,957,400	A:T	0.314	
		rs10987900	129,958,277	C:T	0.225	
GOLGA2	9q34	rs7023913	130,057,783	A:G	0.271	
		rs7871866	130,067,803	G:C	0.164	
		rs2416981	130,081,573	T:C	0.233	

GFI1B	9q34	rs7036106	134,840,395	G:A	0.217
		rs1964196	134,845,241	C:T	0.483
		rs686652	134,845,393	G:A	0.117
		rs8192999	134,848,046	G:C	0.229
		rs2773818	134,849,340	A:C	0.475
		rs3827664	134,850,630	C:G	0.425
		rs2905069	134,850,660	T:C	0.358
		rs2773822	134,850,854	C:T	0.117
		rs649651	134,851,062	C:T	0.108
		rs2073577	134,851,819	T:G	0.375
		rs606141	134,852,300	C:T	0.200
		rs633153	134,852,453	A:G	0.433
		rs8193002	134,852,856	G:A	0.250
		rs8193004	134,855,975	T:C	0.482
		rs667805	134,856,514	T:C	0.150
		rs15906	134,856,788	A:T	0.167
		rs10751499	134,857,246	A:G	0.104
		rs678946	134,857,508	C:T	0.150
		rs1888204	134,857,642	G:C	0.224
		rs621940	134,859,951	G:C	0.167
		rs3011266	134,860,976	C:T	0.205
		rs8193007	134,861,494	C:T	0.127
FAM69B	9q34	rs10117822	138,722,473	T:C	0.442
		rs11793385	138,727,787	C:T	0.225
		rs3739940	138,736,916	C:T	0.119
		rs6874	138,737,657	G:A	0.125
		rs2275161	138,740,516	G:A	0.417
		rs2275160	138,740,989	C:T	0.467
		rs945381	138,741,845	C:T	0.283
TMTC4	13q32	rs8002073	100.056.937	G:T	0,153
		rs12427435	100.057.428	A:G	0,466
		rs9513770	100.057.953	T:C	0,331
		rs9518104	100.060.039	A:G	0,432
		rs1572641	100.062.634	G:A	0,424
		rs1051518	100.073.127	C:T	0,347
		rs9554712	100.075.142	G:A	0,458
		rs9582406	100.078.901	G:C	0,267
		rs7995648	100.079.016	T:A	0,421
		rs17578868	100.079.576	A:G	0,136
		rs9513777	100,101,556	C:G	0.364
		rs17579126	100,104,608	C:T	0.144
		rs1765733	100,108,758	A:G	0.202
		rs9513779	100,112,107	C:T	0.336
		rs9518128	100,119,245	C:T	0.491
		rs9513782	100,122,279	C:A	0.203
		rs1283198	100,122,301	A:G	0.395
		rs946845	100,123,918	A:T	0.102
		rs2791686	100,126,173	C:A	0.475

TTC7B	14q32	rs1294555	90,070,569	C:G	0.367
		rs1294559	90,072,766	T:C	0.283
		rs6575138	90,075,166	T:A	0.233
		rs1294582	90,089,819	T:G	0.433
		rs12889650	90,091,720	C:T	0.482
		rs17793829	90,091,938	C:T	0.144
		rs11620738	90,096,985	G:T	0.433
		rs1286496	90,097,361	G:A	0.333
		rs1286477	90,121,186	G:A	0.127
		rs1286470	90,129,411	A:G	0.203
		rs1286464	90,136,193	T:A	0.331
		rs12432719	90,137,483	T:A	0.125
		rs998338	90,142,590	T:C	0.133
		rs1286459	90,144,903	C:T	0.325
		rs7145137	90,149,063	C:T	0.167
		rs4904708	90,151,119	G:A	0.150
		rs8020106	90,163,647	T:C	0.367
		rs2343	90,164,253	C:T	0.470
		rs9323852	90,165,038	T:C	0.425
		rs942746	90,181,831	T:C	0.458
		rs942748	90,186,623	A:G	0.350
		rs8022840	90,190,570	T:C	0.175
		rs2277510	90,194,640	C:T	0.175
		rs8005454	90,194,687	T:C	0.225
		rs10132961	90,194,863	C:T	0.322
		rs12881399	90,198,242	G:C	0.433
		rs17799388	90,198,527	G:A	0.167
		rs12886545	90,199,131	T:C	0.375
		rs881218	90,200,255	A:G	0.375
		rs1076958	90,201,610	C:T	0.287
		rs10139373	90,205,815	C:T	0.125
		rs4900055	90,207,944	G:A	0.200
		rs4900057	90,208,243	A:G	0.425
		rs753310	90,211,589	G:A	0.415
		rs17799418	90,213,443	T:C	0.200
		rs7152676	90,217,692	G:A	0.331
		rs12147413	90,218,630	A:C	0.183
		rs11629065	90,220,389	G:A	0.142
		rs942738	90,225,904	A:G	0.333
		rs12893100	90,231,816	C:T	0.300
		rs2147829	90,233,622	T:C	0.175
		rs1742100	90,235,359	T:C	0.358
		rs1742098	90,238,170	T:C	0.492
		rs1535321	90,240,579	T:C	0.192
		rs13379124	90,242,791	T:C	0.138
		rs1749715	90,250,913	A:G	0.370
		rs1742084	90,253,595	C:A	0.475
		rs7154098	90,254,691	C:G	0.100
		rs10150862	90,255,198	A:C	0.373
		rs17721032	90,258,121	G:A	0.100
		rs12883490	90,265,125	T:C	0.358

		rs12886812	90,269,540	T:C	0.358	
		rs4904725	90,286,525	C:G	0.442	
		rs9944033	90,296,407	T:C	0.314	
		rs8003019	90,303,488	C:T	0.175	
		rs12894343	90,327,456	A:G	0.297	
		rs1286322	90,338,908	A:G	0.129	
		rs1286305	90,350,817	C:C	0.500	
		rs8016414	90,358,133	G:T	0.308	
		rs2180774	90,360,376	C:G	0.408	
C14orf64	14q32	rs2607049	97,490,243	T:C	0.142	
		rs2604989	97,492,127	A:G	0.158	
		rs754604	97,494,529	A:G	0.356	
		rs899117	97,494,793	G:A	0.439	
		rs3818667	97,496,511	T:C	0.142	
		rs1551540	97,500,909	C:T	0.127	
		rs11628375	97,501,981	C:T	0.283	
		rs2809121	97,504,272	C:T	0.217	
		rs1466439	97,509,237	G:A	0.219	
		rs1038039	97,517,532	G:T	0.125	
PPP2R5C	14q32	rs7143539	101,346,481	G:T	0.110	
		rs2281772	101,365,836	C:G	0.183	
		rs10132483	101,369,173	T:C	0.125	
		rs1678002	101,377,133	T:C	0.100	
		rs3405	101,381,133	C:T	0.442	
		rs1746596	101,385,212	G:A	0.283	
		rs12589350	101,387,005	G:A	0.178	
		rs1678032	101,391,576	C:T	0.110	
		rs1746587	101,394,249	T:C	0.117	
		rs1678019	101,398,698	T:C	0.143	
		rs1677999	101,411,403	T:C	0.200	
		rs1677990	101,418,111	T:C	0.158	
		rs2256537	101,436,766	T:C	0.175	
		rs3783370	101,455,210	C:T	0.100	
		rs11624542	101,468,351	T:C	0.325	
		rs3993391	101,469,047	G:A	0.246	
ANKRD9	14q32	rs3742440	102,040,262	C:T	0.339	
		rs2273905	102,044,752	C:T	0.233	
		rs942024	102,049,216	C:T	0.348	
		rs1007343	102,050,420	C:T	0.092	MAF<0.1
SDC4	20q13	rs6104115	43,383,155	G:A	0.142	
		rs6073708	43,386,291	A:G	0.466	
		rs4599	43,387,821	A:G	0.207	
		rs6073714	43,395,411	G:T	0.217	
		rs2267867	43,399,317	A:G	0.212	
		rs2251252	43,404,771	G:A	0.381	
		rs2267871	43,405,784	A:T	0.227	
		rs2284278	43,406,098	A:G	0.202	
		rs1981430	43,409,081	T:G	0.400	
		rs2072786	43,409,695	C:G	0.300	
		rs1008953	43,414,140	G:A	0.242	

TP53RK	20q13	rs11550540	44,749,176	A:G	0.158	
		rs6012009	44,751,093	C:G	0.092	MAF<0.1
		rs971759	44,754,482	A:G	0.319	
TUBB1	20q13	rs151348	57,022,896	T:C	0.491	
		rs6070696	57,031,040	A:G	0.175	
		rs10485828	57,034,050	G:C	0.200	
		rs151337	57,036,186	C:T	0.322	

MAF: Minor Alele Frequency

These 191 SNPs were genotyped in our complete dataset (Table 4.1). 76% of these SNPs passed our QC requisites and were subjected to statistical analysis (Table 4.16).

The association results and LD plot of genes which displayed significant associations with IS risk (*HEMGN*, *GFI1B*, *TUBB1*, *TMTC4*, *TTC7B* and *SDC4*) are presented in Table 4.17, and in Figures 4.15 to 4.20 (more details in Appendix C: Table C.9). Significant haplotype association results, when exist, are also presented in the same figures and in Appendix C: Table C.10. None of the results, however, survive to the conservative Bonferroni's correction for multiple testing. We did not find any association with IS risk for any SNP or haplotype in the remaining ten genes (*MPL*, *ELOVL7*, *CDC14B*, *LCN2*, *GOLGA2*, *FAM69B*, *C14orf64*, *PPP2R5C*, *ANKRD9* and *TP53RK*), using either χ^2 test or multivariate logistic regressions with backward elimination of risk factors (data not shown). As we studied almost all their tagging SNPs, these negative results suggest that genetic variants in these genes are not important risk factors for IS.

In *HEMGN* and *GFI1B* we found a marginal allelic association (p-value = 0.049) with IS risk for SNPs rs10760017 (upstream of the gene) and rs633153 (intronic), respectively, but no genotypic associations (log-additive model) in unadjusted or adjusted tests for covariates (Figures 4.15 and 4.16; Appendix C: Table C.9). The haplotypes GA and CA defined by block 1 (rs10760017-rs4743146) in *HEMGN* which contains the associated SNP rs10760017, was also moderately associated with IS (GA, p-value = 0.041; CA, p-value = 0.045; Figure 4.15; Appendix C: Table C.10). The SNP rs4743146, with MAF < 0.1 (Table 4.16), has its major allele present in the two combinations. Interestingly, the combination GA is more frequent in cases than in controls (0.640 vs. 0.597), and the combination CA is more frequent in the controls than in cases (0.318 vs. 0.279; Appendix C: Table C.10), in agreement with the allelic association of rs10760017, that suggests that its minor allele (C) confers protection against IS (OR [95% CI] = 0.83 [0.71 - 0.97]; Table 4.17). This is reinforced by the fact that SNP rs10760017 shows a modest unadjusted genotypic association (p-value = 0.023) using the recessive genetic model, with an OR < 1 for the genotype CC (Table 4.17). For the GFI1B gene, the SNP rs633153 shows a modest unadjusted genotypic association with the same p-value (p-value = 0.023) using the recessive genetic model, but with an OR > 1 for the genotype GG. For this SNP the allele G is also the minor allele, but in

this case it confers risk to IS (OR [95% CI] = 1.19 [1.04 - 1.36]; Table 4.17). These generally modest allelic, genotype and haplotype association results obtained for *HEMGN* and *GFI1B*, suggest that their variants may constitute risk factors for IS, but replication in other populations or samples is mandatory.

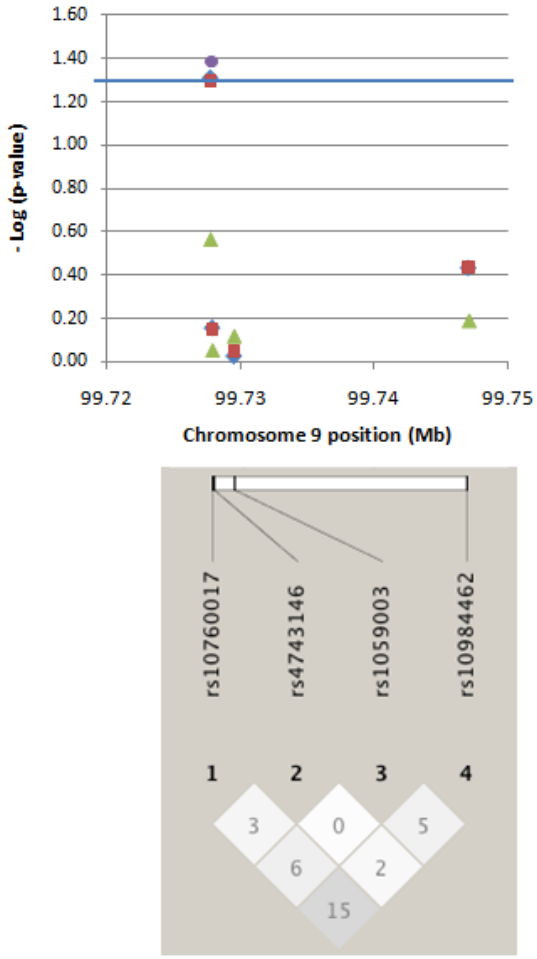


Figure 4.15: Association results and pairwise linkage disequilibrium (LD) for genotyped polymorphisms in *HEMGN*. See legend in Figure 4.1. Associated haplotypes are presented in purple joined circles. The value of the pairwise statistic r^2 is shown inside the diamonds of the LD plot.

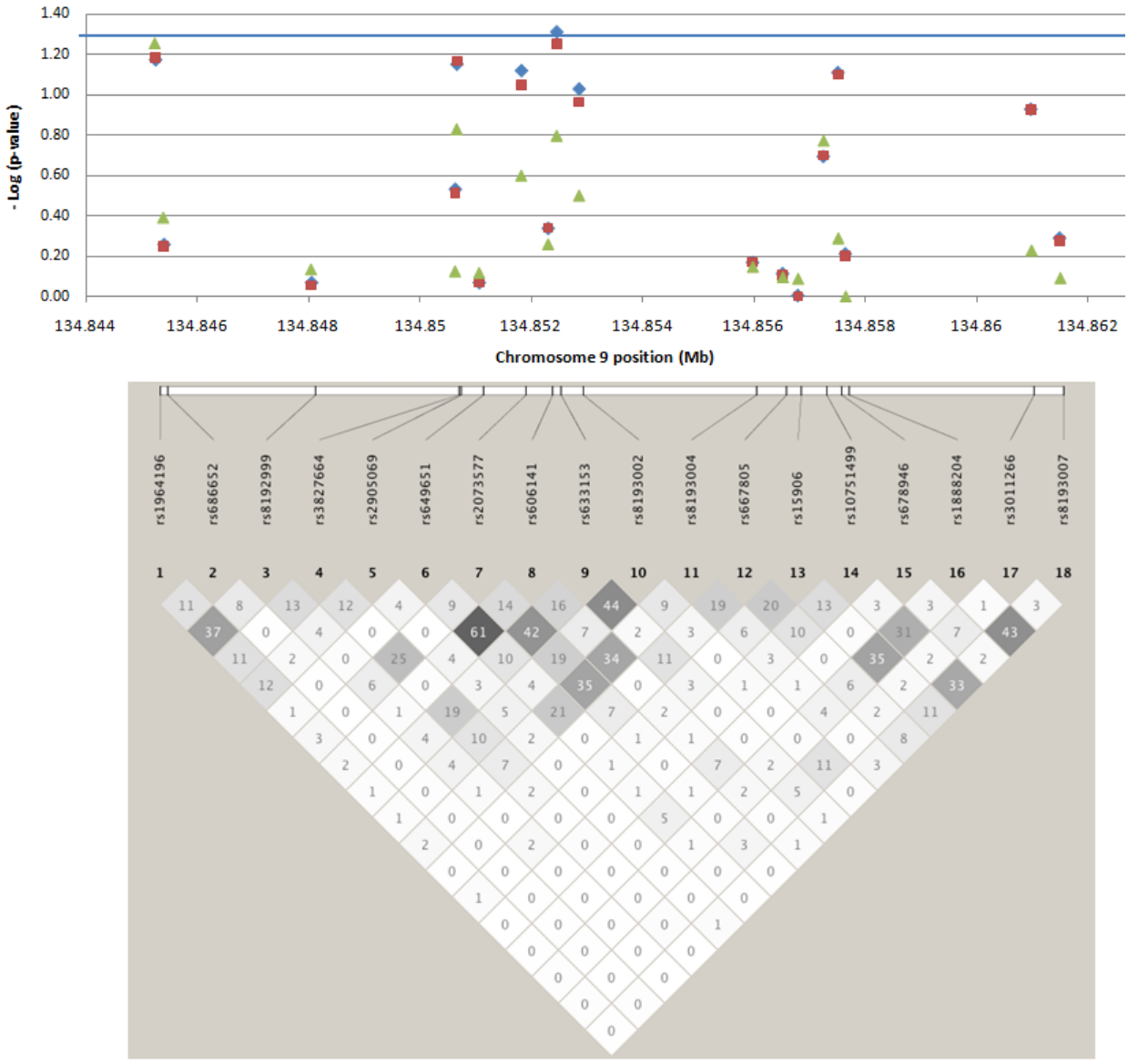


Figure 4.16: Association results and pairwise linkage disequilibrium (LD) for genotyped polymorphisms in *GFI1B*. See legend in Figure 4.1. The value of the pairwise statistic r^2 is shown inside the diamonds of the LD plot.

For *TUBB1*, no association with IS risk was found for single markers in unadjusted tests or haplotypes (Figure 4.17; Appendix C: Table C.9 and Table C.10). However, the upstream SNP rs151348 was modestly associated (log-additive model p-value = 0.021) with IS risk in adjusted genotypic tests for co-variants (Figure 4.17; Appendix C: Table C.9). Additionally, using the dominant genetic model the association persist highly significant in the adjusted test (p-value = 0.002) and marginal (p-value = 0.046) in the unadjusted test for covariates (Table 4.17).

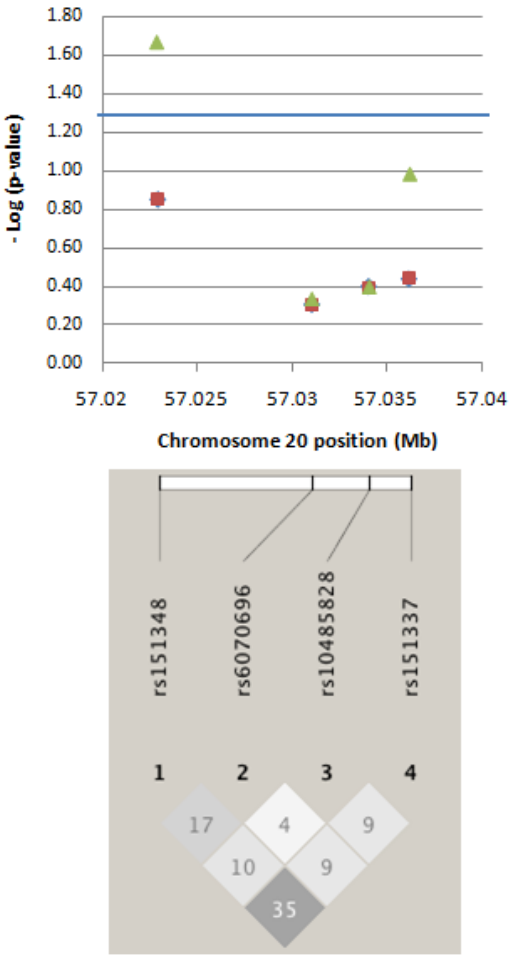


Figure 4.17: Association results and pairwise linkage disequilibrium (LD) for genotyped polymorphisms in *TUBB1*. See legend in Figure 4.1. The value of the pairwise statistic r^2 is shown inside the diamonds of the LD plot.

Intronic SNPs rs9582406 and rs946845 in *TMTC4* were associated with IS risk in unadjusted (allelic p-value = 0.025 and 0.015, and genotypic log-additive model p-value = 0.022 and 0.013, respectively) and adjusted (genotypic log-additive model p-value = 0.050 and 0.047, respectively; Figure 4.18; Appendix C: Table C.9) tests. The genotypic associations remain significant under the recessive model for rs9582406 (p-value = 0.026 and 0.025 for unadjusted and adjusted tests, respectively) and under the dominant model for rs946845 (p-value = 0.015 for unadjusted test; Table 4.17). Analysing the haplotype tests, there is evidence of a modest association of the haplotype CA defined by block 2 (rs9582406-rs7995648) and the AACCC defined by block 3 (rs17578868-rs1765733-rs9513779-rs9518128-rs9513782) with the risk of IS disease (p-value = 0.036 and 0.023, respectively; Figure 4.18; Appendix C: Table C.10). The first haplotype contains the associated SNP rs9582406. Again, the minor

allele of this SNP (C) confers protection against IS (OR [95% CI] = 0.80 [0.68 - 0.94]; Table 4.17), and this allele is present in the associated haplotype that is more frequent in controls than in cases (0.307 vs. 0.266). Haplotypes in the block 2 with the G allele have no association (Appendix C: Table C.10).

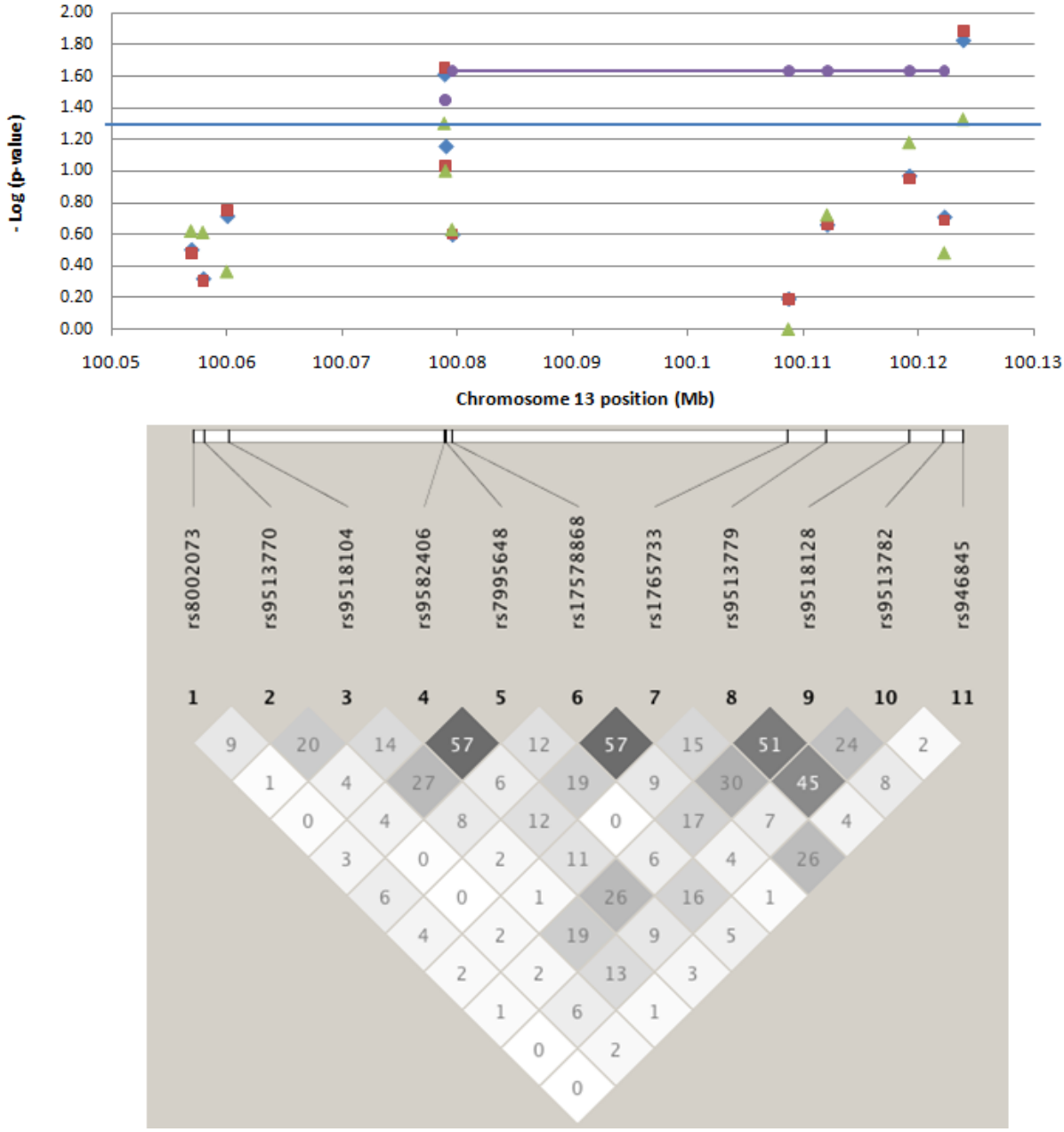


Figure 4.18: Association results and pairwise linkage disequilibrium (LD) for genotyped polymorphisms in *TMTC4*. See legend in Figure 4.1. Associated haplotypes are presented in purple joined circles. The value of the pairwise statistic r^2 is shown inside the diamonds of the LD plot.

Three intronic SNPs in *TTC7B* were associated with IS risk. SNPs rs2343 and rs1535321 are associated in unadjusted tests (allelic p-value = 0.041 and 0.009, and genotypic log-additive model p-value = 0.039 and 0.009, respectively), but not in adjusted genotypic (log-additive model) tests (Figure 4.19; Appendix C: Table C.9). SNP rs11629065 was marginally associated (p-value = 0.050) in the allelic test with IS risk and modestly associated (log-additive model p-value = 0.026) in the adjusted genotypic test (Figure 4.19; Appendix C: Table C.9). As shown in Table 4.17, using the recessive genetic model, the unadjusted genotypic association remains significant for the SNP rs2343 (p-value = 0.027), and both the unadjusted and adjusted genotypic associations become very significant (p-value = 0.004 and 0.008, respectively) for the SNP rs1535321. The association of rs11629065 in the adjusted genotypic test remains modest (p-value = 0.028) using the dominant genetic model, and the unadjusted genotypic test becomes marginal (p-value = 0,046). The SNPs rs11629065 and rs1535321 are also present in the associated haplotype AGG defined by block 7 (rs12147413-rs11629065-rs942738) and TACGTC defined by block 8 (rs12893100-rs1742100-rs1742098-rs1535321-rs13379124-rs7154098) (p-value = 0.015 and 0.012, respectively; Figure 4.19; Appendix C: Table C.10). For rs11629065, the allele A seems to confer risk to IS (OR [95% CI] = 1.28 [1.02 - 1.60]; Table 4.17), which is in agreement with allele G being present in the associated haplotype that is most frequent in controls than in cases (0.286 vs. 0.239; Figure 4.19; Appendix C: Table C.10). For rs1535321, the minor allele G confers risk of IS (OR [95% CI] = 1.33 [1.10 - 1.61]; Table 4.17), and is present in the associated haplotype that is more frequent in cases than in controls (0.221 vs. 0.178; Figure 4.19; Appendix C: Table C.10). The A allele is present in all none associated haplotypes defined by block 8 (Appendix C: Table C.10).

In *SDC4*, we found an association with IS risk for the intronic SNP rs2284278 in all tests performed (allelic p-value = 0.016, unadjusted genotypic log-additive model p-value = 0.020 and adjusted genotypic log-additive model p-value = 0.015), and a modest evidence of association (log-additive model p=0.037) for intronic SNP rs2251252 in the adjusted genotypic test for co-variates (Figure 4.20; Appendix C: Table C.9). For rs2284278, the genotype association results become even more significant using the dominant genetic model (unadjusted p-value = 0.013 and adjusted p-value = 0.006), and for the SNP rs2251252 the observed association for the adjusted genotypic test is also improved using the recessive genetic model (p-value = 0.017; Table 4.17). For this gene, haplotype GGA defined by block 1 (rs6104115-rs6073708-rs4599), that contains different SNPs than the ones that were found associated, was also associated (p-value = 0.011) with the risk of IS (Figure 4.20; Appendix C: Table C.10), being more frequent in cases than in controls (0.211 vs. 0.168; Appendix C: Table C.10).

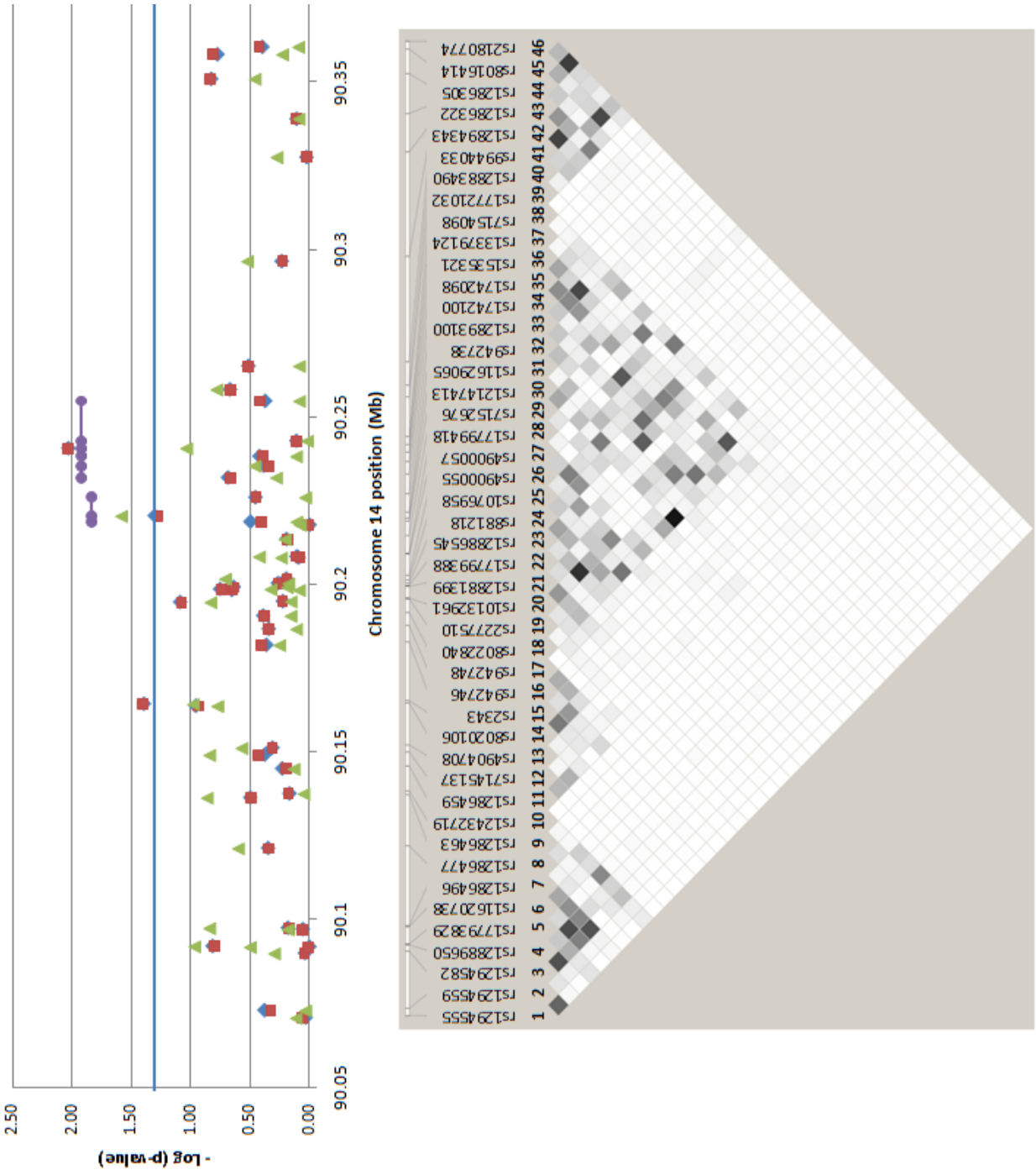


Figure 4.19: Association results and pairwise linkage disequilibrium (LD) for genotyped polymorphisms in *TTC7B*. See legend in Figure 4.1. Associated haplotypes are presented in purple joined circles.

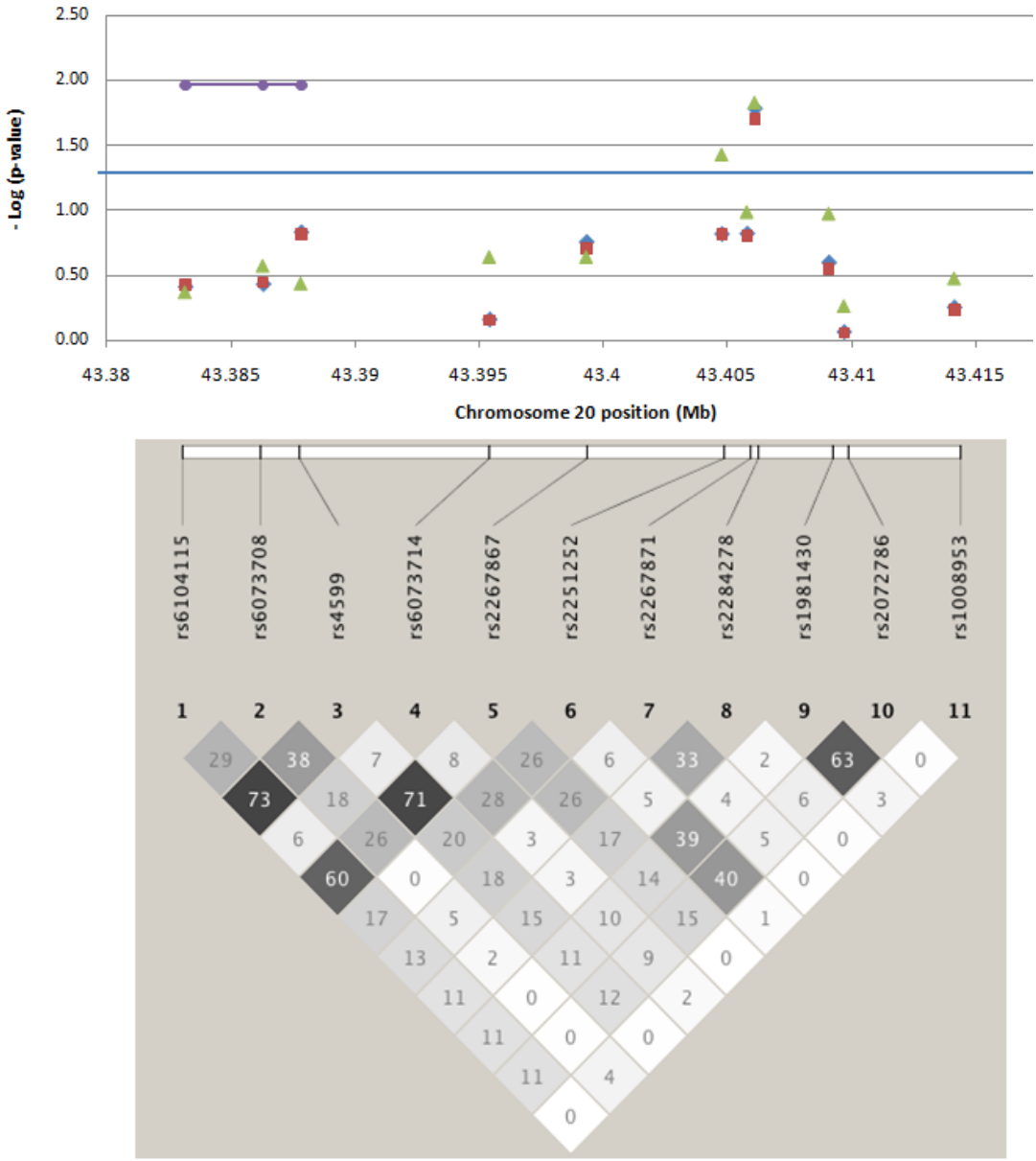


Figure 4.20: Association results and pairwise linkage disequilibrium (LD) for genotyped polymorphisms in *SDC4*. See legend in Figure 4.1. Associated haplotypes are presented in purple joined circles. The value of the pairwise statistic r^2 is shown inside the diamonds of the LD plot.

Overall, these results suggest that genetic variations in *TUBB1*, *TMTC4*, *TTC7B* and *SDC4* may constitute risk factors for IS in the Portuguese population. Given that none of the associations would survive correction for multiple testing, replication in other cohorts must be performed to validate them.

Table 4.17: Detailed association results for associated SNPs in the GC genes. See legend in table 4.3.

Gene	SNP ID/ Model	Unadjusted						Adjusted					
		Cases		Controls		OR [95% CI]	p-value	Cases		Controls		OR [95% CI]	p-value
		N	%	N	%			N	%	N	%		
HEMGN	**rs10760017**												
	Alleles												
	G	782	72.1	698	68.2	1.00	**0.049**						
	C	302	27.9	326	31.8	0.83 [0.71-0.97]							
	Genotypes												
	Dominant												
	G/G	282	52	247	48.2	1.00	0.219	242	51.6	233	49.2	1.00	0.688
	C/G-C/C	260	48	265	51.8	0.86 [0.67-1.09]		227	48.4	241	50.8	0.95 [0.73-1.24]	
	Recessive												
	G/G-C/G	500	92.3	451	88.1	1.00	**0.023**	433	92.3	417	88	1.00	0.076
	C/C	42	7.7	61	11.9	0.62 [0.41-0.94]		36	7.7	57	12	0.66 [0.42-1.05]	
	log-Additive												
	0,1,2					0.83 [0.69-1.00]	0.051					0.89 [0.73-1.09]	0.272
GFI1B	**rs633153**												
	Alleles												
	A	596	55.8	602	60.1	1.00	**0.049**						
	G	472	44.2	400	39.9	1.19 [1.04-1.36]							
	Genotypes												
	Dominant												
	A/A	179	33.5	183	36.5	1.00	0.311	154	33.2	170	36.5	1.00	0.333
	A/G-G/G	355	66.5	318	63.5	1.14 [0.88-1.47]		310	66.8	296	63.5	1.15 [0.87-1.53]	
	Recessive												
	A/A-A/G	417	78.1	419	83.6	1.00	**0.023**	367	79.1	391	83.9	1.00	0.164
	G/G	117	21.9	82	16.4	1.43 [1.05-1.96]		97	20.9	75	16.1	1.28 [0.90-1.81]	
	log-Additive												
	0,1,2					1.18 [1.00-1.40]	0.056					1.15 [0.95-1.38]	0.159
TMTC4	**rs9582406**												
	Alleles												
	G	810	73.6	714	69.1	1.00	**0.025**						
	C	290	26.4	320	30.9	0.80 [0.68-0.94]							
	Genotypes												
	Dominant												
	G/G	300	54.5	255	49.3	1.00	0.088	261	54.8	238	49.7	1.00	0.200
	G/C-C/C	250	45.5	262	50.7	0.81 [0.64-1.03]		215	45.2	241	50.3	0.84 [0.64-1.10]	
	Recessive												
	G/G-G/C	510	92.7	459	88.8	1.00	**0.026**	441	92.6	425	88.7	1.00	**0.025**
	C/C	40	7.3	58	11.2	0.62 [0.41-0.95]		35	7.4	54	11.3	0.59 [0.37-0.94]	
	log-Additive												
	0,1,2					0.81 [0.67-0.97]	**0.022**					0.82 [0.67-1.00]	**0.050**
	rs946845												
	Alleles												
	T	1018	90,89	967	93,7	1,00	**0,015**						
	A	102	9,107	65	6,298	1.49 [1.09-2.03]							
	Genotypes												
	Dominant												
	T/T	461	82,3	452	87,6	1,00	**0,015**	402	82,7	418	87,4	1,00	0,058
	T/A-A/A	99	17,7	64	12,4	1.52 [1.08-2.13]		84	17,3	60	12,6	1.43 [0.99-2.08]	
	Recessive												
	T/T-T/A	557	99,5	515	99,8	1,00	0,345	483	99,4	477	99,8	1,00	0,323
	A/A	3	0,5	1	0,2	2.77 [0.29-26.69]		3	0,6	1	0,2	3.04 [0.29-31.89]	
	log-Additive												
	0,1,2					1.51 [1.09-2.1]	**0,013**					1.44 [1.00-2.06]	**0,047**

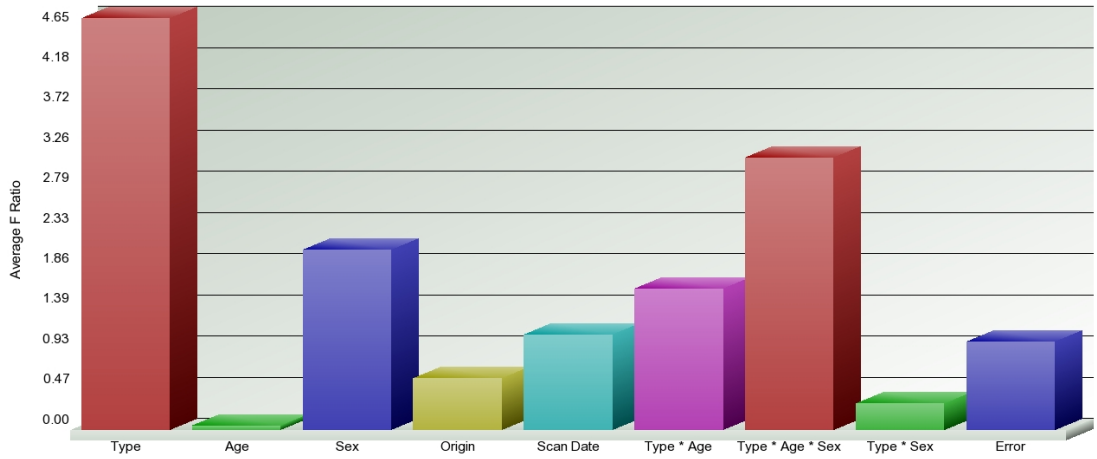
TTC7B	**rs2343**												
	Alleles												
	C	714	64.9	612	60.5	1.00	**0.041**						
	T	386	35.1	400	39.5	0.83 [0.72-0.95]							
	Genotypes												
	Dominant												
	C/C	233	42.4	194	38.3	1.00	0.183	199	41.6	180	38.5	1.00	0.319
	C/T-T/T	317	57.6	312	61.7	0.85 [0.66-1.08]		279	58.4	288	61.5	0.87 [0.66-1.14]	
	Recessive												
	C/C-C/T	481	87.5	418	82.6	1.00	**0.027**	416	87	384	82.1	1.00	0.071
	T/T	69	12.5	88	17.4	0.68 [0.48-0.96]		62	13	84	17.9	0.71 [0.49-1.03]	
	log-Additive												
	0,1,2					0.83 [0.70-0.99]	**0.039**					0.85 [0.71-1.03]	0.105
	rs11629065												
	Alleles												
	G	915	84,1	885	87,11	1,00	**0,050**						
	A	173	15,9	131	12,89	1.28 [1.02-1.60]							
	Genotypes												
	Dominant												
	G/G	386	71	388	76,4	1,00	**0,046**	333	70,6	360	76,6	1,00	**0,028**
	A/G-A/A	158	29	120	23,6	1.32 [1.00-1.74]		139	29,4	110	23,4	1.41 [1.04-1.91]	
	Recessive												
	G/G-A/G	529	97,2	497	97,8	1,00	0,536	458	97	459	97,7	1,00	0,296
	A/A	15	2,8	11	2,2	1.28 [0.58-2.82]		14	3	11	2,3	1.56 [0.68-3.59]	
	log-Additive												
	0,1,2					1.27 [1.00-1.62]	0,053					1.35 [1.04-1.76]	**0,026**
	rs1535321												
	Alleles												
	A	844	77,57	838	82,16	1,00	**0,009**						
	G	244	22,43	182	17,84	1.33 [1.10-1.61]							
	Genotypes												
	Dominant												
	A/A	333	61,2	341	66,9	1,00	0,056	294	62,4	317	66,9	1,00	0,373
	A/G-G/G	211	38,8	169	33,1	1.28 [0.99-1.65]		177	37,6	157	33,1	1.14 [0.86-1.50]	
	Recessive												
	A/A-A/G	511	93,9	497	97,5	1,00	**0,004**	442	93,8	462	97,5	1,00	**0,008**
	G/G	33	6,1	13	2,5	2.47 [1.28-4.75]		29	6,2	12	2,5	2.55 [1.25-5.22]	
	log-Additive												
	0,1,2					1.33 [1.07-1.64]	**0,009**					1.22 [0.97-1.55]	0,092
SDC4	**rs2251252**												
	Alleles												
	G	646	57.9	567	54.8	1.00	0.151						
	A	470	42.1	467	45.2	0.88 [0.78-1.00]							
	Genotypes												
	Dominant												
	G/G	182	32.6	159	30.8	1.00	0.512	164	33.9	146	30.5	1.00	0.253
	A/G-A/A	376	67.4	358	69.2	0.92 [0.71-1.19]		320	66.1	333	69.5	0.85 [0.64-1.13]	
	Recessive												
	G/G-A/G	464	83.2	408	78.9	1.00	0.076	407	84.1	378	78.9	1.00	**0.017**
	A/A	94	16.8	109	21.1	0.76 [0.56-1.03]		77	15.9	101	21.1	0.66 [0.47-0.93]	
	log-Additive												
	0,1,2					0.88 [0.74-1.05]	0.153					0.82 [0.68-0.99]	**0.037**

	rs2284278												
	Alleles												
	A	855	77,59	753	73,11	1,00	**0,016**						
	G	247	22,41	277	26,89	0.79 [0.66-0.93]							
	Genotypes												
	Dominant												
	A/A	340	61,7	279	54,2	1,00	**0,013**	303	63,5	260	54,4	1,00	**0,006**
	G/A-G/G	211	38,3	236	45,8	0.73 [0.57-0.94]		174	36,5	218	45,6	0.68 [0.52-0.90]	
	Recessive												
	A/A-G/A	515	93,5	474	92	1,00	0,368	447	93,7	441	92,3	1,00	0,507
	G/G	36	6,5	41	8	0.81 [0.51-1.29]		30	6,3	37	7,7	0.84 [0.50-1.41]	
	log-Additive												
	0,1,2					0.8 [0.66-0.97]	**0,020**					0.77 [0.62-0.95]	**0,015**
TUBB1	**rs151348**												
	Alleles												
	C	556	49,64	544	52,82	1,00	0,142						
	T	564	50,36	486	47,18	1.14 [1.01-1.28]							
	Genotypes												
	Dominant												
	C/C	131	23,4	148	28,7	1,00	**0,046**	109	22,4	143	30	1,00	**0,002**
	C/T-T/T	429	76,6	367	71,3	1.32 [1.00-1.74]		377	77,6	334	70	1.60 [1.18-2.16]	
	Recessive												
	C/C-C/T	425	75,9	396	76,9	1,00	0,700	366	75,3	367	76,9	1,00	0,490
	T/T	135	24,1	119	23,1	1.06 [0.80-1.40]		120	24,7	110	23,1	1.12 [0.82-1.52]	
	log-Additive												
	0,1,2					1.14 [0.96-1.35]	0,140					1.25 [1.03-1.51]	**0,021**

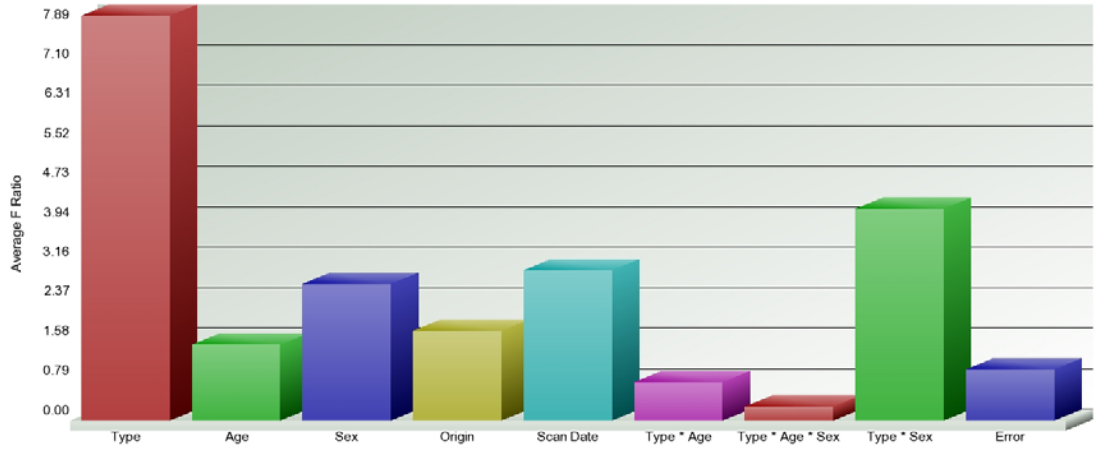
OR: Odds ratio; CI: Confidence interval

It is also interesting that, all of the 6 GC genes that we found associated, had, in our expression studies, a major contribution of the factor type (the affection status) to the overall variance among IS cases and controls, even before we removed the geographic origin and the scan date batch effects (Figure 4.21). All of these genes presented high values in the ANOVA's F ratio for the factor type. Only for the *SDC4*, the F ratio of the geographic origin factor is comparable of the F ratio of the type (Figure 4.21 e)). In Figure 4.22 are shown the expression results obtained for these genes, discriminated by type (cases and controls).

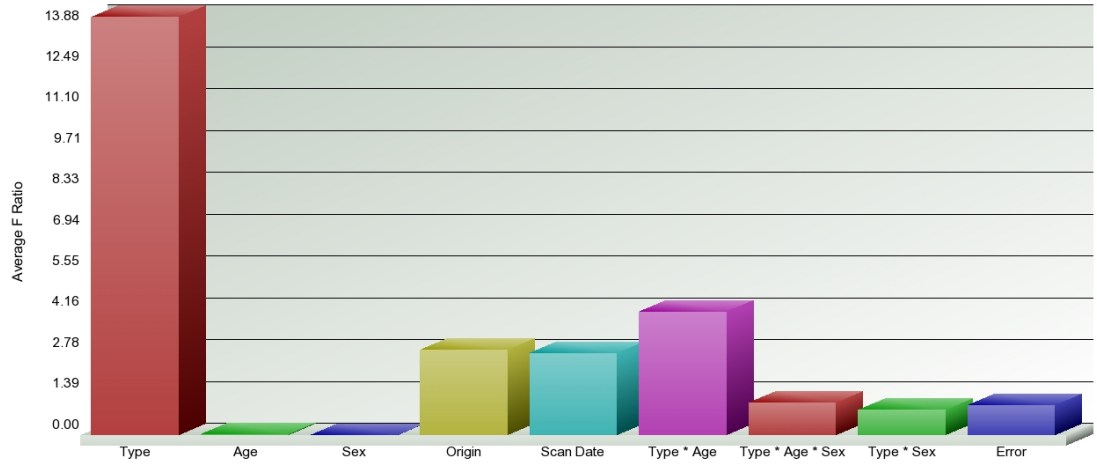
a)


Average F Ratio
4.65
4.18
3.72
3.26
2.79
2.33
1.86
1.39
0.93
0.47
0.00
Type
Age
Sex
Origin
Scan Date
Type * Age
Type * Age * Sex
Type * Sex
Error

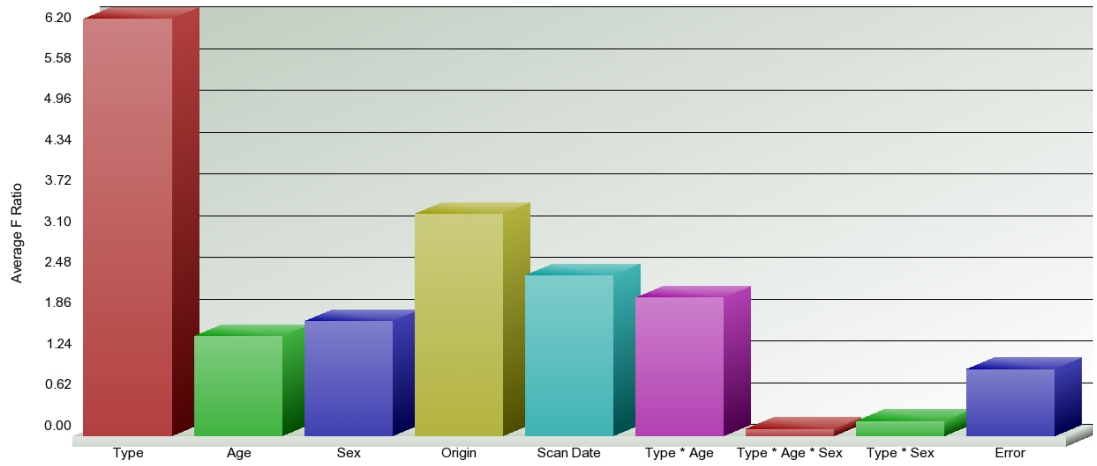
b)


Average F Ratio
7.89
7.10
6.31
5.52
4.73
3.94
3.16
2.37
1.58
0.79
0.00
Type
Age
Sex
Origin
Scan Date
Type * Age
Type * Age * Sex
Type * Sex
Error

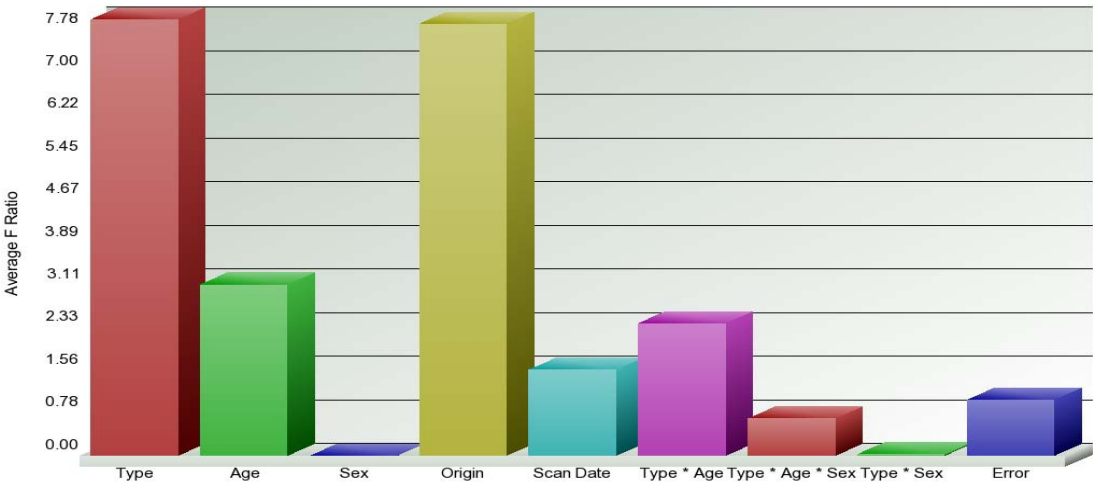
c)


Average F Ratio
13.88
12.49
11.10
9.71
8.33
6.94
5.55
4.16
2.78
1.39
0.00
Type
Age
Sex
Origin
Scan Date
Type * Age
Type * Age * Sex
Type * Sex
Error

d)



e)



f)

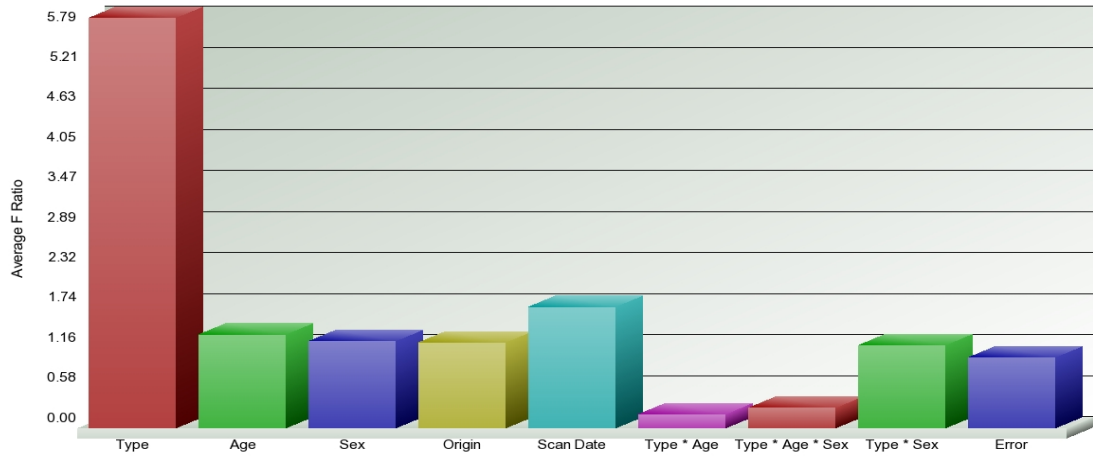


Figure 4.21: Analysis of variance of CG genes. These plots display the average F ratio of the different ANOVA factors (origin, type, age, sex, scan date, and combinations among type, sex and age), before the batch-remove step (origin and scan-date effects), for a) *HEMGN*, b) *GFI1B*, c) *TMTC4*, d) *TTC7B*, e) *SDC4*, and f) *TUBB1*. The plots were constructed using the Partek software.

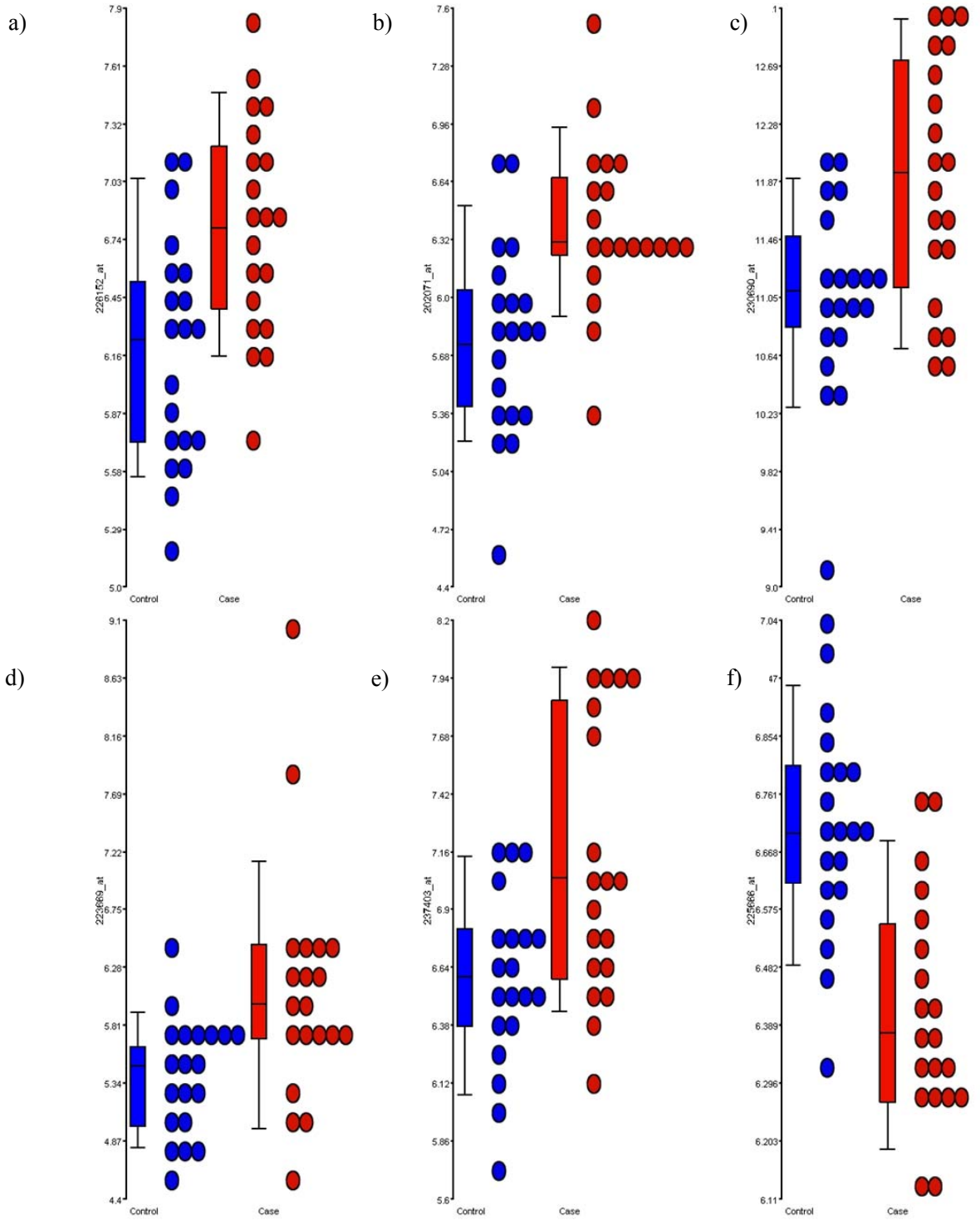


Figure 4.22: Expression results of GC genes. Log_2 transformed expression values obtained for the a) *HEMGN*, b) *GFI1B*, c) *TMTC4*, d) *TTC7B*, e) *SDC4*, and f) *TUBB1* in the microarray profiling experiment. Controls are in blue and cases in red. Each circle represents an individual and box-and-whisker graphs are presented for cases and controls. Graphs were constructed using the Partek software.

4.5. Follow up of the association findings on GC genes in a Spanish cohort

Since our positive association findings do not withstand the conservative Bonferroni's multiple testing correction, it is mandatory to confirm them in an independent population to support the role of these variants and their respective genes in the study of the IS disease. To validate our positive association findings in GC genes, we tested the association in a Spanish cohort of SNPs with single low-stringency significance association (p-value < 0.05), and some SNPs defining associated haplotypes.

In the studied Spanish case-control sample (Table 4.18), gender, hypertension, diabetes, dyslipidemic status, and cigarette smoking were observed at a significantly different frequency among IS cases and controls. On the other hand, age did not differ significantly between the two groups. Since the gender and cigarette smoking are correlated (correlation factor near 0.5; data not shown), the covariates in the adjusted analyses in this dataset were hypertension, diabetes, cigarette smoking, and dyslipidemic status.

Table 4.18: Spanish sample characterization. Principal demographic, clinical and lifestyle characteristics of the IS case-control study sample.

Characteristic	Cases	Controls	p-value[a]
N	570	387	
Sex (n/N, % male)	309 (54.2 %)	183 (47.3 %)	0.035
Age-at-examination (mean ± SD, years)	71.9 ± 12.5	71.7 ± 7.0	0.329
Totast (n/N[b], %)			
Cardioembolic	216/561 (38.5 %)	-	-
Atherothrombotic	171/561 (30.5 %)	-	-
Lacunar	173/561 (30.8 %)	-	-
Undetermined	1/561 (0.2 %)	-	-
Other	0/561 (0.0 %)	-	-
Risk factors (n/N[b], %)			
Hypertension (> 140-85 mmHg)	336/568 (59.2 %)	171/387 (44.2 %)	$< 10^{-4}$
Diabetes	154/568 (27.1 %)	77/387 (19.9 %)	0.011
Ever smoking	86/568 (15.1 %)	147/387 (37.9 %)	$< 10^{-4}$
Dyslipidemia	135/568 (23.8 %)	114/387 (29.5 %)	0.049

[a]p-value of an unpaired Student's t test or a χ^2 test for a quantitative and qualitative data, respectively

[b]Number of individuals for whom the data was available

The Spanish cases were classified according to the TOAST subtype classification system, and 38.5% are cardioembolic, 30.5% are atherothrombotic, 30.8% are lacunar, and 0.2% were undetermined. Among the twenty genotyped SNPs (Appendix C: Table C.11), the genotyping procedures for the SNP rs1535321 of the *TTC7B*, and the SNP rs151348 of the *TUBB1*, did not work properly and consequently their results are not presented.

We found significant associations with IS risk in the *TMTC4*, *TTC7B* and *SDC4* for allelic and genotypic tests as presented in Table 4.19 (and Appendix C: Table C.11). Significant haplotype

association results are also presented in the Table C.13 of the Appendix C. These results do not withstand Bonferroni's correction to multiple testing.

We did not find any significant association with overall IS risk for the studied variants in *HEMGN* or *GFI1B* (Appendix C: Table C.11). The SNP rs10760017 in *HEMGN* presented an adjusted genotypic dominant test p-value of 0.048 for cardioembolic stroke, and the SNP rs633153 of the *GFI1B* an adjusted genotypic dominant test p-value of 0.047 for atherothrombotic stroke in the Spanish sample, however, these results are also very marginal and the same association is not verified even for the allelic tests (data not shown). The fact that none of the variants that were found marginally associated with IS in the Portuguese population for the *HEMGN* and *GFI1B* (Table 4.17 and Figures 4.15 and 4.16) were replicated in the Spanish population for any of the tests performed, suggests that these variants may not be risk factors for IS and that our findings were false-positives.

On the other hand, the SNP rs9582406 in *TMTC4* was associated with IS risk in unadjusted tests (allelic p-value and genotypic log-additive model p-value = 0.019) in the Spanish case-control sample (Table 4.19; Appendix C: Table C.11). However, its minor allele (C) that confers protection against IS in the Portuguese population (OR [95% CI] = 0.80 [0.68 – 0.94]; Table 4.17), now confers risk of IS in the Spanish population (OR [95% CI] = 1.28 [1.07 - 1.54; Table 4.19). The genotypic association remains modest using the dominant genetic model (p-value = 0.033 for the unadjusted test) in this last sample (Table 4.19). Regarding the analysis of the subtypes of stroke in the Spanish population, the SNP rs9582406 also presents significant association results with the risk of atherothrombotic and lacunar forms of stroke for allelic and genotypic tests ($0.011 < \text{p-value} < 0.049$; Appendix C: Table C.12). These results reinforce the importance of this variant in *TMTC4* (or of other variant(s) in high LD with it), in the IS risk, at least in the Iberian population. The observed opposite direction of the associated effects of the SNP rs9582406 may be due its interaction with other risk factors, that can be genetic or environmental.

For *TTC7B* and *SDC4*, none of the single SNPs associated with IS in the Portuguese population were replicated in the Spanish one. However, new significant associations with the risk of IS appear in the *TTC7B* SNPs rs1742098 and rs7154098 in adjusted genotypic (log-additive model) test (p-value = 0.031) and unadjusted tests (allelic p-value = 0.033, and genotypic log-additive model p-value = 0.038) respectively (Table 4.19; Appendix C: Table C.11). The same is verified for the *SDC4* SNP rs6073708 in all tests performed ($0.027 < \text{p-value} < 0.029$; Table 4.19; Appendix C: Table C.11). Both rs7154098 (unadjusted p-value = 0.022) and rs6073708 (unadjusted p-value = 0.006 and adjusted p-value = 0.017) become even more significant using the recessive genetic model (Table 4.19). These last two SNPs present also significant association results for the risk of cardioembolic stroke in the Spanish population for unadjusted genotypic (log-additive model) test (p-value = 0.029), and for all tests performed ($0.008 < \text{p-value} < 0.016$) respectively (Appendix C: Table C.12). SNPs rs1742098, rs7154098 and rs6073708 are part of the haplotype TACGTC defined by block 8 (rs12893100-rs1742100-rs1742098-rs1535321-

rs13379124-rs7154098) in *TTC7B*, and haplotype GGA defined by block 1 (rs6104115-rs6073708-rs4599) in *SDC4* that were found associated in the Portuguese population, being more frequent in cases than in controls (Figures 4.19 and 4.20; Appendix C: Table C.10). Nevertheless, the reconstructed haplotypes (for the *TTC7B*, the block 8* was reconstructed without the SNP rs1535321) are not significantly associated with IS in the Spanish population (Appendix C: Table C.13). Only different allelic combinations for the same blocks present modest associations. The haplotype CATTG defined by block 8* (rs12893100-rs1742100-rs1742098-rs13379124-rs7154098) (p-value = 0.034) is more frequent in controls than in cases (0.130 vs. 0.099), and the GAA defined by block 1 (rs6104115-rs6073708-rs4599) (p-value = 0.032) is more frequent in cases than in controls (0.608 vs. 0.559; Appendix C: Table C.13).

Finally, although they are not associated with IS in the Spanish population, is interesting to refer that the SNP rs2343 and rs12893100 of the *TTC7B* present significant associations with the risk of atherothrombotic stroke and cardioembolic stroke respectively for some of the tests performed, and that SNPs rs2251252 and rs9582406 of the *SDC4* present significant associations with the risk of atherothrombotic stroke and lacunar stroke respectively for all tests performed ($0.011 < \text{p-value} < 0.049$; Appendix C: Table C.12). This heterogeneity of the significantly associated SNP variants for *TTC7B* and *SDC4* among the Portuguese and the Spanish populations highlights the importance of the most detailed study of these genes and surrounding regions.

Table 4.19: Detailed association results for SNPs in GC genes associated with IS in the Spanish population. The allelic and genotype frequencies in cases and controls as well as the association results for the allelic and genotype association tests are shown here. Unadjusted (without co-variates) and adjusted (for hypertension, diabetes, dyslipidemic status and cigarette smoking) genotype association testing was performed using different models (dominant, recessive and log-additive). Significant p-values are bolded and the respective odds ratio and 95% confidence intervals are indicated.

Gene	SNP ID/ Model	Unadjusted						Adjusted					
		Cases		Controls		OR [95% CI]	p-value	Cases		Controls		OR [95% CI]	p-value
		N	%	N	%			N	%	N	%		
TMTC4	**rs9582406**												
	Alleles												
	G	802	71.0	587	75.8	1.00	**0.019**						
	C	328	29.0	187	24.2	1.28 [1.07-1.54]							
	Genotypes												
	Dominant												
	G/G	286	50.6	223	57.6	1.00	**0.033**	285	50.6	222	58	1.00	0.086
	C/G-C/C	279	49.4	164	42.4	1.33 [1.02-1.72]		278	49.4	161	42	1.27 [0.97-1.67]	
	Recessive												
	G/G-C/G	516	91.3	364	94.1	1.00	0.113	514	91.3	360	94	1.00	0.323
	C/C	49	8.7	23	5.9	1.50 [0.90-2.51]		49	8.7	23	6	1.30 [0.77-2.22]	
	log-Additive												
	0,1,2					1.28 [1.04-1.58]	**0.019**					1.22 [0.98-1.52]	0.076

TTC7B	**rs1742098**												
	Alleles												
	T	679	59.7	491	63.4	1.00	0.097						
	C	459	40.3	283	36.6	1.17 [1.01-1.36]							
	Genotypes												
	Dominant												
	T/T	209	36.7	157	40.6	1.00	0.231	209	36.9	155	40.5	1.00	0.091
	C/T-C/C	360	63.3	230	59.4	1.18 [0.90-1.53]		358	63.1	228	59.5	1.27 [0.96-1.68]	
	Recessive												
	T/T-C/T	470	82.6	334	86.3	1.00	0.122	468	82.5	330	86.2	1.00	0.059
	C/C	99	17.4	53	13.7	1.33 [0.92-1.91]		99	17.5	53	13.8	1.44 [0.98-2.11]	
	log-Additive												
	0,1,2					1.17 [0.97-1.40]	0.102					1.24 [1.02-1.51]	**0.031**
	rs7154098												
	Alleles												
	C	1024	90.1	677	87.0	1.00	**0.033**						
	G	112	9.9	101	13.0	0.73 [0.56-0.96]							
	Genotypes												
	Dominant												
	C/C	461	81.2	299	76.9	1.00	0.108	460	81.3	296	76.9	1.00	0.201
	C/G-G/G	107	18.8	90	23.1	0.77 [0.56-1.06]		106	18.7	89	23.1	0.80 [0.58-1.12]	
	Recessive												
	C/C-C/G	563	99.1	378	97.2	1.00	**0.022**	561	99.1	374	97.1	1.00	0.058
	G/G	5	0.9	11	2.8	0.31 [0.11-0.89]		5	0.9	11	2.9	0.35 [0.11-1.07]	
	log-Additive												
	0,1,2					0.74 [0.56-0.98]	**0.038**					0.78 [0.58-1.04]	0.095
SDC4	**rs6073708**												
	Alleles												
	A	693	60.9	436	55.9	1.00	**0.029**						
	G	445	39.1	344	44.1	0.81 [0.71-0.94]							
	Genotypes												
	Dominant												
	A/A	201	35.3	125	32.1	1.00	0.292	200	35.3	125	32.4	1.00	0.189
	A/G-G/G	368	64.7	265	67.9	0.86 [0.66-1.14]		367	64.7	261	67.6	0.82 [0.62-1.10]	
	Recessive												
	A/A-A/G	492	86.5	311	79.7	1.00	**0.006**	490	86.4	308	79.8	1.00	**0.017**
	G/G	77	13.5	79	20.3	0.62 [0.44-0.87]		77	13.6	78	20.2	0.64 [0.45-0.92]	
	log-Additive												
	0,1,2					0.81 [0.67-0.98]	**0.027**					0.80 [0.66-0.98]	**0.028**

For *TTC7B* a joint analysis of the Portuguese and Spanish datasets strengthens the previous findings when the recessive model is used (Figure 4.23; Appendix C: Table C.14). In the joint analysis with both datasets, the co-variates were hypertension, diabetes, ever smoking, and sample origin. SNP rs7154098 was out of HWE in the combined controls (p-value = 0.010) and was not tested further. However, SNPs rs12893100 and rs1742098 were associated with IS in the unadjusted (p-value = 0.013 and p-value = 0.020, respectively) and adjusted (p-value = 0.003 and p-value = 0.007, respectively) recessive models (Figure 4.23; Appendix C: Table C.14), and the ACT haplotype defined by block 8** (rs1742100-rs1742098-rs13379124) was also associated (p-value = 0.046; Figure 4.23; Appendix C: Table C.15).

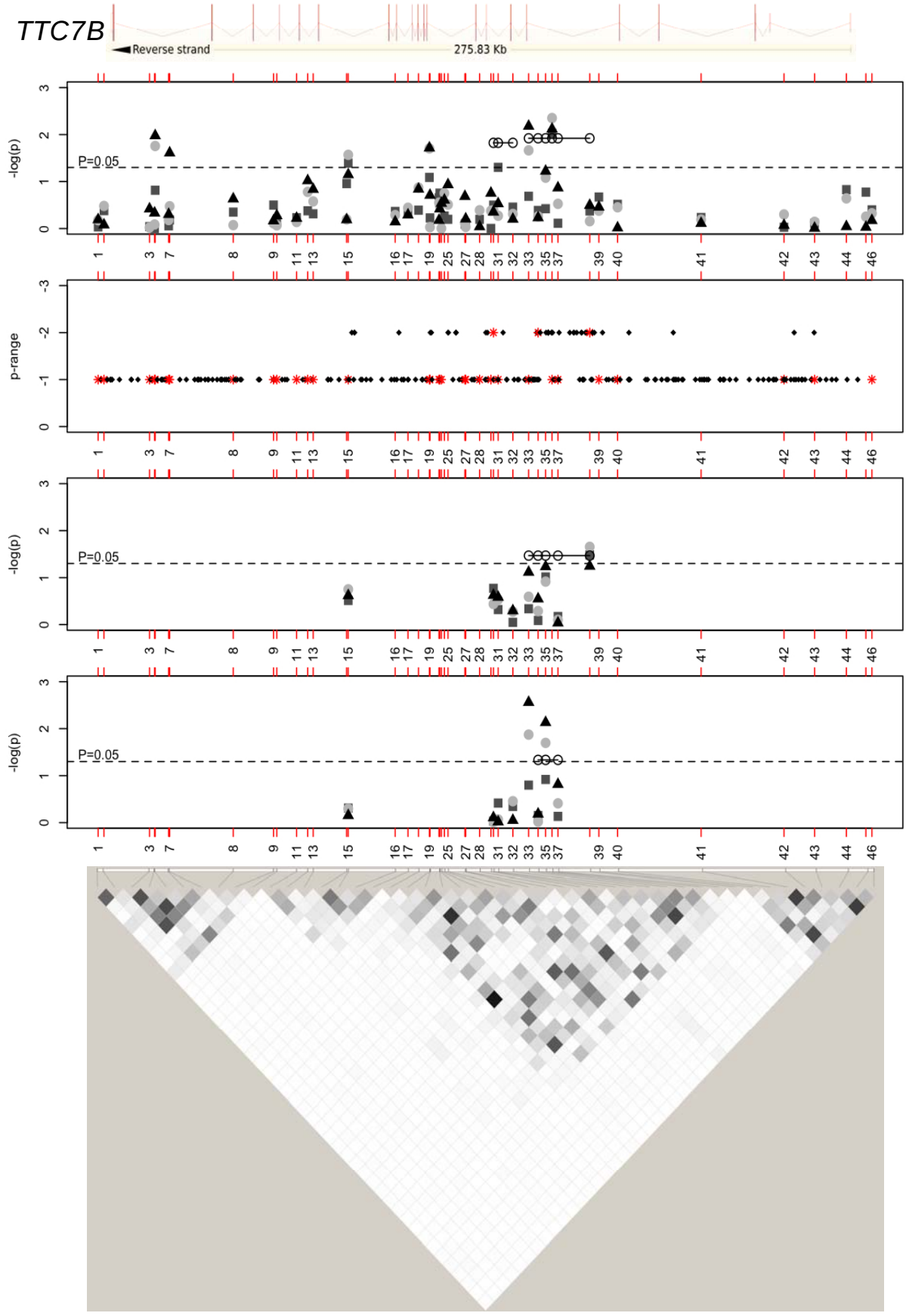

TTC7B
Reverse strand
275.83 Kb
-log(p)
P=0.05
p-range
-log(p)
P=0.05
-log(p)
P=0.05

Figure 4.23: Association results and pairwise linkage disequilibrium (LD) among all genotyped polymorphisms for TTC7B. The diagram on top schematically represents the introns and exons of the TTC7B gene (ENST00000357056 transcript) relative to the genotyped polymorphisms. The first, third and fourth plots display the association results for the recessive model in the Portuguese, Spanish, and combined cohorts, respectively. Allelic (grey squares), crude (light grey discs) and adjusted (black triangles) genotypic (recessive model), as well as the top haplotype (open circles) association results are shown. The second plot depicts the association (p-range) of all SNPs studied in the Ikram *et al.* (2009) GWAS. The stars indicate polymorphisms that were investigated in the Portuguese cohort. The LD plot at the bottom represents the LD between genotyped SNPs in the Portuguese dataset (white-to-black gradient shading within each diamond represents the magnitude of LD using the pairwise statistic r^2).

It would be important to genotype in a second stage the SNP rs1535321 of the *TTC7B* that was the most significantly associated SNP in this gene in the Portuguese population, and the SNP rs151348 of the *TUBB1* to follow up our findings in this gene also.

4.6. Follow up of the association findings on GC genes in the Ikram and colleagues IS GWAS

None of the SNPs studied in our GC genes present a clear association (p-value < 0.01) with IS in the IS GWAS performed by Ikram *et al.* (2009). However, the rs633153 of the *GFI1B* and the rs2284278 of the *SDC4* that were positively associated in our Portuguese sample (Figures 4.16 and 4.20 and Table 4.17), display a marginal association (0.01 < p-value < 0.1) with IS in this GWAS (data not shown). SNPs rs12147413, rs1742100 and rs7154098 of the *TTC7B* that are part of its associated haplotypes shown a marginal association in the IS GWAS performed by Ikram and colleagues (2009) too (Figure 4.23).

More precisely, analysis of the 259 polymorphisms investigated in the Ikram *et al.* (2009) IS GWAS and lying in the TTC7B gene region, reveals a marginal association (0.01 < p-value < 0.1) of 42 SNPs localized between SNPs rs2343and rs1286322, with a higher density (24 SNPs) between SNPs rs12893100 and rs12883490 (Figure 4.23). This region of increased association (between SNPs rs2343and rs1286322) in both the Portuguese, Spanish and multinational cohort of Ikram *et al.* (2009) shows an increased level of LD (Figures 4.19 and 4.23) and maps to the central region of the *TTC7B* locus.

Any of the genes prioritized by GC that were not associated with IS in our Portuguese population (*MPL*, *ELOVL7*, *CDC14B*, *LCN2*, *GOLGA2*, *FAM69B*, *C14orf64*, *PPP2R5C*, *ANKRD9* and *TP53RK*), display also significant associations with IS reported by Ikram and co-authors (2009) for any of their SNPs (data not shown).

5. DISCUSSION

The discovery of novel genes that interact with each other and with environmental factors to produce the final stroke phenotype is both timely and essential to reduce stroke's huge public health burden. The recent development of new genetic, molecular, statistical and bioinformatics tools makes possible an in depth study of stroke genetics and genomics. Candidate gene association studies have led to many unreplicated and, in many cases, spurious associations that have being published. Consequently, a number of guidelines should be followed in the study of complex disorders. The sample size of the biobanks is of major importance, as well as the replication of the significant results in an independent sample. Most individual genetic effects in complex diseases are small, emphasising the need of these guidelines to detect them consistently (Dahlman *et al.* 2002). Very few genes are capable of exerting large effects.

We tested the association of some novel biological candidate genes and genes derived from the unbiased and multi-faceted GC approach in the Portuguese population and followed up our significant results in a Spanish case-control sample of similar size in a second stage. We also compared our results with those of recent and well-powered GWAS conducted on four European and American cohorts (Ikram *et al.* 2009). Additionally, we tried to replicate the association of *PDE4D* and *ALOX5AP* for the first time in the Portuguese population.

5.1. Advantages of Portuguese and Spanish populations in genetic association studies

Unlike rare Mendelian disorders, complex diseases affect a substantial proportion of the general population, and therefore heterogeneity of the selected samples has been an issue in genetic association studies. The human population is quite large and relatively old, has undergone a number of stratifications and divisions and is settled in environments that vary widely. When population stratification occurs, the genes under study could show marked variation in allele frequency across subgroups of the population. The forces that dictate population subdivisions, environmental changes, migrations, and other forces are often difficult to characterize and are therefore usually passed off as random or stochastic factors that affect populations (Schork et al 1997, Thomas *et al.* 2002). Consequently, the risk of stroke could vary over different populations and ethnic groups as it varies with established environmental factors (Flossmann *et al.* 2004).

The Portuguese and Spanish populations are *per si* relatively homogenous providing advantages in the genetic study of stroke. Portugal and Spain are relatively small and isolated countries that have suffered mostly emigration events. For the purposes of candidate gene association studies, Caucasian of

European origin are generally regarded as having a relatively homogeneous gene pool (Cavalli-Sforza *et al.* 1994). Concordantly, in our results, as referred for the additionally genotyped SNPs of the *PDE4D* and *ALOX5AP*, we always get similar allelic frequencies in the control groups of samples comparing to the presented ones in the HapMap project for the European population, as well as similar association results of association among SNPs that are described in almost complete LD in the HapMap project too. This makes possible that, although we were attempting to identify stroke risk factors in the Portuguese population, our results are expected to be applicable, at least, to other Caucasian populations. The same has previously been observed in other diseases such as AD where polymorphisms in the ApoE gene were initially found to be associated with increased risk in the American population (Corder *et al.* 1993), and subsequently these results were confirmed in all other tested populations (Liddell *et al.* 1994; van Duijn *et al.* 1994; Dai *et al.* 1994). This avoids some disadvantages of focusing on small or relatively isolated populations in which the biological relevance to larger populations may be compromised. In addition, in very small populations or those with relatively few founders, the genetic basis of disease susceptibility is likely to have a low number of high frequency polymorphisms or mutations contributing to disease susceptibility (McVean 2002). Another advantage in using a Caucasian population is that the LD extends over large genomic regions, even up to the scale of 100 kb and, consequently, the number of studied polymorphisms per loci needed to capture all the existing variation is lower than in other populations (Huttley *et al.* 1999, Abecasis *et al.* 2001, Reich *et al.* 2001).

There are also several methods that have been developed to test and correct for potential population stratification in association studies using genetic markers to obtain information on population diversity among cases and controls (Lewis *et al.* 2002, Thomas *et al.* 2002). However, in candidate gene studies, correcting for population structure (independently of self-reported origin) requires that a manageable set of "ancestry-informative markers" has been determined *a priori*. In Portuguese datasets there are no genome-wide association scans or population genetic studies with very large numbers of autosomal markers that have been performed so far allowing this kind of correction. We did not consider population substructure to be an issue in this project.

5.2. The design of the studied case-control samples

Even that Portuguese and Spanish populations are relatively homogenous, it continues to be of the major importance to decrease the genetic heterogeneity of the studied case-control samples to avoid false positive results in association studies. Well designed case-control studies are mandatory.

Consequently, to decrease the genetic heterogeneity in our sample and increase our power to detect susceptibility genes, we only collected samples from self-described Portuguese Caucasian

participants, and we focused on the more common form of stroke, IS, which represents 80 - 90% of all strokes. The studied IS patients have no signs of any simple Mendelian disorders predisposing to IS. For the construction of the used Spanish sample, only Spanish Caucasian participants and patients with IS were collected too. The samples from the Spanish cohort were originated mostly in Catalonia.

It was verified that in our Portuguese sample, the mtDNA haplogroup distribution in the control group was in agreement with previously published data on a similar Portuguese normal population, with 8.3–9.9% of the individuals having mtDNA haplogroups characteristic of African populations (L and U6). The fact that almost all haplogroup categories are present in equivalent proportions in cases and controls further suggests that our dataset is well matched for ethnicity and lacks significant substructure (Rosa *et al.* 2008).

It is of the major importance to ensure a good match between the genetic background of cases and controls, so that the genetic differences between them are related to the studied disease and not to biased sampling. Besides the same ethnic group, more subtle genetic differences can be guarded against by collecting more information, such as the status for the well established risk factors (Tables 4.1 and 4.18) about the controls, to account the observed effects in the analysis of the results.

As referred, in our Portuguese sample, given that stroke is such a prevalent disease and TIAs are silent and to avoid mis-classification bias, we selected older controls and, consequently, the age of the participants among the cases and the controls is not balanced. Therefore, age was not included as a co-variate in the association tests as it was biased by design and not as a result of random selection of participants. Inclusion of age as a co-variate would erroneously correct the association results as it seems to be a protective factor in our cohort. Conversely, we selected the cases as young as possible since the early-onset forms of the disease are more likely to be genetic in origin as opposed to being merely a consequence of aging and/or exposure to environmental stimuli (Schork *et al.* 1997, Flossmann *et al.* 2004). We assume that the studied SNPs are not age-dependent. In the association analyses, we did not correct the results for sex because it is strongly correlated with the ever smoking status. We used all the available male controls in our biobank that fullfiled our inclusion criteria, but unfortunately these were not enough to match the proportion of male cases (Table 4.1).

Other common well established risk factors for stroke that are usually used as confounding variables like the body mass index, anticoagulant medication, cardiac disease, and family history, were not included in the analyses due their incomplete characterization in the studied biobanks. Furthermore, the use of many co-variates also increases the number of degrees of freedom of the analyses, being disadvantageous. Therefore, for the Portuguese sample and Spanish samples, we focused our attention on the best characterized and most common stroke risk factors.

It is also important to guarantee the random selection of the collected samples by consecutive collection. As a QC measure, control genotypes should also be in HWE, proving that the selection of the

controls was performed randomly and that they are sufficiently large in size. The fact that the great majority of studied SNPs in our studied Portuguese and Spanish biobanks are in HWE, suggests this random selection of the controls. At the same time, and although this test has low power to predict that, it can indicate that we have not a heterogeneous population, otherwise the controls might have different allele frequencies of their genotypes (Lewis *et al.* 2002).

5.3. The power of the case-control studies and the genotyping errors

In the context of diseases with polygenic basis, well designed case-control association studies provide more statistical power, than family-based linkage studies (Gershon et al 1986, Risch *et al.* 1996). Beside the sample size, the power of association studies is strongly dependent on the size of their effect, and on the frequency of the polymorphisms studied. That is why we selected SNPs with MAF in HapMap CEU higher than 0.1.

It is also of the major importance to ensure the quality of the genotyping assays. To control for potential genotyping errors, extensive quality control measures were implemented. However, although we used duplicate individuals and observed Mendelian inconsistencies on three large families (Figure 3.6), systematic genotyping error can occur and the error rates are underestimated if they are not detected. The detection of Mendelian inconsistencies decreases, for instance, with increasing of SNPs heterozygoty. The genotyping errors can originate false results as well as bias in the LD measurements (Akey *et al.* 2001). A possible way to identify and overcome these genotyping problems is looking for the HWE of the studied variants. Although it has generally low power to predict genotyping errors, its power increases for high rates of genotyping error and/or large sample sizes (Leal *et al.* 2005).

Assuming the random selection of the controls in our studied case-control samples (that is proved by the verification of the HWE in great majority of studied SNPs), the fact that some SNPs that are not in HWE in the controls could reflect genotyping problems, such as genotypes consistently mis-called or specific genotypes giving missing values. For these reason, we eliminated these SNPs from our subsequent analyses in our QC procedures. However, the deviation of the HWE could also result from random chance (1 of every 20 tested SNPs will give a p-value < 0,05 by chance) (Lewis *et al.* 2002). Critical SNPs out of HWE in controls can be genotyped in a different platform to exclude potential genotyping problems, as we proceed for the SNP rs13075202 of the *KALRN* (Table 4.4). As stated by Shoemaker *et al.* (1998), a population will never be exactly in HWE. The assumptions underlying HWE, including random mating, lack of selection according to genotype, and absence of mutation or migration, are rarely completely met in human populations (Shoemaker *et al.* 1998, Ayres *et al.* 1999, Zou et al 2006).

5.4. The *PDE4D* and *ALOX5AP* association with IS risk

PDE4D and *ALOX5AP* have been controversially implicated in the risk of IS. In this project we assessed their association with IS for the first time in the Portuguese population. We presented one of the most comprehensive association studies conducted on these two genes after the original reports, covering several of the SNPs that were initially found associated with stroke in the Icelandic population and that have been replicated in some other cohorts, as well as many tagging SNPs in Caucasians mapping to the most associated regions of these genes. Additionally, although the data was not shown here, in collaboration with Dr. Joan Montaner, we also tested the association of top SNPs in *PDE4D* and *ALOX5AP* in the Portuguese and Spanish datasets jointly (1,092 IS patients and 781 healthy controls). Theses genes had never been tested for association in the Iberian population (Domingues-Montanari *et al.* 2010). Only a Japanese study from Nakayama *et al.* (2006) and an American study from Song *et al.* (2006) have also examined *PDE4D* more extensively (Table 1.2).

5.4.1. *PDE4D*

In the Portuguese dataset, among all single SNPs studied in the *PDE4D*, only SNP rs7442640 was found associated in genotypic tests adjusted for covariates, and none of the positive associations published in original report were replicated (Figure 4.1 and Table 4.3). However, our analyses were limited to markers in the 5' region of the gene, since this was where previous association signals were detected (Gretarsdottir *et al.* 2003, Song *et al.* 2006). We cannot exclude the possibility of association with variants in other regions of the *PDE4D* or in untested variants.

A joint analysis of the Portuguese and Spanish datasets, detected no association of five SNPs (SNP41, SNP45, SNP56, SNP87, and SNP89) with the risk of IS, even if some SNPs were not in HWE (Domingues-Montanari *et al.* 2010). The same was verified in the results of the meta-analysis we also performed in collaboration with Dr. Joan Montaner, which includes new recently published studies and our Iberian data (Domingues-Montanari *et al.* 2010). Bevan *et al.* (2008a) also showed in a meta-analysis that none of their studied genetic variants presented a strong and consistent association to IS (Table 1.2), suggesting that the reported associations in *PDE4D* may be restricted to specific populations. All significant associations they obtained became non significant when excluding the results from the original study. When analysing only white individuals, no associations with stroke were found. In the same way, Rosand *et al.* (2006) examined all 11 reported nominally significant replicated SNPs in *PDE4D* until 2006 and examined their correlation with the at-risk haplotype G0 in HapMap CEPH trios. They found that no SNP was strongly correlated with the G0 haplotype and concluded that reported associations were

unlikely to be true replication given this lack of correlation. Although the data presented here can not totally exclude a link between *PDE4D* and IS, it suggests that any association that may exist is likely to be marginal and possibly restricted to specific populations.

We did not study the association with stroke subtypes as performed in the original report and in some replication studies (Table 1.2) because we have not this classification for the majority of Portuguese IS samples. In general, although the findings were not always consistent, all studies that examined the subgroups of stroke type found significant p-values even when no associations were found for overall IS. Specifically, in Iceland (Gretarsdottir *et al.* 2003), the stronger associations were found for carotid and cardiogenic subtypes of stroke (Table 1.2). These findings implicated *PDE4D* in the atherosclerotic process associated with these subtypes of stroke, although recent publications do not support this hypothesis. Song *et al.* (2006), for instance, have demonstrated an effect of the *PDE4D* locus on stroke risk in young adults, a population with a low prevalence of atherosclerotic disease. Their most strongly associated SNP, rs918592, was associated with multiple IS subtypes (Table 1.2), and other significant findings were found not only for atherosclerotic stroke, but also for lacunar stroke and non-lacunar stroke of undetermined aetiology which includes large artery strokes without evidence for a significant degree of proximal atherosclerosis.

5.4.2. *ALOX5AP*

Regarding *ALOX5AP*, the present study suggests that there are sequence variants in the gene significantly associated with the risk of IS that can constitute risk factors for the disease in the Portuguese population (Figure 4.2 and Table 4.3) even though they are not the same as the variants reported originally by Helgadottir and colleagues (2004). In the same way, different at-risk haplotypes were also found between the British and Icelandic study populations (Helgadottir *et al.* 2004). These findings suggest a modest but significant association of the *ALOX5AP* in the Portuguese sample, even if none of the found positive associations would survive to Bonferroni's correction for multiple testing.

In the joint analysis of Portuguese and Spanish datasets, we found that the T allele of the SNP rs10507391 (or SG13S114), which contributes to the definition of the associated HapA haplotype, was associated with IS risk (Domingues-Montanari *et al.* 2010) as described in the original report (Helgadottir *et al.* 2004; Table 1.3). This SNP was out of the HWE in the Portuguese population (Table 4.2), but becomes in the equilibrium with the enlargement of the dataset. Identical results have been described in the German population with the T allele of this polymorphism associated with a higher risk of IS (Lohmussaar *et al.* 2005, Table 1.3).

For other side, Kaushal *et al.* (2007) also report an association of the *ALOX5AP* with the risk of stroke, but do not replicate in the American white population the SNP rs10507391 (Table 1.3). The MAF

of this SNP in their population was quite different, indicating that the white American and the white European population frequencies are not comparable for this marker. This could also explain the fact that two other studies in American white populations did not observe an association of the SNP rs10507391 with IS (Meschia *et al.* 2005, Zee *et al.* 2006b, Table 1.3).

Other studies published a replication of the association of *ALOX5AP* with IS, but did not study or publish the results obtained for this single SNP rs10507391, investigating in some cases only the effect of the haplotype HapA (Helgadottir *et al.* 2005, Kostulas *et al.* 2007, Bevan *et al.* 2008b, Lövkvist *et al.* 2008 Table 1.3). In order to definitively elucidate the contribution of the variant rs10507391 on the risk of stroke in white populations, a new meta-analysis was also performed, including all data available in the literature of white ancestry case-control studies and our results for the Portuguese and Spanish populations (3,318 case alleles and 2,923 control alleles, respectively), confirming that the SNP rs10507391 is independently associated with IS, in concordance with the results obtained in our population (Domingues-Montanari *et al.* 2010).

All these findings suggest that variants in *ALOX5AP* may constitute genetic risk factors for IS. Additionally, Bevan *et al.* (2008b) showed that the associated variants may contribute differently to the several IS subtypes.

5.4.3. Justification of the published conflicting results

The presented results in this project and in other publications suggest that Icelandic population may not be typical of other populations and highlight the importance of the replication of the association results in independent populations to exclude spurious findings. The Icelandic population is a relatively isolated population which may have strong founder effects. Population differences in allele and haplotype frequencies as well as LD structure contribute to the observed discrepancy between the published studies, as among the Portuguese and Icelandic ones for the *PDE4D*. Lohmussaar el al., for instance, observed less pronounced LD between SNPs in their European population and the SNPs in the Icelandic population (Lohmussaar *et al.* 2005). These results were confirmed by Kuhlenbäumer *et al.* (2006) that found a close concordance of the allele frequencies for a common SNP also in another German cohort. Consequently, the PDE4D gene variation may play a substantial role in stroke in Iceland, but a lesser or a absent role in non-Icelandic populations.

For both *PDE4D* and *ALOX5AP* genes, most of the published results have differed in the positively associated SNP(s), and the associated allele is sometimes opposite to the allele associated in the Icelandic population. Not all published studies presented information on allele, genotype and at-risk haplotype frequencies, LD structure, and risk estimates, thus making a direct comparison and informative interpretation across studies difficult. Additionally, the interaction with other risk factors (genetic or

environmental) could be significantly different across the studied populations. Apart from the possibility of publication bias, the divergence regarding the studies might reflect unknown gene–gene or gene–environment interactions that differ between populations. This may be explained by variations between the study populations, including age, male and females ratio, hypertension status, other vascular risk factors, and cultural and ethnic factors that could be important to take into account. The non-replication of results in different populations may also be exacerbated by an insufficient number of SNPs studied per gene of interest, or by studies that ascertain patients using different clinical criteria.

5.5. The association of the biological candidate genes with IS risk

Nowadays, with the power of the novel GWAS, it starts to make no sense to select candidate genes only regarding their function or association with other diseases or conditions. It seems to be much more reasonable to try to replicate the most significant associations obtained in published GWAS for the studied diseases in different none studied populations to further study the involvement of the corresponding genes to the risk of developing them. Nevertheless, in this project, we selected and analysed our candidate genes before the publication of the IS GWAS performed by Ikram *et al.* (2009).

5.5.1. *KALRN*

We highlight for the first time an association of the Kalirin gene with risk for IS (Krug *et al.* 2010). Given that *KALRN* has already been implicated in susceptibility to cardiovascular disorders and some of its risk factors (e.g. T2D) (Wang *et al.* 2007, Rudock *et al.* 2008, Ikram *et al.* 2009), we propose that Kalirin may constitute a novel genetic risk factor for vascular phenotypes.

Kalirin is an extremely complex gene generating many alternative transcripts under the control of several promoters in a tissue-specific and developmentally regulated manner (McPherson *et al.* 2004, Johnson *et al.* 2000, Ma *et al.* 2001), encoding multidomain and multifunctional proteins (McPherson *et al.* 2002). A number of recent studies have demonstrated that Kalirin-7, the most abundant isoform in adult brain (Penzes *et al.* 2000), is involved in dendritic spine development, plasticity and stability (Penzes *et al.* 2008). This potentially implicates Kalirin-7 in psychiatric and neurodegenerative disorders such as schizophrenia and AD. In fact, Kalirin-7 gene transcripts are underexpressed in AD hippocampal specimens (Youn *et al.* 2007a), and underexpression of Kalirin-7 increases iNOS activity in cultured cells and correlates with increased iNOS activity in AD hippocampus (Youn *et al.* 2007b). Interestingly, Kalirin is known to associate with iNOS in vitro and in vivo and to prevent iNOS dimerization, and therefore iNOS activity (Ratovitski *et al.* 1999). NO, a potent cell-signalling, effector and vasodilator molecule

characterized by its strong reactivity and diffusibility, is produced in a complex and tightly controlled process by three NO synthases: iNOS, nNOS, and eNOS. Cytokines, endotoxin, or other proinflammatory stimuli induce iNOS expression in virtually all tissues, while nNOS and eNOS are constitutively expressed primarily in neural tissue and endothelium, respectively, which are the principal tissues involved in stroke. Several published case-control studies describe an association of genetic variants in *eNOS* with IS risk (Berger *et al.* 2007). Kalirin-7 is involved in the regulation of ischemic signal transduction (Beresewicz *et al.* 2008) and may play a neuroprotective role during inflammation of the CNS by inhibiting iNOS activity (Ratovitski *et al.* 1999). It would therefore be interesting to assess the role played by genetic variation in Kalirin in stroke severity and recovery, as well as epistatic effects between *KALRN* and NO synthases in IS risk.

Several independent *KALRN* polymorphisms have now been associated in previous studies (Wang *et al.* 2007, Rudock *et al.* 2008, Ikram *et al.* 2009) with vascular phenotypes (Table 4.4 and Figure 4.3), and therefore it is not clear which is(are) the exact susceptibility variant(s). Kalirin appears to be a pleiotropic protein, and it is possible that different variants in the same gene cause related phenotypes. Alternatively, the causal polymorphism may not have been directly studied yet, as suggested by our imputation approach. Given that we had genotyped a relatively high number of haplotype tagging SNPs in the KALRN gene region, we were able to impute hundreds of neighbouring polymorphisms with high confidence and identified many SNPs with stronger association than those directly genotyped (Figure 4.4). Genome-wide imputation is currently performed routinely in whole-genome association studies to increase power, but "local" imputation may also be a powerful approach to detect novel associations when there is a dense map of observed genotypes. We validated this local imputation approach in our Portuguese dataset through direct genotyping (Table 4.6). Another possibility is that *KALRN* demonstrates an even stronger association with a pathogenic mechanism (e.g. inflammation) or subtype (e.g. small vessel) of vascular disease that has not yet been tested directly or corrected for in the statistical analyses. This hypothesis is supported by the significant increase in the evidence for linkage at 3q13 (LOD=5.10, p=0.008) in GENECARD families with lower-risk lipid profiles and fewer known risk factors (Shah *et al.* 2006). We did not test the association of *KALRN* with stroke subtypes in our Portuguese population due the incomplete information about the IS subtype of our samples.

Although our results would not survive the most conservative Bonferroni's correction, replication in multiple independent datasets remains the gold-standard of association studies, even for modest associations. Our study highlights the importance of replicating associations that fall below the genome-wide significance threshold (Krug *et al*. 2010). These variants may account for part of the missing heritability. The *KALRN* association is promising because of the multiple independent lines of evidence for association, namely the association of several uncorrelated SNPs, the replication of the most associated SNP results (in particular the SNP rs11712039) on the Ikram *et al.* (2009) GWAS published study, and the

previous reports of association with cardiometabolic syndrome (Wang *et al.* 2007, Rudock *et al.* 2008). Additional replication studies in other independent datasets with stroke and vascular pathologies are needed to further confirm *KALRN*'s role in vascular disease susceptibility.

5.5.2. *CFH*

For *CFH*, our results suggest that the Y402H polymorphism is a risk factor for IS in the Portuguese population. Conversely, Zee *et al.* (2006c) reported that no association was found between this SNP and risk of IS (p-value = 0.52) in a male sample of 235 cases and 235 controls from the US. Since only one polymorphism in the CFH gene was studied, it is possible that other SNPs or haplotypes are involved in stroke susceptibility. Using the International HapMap Project data (NCBI B35, 30 CEU family trios), there are 15 tagging SNPs (rs1329421, rs11799595, rs403846, rs529825, rs1329428, rs12127759, rs2019727, rs12405238, rs482934, rs10922096, rs424535, rs7524776, rs6695321, rs1576340, rs419137) with minor allele frequency of 0.10 (pairwise tagging with $r^2>0.8$) in *CFH* that could be tested in more detail. We focused our attention on rs1061170 because extensive haplotype analyses by other groups (Klein *et al.* 2005, Haines *et al.* 2005, Edwards *et al.* 2005) suggested that this SNP was the most strongly associated with AMD, but it is possible that another SNP in this gene is a risk factor for IS. However, it is important to take in account that, although some of the presented haplotype tagging SNPs were marginally associated in the IS GWAS performed by Ikram *et al.* (2009), none of them show a substantial association that justifies a more in depth study of this gene.

5.5.3. *EPO*, *HO2* and *KLK1*

The *EPO*, *HO2* and *KLK1* are fairly small and we studied all their tagging SNPs (as determined from HapMap data on CEU individuals). Consequently, our results suggest that these neuroprotective genes are not risk factors for IS in the Portuguese population. Additionally, none of the studied tagging SNPs was found associated with IS risk in the GWAS performed by Ikram *et al.* (2009), reinforcing our earlier findings.

5.6. Discovering of new susceptibility genes for IS risk using the CG approach

Face to the conflicting results that typically emerge from candidate gene association studies, to identify novel stroke susceptibility genes we used the multifactorial GC approach (Hauser *et al.* 2003) for the first time in the field of stroke genetics. Combining genome wide linkage studies, expression analysis

and association studies, this approach has the tremendous advantage of being unbiased. Although it is now commonly accepted that the best strategy to search for variants associated with complex diseases is a GWAS followed by resequencing of the regions of interest, the cost and requirement of very large number of participants makes this endeavour possible only within the frame of an international collaboration. Determining the inherited component of IS risk requires converging lines of evidence from various methodologies, from the most recent GWAS to the traditional expression analyses and association studies.

5.6.1. The use of PBMCs for the study of stroke

The hypothesis underlying expression studies is that genes differentially expressed among cases and controls in a relevant tissue suggest their involvement in the disease process. Blood constitutes a clinically relevant tissue since it is readily accessible. It is biologically relevant for stroke since the complex immune and homeostatic responses to the vascular injury that cause the stroke event are likely to be reflected in the expression profiles of circulating PBMCs. Although there are significant interindividual variations of gene expression patterns in PBMCs (Whitney *et al.* 2003), these differences are expected to be smaller than differences between patients and healthy controls. PBMCs are the key players in the development of innate and adaptive immune responses.

According to Frank *et al.* (2003), gene expression profiles of circulating leukocytes might be used as surrogate markers for diseases that are not primarily associated with peripheral blood. Tang *et al.* (2001), for instance, propose that the blood genomic response could serve as a finger-print for several medical and neurological diseases. B lymphocytes had already been used to verify that the expression levels of PDE4D were altered in stroke patients relatively to controls (Gretarsdottir *et al.* 2003). Expression studies in blood samples have been performed for other diseases such as sickle cell disease (Jison *et al.* 2004), rheumatoid arthritis (Bovin *et al.* 2004, Olsen *et al.* 2004), multiple sclerosis and systemic lupus erythematosus (Bennett *et al.* 2003, Baechler *et al.* 2003, Han *et al.* 2003, Bomprezzi *et al.* 2003, Iglesias *et al.* 2004, Mandel *et al.* 2004), and lead to novel results. Furthermore, large epidemiologic studies have demonstrated correlations between increased production of markers of systemic inflammation in blood samples and future cardiovascular events, including MI, CHD (Danesh *et al.* 2000) and stroke (Di Napoli *et al.* 2001).

5.6.2. The time of collection of the IS case blood samples

For our expression analysis, the affected individuals were not sampled in the acute phase of their last IS stroke event, but rather long after the stroke has been resolved. The genetic risk factors for another

stroke still exist. Unlike previous studies published by Moore *et al.* (2005a and 2005b) and Tang *et al.* (2006) that analysed the expression patterns of the PBMCs in the acute phase of IS or in the first months after the stroke event, addressing the severity and/or recovery mechanisms more than the risk of a stroke event in the first place (also highlighting the potential of these cells for the study of stoke), we intend to understand which genetic specifically increase the risk for a stroke event.

Consequently, as we intend to assess only the risk of IS, we collected the patient blood samples at least 6 months after the first and only IS to minimise the possibility of the PBMCs expression be affected by the acute ischemic attack. We estimated that 6 months may be a sufficient time window to restore the expression profiles of the PBMCs after the stroke event and its characteristic subsequent remodelling and repair. Moreover, with this time window, we should avoid any kind of ischemic precondictioning that could be reflected in the expression profile of the PBMCs of the affected participants in the first days or weeks after the stroke event, masking the expression of some genes that can confer risk of suffering IS. Ischemic precondictioning is reported as the process of tolerance in the brain by which exposure to sub-threshold insults of ischemia provides protection, or tolerance, against the injurious effects of a strong period of ischemia. Is a powerful adaptive defence that involves an endogenous program of neuroprotection (Stenzel-Poore *et al.* 2003, Stenzel-Poore *et al.* 2007). Studies in animal models suggest that after precondictioning, the transcriptional response of the cortical region of the brain to ischemic injury was largely suppressed and, consequently, previous TIAs are associated with better clinical outcome after subsequent stroke. Gene expression changes leading to an analogous state that is refractory to ischemic injury seem to be elicited (Stenzel-Poore *et al.* 2003, Stenzel-Poore *et al.* 2007). If existing in humans after an IS, this ischemic tolerance should result of a cell reprogramming that could be reflected in the PBMCs of the IS patients, but that should also have a peak that diminishes over the time as in the brain of animal models. In animals this neuroprotection peaks at 3 days and diminishes over the course of 1 week (Stenzel-Poore *et al.* 2007).

5.6.3. Problems on sample manipulation in gene expression studies

Several reported studies describe that technical aspects of blood sampling, isolation of cellular components, and RNA isolation techniques, could significantly affect the obtained gene expression patterns. Consequently, in this project, several measures had been taken in account to avoid these variations, such the use of the same techniques from the blood sampling to RNA isolation, as well as the same equipments whenever it is possible.

Additionally, it is reported that time delay to analysis of the gene profiles reveal striking impact on the obtained patterns. Samples with a higher time of delay not always present RNA degradation, but are characterized by a cellular response to stress conditions such hypoxia and hypothermia, and if these

changes differ among the case and control groups, significant biological differences related with the studied diseases might be pronounced or masked (Debey *et al.* 2004). Baechler *et al.* (2004), for instance, have explored the use of peripheral blood cells as a readily source of material for gene expression analyses in human diseases, focusing their study in rheumatoid arthritis and systemic lupus erythematosus. Their experiments identify many genes that are sensitive to *ex vivo* incubation of the samples, suggesting caution in their handling and in the treatment of the results. Variants as the time of incubation, but also the differences in peripheral blood cell populations and circadian influences on gene expression, have significant influence on the expression data. In delayed analysed samples the authors notice clusters of genes that were strongly up- and down-regulated and belong to a variety of biological pathways, such as stress-induced pathways involving immediate early genes, early growth response genes, heat shock proteins, among others. Transcriptional regulation, cell cycle progression, and apoptosis can be altered too. The reported results also suggest that significant stress responses occur as early as 3 hours after blood draw influencing the gene expression data. Consequently, in our expression study, we ensured the stabilization of the RNA from the PBMCs in the first 3 hours after the blood collection.

There are methods for RNA stabilization of peripheral blood cells, such as the PAXgene tube system (Qiagen/BD company, Germany/USA), which allow an immediate stabilization of the RNA from whole blood (Rainen *et al.* 2002). However, the use of all peripheral blood cells comparing with the use of PBMCs increases the heterogeneity of the sample. On the other hand, the use of total RNA from PBMCs has the advantage of increase detection sensitivity when compared to the use of total RNA from whole blood due to the very high levels of globin mRNA in erythrocytes, which represent approximately 95% of all blood cells (Affymetrix Technical Note 2003, Tang et al 2003). Generally, PBMCs consist of T cell lymphocytes (60%), B cell lymphocytes (10%), monocytes (15%), and natural killer cells (15%) (Baird *et al.* 2007).

5.6.4. The importance of the experimental design and of the batch effects in ANOVA

The genetic and environmental heterogeneity existing among individuals gives rise to a new level of complexity in human studies that is usually not encountered in well-controlled animal experiments. Additionally, the transcriptome is also shaped by disease treatment. Separating the effects of the disease from the effects of treatment or the environmental factors remains one of the most challenging aspects of human research. However, non-biological variations can be reduced through careful experimental design. In this project, special care was taken to control for experimental errors and variables (e.g. processing experimental and control samples in parallel, collecting information about different factors of interest and matching them among cases and controls). In our expression studies, there was no significant difference in

age, sex ratio, hypertension status, or other established risk factors between the cases and controls. Only the frequency of diabetes was significantly higher in IS patients than in controls (p-value = 0.035), but all the patients with diabetes are controlled by medication.

For the statistical analysis of the results we need to choice the factors that allow to increase the statistical power of the ANOVA by distinguishing personal biological characteristics from the experimental noise. The noise needs to be taken in account to be clearly distinguished of the biological patterns. Consequently, in our expression analysis, the geographic origin and scan date factors needed to be included as batch factors (the noise) in ANOVA to be accounted for (Figures 4.8 to 4.11). The ANOVA allows the partition of the variance in data into separate factors. The total variance is partitioned into variance due to influencing factors and the rest is assumed to be random error (also noise). Since the log2 transformed expression data is generally normally distributed, the balanced experiments have equal variances within different groups, and the study participants are all independent, all the assumptions of the ANOVA are met.

In studies with numerous samples, RNA isolation batches and microarray hybridization batches are frequent examples of technical/processing inconsistencies in expression studies. They were included as random factors in our analysis. In general, taking a random effect as a fixed effect will produce an over-optimistic p-value, leading to higher false discoveries. It is interesting to highlight that, all the 6 GC genes we found with significant associations, have a major contribution of the factor type (the main factor of interest) to their overall variance among IS cases and controls (Figure 4.21), supporting the adequacy of the statistical analysis performed.

When we started collecting samples, our initial aim was also to compare the expression profile among cases and controls in several subgroups in a sequence of the prospective study of stroke conducted in Portugal in 2004 (Correia *et al.* 2004). Besides the comparison among the old cases and controls and among the young cases and controls separately, we intended to compare the samples according to the geographic origin of the participants. Given the discrepancy in the incidence of stroke between individuals from Porto and Trás-os-Montes (Mirandela and Vila Real) with ages between the 75 and 84 years old, it would be interesting to identify the susceptibility genes for stroke that could be responsible for this difference. However, since we needed to remove the geographic origin batch effect from our results, we could not analyse these differences among urban and rural regions.

5.6.5. Differentially expressed genes among IS cases and controls

This study demonstrates an altered gene expression profile in PBMCs of IS patients sampled at least six months after the first and only stroke episode, relative to controls. After a sensitivity analysis, we opted to use a threshold of 1.2 fold-change and a p-value < 0.05 to obtain a less stringent list of

differentially expressed genes to use in our GC approach (Figure 4.12). It is also appropriate for a late-onset disease such as stroke where small changes in expression over a long period of time are expected to result in the phenotype. This can explain why we only obtained 1,675 probe sets differentially expressed with a threshold of 1.2 fold-change, representing 3,5% of the transcripts on the arrays. The increase of this threshold reduce even more the number of obtained differentially expressed genes, and, additionally, eliminate genes that according our analysis seem to be pertinent to the distinction among cases and controls and may have a real biological effect (Figures 4.12 and 4.13). The elimination of biologically true expression changes from further consideration leads to a large increase of false-negative (type-II) errors. However, decreasing the threshold, we need to be aware that we could increase the number of false-positives (type-I) errors too.

When correcting for multiple measurements, all probe sets with a threshold of 1.2 fold-change and p-value < 0.05 remain differentially expressed using the q-value method to determine FDR (q-value < 0.05), but none of them would resist to the Bonferroni's correction. Bonferroni's correction, however, is generally considered overly conservative for two reasons: it assumes all the tests are independent, which may not be true; and it protects against even a single false positive, which may be too strict if thousands of tests are being conducted. On the other hand, FDR is the most lenient multiple test adjustment. Instead of calculating the change of any false positives, it controls the predictable proportion of false positives among all positives, being more appropriate to apply to microarrays. The FDR threshold is determined from the observed p-value distribution and consequently is adaptive to the data.

The role of several differentially expressed genes has been investigated before in stroke or other related diseases. *TTC7B* and *RAD51L1* (RAD51-like 1; Appendix D: Table D.1) are among the most associated genes with major CVD and AF, respectively, in the GWAS for cardiovascular diseases of the Framingham Heart Study 100K project (Larson *et al.* 2007). The fact they appear differentially expressed in our results further supports their role in IS aetiology and strengthen our expression findings. In the Framingham Heart Study 100K project, the major CVD phenotype includes MI, coronary insufficiency, CHD death, and atherothrombotic stroke diseases. Also differentially expressed in our results (Appendix D: Table D.1) are *TPM1* (tropomyosin 1 (alpha)) and *LMNA* (lamin A/C), which have been tested for association in the Framingham project as pre-specified candidate genes for heart failure. Two genetic variants in *TPM1* and none in *LMNA* were found associated with the heart failure disease (Larson *et al.* 2007).

Additionally, *SELP*, *F13A1* (coagulation factor XIII, A1 polypeptide), and *TUBB1* were differentially expressed (Appendix D: Table D.1). Some studies have investigated the association of *SELP* with stroke. Berger and colleagues (2007) reported a significant association of *SELP* with IS in the German population, as well as the integrin alpha 2 gene, in their first case-control analysis. Other findings suggest that *F13A1* is associated with the efficacy of the thrombolytic therapy in IS patients (González-

Conejero *et al.* 2006), and that *TUBB1* is associated with hemorrhagic stroke in men (Navarro-Núñez *et al.* 2007). Also, the *API5* (apoptosis inhibitor 5), which is differentially expressed in our study, is close to SNP rs10837576 which is one of the most significant associations with IS in the GWAS conducted by Ikram *et al.* (2009).

Hierarchical clustering showed that IS patients were almost perfectly separated from the controls based on their gene expression profile (Figure 4.13). In fact, consistent with the model of a multifactorial disease, neither single genes nor a fixed group of genes are expected to function as perfect classifiers. Besides the interindividual variations, the heterogeneous gene expression profiles within the IS the cases and control clusters, demonstrated by different branches, can also result from characteristic signatures of gene expression in IS patients with clinically different subtypes, although there are no clear tendencies. Moreover, as proposed by Bomprezzi *et al.* (2003) for multiple sclerosis, there is a distinction between the genes that predispose to the disease (the susceptibility genes) and those whose expression is altered in an affected individual as consequence of other genes. Subsets of genes account for direct genotype-phenotype correlation, whereas others may function as indicators of the disease status, and by identifying the former we may dissect the interaction of multiple genes each weakly contributing to disease susceptibility.

It need to be taken in account that the abundance of the mRNA coding for a specific protein may be poorly correlated with protein abundance and consequently the interpretation of the obtained results required the respective care. It should be also interesting to determine how specific our results are for IS and whether could be due to unknown non-stroke factors. Unspecific results obtained as consequence from nonspecific stress on gene expression in PBMCs could be determined using samples from patients with different diseases as positive disease controls to evaluate the effects. For instance, groups of persons with hypertension, hypercholesterolemia or with cardiac diseases, could be included in the analysis. Moore and colleagues (2005a), to address the specificity of their expression studies, compared the differentially expressed gene list obtained for stroke, with the lists from PBMCs in sickle cell disease, a hematologic disorder characterized by chronic inflammation and ischemic crises, and multiple sclerosis.

5.6.6. Pathways analyses in IS disease

One limitation of this type of expression profiling studies is that, as referred previously, even though differentially expressed genes may constitute good indicators of affection status and of the downstream disease mechanisms, they may not be the susceptibility genes that directly account for the initial phenotype under investigation. One strategy to address this issue is to study groups of genes and/or pathways significantly over-represented among the list of differentially expressed genes.

The significant over-representation of genes related with antigen binding, immune and

inflammatory responses, platelet alpha granule membrane, and other cellular functions, among the differentially expressed genes, suggest an active role of the PBMCs of the IS patients in the complex immune responses to the vascular injuries that cause the ischemic events. Moreover, the fact that several differentially expressed genes are related with cell death, cellular growth and proliferation, cellular movement, cell-to-cell signalling and interaction and cellular function and maintenance, supports the relevance of using of PBMCs to the study of stroke as well. Several of these genes have been associated with other disorders and conditions such as haematological diseases, ischemia, thrombosis of blood vessel, and injury of tissues (Appendix D: Table D.4), also evidencing that complex disorders, resulting from the interplay of genetics and environment, can share many genetic risk factors.

Although complex diseases result from the cumulative effects of many different genes, it can be assumed there are often only a limited number of pathways that contribute to disease aetiology. The most significant associated pathway in our analysis was the cell adhesion molecules pathway (Figure 4.14 and Table 4.14) includes the SDC4 gene. This pathway appears could be related with the inflammatory process of the stroke disease and consequently with the migration and interaction of the studied PBMCs with the EC of the blood vessels. Cytokine-cytokine receptor interaction, gap junction, focal adhesion, complement and coagulation cascades, adherens junction, antigen processing and presentation pathways (Figure 4.14), on the other hand, highlight the importance of the inflammatory process in the aetiology of stroke.

5.6.7. Gene annotations in the Affymetrix GeneChip microarrays

The use of Affymetrix GeneChip microarrays, despite their several highly advantages, could also introduce another limitation in our study. Its selection of probes relied on earlier genome and transcriptome annotation which is significantly different from current knowledge. The original annotation assignments could be, for some genes, out-dated and suboptimal, and the resultant problems could have an important impact on analysis and interpretation of the data (Zhang *et al.* 2005, Gautier *et al.* 2004, Harbig *et al.* 2005). The use of wrong probe set definitions could mask important genes for the understanding of the mechanisms of the studied diseases. Only for the ones we focus our attention, if they are not too many, we can confirm the correct probeset definitions, but in practice, downstream results can be affected, especially with regards to differential expression and gene-set based analyses (Lu *et al.* 2007).

Consequently, several groups have been working on reorganizing probes on the most popular used GeneChips into gene-, transcript- and exon-specific probe sets, under the assumption that current genome and transcriptome databases are more accurate than those used for GeneChip design. In an ideal situation, a gene-specific probeset should only contain probes whose sequence will be present on the shared sequences of all splicing products from the same gene, as the signal level of such probe set should

not be influenced by the alternative splicing in different tissues or individuals. However, even nowadays, for most genes, current knowledge of potential alternative splicing products is far from complete. Potential alternative splicing events can conceivably be explored by the transcript- or exon-specific probesets (Dai *et al.* 2005, Lu *et al.* 2007, Ferrari *et al.* 2007).

It will beneficial to analyse some of the new differentially expressed genes that comes from the re-analyses of our GeneChip expression results with published updated probeset definitions. Custom CDF files for the specific type of Affymetrix GeneChip microarrays that we used can be selected on public databases and used instead of the Affymetrix provided one.

5.6.8. Genomic convergence with stroke linkage screens on non-Portuguese populations

Following the GC approach, we converged our expression results with those from the published stroke linkage screens on Icelandic and Swedish populations (Gretarsdottir *et al.* 2002, Nilsson-Ardnor *et al.* 2005 and 2007) to prioritize genes for association studies. This could potentially be a limitation for the study of stroke in the Portuguese population. However, previous linkage studies for complex traits such as bipolar disorder, schizophrenia or multiple sclerosis in a Portuguese sample have identified linkage to loci strongly implicated in other populations as well (Martins Silva *et al.* 2003, Middleton *et al.* 2004, Sklar *et al.* 2004), suggesting that the Portuguese population is not so isolated that only Portuguese-specific loci are uncovered. In addition, even nonsignificant linkage results in initial genome-wide scans of homogeneous populations have previously been used for the successful identification of candidate genes for complex diseases, with a strong association to the disease in more heterogeneous populations (Grant *et al.* 2006, Helgadottir *et al.* 2005, Helgadottir *et al.* 2006). Several studies report that the types of stroke of the Icelandic population do not reflect a rare stroke form of the disease or a specific form to Iceland. Rather, the diverse phenotypes in Icelanders, as well as known risk factors for stroke, are similar to those of most other white populations (Sveinbjornsdottir *et al.* 1998, Valdimarsson *et al.* 1998, Eliasson *et al.* 1999), as confirmed also by the replication of the linkage peak 5q12 in the northern Sweden population (Gretarsdottir *et al.* 2002, Nilsson-Ardnor *et al.* 2005). In the same way, in Sweden, low genetic variation in the population had previously been taken advantage of in successful linkage studies of complex diseases (Lindholm *et al.* 2001, Venken *et al.* 2005).

We did not conduct a whole-genome linkage screen in the Portuguese population due the difficulty in collecting large enough sample sizes from multiplex families to have enough power to detected linkage.

5.6.9. The relevance of the GC prioritized genes for association studies

Although one of the main goals of the GC approach is to be unbiased by preconceived models of disease, it is interesting that several of the prioritized genes for association studies were previously reported to be associated with some other genetic disorder (Appendix D: Table D.4) or had functions that could be easily associated with stroke. As mentioned before, *TTC7B* was one of the top associated genes with major CVD (which also includes stroke) in the GWAS for cardiovascular diseases of the Framingham Heart Study 100K project (Larson *et al.* 2007), and *TUBB1* was previously associated with hemorrhagic stroke in men (Navarro-Núñez *et al.* 2007). *TUBB1* is also part of the significantly affected gap junction pathway, and *SDC4* is involved in the cell adhesion molecule, and ECM-receptor interaction pathways (Table 4.14). These findings help to support the used approach and the quality and relevance of the analyses we performed.

We identified yet a set of differentially expressed genes orthologous to rat genes (data not shown) mapping to chromosomes 1 and 5 linkage peaks in rat studies (Rubattu *et al.* 1996, Jeffs *et al.* 1997 and Kato *et al.* 2003), which could also be followed-up in association studies.

5.6.10. The association of the GC prioritized genes with IS risk

For the association studies of GC genes, we genotyped all their haplotype tagging SNPs in our entire Portuguese case-control biobank, taking into account the well known co-variates in the analyses. This procedure allows a representative selection of SNPs of the genes tested in association studies. Frequent limitations of the association studies such as the poor phenotyping, low statistical power, poor and/or inadequate matching of controls, genetic heterogeneity, among others, have been minimized here.

The multiple testing issue has been one of the most discussed problems in association studies. Some authors report that a Bonferroni's correction of the p-values is overly conservative, because the SNPs may be in some LD and consequently they are not independent (Lewis *et al.* 2002). Additionally, there is recent evidence that the replication of a nominal p-value in a second cohort is less stringent and is a better alternative to the conservative multiple-testing corrections (Shephard *et al.* 2005). Therefore, we sought to validate our most promising associations in the Spanish case-control dataset, to filter out the false-positive associations.

The follow up of the association findings in a different population reduces the total number of tests performed by removing unassociated loci in the first population. In our replication study in a Spanish cohort, we tested the association of gene variants with stroke subtypes. Although there is no clear evidence suggesting that the IS subtype correlates with heterogeneity in genetic risk factors, it is

reasonable to assume that patients with more similar clinical features are under the influence of more similar pathophysiologic mechanisms (Schork et al 1997). According to Flossmann *et al.* (2004), not all stroke types are likely to be equally heritable, and mixture of different stroke subtypes is therefore likely to negatively influence the results. Large- and small-vessel disease seems to be more strongly associated with a positive family history of stroke than the remainder types of IS. However, the authors also defend that the exact stroke phenotype is sometimes difficult to ascertain, and this could mask stronger associations of less frequent stroke subtypes.

5.6.10.1. *TUBB1*

Navarro-Núñez *et al.*, in 2007, associated the *TUBB1* Q43P platelet polymorphism with hemorrhagic stroke in men. This polymorphism, characterized by a double nucleotide substitution (AG > CC), causes a lower reactivity in platelets carrying the variant form of b1-tubulin. It affects a highly conserved residue within a region implicated in the binding of TUBB1 with other isoforms of tubulin. The Q43P polymorphism also seems to have a protective effect on arterial thrombosis in a Belgian cohort (Freson *et al.* 2005).

Even that our association results suggest that the SNP rs151348, upstream of *TUBB1*, may constitute a risk factor for IS in the Portuguese population (Figure 4.17 and Table 4.17), the obtained p-value does not survive to the Bonferroni's correction for multiple testing, and this SNP could not be genotyped in the Spanish population for validation. On the other hand, it could also be observed that we have not a significant association for the allelic test performed for the SNP rs151348 (Table 4.17). In general, analysing the data by alleles breaks down genotypes to compare the total number of each allele in cases and controls regardless of the genotypes from which these alleles are constructed. In this analysis, single alleles do not act independently, but provides a powerful method of testing under a multiplicative model, where the risk of developing stroke increases for each risk allele carried (Lewis *et al.* 2002). We propose that this gene should be further studied in other datasets to validate and understand its role in IS.

TUBB1 is one of the β-isoforms essential to form heterodimers with α-isoforms. The distinct α- and β-tubulin isoforms contribute to the functional diversity of microtubules either through their differential polymerization into high-ordered microtubule polymers, or by virtue of unique interactions with distinct microtubule-associated proteins (Italiano *et al.* 1999, Schwer *et al.* 2001, Patel *et al.* 2005). In general, microtubules are an important component of the surface membrane cytoskeleton and, together with the cytoplasmic, actin-rich cytoskeleton, are responsible for intracellular transport of vesicles, cell morphogenesis, and chromosome segregation during cell division in all eukaryotes (Sullivan *et al.* 1988). To our knowledge, few studies have addressed the importance of genetic changes in TUBB1 in ischemic risk.

5.6.10.2. *HEMGN and GFI1B*

Related with haematopoiesis, and with leukaemia and lymphoma, respectively, *HEMGN* and *GFI1B* were also prioritized following our GC approach. *HEMGN* (also called *EDAG* in humans, embryonic development associated gene) encodes a nuclear protein that is highly expressed in hematopoietic tissues and acute leukaemia. It is involved in the regulation of proliferation, differentiation and apoptosis of hematopoietic cells (Yang *et al.* 2006). GFI1B is a transcription factor essential for the development and differentiation of erythroid and megakaryocytic lineages, but its role in haematopoiesis has not been well characterized. GFI1B is a zinc finger protein that has a highly conserved transcriptional repressor snail-Gfi-1 domain and 6 zinc finger domains at the N- and C-terminus, respectively (Osawa *et al.* 2002). To our knowledge, there are no studies addressing the importance of these genes in ischemic risk.

With a generally marginal allelic, genotype and haplotype association results (Figures 4.15 and 4.16 and Table 4.17), we conclude that variants in *HEMGN* and *GFI1B* may constitute risk factors for IS in the Portuguese population. However, none of these variants were associated in the Spanish population in any test performed. This suggests that our findings in the Portuguese population may be false-positives, or that the variants on these genes are not important risk factors for IS in the Spanish population (although there are marginal associations verified for cardioembolic and atherothrombotic stroke). Due the marginal significance of the associations it would be important to enlarge our Portuguese sample to verify if our positive results withstand with the increase in power.

5.6.10.3. *TMTC4*

The function of the protein encoded by *TMTC4* is unknown. We verify that the SNP rs9582406 in this gene was associated in the Portuguese dataset (Figure 4.18 and Table 4.17) and was replicated in the Spanish case-control sample in unadjusted tests performed for alleles and genotypes (Table 4.19). Additionally, this association appear to be most significant in the atherothrombotic and lacunar stroke. However, the effect of the associated allele for the IS risk seems to be opposite in the Portuguese and the Spanish cohorts. As explained before for *PED4D* and *ALOX5AP*, the observed opposite direction of the associated effects of the SNP rs9582406 may be due its interaction with other genetic or environmental risk factors. This can also suggest that probably, the true susceptibility variant(s) have not been studied yet and its/their level of LD with the SNP rs9582406 may vary among the tested populations. However, no SNPs in this gene were associated with IS in the GWAS performed by Ikram *et al.* (2009).

5.6.10.4. *TTC7B and SDC4*

For *TTC7B* and *SDC4* no single SNPs or haplotypes are consistently associated with IS in the Portuguese or in the Spanish samples (Figures 4.19 and 4.20 and Tables 4.17 and 4.19). Since there are also no SNPs on the *TTC7B* and *SDC4* (even beyond the haplotype tagging SNPs; Figure 4.23) that present a substantial association in the IS GWAS performed by Ikram *et al.* (2009), it is probable that the risk variants of these genes, with which the variants we studied should be in high LD, remains elusive. This heterogeneity highlights the need for a more detailed study of these genes, possibly involving the investigation of other types of genetic markers (e.g. insertion/deletion, copy number variations) or less common non-tagging variants (using next-generation sequencing).

There are no reports to date on the function of *TTC7B*, however it is a member of the tetratricopeptide repeat (TPR) gene family. TPRs consist of tandem arrays of highly degenerate 34 amino acid repeats predicted to form extended superhelical arrangements. These TPR domains function as protein-protein interaction modules for macromolecular complexes involved in numerous cellular processes, including transcriptional regulation, mRNA processing, protein folding and translocation (Krachler *et al.* 2010). The SNPs and haplotypes that we found associated with IS are located in the intron 5 through intron 6 region of *TTC7B*.

SDC4 is an ubiquitous transmembrane heparan sulfate proteoglycan that localizes to the focal adhesions of adherent cells and binds to a range of extracellular ligands, including growth factors and extracellular-matrix proteins. SDC4 is essential for adhesion formation in cells adhering via certain integrins, and for cell proliferation and migration in response to growth factors (Woods *et al.* 1994, Bass *et al.* 2002). The protein is found as a homodimer and is a member of the syndecan proteoglycan family. Both extracellular and cytoplasmic domains of the protein are necessary for regulated activation of associated transmembrane receptors. Its cytoplasmic domain interacts with a number of signalling and structural proteins, and, in particular, the C-terminus of the cytoplasmic domain coordinates clustering of receptors and the connection to the actin cytoskeleton when binded to PDZ domain-containing scaffold proteins. These characteristics makes the SDC4 as one of the most influential receptors in the cell. (Bass *et al.* 2002, Gao *et al.* 2000).

The fact that some of the SNPs we found associated in our Portuguese sample (SNPs rs2343 of the *TTC7B*, and SNPs rs2251252 of the *SDC4*; Figures 4.19 and 4.20 and Table 4.17), present also significant associations with the risk of atherothrombotic and cardioembolic stroke for some or all of the tests performed, support the role of these genes in stroke risk. Additionally, as referred before, *TTC7B* was one of the top associated genes with major CVD (that includes stroke) in the GWAS for cardiovascular diseases of the Framingham Heart Study 100K project (Larson *et al.* 2007).

Given the discrepancies in the findings among GWASs, validation of their most significant results

(above and below genome-wide significance level) in independent datasets, as well as other approaches such as the one proposed here (combination of genetics and genomic profiling), must be undertaken to pinpoint the real genetic players in stroke aetiology. Our findings in the Portuguese and Spanish cohorts through an independent approach together with the previous GWAS associations (Figure 4.23) strongly support the involvement of *TTC7B* in stroke aetiology.

Future work must be directed toward elucidating the biochemical functions of *TTC7B*, and the functional relevance of their SNPs. Further investigation of this region may include deep-sequencing to identify rare SNPs and subsequent testing for association.

5.7. The putative function of the associated SNP variants

The majority of haplotype tagging SNPs for which we found an association are intronic. It is not of major importance to investigate the function of these SNPs at this point of the project because they can be only representing the true susceptibility SNPs. Additionally, very little is known about the function that specific intronic sequences have with regard to the secondary structure, protein binding, stability, and splicing efficiency of heterogeneous nuclear RNA.

The true susceptibility SNPs of the GC genes may affect, for instance, their expression, and could be further investigated in future studies by testing for case-control differences in transcript structures, spatial distribution, and developmental distribution. They may also affect the transcription, the splicing, or the structure or activity of their receptors. According to FastSNP (http://fastsnp.ibms.sinica.edu.tw), a Web server that allows identification and prioritization of high-risk SNPs, the SNPs rs946845 in *TMTC4*, rs11629065 in *TTC7B*, and rs11712619, rs6438833, rs11712039 and rs7620580 in *KALRN* could have a possible functional effect as intronic enhancers, whereas the functions of the other studied intronic tagging SNPs are not known (data not shown). Additionally, it is also possible that the intronic SNPs result in changes in the structure of either the DNA or RNA molecules. For instance, a change in the ability of the DNA to wrap itself into nucleosome around histones, thus affecting the access of regulatory proteins to their binding sites (Segal *et al.* 2006).

5.8. The study of rare variants

Although we focused our study of IS on common genetic variants with MAF greater than 0.1, the heritability of a complex trait such as stroke could result from both common and rare variants. The motivation for studying common variants was greater than for rare variants, because they are easier to identify, have more power in association studies, and they may be of public health importance by

identifying subpopulations at increased risk of disease (Carlson *et al.* 2004). However, nowadays the focus has now shifted towards rare variants with the development of increasingly more affordable high-throughput sequencing technologies.

There are two different hypothesized genetic models for complex disorders that had been important conceptual developments. The common disease/common variant proposed that the genetic basis of common genetically complex disorders is principally due to genetic variants that are common in the population, as the ones we studied for the risk of IS. The disease susceptibility is then influenced by a number of loci, each of which has a single major allele contributing to the phenotype (Lander *et al.* 1994, Reich *et al.* 2001). Each allele may have low penetrance, but if enough data can be gathered, associations should be detectable. In contrast, the common disease/rare variant model argued that in complex disorders there is a significant contribution from rare variants, which include most of those with the most significant individual effects (Pritchard 2001, Wright *et al.* 2003). In this cases, association studies as the performed ones in this project have a little power. A possible way to mitigate this problem is performing multimarkers tests as we did constructing haplotypes and studding them for association with the IS risk. This approach substantially increases tagging efficiency relative to single-marker approaches, without loss of power. Multimarker tests improve power for less common casual alleles although they are neutral or reduce power when the casual SNP is common (Bakker *et al.* 2005).

Stroke, like other complex diseases, most likely results from a combination of these two models. Many genes with variants showing small and peripheral effects on disease and a smaller number of genes with variants showing moderate large effects could be related with the disease aetiology (Wright *et al.* 2003). Individuals with similar clinical features of the disease can possess different dysfunctional genes in the origin of those common clinical features (Schork *et al.* 1997).

6. FUTURE WORK

A more detailed investigation is necessary at several levels to clarify and confirm our findings. The enlargement of our sample and the replication of the most significant association results in independent datasets belonging to different populations is mandatory. It is generally accepted that no candidate gene association should be considered "confirmed" until replicated by well-designed studies in different populations where any effects of population stratification or other methodological biases are unlikely to act in a consistent manner. Furthermore, a more in depth study of the influence of the associated genetic variants with specific subtypes of IS must be performed in much larger samples than the one used here in order to have enough power to test it. For *KALRN*, for instance, it is expected that the association is stronger for large-vessel stroke and we could not test it.

It is also of the major interest to replicate in independent datasets the most significant results (above and below genome-wide significance) of the recent well-powered GWAS performed for stroke by Ikram *et al.* (2009) covering approximately 2.5 millions of SNPs on four European and American cohorts. In a near future, it should be made an effort to perform well powered GWASs for the different stroke subtypes.

Also microsatellites or CNVs (copy number variations), among other genetics variants, can affect the risk of IS and could be studied in the chromosome regions that seem to present higher evidence to the disease. On the other hand, once bonafide susceptibility loci are found, deep sequencing may have to be used to precisely identify the causal genetic variants in strong LD with the studied ones, or rare variants that the association studies have no power to detect.

Finally, the study of the gene-environmental interactions and gene-gene interactions (or epistasis) in a well powered population could capture more information about stroke risk than analysis of single susceptibility genes allowing the discovery of genes with smallest effects that otherwise are kept unknown. Nonlinearities between genotype and phenotype depend, for instance, of the influence of these interactions (Moore *et al.* 2006), and the analysis of one gene at a time may be confused by unstudied genes and/or environmental factors which influence the phenotype. However, the identification and characterization of the susceptibility gene-gene and/or gene-environmental interactions have been limited by small sample size of the studies performed and for the large number of potential interactions between genes.

Several methods have been proposed to detect gene-gene or gene-environmental interactions in a case-control study, as the penalized logistic regression that has been used in several publications as a parametric approach to detect gene-gene interactions, and the Multifactor Dimensionality Reduction (MDR) that is nonparametric and genetic model-free approach to detect genotype combinations associated

with disease risk (He *et al.* 2009, Ritchie *et al.* 2001). They focus on the statistical epistasis in human populations.

We propose that the modelling of gene-gene and gene-environmental interactions contributing to the IS risk could be carried out using the MDR method in a well-powered dataset. It is mandatory to study with detail gene-environment interactions, which have been mostly neglected until now. We did not performed this modelling analysis in our Portuguese sample since it is underpowered for the number of tests that need to be performed. Depending of the strength of the interactions and of the allele frequencies, sample size requirements to detect statistically significant gene-gene and/or gene-environmental interactions may be substantially larger than the sample sizes to identify a single genetic or environmental marginal effect. The correction for multiple testing requires much smaller levels of significance and thus much larger samples (Dempfle *et al.* 2008). Consequently, inadequate sample sizes could lead to underpowered studies that give rise to both false-negative and false-positive findings especially at the hypothesis-generating stage.

After a gene that is thought to influence a complex disease has been identified, one could then study its context dependency through laboratory manipulations, for instance, conducting molecular biology and cellular assays *in vitro* or *in vivo* (using animal models) studies. The role of the identified susceptibility polymorphisms should be studied to understand how they influence the risk of the disease and the respective gene function. Functional studies may be required to pinpoint the true variant among those that are in strong LD.

The animal models have the advantage of providing a high genetic homogeneity not possible with human populations. They can be very useful by allowing one to study the molecular pathways in which they are involved, but also to identify potential homologues of the multiple genes involved in the stroke disease.

7. CONCLUSION

The "genetic revolution" continues to open many new directions in disease research, especially in genetically complex diseases such as stroke. This project represents a collaboration, involving groups from many institutions (hospitals, and research institutes) with different fields of expertise (clinical, genetic, and statistical). The genetic contribution to stroke is well established, and even though some progress has been made regarding the identification of susceptibility genes, the complexity of the phenotype and likely genetic complexity has hindered the unequivocal identification of genetic factors.

Regarding *PDE4D*, we did not find any clear association of the studied variants with IS risk in the Portuguese population indicating that this gene may not constitute a major risk factor for IS risk in Portugal. None of the studied variants in the Spanish population showed also a significant association. However, our findings suggest that variants in *ALOX5AP* may constitute genetic risk factors for IS, underlining the importance of this gene in the risk of IS in the overall population. A joint analysis of Portuguese and Spanish datasets and a new meta-analysis in white populations suggest that SNP rs10507391 (or SG13S114), that is part of the associated HapA haplotype in the original report (Helgadottir *et al.* 2004), was associated with IS risk.

This project was also designed to take advantage of various complementary approaches, including whole genome linkage screens, gene expression and candidate gene analysis, in a large population of patients and controls, to identify novel susceptibility genes. We suggest that this project led to the identification of novel susceptibility genes for IS risk, such as the *KALRN*, but probably also the *TUBB1*, *TMTC4*, *TTC7B, SDC4*, in different chromosomic regions. The association of *KALRN* is reinforced by the multiple independent lines of evidence, namely the association of several uncorrelated SNPs and the replication of the most associated SNP results on the GWAS published by Ikram *et al.* (2009) for IS performed on four European and American cohorts. On the other hand, *TMTC4* seems to contribute to the IS risk in the Iberian population but is not replicated in the referred GWAS. *TTC7B* and *SDC4* presented a significant heterogeneity in the associated variants among the Portuguese and Spanish cohorts, and the GWAS, being probable that the putative risk variants in these genes were not studied.

Looking for the overall results, it is important to underline that any significant association result has several possible interpretations. The tested variants may be not the causative disease susceptibility SNPs that directly affects the risk of suffering stroke, but the SNPs may merely be in LD with the true genetic variants and the significant results arises because the studied SNP and the true variant are co-inherited in the studied populations being rarely separated by recombinations. Consequently, further studies in our suggestive genes need to be performed. If our findings could be confirmed in independent datasets and a most complete study of other possible genetic variants (such as rare SNP variants,

microsatellites, and CNVs, that could also be in LD with the studied SNPs) in the loci of interest could be performed, we think that in a recent future functional studies of the genes and of their causative variants may allow a significant improvement of our knowledge on stroke pathogenesis. Understanding the function of these proteins in the aetiology of the disease could lead to new innovative strategies for primary and secondary prevention of stroke. We also identified several affected biological pathways in stroke patients, such as the highly significant associated cell adhesion molecules pathway, that can be studied as new targets for drug development. The development of more efficient therapies for stroke in an ageing world population would have large repercussions for the world economy.

Finally, it is of the major importance to analyse gene-gene and/or gene-environmental variants as susceptibility factors for the IS risk. Nowadays, the most recent efforts for the study of complex diseases has been to developed computationally efficient methods to simultaneously analyse several genetic variants or other factors at the same time. This will avoid ignoring the effects of all other variants when a single one is being studied, improving the performance of the single tests. We suggest that, identifying accurately the genetic determinants of stroke using different strategies and populations, and analysing them in an integrated view as we proposed in this project, is a most complete form to study the stoke disease in order to improve its diagnosis and treatment strategies, and to reduce the stroke's huge public health burden.

8. REFERENCES

Abecasis G, Noguchi E, Heinzmann A et al.: Extent and distribution of linkage disequilibrium in three genomic regions. Am J Hum Genet. 68:191-197 (2001).

Adams HP Jr, Bendixen BH, Kappelle LJ *et al.*: Classification of subtype of acute ischemic stroke. Definitions for use in a multicenter clinical trial. TOAST. Trial of Org 10172 in Acute Stroke Treatment. Stroke. 24:35-41 (1993).

Affymetrix Technical Note: An analysis of blood processing methods to prepare samples for GeneChip expression profiling (2003).

Agerholm-Larsen B, Nordestgaard BG and Tybjaerg-Hansen A: ACE gene polymorphism in cardiovascular disease: meta-analyses of small and large studies in whites. Arterioscler Thromb Vasc Biol. 20:484-492 (2000).

Allison DB, Cui X, Page GP *et al.*: Microarray data analysis: from disarray to consolidation and consensus. Nat Rev Genet. 7:55-65 (2006).

Akey JM, Zhang K, Xiong M *et al.*: The effect that genotyping errors have on the robustness of common linkage-disequilibrium measures. Am J Hum Genet. 68:1447-1456 (2001).

Appelros P, Nydevik I, Seiger A *et al.*: High incidence rates of stroke in Orebro, Sweden: Further support for regional incidence differences within Scandinavia. Cerebrovasc Dis. 14:161-168 (2002).

Archacki SR, Angheloiu G, Tian XL *et al.*: Identification of new genes differentially expressed in coronary artery disease by expression profiling. Physiol Genomics. 15:65-74 (2003).

Ariga M, Neitzert B, Nakae S *et al.*: Nonredundant function of phosphodiesterases 4d and 4b in neutrophil recruitment to the site of inflammation. J Immunol. 173:7531-7538 (2004).

Ariyaratnam R, Casas JP, Whittaker J *et al.*: Genetics of ischaemic stroke among persons of non-European descent: a meta-analysis of eight genes involving approximately 32,500 individuals. PLoS Med. 4:e131 (2007).

Asplund K, Stegmayr B and Peltonen M: From the twentieth to the twenty-first century: A public health perspective on stroke. In: Ginsberg MD, Bogousslavsky J, eds. Cerebrovascular disease pathophysiology, diagnosis, and management. Blackwell Science. 2:901-918 (1998).

Ayres KL and Overall ADJ: Allowing for within-subpopulation inbreeding in forensic match probabilities. Forensic Science International. 103:207-216 (1999).

Baechler EC, Batliwalla FM, Karypis G *et al.*: Expression levels for many genes in human peripheral blood cells are higly sensitive to ex vivo incubation. Genes and Immunity. 5:347-353 (2004).

Baechler EC, Batliwalla FM, Karypis G *et al.*: Interferon-inducible gene expression signature in peripheral blood cells of patients with severe lupus. Proc Natl Acad Sci USA. 100:2610-2615 (2003).

Baird AE: Blood genomics in human stroke. Stroke. 38:694-698 (2007).

Bak S, Gaist D, Sindrup SH *et al.*: Genetic liability in stroke: a long-term follow-up study of Danish twins. Stroke. 33:769-774 (2002).

Baker WH, Howard VJ, Howard G *et al.*: Effect of contralateral occlusion on long-term efficacy of endarterectomy in the Asymptomatic Carotid Atherosclerosis Study (ACAS). Stroke. 31:2330-2334 (2000).

Bakker PIW, Yelensky R, Pe'er I *et al.*: Efficiency and power in genetic association studies. Nature Genetics. 37:1217-1223 (2005).

Bamford J, Sandercock P, Dennis M, *et al.*: Classification and natural history of clinically identifiable subtypes of cerebral infarction. Lancet. 337:1521-1526 (1991).

Banerjee I, Gupta V and Ganesh S: Association of gene polymorphism with genetic susceptibility to stroke in Asian populations: a meta-analysis. J Hum Genet. 52:205-219 (2007).

Baranano DE and Snyder SH: Neural roles for heme oxygenase: contrasts to nitric oxide synthase. Proc Natl Acad Sci USA. 98:10996-11002 (2001).

Barrett JC, Fry B, Maller J *et al.*: Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics. 21:263-265 (2005).

Bass MD and Humphries MJ: Cytoplasmic interactions of syndecan-4 orchestrate adhesion receptor and growth factor receptor signalling. Biochem J. 368:1-15 (2002)

Bennett L, Palucka AK, Arce K *et al.*: Interferon and granulopoiesis signatures in systemic lupus erythematosus blood. J Exp Med. 197:711-723 (2003).

Beresewicz M, Kowalczyk JE and Zabłocka B: Kalirin-7, a protein enriched in postsynaptic density, is involved in ischemic signal transduction. Neurochem Res. 33:1789-1794 (2008).

Berger K, Stögbauer F, Stoll M *et al.*: The glu298asp polymorphism in the nitric oxide synthase 3 gene is associated with the risk of ischemic stroke in two large independent case-control studies. Hum Genet. 121:169-178 (2007).

Bernaudin M, Marti HH, Roussel S *et al.*: A potential role for erythropoietin in focal permanent cerebral ischemia in mice. J Cereb Blood Flow Metab. 19:643-651 (1999).

Bevan S and Markus H: The genetics of stroke. ACNR. 4:8-11 (2004).

Bevan S, Dichgans M, Gschwendtner A *et al.*: Variation in the PDE4D gene and ischemic stroke risk: a systematic review and meta-analysis on 5200 cases and 6600 controls. Stroke. 39:1966-1971 (2008a).

Bevan S, Dichgans M, Wiechmann HE *et al.*: Genetic variations in members of the leukotriene biosynthesis pathway confer an increased risk of ischemic stroke: a replication study in two independent populations. Stroke. 39:1109-1114 (2008b).

Bevan S, Porteous L, Sitzer M *et al.*: Phosphodiesterase 4D gene, ischemic stroke, and asymptomatic carotid atherosclerosis. Stroke. 36:949-953 (2005).

Birk S, Edvinsson L, Olesen J *et al.*: Analysis of the effects of phosphodiesterase type 3 and 4 inhibitors in cerebral arteries. Eur J Pharmacol. 489:93-100 (2004)

Bomprezzi R, Ringnér M, Kim S *et al.*: Gene expression profile in multiple sclerosis patients and healthy controls: identifying pathways relevant to disease. Hum Mol Genet. 12:2191-2199 (2003).

Bovin LF, Rieneck K, Workman C *et al.*: Blood cell gene expression profiling in rheumatoid arthritis discriminative genes and effect of rheumatoid factor. Immunology letters. 93:217-226 (2004).

Bowden DW, Rudock M, Ziegler J *et al.*: Coincident linkage of type 2 diabetes, metabolic syndrome, and measures of cardiovascular disease in a genome scan of the diabetes heart study. Diabetes. 55:1985-1994 (2006).

Brass LM, Isaacsohn JL, Merikangas KR *et al.*: A study of twins and stroke. Stroke. 23:221-223 (1992).

Brazma A, Hingamp P, Quackenbush J *et al.*: Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat Genet. 29:365-371 (2001).

Brem RB, Yvert G, Clinton R *et al.*: Genetic dissection of transcriptional regulation in budding yeast. Science. 296:752-755 (2002).

Brezinski DA, Nesto RW and Serhan CN: Angioplasty triggers intracoronary leukotrienes and lipoxin A4.

Impact of aspirin therapy. Circulation. 86:56-63 (1992).

Brines ML, Ghezzi P, Keenan S *et al.*: Erythropoietin crosses the blood-brain barrier to protect against experimental brain injury. Proc Natl Acad Sci USA. 97:10526-10531 (2000).

Brophy VH, Ro SK, Rhees BK *et al.*: Association of phosphodiesterase 4D polymorphisms with ischemic stroke in a US population stratified by hypertension status. Stroke. 37:1385-1390 (2006).

Caplan LR: Caplan's stroke: a clinical approach, 3d ed. Butterworth-Heinemann, Stoneham, MA, pp1-556 (2000).

Caplan LR and Manning WJ: Brain Embolism. New York: Informa Healthcare (2006).

Cardon LR and Abecasis GR: Using haplotype blocks to map human complex trait loci: Trends Genet. 19:135-140 (2003).

Carlson CS, Eberle MA, Kruglyak L *et al.*: Mapping complex disease loci in whole-genome association studies. Nature. 429:446-452 (2004).

Carswell HV, McBride MW, Graham D *et al.*: Mutant animal models of stroke and gene expression: the stroke-prone spontaneously hypertensive rat. Methods in molecular medicine. 104:49-74 (2005).

Carter AM, Catto AJ, Bamford JM *et al.*: Gender-specific associations of the fibrinogen b 448 polymorphism, fibrinogen levels, and acute cerebrovascular disease. Arterioscler thromb Vasc Biol. 17:589-594 (1997).

Casas JP, Hingorani AD, Bautista LE *et al.*: Meta-analysis of genetic studies in ischemic stroke: thirty-two genes involving approximately 18,000 cases and 58,000 controls. Arch Neurol. 61:1652-1661 (2004).

Cattaneo M, Chantarangkul V, Taioli E *et al.*: The G20210A mutation of the prothrombin gene in patients with previous first episodes of deep-vein thrombosis: prevalence and association with factor V G1691A, methylenetetrahydrofolate reductase C677T and plasma prothrombin levels. Thromb Res. 93:1-8 (1999).

Cavalli-Sforza LL, Menozzi P and Piazza A: The history and geography of human genes. Princeton, NJ: Princeton University Press (1994).

Chiodini BD and Lewis CM: Meta-analysis of 4 coronary heart disease genome-wide linkage studies confirms a susceptibility locus on chromosome 3q. Arterioscler Thromb Vasc Biol. 23:1863–1868 (2003).

Chomczynski P and Sacchi N: Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. Anal. Biochem. 162:156-159 (1987).

Corder EH, Saunders AM, Strittmatter WJ *et al.*: Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. Science. 261:921-923 (1993).

Correia M, Silva MR, Matos I *et al.*: Prospective community-based study of stroke in Northern Portugal: incidence and case fatality in rural and urban populations. Stroke. 35:2048-2053 (2004).

Dahlback B: New molecular insights into the genetics of thrombophilia: resistance to activated protein C caused by Arg506 to Gln mutation in factor V as a pathogenic risk factor for venous thrombosis. Thromb Haemost. 74:139-148 (1995).

Dahlen SE, Bjork J, Hedqvist P *et al.*: Leukotrienes promote plasma leakage and leukocyte adhesion in postcapillary venules: in vivo effects with relevance to the acute inflammatory response. Proc Natl Acad Sci USA. 78:3887-3891 (1981).

Dahlman L, Eaves IA, Kosoy R *et al.*: Parameters for reliable results in genetic association studies in common disease. Nat Genet. 30:149-150 (2002).

Dai M, Wang P, Boyd AD *et al.*: Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. Nucleic Acids Resarch. 33:e175 (2005).

Dai XY, Nanko S, Hattori M *et al.*: Association of apolipoprotein E4 with sporadic Alzheimer's disease is more pronounced in early onset type. Neurosci Lett. 175:74-76 (1994).

Danesh J, Whincup P, Walker M *et al.*: Low grade inflammation and coronary heart disease: prospective study and up-dated meta analyses. BMJ. 321:199-204 (2000).

Debey S, Schoenbeck U, Hellmich M *et al.*: Comparison of different isolation techniques prior gene expression profiling of blood derived cells: impact on physiological responses, on overall expression and the role of different cell types. Pharmacogenomics J. 4:193-207 (2004).

Dempfle A, Scherag A, Hein R *et al.*: Gene-environment interactions for complex traits: definitions, methodological requirements and challenges. European Journal of Human Genetics. 16:1164-1172 (2008).

DGS, Direcção Geral da Saúde: Risco de morrer em Portugal. Lisboa, Portugal (1999).

DGS, Direcção Geral da Saúde: Unidades de AVC. Lisboa, Portugal (2001).

Dhodda VK, Sailor KA, Bowen KK *et al.*: Putative endogenous mediators of preconditioning-induced ischemic tolerance in rat brain identified by genomic and proteomic analysis. J Neurochem. 89:73-89 (2004).

Di Napoli M, Papa F and Bocola V: C-reactive protein in ischemic stroke: an independent prognostic factor. Stroke. 32:917-924 (2001).

Dichgans M: Genetics of ischaemic stroke. Lancet Neurol. 6:149-161 (2007).

Dixon RA, Diehl RE, Opas E *et al.*: Requirement of a 5-lipoxygenase-activating protein for leukotriene synthesis. Nature. 343:282-284 (1990).

Domingues-Montanari S, Fernández-Cadenas I, del Rio-Espinola A *et al.*: Association of a genetic variant in the *ALOX5AP* with higher risk of ischemic stroke: a case-control, meta-analysis and functional study. Cerebrovascular diseases. 29:528-537 (2010).

Domingues-Montanari S, Mendioroz M, del Rio-Espinola A *et al.*: Genetics of stroke: a review of recent advances. Expert Rev Mol Diagn. 8:495-513 (2008).

Dominiczak AF and McBride MW: Genetics of common polygenic stroke. Nat Genet. 35:116-117 (2003).

Dong Y, Hassan A, Zhang Z *et al.*: Yield of screening for CADASIL mutations in lacunar stroke and leukoaraiosis. Stroke. 34:203-205 (2003).

Doré S, Goto S, Sampei K *et al.*: Heme oxygenase-2 acts to prevent neuronal death in brain cultures and following transient cerebral ischemia. Neuroscience. 99:587-92 (2000).

Dubinsky R and Lai SM: Mortality of stroke patients treated with thrombolysis: analysis of nationwide inpatient sample. Neurology. 66:1742-1744 (2006).

Duerr RH, Taylor KD, Brant SR *et al.*: A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. Science. 314:1461-1463 (2006).

Eaves IA, Wicker LS, Ghandour G *et al.*: Combining mouse congenic strains and microarray gene expression analyses to study a complex trait: the NOD model of type 1 diabetes. Genome Res. 12:232-243 (2002).

Edwards AO, Ritter R 3rd, Abel KJ *et al.*: Complement factor H polymorphism and age-related macular degeneration. Science. 308:421-424 (2005)

Eliasson JH, Valdimarsson EM and Jakobsson F: Case fatality after acute stroke at the Reykjavik Hospital in 1996-1997. Læknablaȝiȝ 85:517-525 (1999).

Ellekjaer H, Holmen J, Indredavik B *et al.*: Epidemiology of stroke in Innherred, Norway, 1994 to 1996. Incidence and 30-day case-fatality rate. Stroke. 28:2180-2184 (1997).

Emami N and Diamandis EP: Human tissue kallikreins: a road under construction. Clin Chim Acta. 381:78-84 (2007).

Emanueli C, Minasi A, Zacheo A *et al.*: Local delivery of human tissue kallikrein gene accelerates spontaneous angiogenesis in mouse model of hindlimb ischemia. Circulation. 103:125-132 (2001).

Evans DM, Cardon LR and Morris AP: Genotype prediction using dense map of SNPs. Genet Epidemiol. 27:375-384 (2004).

Evans JC, Frayling TM, Cassell PG *et al.*: Studies of association between the gene for calpain-10 and type 2 diabetes mellitus in the United Kingdom. Am J Hum Genet. 69:544-552 (2001).

Expression profiling - best practices for data generation and interpretation in clinical trials. Nat Rev Genet. 5:229-237 (2004).

Faber BC, Cleutjens KB, Niessen RL *et al.*: Identification of genes potentially involved in rupture of human atherosclerotic plaques. Circ Res. 89:547-554 (2001).

Falcao JM, Ferro JM, Correia M *et al.*: Stroke before age 65: Types, risk factors, clinical features, therapy and outcome. Lisboa, ONSA (http://www.onsa.pt) (2001).

Feigin VL: Stroke epidemiology in the developing world. Lancet. 65:2160-2161 (2005).

Ferrari F, Bortoluzzi S, Coppe A et al: Novel definition files for human GeneChips based on GeneAnnot. BMC Bioinformatics. 8:446 (2007).

Ferraro F, Ma XM, Sobota JÁ *et al.*: Kalirin/Trio Rho guanine nucleotide exchange factors regulate a novel step in secretory granule maturation. Mol Biol Cell. 18:4813-4825 (2007).

Ferro JM: Cardioembolic stroke: an update. Lancet Neurol. 2:177-188 (2003).

Fidani L, Clarimon J, Goulas A *et al.*: Association of phosphodiesterase 4D gene G0 haplotype and ischaemic stroke in a Greek population. Eur J Neurol. 14:745-749 (2007).

Flossmann E, Schulz UG and Rothwell PM: Systematic review of methods and results of studies of the genetic epidemiology of ischemic stroke. Stroke. 35:212-227 (2004).

Fornage M, Swank MW, Boerwinkle E *et al.*: Gene expression profiling and functional proteomic analysis reveal perturbed kinase-mediated signaling in genetic stroke susceptibility. Physiol Genomics. 15:75-83 (2003).

Franco RF, Trip MD, ten Cate H *et al.*: The 20210 G->A mutation in the 3' untranslated region of the prothrombin gene and the risk for arterial thrombotic disease. Br J Haematol. 104:50-54 (1999).

Frank R and Hargreaves R: Clinical biomarkers in drug discovery and development. Nat Rev Drug Discov. 2:566-580 (2003).

Freson K, De Vos R, Wittevrongel C *et al.*: The TUBB1 Q43P functional polymorphism reduces the risk of cardiovascular disease in men by modulating platelet function and structure. Blood. 106:2356-2362 (2005).

Fung HC, Scholz S, Matarin M *et al.*: Genome-wide genotyping in Parkinson's disease and neurologically normal controls: first stage analysis and public release of data. Lancet Neurol. 5:911-916 (2006).

Gao Y, Li M, Chen W *et al.*: Synectin, syndecan-4 cytoplasmic domain binding PDZ protein, inhibits cell migration. J Cell Physiol. 184:373-379 (2000).

Gautier L, Moller M, Friis-Hansen L *et al.*: Alternative mapping of probes to genes for Affymetrix chips. BMC Bioinformatics. 5:111 (2004).

Gershon ES and Godin LR: Clinical methods in psychiatric genetics. I. Robustness of genetic marker investigative strategies. Acta Psychiatr Scand. 74:113-118 (1986).

Gimbrone MA, Brock AF and Schafer AI: Leukotriene B4 stimulates polymorphonuclear leukocyte adhesion to cultured vascular endothelial cells. J Clin Invest. 74:1552-1555 (1984).

Goldstein DB, Ahmadi KR, Weale ME *et al.*: Genome scans and candidate gene approaches in the study of common diseases and variable drug responses. Trends Genet. 19:615-622 (2003).

Goldstein LB and Hanley GJ: Advances in primary stroke prevention. Stroke. 37:317-319 (2006).

González JR, Armengol L, Solé X et al: SNPassoc: an R package to perform whole genome association studies. Bioinformatics. 23:644-645 (2007).

González-Conejero R, Fernández-Cadenas I, Iniesta JA *et al.*: Role of fibrinogem levels and factor XIII V34L polymorphism in thrombolytic therapy in stroke patients. Stroke. 37:2288-2293 (2006).

Goodarzi K, Goodarzi M, Tager AM *et al.*: Leukotriene B4 and BLT1 control cytotoxic effector T cell recruitment to inflamed tissues. Nat Immunol. 4:965-973 (2003).

Granger CV, Dewis LS, Peters NC *et al.*: Stroke rehabilitation: analysis of repeated Barthel index measures. Arch Phys Med Rehabil. 60:14-17 (1979).

Grant SF, Thorleifsson G, Reynisdottir I *et al.*: Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes. Nat Genet. 38:320-323 (2006).

Grau AJ, Weimar C, Buggle F *et al.*: Risk factors, outcome, and treatment in subtypes of ischemic stroke: the German stroke data bank. Stroke. 32:2559-2566 (2001).

Greenberg SM, Rebeck GW, Vonsattel JP *et al.*: Apolipoprotein E epsilon 4 and cerebral hemorrhage associated with amyloid angiopathy. Ann Neurol. 38:254-259 (1995).

Greenberg SM, Vonsattel JP, Segal AZ *et al.*: Association of apolipoprotein E ε2 and vasculopathy in cerebral amyloid angiopathy. Neurology. 50:961-965 (1998).

Gretarsdottir S, Sveinbjornsdottir S, Jonsson HH *et al.*: Localization of a susceptibility gene for common forms of stroke to 5q12. Am J Hum Genet. 70:593-603 (2002).

Gretarsdottir S, Thorleifsson G, Manolescu A *et al.*: Risk variants for atrial fibrillation on chromosome 4q25 associate with ischemic stroke. Ann Neurol. 64:402-409 (2008).

Gretarsdottir S, Thorleifsson G, Reynisdottir ST *et al.*: The gene encoding phosphodiesterase 4D confers risk of ischemic stroke. Nat Genet. 35:131-138 (2003).

Grond-Ginsbach C, Hummel M, Wiest T *et al.*: Gene expression in human peripheral blood mononuclear cells upon acute ischemic stroke. J Neurol. 255:723-731 (2008).

Gu CC, Chang YP, Hunt SC *et al.*: Haplotype association analysis of AGT variants with hypertension-related traits: the HyperGEN study. Hum Hered. 60:164-176 (2005).

Günel M and Lifton RP: Counting strokes. Nat Genet. 13:384-385 (1996).

Gulcher JR, Gretarsdottir S, Helgadottir A *et al.*: Genes contributing to risk for common forms of stroke. Trends Mol Med. 11:217-224 (2005).

Haines JL, Hauser MA, Schmidt S *et al.*: Complement factor H variant increases the risk of age-related macular degeneration. Science. 308:419-421 (2005).

Hampe J, Cuthbert A, Croucher PJ *et al.*: Association between insertion mutation in NOD2 gene and Crohn's disease in German and British populations. Lancet. 357:1925-1928 (2001).

Han GM, Chen SL, Shen N *et al.*: Analysis of gene expression profiles in human systemic lupus erythematosus using oligonucleotide microarray. Genes and Immunity. 4:177-186 (2003).

Harbig J, Sprinkle R and Enkemann SA: A sequence-based identification of the genes detected by probesets on the Affymetrix U133 plus 2.0 array. Nucleic Acids Res. 33:e31 (2005).

Hauser ER, Crossman DC, Granger CB et al: A genome-wide scan for early-onset coronary artery disease in 438 families: the GENECARD study. Am J Hum Genet. 75:436-447 (2004).

Hauser MA, Li YJ, Takeuchi S *et al.*: Genomic convergence: identifying candidate genes for Parkinson's disease by combining serial analysis of gene expression and genetic linkage. Hum Mol Genet. 12:671-677 (2003).

He H, Oetting WS, Brott MJ *et al.*: Power of multifactor dimensionality reduction and penalized logistic regression for detecting gene-gene interaction in a case-controls study. BMC Medical Genetics. 10:127 (2009).

Helgadottir A, Gretarsdottir S, St Clair D *et al.*: Association between the gene encoding 5-lipoxygenase-activating protein and stroke replicated in a Scottish population. Am J Hum Genet. 76:505-509 (2005).

Helgadottir A, Manolescu A, Helgason A *et al.*: A variant of the gene encoding leukotriene A4 hydrolase confers ethnicity-specific risk of myocardial infarction. Nat Genet. 38:68-74 (2006)

Helgadottir A, Manolescu A, Thorleifsson G *et al.*: The gene encoding 5-lipoxygenase activating protein confers risk of myocardial infarction and stroke. Nat Genet. 36:233-239 (2004).

Helgadottir A, Thorleifsson G, Magnusson KP et al: The same sequence variant on 9p21 associates with myocardial infarction, abdominal aortic aneurysm and intracranial aneurysm. Nat Genet. 40:217-224 (2008).

Helgadottir A, Thorleifsson G, Manolescu A et al: A Common Variant on Chromosome 9p21 Affects the Risk of Myocardial Infarction. Science. 316:1491-1493 (2007).

Hiltunen MO, Tuomisto TT, Niemi M *et al.*: Changes in gene expression in atherosclerosis plaques analyzed using DNA array. Atherosclerosis. 165:23-32 (2002).

Hoggart CJ, Whittaker JC, Iorio MD et al: Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. PLoS Genetics. 4:e1000130 (2008).

Horikawa Y, Oda N, Cox NJ et al: Genetic variation in the gene encoding calpain-10 is associated with type 2 diabetes mellitus. Nat Genet. 26:163-175 (2000).

Hugot JP, Chamaillard M, Zouali H *et al.*: Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. Nature. 411:599-603 (2001).

Huttley G, Smith M, Carrington M *et al.*: A scan for linkage disequilibrium across the human genome. Genetics. 152:1711-1722 (1999).

Iglesias AH, Camelo S, Hwang D *et al.*: Microarray detection of E2F pathway and other targets in multiple sclerosis peripheral blood mononuclear cells. Journal of Neroimmunology. 150:163-177 (2004).

Ikram MA, Seshadri S, Bis JC *et al.*: Genome-wide Association Studies of Stroke. N Engl J Med. 360:1718-1728 (2009).

Italiano JE Jr, Lecine P, Shivdasani RA *et al.*: Blood platelets are assembled principally at the ends of proplatelet processes produced by differentiated megakaryocytes. J Cell Biol. 147:1299-1312 (1999).

Ivanova R, Lepage V, Charron D *et al.*: Mitochondrial genotype associated with French Caucasian centenarians. Gerontology. 44:349 (1998).

Jamrozik K, Broadhurst RJ, Anderson CS *et al.*: The role of lifestyle factors in the etiology of stroke. A population-based case-control study in Perth, Western Australia. Stroke. 25:51-59 (1994).

Jeffs B, Clark JS, Anderson NH *et al.*: Sensitivity to cerebral ischaemic insult in a rat model of stroke is determined by a single genetic locus. Nat Genet. 16:364-367 (1997)

Jin K, Mao XO, Eshoo MW *et al.*: Microarray analysis of hippocampal gene expression in global cerebral ischemia. Ann Neurol. 50:93-103 (2001).

Jison ML, Munson PJ, Barb JJ et al: Blood mononuclear cell gene expression profiles characterize the oxidant, hemolytic, and inflammatory stress of sickle cell disease. Blood. 104:270-280 (2004).

Johnson GC, Esposito L, Barratt BJ *et al.*: Haplotype tagging for the identification of common disease genes. Nat Genet. 29:233-237 (2001)

Johnson RC, Penzes P, Eipper BA *et al.*: Isoforms of kalirin, a neuronal Dbl family member, generated through use of different 5'- and 3'-ends along with an internal translational initiation site. J Biol Chem. 275:19324-19333 (2000).

Jones WJ, Williams LS, Redmon G *et al.*: Validating the questionnaire to verify stroke-free status by patient history and physical examination. Abstracts of the International Stroke Conference. Stroke. 32:362-a (2000).

Jousilahti P, Rastenyte D, Tuomilehto J et al: Parental history of cardiovascular disease and risk of stroke. A prospective follow-up of 14371 middle-aged men and women in Finland. Stroke. 28:1361-1366 (1997).

Joutel A, Corpechot C, Ducros A *et al.*: Notch3 mutations in CADASIL, a hereditary adult-onset condition causing stroke and dementia. Nature. 383:707-710 (1996).

Juul K, Tybjaerg-Hansen A, Schnohr P *et al.*: Factor V Leiden and the risk for venous thromboembolism in the adult Danish population. Ann Intern Med. 140:330-337 (2004).

Kannel WB and Sorlie P: Hypertension in Framingham. In Epidmiology and control of hypertension. (ed. Paul O) 553-592 (1975).

Kardys I, Klaver CC, Despriet DD *et al.*: A common polymorphism in the complement factor H gene is associated with increased risk of myocardial infarction: the Rotterdam Study. J Am Coll Cardiol. 47:1568-1575 (2006).

Karp CL, Grupe A, Schadt E *et al.*: Identification of complement factor 5 as a susceptibility locus for experimental allergic asthma. Nat Immunol. 1:221-226 (2000).

Karvanen J, Silander K, Kee F et al: The impact of newly identified loci on coronary heart disease, stroke and total mortality in the MORGAM prospective cohorts. Genet Epidemiol. 33:237-46 (2009).

Kassner SS, Kollmar R, Bonaterra GA *et al.*: The early immunological response to acute ischemic stroke: differential gene expression in subpopulations of mononuclear cells. Neuroscience. 160:394-401 (2009).

Kato N, Nabika T, Liang YQ *et al.*: Isolation of a chromosome 1 region affecting blood pressure and vascular disease traits in the stroke-prone rat model. Hypertension. 42:1191-1197 (2003).

Kaushal R, Pal P, Alwell K *et al.*: Association of ALOX5AP with ischemic stroke: a population-based case-control study. Hum Genet. 121:601-607 (2007).

Keijzer MB, den Heijer M, Blom HJ *et al.*: Interaction between hyperhomocysteinemia, mutated methylenetetrahydrofolateredutase (MTHFR) and inherited thrombophilic factors in recurrent venous thrombosis. Thromb Haemost. 88:723-728 (2002).

Kelly PJ, Rosand J, Kistler JP *et al.*: Homocysteine, MTHFR 677C->T polymorphism, and risk of ischemic stroke: results of a meta-analysis. Neurology. 59:529-536 (2002).

Kessler C, Spitzer C, Stauske D *et al.*: The apolipoprotein E and b-fibrinogen G/A-455 gene polymorphisms are associated with ischemic stroke involving large-vessel disease. Arterioscler Thromb Vasc Biol. 17:2880-2884 (1997).

Khaw KT and Barrett-Connor E: Family history of stroke as an independent predictor of ischemic heart disease in men and stroke in women. Am J Epidemiol. 123:59-66 (1986).

Kiely DK, Wolf PA, Cupples LA *et al.*: Familial aggregation of stroke. The Framingham Study. Stroke. 24:1366-1371 (1993).

Kilic U, Kilic E, Soliz J et al.: Erythropoietin protects from axotomy-induced degeneration of retinal

ganglion cells by activating ERK-1/-2. FASEB J. 19:249-251 (2005).

Kim S and Iwao H: Molecular and cellular mechanisms of angiotensin II-mediated cardiovascular and renal diseases. Pharmacol Rev. 52:11-34 (2000).

Kim YD, Sohn NW, Kang C *et al.*: DNA array reveals altered gene expression in response to focal cerebral ischemia. Brain Res Bull. 58:491-498 (2002).

Kiyohara Y, Kubo M, Kato I *et al.*: Ten-year prognosis of stroke and risk factors for death in a Japanese community: the Hisayama study. Stroke. 34:2343-2347 (2003).

Klein RJ, Zeiss C, Chew EY *et al.*: Complement factor H polymorphism in age-related macular degeneration. Science. 308:385-389 (2005).

Klerk M, Verhoef P, Clarke R *et al.*: MTHFR 677C->T polymorphism and risk of coronary heart disease: a meta-analysis. JAMA. 288:2023-2031 (2002).

Kostulas K, Gretarsdottir S, Kostulas V *et al.*: PDE4D and ALOX5AP genetic variants and risk for Ischemic Cerebrovascular Disease in Sweden. J Neurol Sci. 263:113-117 (2007).

Krachler AM, Sharma A, Kleanthous C: Self-association of TPR domains: Lessons learned from a designed, consensus-based TPR oligomer. Proteins. 78:2131-2143 (2010).

Krug T, Manso H, Gouveia L *et al.*: Kalirin: a novel genetic risk factor for ischemic stroke. Human Genetics. 127:513-523 (2010).

Kruglyak L and Nickerson DA: Variation is the spice of live. Nat Genet. 27:234-236 (2001).

Kubo M, Hata J, Ninomiya T *et al.*: A nonsynonymous SNP in PRKCH (protein kinase C eta) increases the risk of cerebral infarction. Nat Genet. 39:212-217 (2007).

Kubo M, Kiyohara Y, Kato I *et al.*: Trends in the incidence, mortality, and survival rate of cardiovascular disease in a Japanese community: the Hisayama study. Stroke. 34:2349-2354 (2003).

Kuhlenbäumer G, Berger K, Huge A *et al.*: Evaluation of single nucleotide polymorphisms in the phosphodiesterase 4D gene (PDE4D) and their association with ischaemic stroke in a large German cohort. J Neurol Neurosurg Psychiatry. 77:521-524 (2006).

Laberge-le Couteulx S, Jung HH, Labauge P *et al.*: Truncating mutations in CCM1, encoding KRIT1, cause hereditary cavernous angiomas. Nat Genet. 23:189-193 (1999).

Lacombe C and Mayeux P: The molecular biology of erythropoietin. Nephrol Dial Transplant. 14(Suppl 2):22-28 (1999).

Lander ES and Schork NJ: Genetic dissection of complex traits. Science. 265:2037-48 (1994).

Larson MG, Atwood LD, Benjamin EJ et al: Framingham Heart Study 100K project: genome-wide associations for cardiovascular disease outcomes. BMC Medical Genetics. 8:S5 (2007).

Leal SM: Detection of genotyping errors and pseudo-SNPs via deviations from Hardy-Weinberg Equilibrium. Genetic Epidemiology. 29:204-214 (2005)

Lee BC, Ahn SY, Doo HK *et al.*: Susceptibility for ischemic stroke in Korean population is associated with polymorphisms of the interleukin-1 receptor antagonist and tumor necrosis factor-a genes, but not the interleukin-1b gene. Neurosci Lett. 577:33-36 (2004).

Lee M, Huang WY, Weng HH *et al.*: First-ever ischemic stroke in very old Asians: clinical features, stroke subtypes, risk factors and outcome. Eur Neurol. 58:44-48 (2007).

Levy E, Carman MD, Fernandez-Madrid IJ *et al.*: Mutation of the Alzheimer's disease amyloid gene in

hereditary cerebral hemorrhage, Dutch type. Science. 248:1124-1126 (1990).

Lewis CM: Genetic association studies: design, analysis and interpretation. Briefings in bioinformatics. 3:146-153 (2002).

Li YJ, Oliveira SA, Xu P *et al.*: Glutathione S-transferase omega-1 modifies age-at-onset of Alzheimer disease and Parkinson disease. Hum Mol Genet. 12:3259-3267 (2003).

Liao D, Myers R, Hunt S *et al.*: Familial history of stroke and stroke risk. The Family Heart Study. Stroke. 28:1908-1912 (1997).

Liddell M, Williams J, Bayer A *et al.*: Confirmation of association between the e4 allele of apolipoprotein E and Alzheimer's disease. J Med Genet. 31:197-200 (1994).

Lindenstrom E, Boysen G and Nyboe J: Risk factors for stroke in Copenhagen, Denmark. I. Basic demographic and social factors. Neuroepidemiology. 12:37-42 (1993).

Lindholm E, Ekholm B, Shaw S *et al.*: A schizophrenia-susceptibility locus at 6q25, in one of the world's largest reported pedigrees. Am J Hum Genet. 69:96-105 (2001).

Lisabeth L, Smith MA, Brown DL *et al.*: Family history and stroke outcome in a bi-ethnic, population-based stroke surveillance study. BMC Neurology. 5:20 (2005).

Lloyd-Jones D, Adams R, Carnethon M *et al.*: Heart disease and stroke statistics - 2009 update: a report from the American heart association statistics committee and stroke statistics subcommittee. Circulation. 119:e21-e181 (2009).

Lohmussaar E, Gschwendtner A, Mueller JC *et al.*: ALOX5AP gene and the PDE4D gene in a central European population of stroke patients. Stroke. 36:731-736 (2005).

Lövkvist H, Smith JG, Luthman H *et al.*: Ischaemic stroke in hypertensive patients is associated with variations in the PDE4D genome region. Eur J Hum Genet. 16:1117-1125 (2008).

Lu J, Lee JC, Salit ML *et al.*: Transcript-based redefinition of grouped oligonucleotide probe sets using AceView: high-resolution annotation for microarrays. BMC Bioinformatics. 8:108 (2007).

Lyden P, Lu M, Jackson C *et al.*: Underlying structure of the National Institutes of Health Stroke Scale: results of a factor analysis. NINDS tPA Stroke Trial Investigators. Stroke. 30:2347-2354 (1999).

Ma XM, Johnson RC, Mains RE *et al.*: Expression of kalirin, a neuronal GDP/GTP exchange factor of the trio family, in the central nervous system of the adult rat. J Comp Neurol. 429:388-402 (2001).

Maines MD. The heme oxygenase system: a regulator of second messenger gases. Ann Rev Pharmacol Toxicol. 37:517-554 (1997).

Mandel M, Gurevich M, Pauzner R *et al.*: Autoimmunity gene expression portrait: specific signature that intersects or differentiates between multiple sclerosis and systemic lupus erythematosus. Clin Exp Immunol. 138:164-170 (2004).

Maraganore DM, Andrade M, Lesnick TG *et al.*: High-resolution whole-genome association study of Parkinson disease. Am J Hum Genet. 77:685-693 (2005).

Markus HS: An introduction to stroke. In Stroke Genetics, Oxford University press, pp.1-30 (2003).

Marmot MG and Poulter NR: Primary prevention of stroke. Lancet. 339:344-347 (1992).

Marti HH, Wenger RH, Rivas LA *et al.*: Erythropoietin gene expression in human, monkey and murine brain. Eur J Neurosci. 8:666-676 (1996).

Martins Silva B, Thorlacius T, Benediktsson K *et al.*: A whole genome association study in multiple sclerosis patients from north Portugal. J Neuroimmunol. 143:116-119 (2003).

Matarin M, Brown WM, Scholz S *et al.*: A genome-wide genotyping study in patients with ischaemic

stroke: initial analysis and date release. Lancet Neurol. 6:414-420 (2007).

Matarin M, Brown WM, Singleton A *et al.*: Whole genome analyses suggest ischemic stroke and heart disease share an association with polymorphisms on chromosome 9p21. Stroke. 39:1586-1589 (2008).

McPherson CE, Eipper BA and Mains RE: Genomic organization and differential expression of Kalirin isoforms. Gene. 284:41-51 (2002).

McPherson CE, Eipper BA and Mains RE: Kalirin expression is regulated by multiple promoters. J Mol Neurosci. 22:51-62 (2004).

McPherson R, Pertsemlidis A, Kavaslar N et al: A common allele on chromosome 9 associated with coronary heart disease. Science. 316:1488-1491 (2007).

McVean G: Linkage disequilibrium and association mapping. Department of Statistics, University of Oxford (2002).

Médicos Sentinela: Acidente Vascular Cerebral 1992. Direcção Geral da Saúde. 1994:45-46 (1994).

Mehrabian M, Allayee H, Wong J *et al.*: Identification of 5-lipoxygenase as a major gene contributing to atherosclerosis susceptibility in mice. Circ Res. 91:120-126 (2002).

Menken M, Munsat TL, Toole JF: The global burden of disease study - Implications for neurology. Arch Neurol. 57:418-420 (2000).

Meschia JF, Brott TG, Brown RD Jr *et al.*: Phosphodiesterase 4D and 5-lipoxygenase activating protein in ischemic stroke. Ann Neurol. 58:351-361 (2005).

Meschia JF, Brott TG, Chukwudelunzu FE *et al.*: Verifying the stroke-free phenotype by structured telephone interview. Stroke. 31:1076-1080 (2000).

Middleton FA, Pato MT, Gentile KL *et al.*: Genome-wide linkage analysis of bipolar disorder by use of a high-density single-nucleotide-polymorphism (SNP) genotyping assay: a comparison with microsatellite marker assays and finding of significant linkage to chromosome 6q22. Am J Hum Genet. 74:886-897 (2004).

Miro X, Casacuberta JM, Gutierrez-Lopez MD, *et al.*: Phosphodiesterases 4d and 7a splice variants in the response of HUVEC cells to TNF-alpha(1). Biochem Biophys Res Commun. 274:415-421 (2000).

Mohr JP, Caplan LR, Melski JW: The Harvard cooperative stroke registry: a prospective registry. Neurology. 28:754-762 (1978).

Molina CA, Montaner J, Arenillas JF *et al.*: Differential pattern of tissue plasminogen activator-induced proximal middle cerebral artery recanalization among stroke subtypes. Stroke. 35:486-490 (2004).

Montaner J, Fernandez-Cadenas I, Molina CA *et al.*: Poststroke C-reactive protein is a powerful prognostic tool among candidates for thrombolysis. Stroke. 37:1205-1210 (2006).

Moore DF, Li H, Jeffries N *et al.*: Using peripheral blood mononuclear cells to determine a gene expression profile of acute ischemic stroke: a pilot investigation. Circulation. 111:212-221 (2005a).

Moore DF, Wright V, Yu H *et al.*: Gene expression changes during recovery from ischemic stroke: a longitudinal study and data analysis. Ann Neurol. 58:S42-S43 (2005b).

Moore JH, Gilbert JC, Tsai CT *et al.*: A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility. Journal of Theoretical Biology. 241:252-261 (2006).

Munshi A and Kaul S: Stroke genetics – focus on PDE4D gene. International Journal of Stroke. 3:188-182 (2008)

Murat Sumer M and Erturk O: Ischemic stroke subtypes: risk factors, functional outcome and recurrence. Neurol Sci. 22:449-454 (2002).

Nakayama T, Asai S, Sato N *et al.*: Genotype and haplotype association study of the STRK1 region on

5q12 among Japanese: a case-control study. Stroke. 37:69-76 (2006).

Namiranian K, Koehler RC, Sapirstein A *et al.*: Stroke outcomes in mice lacking the genes for neuronal hemeoxygenase-2 and nitric oxide synthase. Curr Neurovasc Res. 2:23-27 (2005).

Natowicz M and Kelley RI: Mendelian etiologies of stroke. Ann Neurol. 22:175-192 (1987).

Navarro-Núñez L, Lozano ML, Rivera J *et al.*: The association of the b1-tubulin Q43P polymorphism with intracerebral hemorrhage in men. Haematologica. 92:513-518 (2007).

Nielsen DM, Ehm MG and Weir BS: Detecting marker-disease association by testing for Hardy-Weinberg disequilibrium at a marker locus. Am J Hum Genet. 63:1531-1540 (1998).

Nilsson-Ardnor S, Janunger T, Wiklund PG *et al.*: Genome-wide linkage scan of common stroke in families from northern Sweden. Stroke. 38:34-40 (2007).

Nilsson-Ardnor S, Wiklund PG, Lindgren P *et al.*: Linkage of ischemic stroke to the PDE4D region on 5q in a Swedish population. Stroke. 36:1666-1671 (2005).

No authors listed 1:Tissue plasminogen activator for acute ischemic stroke. The National Institute of Neurological Disorders and Stroke rt-PA Stroke Study Group. N Engl J Med. 333:1581-1587 (1995).

No authors listed 2: GeneChip Expression Analysis - Data Analysis Fundamentals. Affymetrix P/N 701190 Rev. 4 (2004).

No authors listed 3: GeneChip Expression Analysis Technical Manual. Affymetrix P/N 702232 Rev. 2 (2005-2006).

O'Donnell CJ, Cupples LA, D'Agostino RB et al: Genome-wide association study for subclinical atherosclerosis in major arterial territories in the NHLBI's Framingham Heart Study. BMC Med Genet. 8:S4 (2007).

Ogura Y, Bonen DK, Inohara N *et al.*: A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. Nature. 411:603-606 (2001).

Oksjoki R, Jarva H, Kovanen PT *et al.*: Association between complement factor H and proteoglycans in early human coronary atherosclerotic lesions: implications for local regulation of complement activation. Arterioscler Thromb Vasc Biol. 23:630-636 (2003).

Olsen N, Sokka T, Seehorn CL *et al.*: A gene expression signature for recent onset rheumatoid arthritis in peripheral blood mononuclear cells. Ann Rheum Dis. 63:1387-1392 (2004).

Olson JM, Vongpunsawad S, Kuivaniemi H *et al.*: Search for intracranial aneurysm susceptibility gene(s) using Finnish families. BMC Med Genet. 3:7 (2002).

Onda H, Kasuya H, Yoneyama T *et al.*: Genome-wide - linkage and haplotype-association studies map intracranial aneurysm to chromosome 7q11. Am J Hum Genet. 69:804-819 (2001).

ONSA: De que se morre mais em Portugal: Principais causas de morte em Portugal de 1990 a 1999. Lisboa, (http://www.onsa.pt) (2003).

Osawa M, Yamaguchi T, Nakamura Y *et al.*: Erythroid expansion mediated by the Gfi-1B zinc finger protein: role in normal hematopoiesis. Blood. 100:2769-2777 (2002).

Otero Palleiro MM and Barbagelata López C: Etiologic subtypes of ischemic stroke in young adults aged 18 to 45 years: a study of a series of 93 patients. Rev Clin Esp. 207:158-165 (2007).

Ozaki K, Ohnishi Y, Iida A, *et al.*: Functional SNPs in the lumphotoxin-alpha gene that are associated with susceptibility to myocardial infarction. Nat Genet. 32:650-654 (2002).

Palmer D, Tsoi K, Maurice DH: Synergistic inhibition of vascular smooth muscle cell migration by phosphodiesterase 3 and phosphodiesterase 4 inhibitors. Circ Res. 82:852-861 (1998).

Palsdottir A, Abrahamson M, Thorsteinsson L *et al.*: Mutation in cystatin C gene causes hereditary brain haemorrhage. Lancet. 2:603-604 (1988).

Pan X, Arauz E, Krzanowski JJ *et al.*: Synergistic interactions between selective pharmacological inhibitors of phosphodiesterase isozyme families PDE III and PDE IV to attenuate proliferation of rat vascular smooth muscle cells. Biochem Pharmacol. 48:827-835 (1994).

Patel SR, Hartwig JH and Italiano JE Jr: The biogenesis of platelets from megakaryocyte proplatelets. J Clin Invest. 115:3348-3354 (2005).

Penzes P, Johnson RC, Alam MR *et al.*: An isoform of kalirin, a brain-specific GDP/GTP exchange factor, is enriched in the postsynaptic density fraction. J Biol Chem. 275:6395-6403 (2000).

Penzes P and Jones KA: Dendritic spine dynamics - a key role for kalirin-7. Trends Neurosci. 31:419-427 (2008).

Powles J, Kirov P, Feschieva N *et al.*: Stroke in urban and rural populations in north-east Bulgaria: incidence and case fatality findings from a 'hot pursuit' study. BMC Public Health. 2:24 (2002).

Pritchard JK: Are rare variants responsible for susceptibility to complex diseases? Am J Hum Genet. 69:124-137 (2001).

Purcell S, Neale B, Todd-Brown K et al: PLINK: a toolset for whole-genome association and population-based linkage analysis. Am J Hum Genet. 81:559-575 (2007).

Qian HR and Huang S: Comparison of false discovery rate methods in identifying genes with differential expression. Genomics. 86:495-503 (2005).

Rabiner CA, Mains RE, Eipper BA: Kalirin: a dual Rho guanine nucleotide exchange factor that is so much more than the sum of its many parts. Neuroscientist. 11:148-160 (2005).

Rainen L, Oelmueller U, Jurgensen S *et al.*: Stabilization of mRNA expression in whole blood samples. Clin Chem. 48:1883-1890 (2002).

Randi AM, Biguzzi E, Falciani F *et al.*: Identification of differentially expressed genes in coronary atherosclerotic plaques from patients with stable or unstable angina by cDNA array analysis. J Thromb Haemost. 1:829-835 (2003).

Ratovitski EA, Alam MR, Quick RA et al: Kalirin inhibition of inducible nitric-oxide synthase. J Biol Chem. 274:993-999 (1999).

Reich DE and Lander ES: On the allelic spectrum of human disease. Trends Genet. 17:502-510 (2001).

Reich DE, Cargill M, Bolk S et al: Linkage disequilibrium in the human genome. Nature. 411:199-204 (2001).

Risch N and Merikangas K:The future of genetic studies of complex human diseases. Science. 273:1516-1517 (1996).

Ritchie MD, Hahn LW, Roodi N *et al.*: Multifactor-Dimensionality Reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. Am J Hum Genet. 69:138-147 (2001).

Rodrigues M, Noronha MM, Vieira-Dias M *et al.*: Stroke in Europe: where is Portugal? POP-BASIS 2000 Study. Cerebrovasc Dis. 13:S72 (2000).

Rodriguez de Córdoba S, Esparza-Gordillo J, Goicoechea de Jorge E *et al.*: The human complement factor H:functional roles, genetic variations and disease associations. Mol Immunol. 41:355-367 (2004).

Rosa A, Fonseca BV, Krug T *et al.*: Mitochondrial haplogroup H1 is protective for ischemic stroke in Portuguese patients. BMC Med Genet. 9:57 (2008).

Rosand J, Bayley N, Rost N *et al.*: Many hypothesis but no replication for the association between PDE4D and stroke. Nat Genet. 38:1091-1092 (2006).

Roth A, Gill R and Certa U: Temporal and spatial gene expression patterns after experimental stroke in a rat model and characterization of PC4, a potential regulator of transcription. Mol Cell Neuro. 22:353-364 (2003).

Rothwell PM, Coull AJ, Giles MF et al: Change in stroke incidence, mortality, case-fatality, severity, and risk factors in Oxfordshire, UK from 1981 to 2004 (Oxford Vascular Study). Lancet. 363:1925-1933 (2004).

Rubattu S, Volpe M, Kreutz R *et al.*: Chromosomal mapping of quantitative trait loci contributing to stroke in a rat model of complex human disease. Nat Genet. 13:429-434 (1996);

Rudock ME, Ziegler JT, Lehtinen AB et al: Analysis of kalirin polymorphisms with cardiovascular risk, type 2 diabetes, metabolic syndrome in the Diabetes Heart Study. ASHG meeting 2008, abstract 2018 (2008).

Sacco RL: Newer risk factors for stroke. Neurology. 57:S31-S34 (2001).

Sakanaka M, Wen TC, Matsuda S *et al.*: In vivo evidence that erythropoietin protects neurons from ischemic damage. Proc Natl Acad Sci USA. 95:4635-4640 (1998);

Saleheen D, Bukhari S, Haider SR *et al.*: Association of phosphodiesterase 4D gene with ischemic stroke in a Pakistani population. Stroke. 36:2275-2277 (2005).

Satagopan JM, Verbel DA, Venkatraman ES *et al.*: Two-stage designs for gene-disease association studies with sample size constraints. Biometrics. 60:589-597 (2004).

Schadt EE, Monks SA, Drake TA *et al.*: Genetics of gene expression surveyed in maize, mouse and man. Nature. 422:297-302 (2003).

Schmidt H, Schmidt R, Niederkorn K *et al.*: b-Fibrinogen gene polymorphism (C148->T) is associated with carotid atherosclerosis: results of the Austrian Stroke Prevention Study. Arterioscler thromb Vasc Biol. 18:487-492 (1998).

Scholz H and Wagner N: Cerebral ischemia in heme oxygenase-2-deficient mice: the second filament makes the difference. Am J Physiol Regul Integr Comp Physiol. 285:R28-R29 (2003).

Schork NJ: Genetics of complex disease. Approaches, problems, and solutions. Am J Respir Crit Med. 156:S103-S109 (1997).

Schwer HD, Lecine P, Tiwari S *et al.*: A lineage-restricted and divergent b-tubulin isoform is essential for the biogenesis, structure and function of blood platelets. Curr Biol. 11:579-586 (2001).

Segal E, Fondufe-Mittendorf Y, Chen L *et al.*: A genomic code for nucleosome positioning. Nature. 442:772-778 (2006).

Shah SH, Kraus WE, Crossman DC et al: Serum lipids in the GENECARD study of coronary artery disease identify quantitative trait loci and phenotypic subsets on chromosomes 3q and 5q. Ann Hum Genet. 70:738-748 (2006).

Shariat-Madar Z, Mahdi F and Schmaier AH: Assembly and activation of the plasma kallikrein/kinin system: a new interpretation. Int Immunopharmacol. 2:1841-1849 (2002).

Sharma P: Genes for ischaemic stroke: strategies for their detection. J Hypertens. 14:277-285 (1996).

Sheehan JJ, Tsirka SE: Fibrin-modifying serine proteases thrombin, rt-PA, and plasmin in ischemic stroke: a review. Glia. 50:340-350 (2005).

Shen CD, Zhang WL, Sun K *et al.*: Interaction of genetic risk factors confers higher risk for thrombotic stroke in male Chinese: a multicenter case-control study. Ann Hum Genet. 71:620-629 (2007).

Shephard N, John S, Cardon L *et al.*: Will the real disease gene please stand up? BMC Genet. Suppl 6:S66 (2005).

Shoemaker J, Painter I and Weir BS: A Bayesian characterization of Hardy-Weinberg disequilibrium. Genetics. 149:2079-2088 (1998).

Sklar P, Pato MT, Kirby A *et al.*: Genome-wide scan in Portuguese Island families identifies 5q31-5q35 as a susceptibility locus for schizophrenia and psychosis. Mol Psychiatry. 9:213-218 (2004).

Slowik A, Turaj W, Dziedzic T *et al.*: DD genotype of ACE gene is a risk factor for intracerebral hemorrhage. Neurology. 63:359-361 (2004).

Song Q, Cole JW, O'Connell JR *et al.*: Phosphodiesterase 4D polymorphisms and the risk of cerebral infarction in a biracial population: the Stroke Prevention in Young Women Study. Hum Mol Genet. 15:2468-2478 (2006).

Spanbroek R, Gräbner R, Lötzer K *et al.*: Expanding expression of the 5-lipoxygenase pathway within the arterial wall during human atherogenesis. Proc Natl Acad Sci USA. 100:1238-1243 (2003).

Staton JM, Sayer MS, Hankey GJ *et al.*: Association between phosphodiesterase 4D gene and ischaemic stroke. J Neurol Neurosurg Psychiatry. 77:1067-1069 (2006).

Stenzel-Poore MP, Stevens SL, King JS *et al.*: Preconditioning reprograms the response to ischemic injury and primes the emergence of unique endogenous neuroprotective phenotypes: a speculative synthesis. Stroke. 38:680-685 (2007).

Stenzel-Poore MP, Stevens SL, Xiong Z *et al.*: Effect of ischaemic preconditioning on genomic response to cerebral ischaemia: similarity to neuroprotective strategies in hibernation and hypoxia-tolerant states. Lancet. 362:1028-1037 (2003).

Stoclet JC, Keravis T, Komas N *et al.*: Cyclic nucleotide phophodiesterases as therapeutic targets in cardiovascular diseases. Expert Opin Investig Drugs. 4:1081C-1100C (1995).

Storey JD: A direct approach to false discovery rate. J R Stat Soc B, 64:479-498 (2002).

Storini C, Bergamaschini L, Gesuete R *et al.*: Selective inhibition of plasma kallikrein protects brain from reperfusion injury. J Pharmacol Exp Ther. 318:849-854 (2006).

Sullivan KF: Structure and utilization of tubulin isotypes. Annu Rev Cell Biol. 4:687-716 (1988).

Sun Y, Huang Y, Chen X *et al.*: Association between the PDE4D gene and ischaemic stroke in the Chinese Han population. Clin Sci (Lond). 117:265-272 (2009).

Sveinbjornsdottir S, Einarsson G, Magnusdottir S *et al.*: Systematic registration of patients with stroke and TIA admitted to the National University Hospital, Reykjavik, Iceland, in 1997. Læknablaȥiȝ. Suppl 36:54 (1998).

Tanaka M, Gong JS, Zhang J *et al.*: Mitochondrial genotype associated with longevity. Lancet. 351:185-186 (1998).

Tang Y, Lu A, Aronow BJ *et al.*: Blood genomic responses differ after stroke, seizures, hypoglycemia, and hypoxia: blood genomic fingerprints of disease. Ann Neurol. 50:699-707 (2001).

Tang Y, Nee AC, Lu A *et al.*: Blood genomic expression profile for neuronal injury. J Cereb Blood Flow Metab. 23:310-319 (2003).

Tang Y, Xu H, Du X *et al.*: Gene expression in blood changes rapidly in neutrophils and monocytes after ischemic stroke in humans: a microarray study. J Cereb Blood Flow Metab. 26:1089-1102 (2006).

Terent A: Increasing incidence of stroke among Swedish women. Stroke. 19:598-603 (1998).

Thakkinstian A, Han P, McEvoy M *et al.*: Systematic review and meta-analysis of the association between complement factor H Y402H polymorphisms and age-related macular degeneration. Hum Mol Genet. 15:2784-2790 (2006).

The International HapMap Project. Nature. 426:789-796 (2003).

Thomas D, Xie RR and Gebregziabher M: Two-stage sampling designs for gene association studies. Genet Epidemiol. 27:401-414 (2004).

Thomas DC and Witte JS: Point: population stratification: a problem for casa-control studies of candidate-gene associations? Cacer Epidemiol Biomarkers Prev. 11:505-512 (2002).

Thompson DW and Furlan Aj: Clinical epidemiology of stroke. Neurosurg Clin N Am. 8:265-269 (1997).

Tiret L, Rigat B, Visvikis S *et al.*: Evidence, from combined segregation and linkage analysis, that a variant of the angiotensin I-converting enzyme (ACE) gene controls plasma ACE levels. Am J Hum Genet. 51:197-205 (1992).

Tonk M and Haan J: A review of genetic causes of ischemic and hemorrhagic stroke. J Neurol Sci. 257:273-279 (2007).

Tournier-Lasserve E: New players in the genetics of stroke. N Engl J Med. 347:1711-1712 (2002)

Towfighi A, Greenberg SM and Rosand J: Treatment and prevention of primary intracerebral hemorrhage. Semin Neurol. 25:445-452 (2005).

Valdimarsson EM, Sigurdsson G and Jakobsson F: Etiology and treatment of cerebral ischemia at the Department of Neurology and Rehabilitation Medicine at Reykjavik City Hospital. Læknablaȝiȝ. 84:921-927 (1998).

van Duijn CM, de Knijff P, Cruts M *et al.*: Apolipoprotein E4 allele in a population-based study of early-onset Alzheimer's disease. Nat Genet. 7:74-78 (1994).

van Rijn MJ, Slooter AJ, Schut AF *et al.*: Familial aggregation, the PDE4D gene, and ischemic stroke in a genetically isolated population. Neurology. 65:1203-1209 (2005).

van Swieten JC, Koudstaal PJ, Visser MC *et al.*: Interobserver agreement for the assessment of handicap in stroke patients. Stroke. 19:604-607 (1988).

Vance JM and Ben Othmane K: Methods of Genotyping. In: Haines JL, Pericak-Vance MA, eds. Approaches to Gene Mapping in Complex Human Diseases. New York, NY: Wiley-Liss, pp.213-228 (1998).

Vemuganti R and Dempsey RJ: Carotid atherosclerotic plaques from symptomatic stroke patients share the molecular fingerprints to develop in a neoplastic fashion: a microarray analysis study. Neuroscience. 131:359-374 (2005).

Venken T, Claes S, Sluijs S *et al.*: Genome-wide scan for affective disorder susceptibility loci in families of a northern Swedish isolated population. Am J Hum Genet. 76:237-248 (2005).

Vidal R, Frangione B, Rostagno A *et al.*: A stop-codon mutation in the BRI gene associated with familial British dementia. Nature. 399:776-781 (1999).

Wald DS, Law M and Morris JK: Homocysteine and cardiovascular disease: evidence on causality from a meta-analysis. BMJ. 325:1202-1206 (2002).

Wang L, Hauser ER, Shah SH *et al.*: Peakwide mapping on chromosome 3q13 identifies the kalirin gene as a novel candidate gene for coronary artery disease. Am J Hum Genet. 80:650-663 (2007).

Wang WYS, Barratt BJ, Clayton DG *et al.*: Genome-wide association studies: theoretical and practical concerns. Nat Genet. 6:109-118 (2005).

Wardlaw JM, Warlow CP and Counsell C: Systematic review of evidence on thrombolytic therapy for acute ischaemic stroke. Lancet. 350:607-614 (1997).

Weiss KM and Terwilliger JD: How many diseases does it take to map a gene with SNPs? Nat Genet. 26:151-157 (2000).

Whitney AR, Diehn M, Popper SJ *et al.*: Individuality and variation in gene expression patterns in human blood. Proc Natl Acad Sci USA. 100:1896-1901 (2003).

WHO: World Health Organization, Health for all database (1999).

WHO: World Health Organization, The World Heath Report 2002: Reducing Risks, Promoting Healthy Life. Geneva, Switzerland: World Health Organization (2002).

WHO: World Health Organization, The Atlas of Heart Disease and Stroke (2004).

WHO MONICA Project Principal Investigators: The World Health Organization MONICA Project (monitoring trends and determinants in cardiovascular disease): a major international collaboration. J Clin Epidemiol. 41:105-114 (1988).

Woo D, Kaushal R, Chakraborty R *et al.*: Association of apolipoprotein E4 and haplotypes of the hapoliprotein E gene with lobar intracerebral hemorrhage. Stroke. 36:1874-1879 (2005b).

Woo D, Kaushal R, Kissela B *et al.*: Association of Phosphodiesterase 4D with ischemic stroke: a population-based case-control study. Stroke. 37:371-376 (2006).

Woo D, Sekar P, Chakrabortly R *et al.*: Genetic epidemiology of intracerebral hemorrhage. J Stroke Cerebrovasc Dis. 14:239-243 (2005a).

Woods A and Couchman JR: Syndecan 4 heparan sulfate proteoglycan is a selectively enriched and widespread focal adhesion component. Molecular Biology of the Cell. 5:183-192 (1994).

Woodside KJ, Hernandez A, Smith FW *et al.*: Differential gene expression in primary and recurrent carotid stenosis. Biochem Biophys Res Commun. 302:509-514 (2003).

Wright A, Charlesworth B, Rudan I *et al.*: A polygenic basis for late-onset disease. Trends Genet. 19:97-106 (2003).

Wuttge DM, Sirsjo A, Eriksson P, *et al.*: Gene expression in atherosclerotic lesion of ApoE deficient mice. Mol Med. 7:383-392 (2001).

Xia CF, Yin H, Yao YY *et al.*: Kallikrein protects against ischemic stroke by inhibiting apoptosis and inflammation and promoting angiogenesis and neurogenesis. Hum Gene Ther. 17:206-219 (2006).

Xu H, Tang Y, Liu DZ *et al.*: Gene expression in peripheral blood differs after cardioembolic compared with large-vessel atherosclerotic stroke: biomarkers for the etiology of ischemic stroke. Journal of Cerebral Blood Flow & Metabolism. 28:1320-1328 (2008).

Xue H, Wang H, Song X *et al.*: Phosphodiesterase 4D gene polymorphism is associated with ischaemic and haemorrhagic stroke. Clin Sci (Lond). 16:335-340 (2009).

Yamori Y, Horie R, Handa H *et al.*: Pathogenetic similarity of strokes in stroke-prone spontaneously hypertensive rats and humans. Stroke. 7:46-53 (1976).

Yang LV, Wan J, Ge Y *et al.*: The GATA site-dependent hemogen promoter is transcriptionally regulated by GATA1 in hematopoietic and leukemia cells. Leukemia. 20:417-4125 (2006).

Ye S, Willeit J, Kronenberg F *et al.*: Association of genetic variation on chromosome 9p21 with susceptibility and progression of atherosclerosis: a population-based, prospective study. J Am Coll Cardiol. 52:378-384 (2008).

Youn H, Jeoung M, Koo Y et al: Kalirin is under-expressed in Alzheimer's disease hippocampus. J Alzheimers Dis. 11:385-397 (2007a).

Youn H, Ji I, Ji HP *et al.*: Under-expression of Kalirin-7 Increases iNOS activity in cultured cells and correlates to elevated iNOS activity in Alzheimer's disease hippocampus. J Alzheimers Dis. 12:271-281 (2007b).

Youssef MY, Mojiminiyi AO and Abdella NA: Plasma concentrations of C-reactive protein and total homocysteine in relation to the severity and risk factors for cerebrovascular disease. Transl Res. 150:158-163 (2007).

Zee RY, Brophy VH, Cheng S *et al.*: Polymorphisms of the phosphodiesterase 4D, cAMP-Specific (PDE4D) gene and risk of ischemic stroke: a prospective, nested case-control evaluation. Stroke. 37:2012-2017 (2006a).

Zee RY, Cheng S, Hegener HH *et al.*: Genetic variants of arachidonate 5-lipoxygenase-activating protein, and risk of incident myocardial infarction and ischemic stroke: a nested case-control approach. Stroke. 37:2007-2011 (2006b).

Zee RY, Diehl KA, Ridker PM: Complement factor H Y402H gene polymorphism, C-reactive protein, and risk of incident myocardial infarction, ischaemic

stroke, and venous thromboembolism: a nested case-control study. Atherosclerosis. 187:332-335 (2006c).

Zhang J, Finney RP, Clifford RJ *et al.*: Detecting false expression signals in high-density oligonucleotide arrays by an in silico approach. Genomics. 85:297-308 (2005).

Zhang WL, Yang XM, Shi J *et al.*: Polymorphism of SG13S114T/A in the ALOX5AP gene and the risk for stroke in a large Chinese cohort. Yi Chuan Xue Bao. 33:678-684 (2006).

Zintzaras E, Rodopoulou P and Sakellaridis N: Variants of the arachidonate 5-lipoxygenase-activating protein (ALOX5AP) gene and risk of stroke: a huge gene-disease association review and meta-analysis. Am J Epidemiol. 169:523-532 (2009).

Zondervan KT and Cardon LR: The complex interplay among factors that influence association. Nat Rev Genet. 5:89-100 (2004).

Zou GY and Donner A: The merits of testing Hardy-Weinberg Equilibrium in the analysis of unmatched case-control data: a cautionary note. Annals of Human Genetics. 70:923-933 (2006).

APPENDIX A – STUDY MANUAL AND CONSENT FORM

We developed a study manual for this project describing the study objectives, inclusion and exclusion criteria for subject participation, the diagnostic criteria and the evaluation procedures for patients and controls. To collect the extensive information we need, we especially developed information collection forms and other documents such as the information leaflet and the consent form. In this Appendix we present all these documents (Appendix A.1 to A.7) that were approved by the ethics committees of several participating institutions and of the INSARJ.

Appendix A.1. – Information leaflet

Folha de Informação

Genética de AVC Individual

Os AVCs representam a principal causa de morte em Portugal. Os défices neurológicos e físicos causados por AVCs têm um impacto financeiro e socioeconómico substancial.

Os AVCs são uma doença complexa que resulta de factores genéticos e ambientais. Os principais factores de risco incluem história familiar, idade, hipertensão, hipercolesterolémia, diabetes, doença cardiovascular, consumo de tabaco e álcool.

A identificação de genes que aumentem a susceptibilidade de desenvolver AVCs teria um elevado impacto na saúde pública, desde aumentar a motivação para modificar hábitos e estilos de vida em indivíduos susceptíveis até fornecer informação biológica e clínica básica sobre o desenvolvimento, prevenção e tratamento de AVCs.

A componente genética dos AVCs já foi demonstrada, mas muito poucos genes de susceptibilidade para as formas mais comuns de AVCs foram até ao momento identificados.

Propomos recolher amostras de sangue de 500 indivíduos afectados por AVCs, 500 familiares não afectados, e 500 controlos (geneticamente não relacionados com os doentes e não afectados). As amostras serão recolhidas em diversos hospitais Portugueses e conservadas e processadas no laboratório de Genética Humana no IGC (responsável: Dra. Sofia Oliveira).

Pode encontrar mais informação sobre este projecto na nossa página na internet (http://www.igc.gulbenkian.pt/indexpt.php).

Appendix A.2. – Consent form

Folha de consentimento

Genética de AVC Individual

Nome do participante no estudo: _____

O objectivo deste estudo, os procedimentos a serem seguidos, os riscos e benefícios foram-me explicados. Tive oportunidade de colocar questões adicionais que foram respondidas satisfatoriamente. Fui informado que posso contactar o Prof. José Ferro para responder a qualquer dúvida que tenha em qualquer momento sobre a investigação. Consinto em participar no estudo sabendo que posso desistir a qualquer momento sem que isso interfira com o meus cuidados médicos. Foi-me entregue uma cópia desta folha de consentimento.

_____

Assinatura do participante ou do responsável legal

_____

Nome do responsável legal e relação com o participante

_____ _____

Assinatura da pessoa que obteve o consentimento Data

A sua participação neste estudo rege-se pelos seguintes princípios:

1. Participação: a participação neste estudo é voluntária. Embora a sua participação possa não lhe trazer benefícios imediatos, o conhecimento gerado por este estudo poderá um dia beneficiá-lo a si, à sua família ou a outros. Mediante um pedido por escrito ao médico que o recrutou para este estudo, poderá solicitar a qualquer momento para ser excluído do estudo e as respectivas amostras biológicas destruídas.

2. Duração e objectivo do estudo: o estudo continuará até se identificarem e caracterizarem os factores genéticos que contribuem para a susceptibilidade a AVCs.

3. Uso das amostras biológicas: as amostras biológicas (3 tubos de sangue dos quais se poderá extrair ADN e/ou outros componentes do sangue) serão usadas apenas para as finalidades expressas no âmbito deste estudo e serão destruídas quando o estudo terminar. As amostras serão usadas apenas para investigação. Não serão guardadas amostras para testes clínicos ou de diagnóstico, nem para depósito pessoal ou comercial.

4. Confidencialidade: será mantida a confidencialidade dos dados de modo a evitar discriminação ou uso da informação prejudicial aos participantes. O acesso à informação com identificadores pessoais (por exemplo nome, número de processo clínico) será restrita à equipa clínica do projecto. Por outro lado, a equipa clínica não terá acesso a dados individuais produzidos pela equipa laboratorial. A informação sobre os participantes terá acesso restrito aos investigadores, codificada, e sem identificadores pessoais. A apresentação dos resultados da investigação em ambientes profissionais (por exemplo revistas, conferências) será feita sem identificadores pessoais.

5. Resultados da investigação: este estudo não fornecerá resultados individualizados aos participantes ou aos seus médicos dado que os resultados da investigação são geralmente preliminares e as suas implicações são desconhecidas durante vários anos. Exceptuam-se os resultados que conduzirem à descoberta de condições evitáveis ou tratáveis. Os participantes receberão anualmente um resumo dos progressos científicos da investigação realizada no âmbito deste projecto.

Appendix A.3. – Sample acquisition form

Folha de aquisição de amostra

Hospital ☐☐☐ **Estudo** { ☐ AVC - Famílias ☐ AVC - Indivíduos

Família ☐☐☐☐ Data de nascimento ☐☐/☐☐/☐☐☐☐ (Dia / Mês / Ano)

Indivíduo ☐☐☐☐ Sexo ☐ M ☐ F

Número do processo clínico _____

Nome do indivíduo _____

Formulário de consentimento obtido? ☐ Não ☐ Sim **Data de assinatura do consentimento** ☐☐/☐☐/☐☐☐☐ (Dia / Mês / Ano)

Data de colheita da amostra ☐☐/☐☐/☐☐☐☐ (Dia / Mês / Ano) **Amostra recolhida por:** _____

Número e tipo de tubos recolhidos

☐ BD Vacutainer CPT 8.0 ml: Quantos? _____

☐ EDTA S-Monovette 7.5 ml Quantos? _____

☐ Serum Gel S-Monovette 4.9 ml Quantos? _____

☐ EDTA Hemograma: Quantos? _____

☐ Outros: _____ Quantos? _____

Número de aquisição de amostra (código de barras) ☐

Appendix A.4. – Participant Report Form (PRF)

Genética de AVCs - Participant Report Form

Campos assinalados com "*" são de preenchimento obrigatório.

1. IDENTIFICAÇÃO DO PARTICIPANTE **(TODOS OS PARTICIPANTES)**

***Estudo:** □ AVC – Famílias □ AVC - Indivíduos □ Outro

***Hospital:**
□ H. Distrital de Mirandela
□ H. de São Pedro
□ H. Santo António
□ Outro

***Família:** __ __ __ __ *[sempre 4 dígitos]*

***Indivíduo:** __ __ __ __ *[sempre 4 dígitos]*

***Tipo de participante:**
□ Doente
□ Controlo
□ Outro

***Iniciais do médico que preencheu este formulário:** __ __ __ *[3 letras]*

***Data de observação:** (dia) __ __ / (mês) __ __ / (ano) __ __ __ __ *[2/2/4 dígitos]*

***Dados fornecidos pelo:** □ Participante □ Outro

***Iniciais do participante:** __ __ __ *[3 letras]*

Morada: _____ *[texto até 300 letras]*

_____

Telefone: (__ __) __ __ __ __ __ __ __ __ *[9 dígitos]*

***Data de nascimento:** (dia) __ __ / (mês) __ __ / (ano) __ __ __ __ *[2/2/4 dígitos]*

***Etnia:** □ Caucasiano □ Outro

***Género:** □ M □ F

Número de anos de escolaridade: __ __ *[max 2 dígitos]*

Profissão actual ou última conhecida: _____ *[texto até 100 letras]*

_____

***Nº da amostra de sangue:** __ __ __ __ __ __ *[2 letras e 4 dígitos]*

2. ANTECEDENTES PESSOAIS **(TODOS OS PARTICIPANTES)**

***Tem ou teve TA alta (>140-85 mmHg) pelo menos duas vezes na vida ou toma medicamentos para baixar a tensão?**

□ Sim Nº de anos com HTA __ __ *[max 2 dígitos]*
□ Não
□ N/S

***É diabético ou toma insulina ou medicamentos para a diabetes?**

☐ Sim Nº de anos com diabetes __ __ *[max 2 dígitos]*
☐ Não
☐ N/S

***Tem colesterol alto (>200 mg/dl) ou toma medicamentos para o colesterol?**

☐ Sim
☐ Não
☐ N/S

***Fuma ou já fumou diáriamente (antes do 1º AVC)?**

☐ Sim Nº médio de cigarros por dia __ __ *[max 2 dígitos]*
☐ Não
☐ N/S

***Número médio de copos/dia de qualquer bebida alcoólica (antes do 1º AVC):**

Copos/dia __ __ *[max 2 dígitos]*

***Usava anticoncepcionais orais antes do 1º AVC?**

☐ Sim
☐ Não
☐ N/S

***Usava terapêutica hormonal de substituição antes do 1º AVC?**

☐ Sim
☐ Não
☐ N/S

***Número de horas de actividade física por semana no último ano:**

Horas __ __ *[max 2 dígitos]*

Doença cardíaca:

*Enfarte do miocárdio ☐ Sim ☐ Não ☐ N/S
*Fibrilhação auricular ☐ Sim ☐ Não ☐ N/S
*Outra doença cardíaca ☐ Sim Qual? _____
☐ Não *[texto até 100 letras]*
☐ N/S

***Outra doença vascular?**

☐ Sim Qual? _____ *[texto até 100 letras]*
☐ Não
☐ N/S

***Já fez alguma cirurgia?**

☐ Sim Qual? _____ *[texto até 100 letras]*
☐ Não
☐ N/S

***Outros antecedentes relevantes?**

☐ Sim Qual? _____ *[texto até 100 letras]*
☐ Não
☐ N/S

***Peso (actual ou último conhecido):**

Peso (kg) __ __ __ *[max 3 dígitos]*

***Altura (actual ou última conhecida):**

Altura (cm) __ __ __ *[max 3 dígitos]*

3. CARACTERIZAÇÃO DO AVC **(TODOS OS DOENTES)**

Nº total de AVCs: __ □ N/S *[max 1 dígito]*

Idades em que ocorreram os AVCs: __ __ __ __ __ __ __ __ □ N/S *[max 2/2/2/2 dígitos]*

Tipo etiológico do primeiro AVC:
□ AVC isquémico
□ AVC hemorrágico
□ N/S

Pontuação na NIHSS na admissão: __ __ □ N/S *[max 2 dígitos]*

Tipo clínico de AVC: classificação de Oxfordshire

a. Disfunção cerebral superior "de novo" (ex: disfasia ou dyscalculia ou alteração visuo-espacial)	□ S	□ N	□ N/S
b. Défice motor e/ou sensitivo ipsilateral afectando pelo menos uma das três áreas: face, membro superior ou membro inferior	□ S	□ N	□ N/S
c. Alteração no campo visual homónimo (hemianópsia)	□ S	□ N	□ N/S
d. Défice motor puro unilateral afectando pelo menos duas das três áreas: face, membro superior ou membro inferior	□ S	□ N	□ N/S
e. Défice sensitivo puro unilateral afectando pelo menos duas das três áreas: face, membro superior ou membro inferior	□ S	□ N	□ N/S
f. Défice sensitivo motor unilateral afectando pelo menos duas das três áreas: face, membro superior ou membro inferior	□ S	□ N	□ N/S
g. Hemiparésia atáxica ipsilateral	□ S	□ N	□ N/S
h. Sinais inequívocos de disfunção do tronco cerebral/cerebelo	□ S	□ N	□ N/S

Tipo etiológico de AVC: classificação TOAST

AVC isquémico
- □ Cardioembólico (cardiopatia de alto-médio risco embolígena)
- □ Aterotrombótico, grandes vasos (estenose >50% na artéria sintomática)
- □ Pequenos vasos (lacunar, motor puro, sensitivo puro, sensitivo-motor, hemiparésia atáxica e disartria-mão desajeitada)
- □ Isquémico de outra etiologia
- □ Isquémico de causa desconhecida, investigação incompleta
- □ Isquémico de causa desconhecida, investigação completa
- □ Isquémico de causas múltiplas

AVC hemorrágico

Hemorragia intracerebral
- □ aneurisma
- □ MAV
- □ cavernoma
- □ outro
- □ N/S

Hemorragia subacnóideia
- □ aneurisma
- □ outro
- □ N/S

□ Outro Qual? _____ *[texto até 100 letras]*

☐ Desconhecido (sem TAC/RM ou TAC>1 mês depois do AVC)

Meios complementares de diagnóstico efectuados

TAC	☐ S	☐ N	☐ N/S
RM	☐ S	☐ N	☐ N/S
Avaliação da circulação cerebral extracraniana por eco-Doppler ou angiografia (qualquer tipo)	☐ S	☐ N	☐ N/S
Avaliação da circulação cerebral intracraniana por Doppler transcraniano ou angiografia (qualquer tipo)	☐ S	☐ N	☐ N/S
Avaliação cardioembólica: ECG ou Holter	☐ S	☐ N	☐ N/S
Ecocardiograma TT ou TE	☐ S	☐ N	☐ N/S

Outros: _____ *[texto até 100 letras]*

Pontuação actual na Escala Modificada de Rankin (na alta)

☐ Sem sintomas

☐ Sintomas menores

☐ Alguma limitação no estilo de vida, mas independente

☐ Limitação importante no estilo de vida ou necessidade de alguma ajuda

☐ Dependente, mas sem requerer atenção constante

☐ Dependência completa, necessita de atenção dia e noite

☐ Falecido

☐ N/S

Observações

Outros comentários: _____

_____

_____

[texto até 500 letras]

Appendix A.5. – Supplement of the PRF

Suplemento Participant Report Form

4. OUTRAS INFORMAÇÕES

(TODOS OS PARTICIPANTES)

Medicamentos que se encontra a tomar:

Nome	Objectivo[+]	A tomar desde	Última vez que foi tomado		
			Dia	Hora	Dose

[+]Razão pela qual o medicamento foi prescrito (por exemplo, controlar a tensão, diabetes, colesterol).

Observações

Outros comentários: _____

_____

_____

[texto até 500 letras]

TODOS OS DOENTES

Pontuação actual na Escala Modificada de Rankin

- ☐ Sem sintomas
- ☐ Sintomas menores
- ☐ Alguma limitação no estilo de vida, mas independente
- ☐ Limitação importante no estilo de vida ou necessidade de alguma ajuda
- ☐ Dependente, mas sem requerer atenção constante
- ☐ Dependência completa, necessita de atenção dia e noite
- ☐ Falecido
- ☐ N/S

Pontuação actual na Escala de Barthel (ver página seguinte): _____ *[0-100]*

Índice de Barthel

ALIMENTAÇÃO _____
10 Independente.
5 Necessita de ajuda, por exemplo, para utilizar o talher.
0 Incapaz.

BANHO _____
5 Toma banho sem ajuda.
0 Dependente.

TOILETTE PESSOAL _____
5 Lava a cara, penteia-se, lava os dentes, barbeia-se independentemente.
0 Necessita de ajuda.

VESTIR _____
10 Independente. Aperta os laços dos sapatos, aperta os cintos, abotoa-se.
5 Necessita de auxílio por incapacidade.
0 Totalmente dependente.

CONTROLO INTESTINAL _____
10 Sem complicações. Capaz de administrar enema ou supositório se necessário.
5 Complicações ocasionais ou necessita de ajuda quando administra enema ou supositório.
0 Incontinente.

CONTROLO DA BEXIGA _____
10 Sem complicações. Capaz de recolher a arrastadeira no caso de a utilizar.
5 Complicações ocasionais ou necessita de ajuda com a utilização de arrastadeira.
0 Incontinente.

DESLOCAÇÃO PARA O QUARTO DE BANHO _____
10 Independente.
5 Necessita de ajuda para se equilibrar, para o manuseamento de roupas ou de papel higiénico.
0 Dependente.

DESLOCAÇÃO CADEIRA / CAMA _____
15 Independente, capaz de manusear o travão da cadeira de rodas e levantar o descanso para os pés.
10 Assistência mínima ou supervisão.
5 Capaz de se sentar embora necessite de total assistência para se deslocar.
0 Incapaz.

DEAMBULAÇÃO _____
15 Independente durante um percurso de 50m. Pode utilizar apoios na escada rolante.
10 Precisa de auxílio para percorrer 50m.
5 Independente, em cadeira de rodas, para um percurso de 50m.
0 Imóvel.

SUBIR ESCADAS _____
10 Independente. Pode utilizar apoios.
5 Necessita de auxílio ou supervisão.
0 Incapaz.

TOTAL _____

Appendix A.6. – Pedigree sheet

Folha da árvore genealógica

Estudo ☐ AVC - Famílias ☐ AVC - Indivíduos

Hospital ☐☐☐ **Família** ☐☐☐☐

Preenchido por _____ **Data** ☐☐/☐☐/☐☐☐☐ (dia) (mês) (ano)

Árvore genealógica

Indivíduo						
Ano de nascimento						
Ano de morte						
Causa da morte						
Principais doenças						

Indivíduo						
Ano de nascimento						
Ano de morte						
Causa da morte						
Principais doenças						

Appendix A.7. – Questionnaire for Verifying Stroke-Free Status

Questionário para Verificar o Estado Livre de AVC (QVSFS)

Hospital ☐☐☐ **Estudo** ☐ AVC - Famílias ☐ AVC - Indivíduos

Família ☐☐☐☐

Data de nascimento ☐☐/☐☐/☐☐☐☐ (Dia / Mês / Ano)

Indivíduo ☐☐☐☐ **Sexo** ☐ M ☐ F

Nome do indivíduo _____

Iniciais de quem preencheu _____

1. Alguma vez um médico lhe disse que tinha tido um AVC, uma trombose, uma embolia ou um derrame cerebral?
   Sim ☐ Não ☐ desconhecido ☐

2. Alguma vez um médico lhe disse que tinha tido um AIT, um pequeno AVC (mini AVC), um acidente isquémico transitório ou um espasmo cerebral?
   Sim ☐ Não ☐ desconhecido ☐

3. Alguma vez teve uma paralisia súbita e sem dor num dos lados do corpo?
   Sim ☐ Não ☐ desconhecido ☐

4. Alguma vez teve perda de sensibilidade súbita num dos lados do corpo ou sentiu um dos lados do corpo como morto ou adormecido?
   Sim ☐ Não ☐ desconhecido ☐

5. Alguma vez teve uma perda súbita e sem dor da visão de um ou dos dois olhos?
   Sim ☐ Não ☐ desconhecido ☐

6. Alguma vez perdeu subitamente metade da visão?
   Sim ☐ Não ☐ desconhecido ☐

7. Alguma vez perdeu subitamente a capacidade de perceber o que as pessoas lhe diziam?
   Sim ☐ Não ☐ desconhecido ☐

8. Alguma vez perdeu subitamente a capacidade de falar ou de escrever?
   Sim ☐ Não ☐ desconhecido ☐

APPENDIX B – GENOTYPING PRIMERS

We genotyped 345 of the selected SNPs in a 384-well format using Sequenom´s (Sequenom, USA) iPlex assays. In this Appendix are presented for each SNP the primer sequences that were designed in house using Sequenom's MassARRAY® Assay Design 3.0 software according to the Cambridge reference sequence (http://wwww.mitomap.org/mitoseq.html) (Table B.1). It is indicated which forward and reverse primers were used for PCR amplification and which is the respective extension primer. In the Table B.2 are presented the same information for the three SNPs that were genotyped to validate our imputed results for the KALRN gene.

Table B.1: PCR (F and R) and extension (E) primer sequences designed to genotyped selected SNPs using Sequenom's iPlex assays. Genes are in the same order than they are presented in the materials and methods and SNPs ordered by name. The plex number where they were included is indicated.

Gene	SNP	F Primer	R Primer	E Primer	# Plex
PDE4D	rs10051720	AATGATTTCATCCCTCCCGC	CTTTGCTGTAGTGTGGTGAG	CAAACAAATTGGTTCACAGG	1
	rs10066510	TTGTCTTTCCTGCTAGCAAG	TTGAACGAGGCTGAAAGATG	GCTAGCAAGGCCTACC	5
	rs10471476	GAAACAGAGAATGCAGGAGG	GTCACCTCAAATTCAGTGTG	GTTCAGTGTGCCCAGAC	8
	rs10471477	ATGTCCACTCAGTGACAAGC	GGCATGCTCATCTATTCCAG	TAAACACTCATGTATTGATCACTCC	2
	rs10471477	ATGTCCACTCAGTGACAAGC	GGCATGCTCATCTATTCCAG	ACTCATGTATTGATCACTCC	10
	rs10514895	CCATTAGATTATCATACAGTC	CAGCAAATTCCAAGTACGGC	GAGATTATCATACAGTCTGACTTTCAA	2
	rs10939851	CAGGAAGTTAGACTTGAAAG	AGCAACTGGCAGGTTCAAG	GTGAAAGGCTTTTTGAAACAA	1
	rs11739760	TGCTCAAGTGAGGAGCAAAC	GCTGTGGGCACATTTGGTTT	AGGAGCAAACCATAGCGAGC	2
	rs11745887	ATTGCAGACGCTCAACTTCC	CTCTCTTCTCAAAGAGAAAAC	TTCTCAAAGAGAAAACATTTTATTT	10
	rs11746901	AAGTGATGGAAGTCCCAGAG	GCTCAAGGAAATTCTGGTGG	ATAGGAGTATTTACGTCTTGCATC	2
	rs11951422	GATCACTTCCTTCAAGGAGC	GTGATTGCTGTTCATGGAAG	CCCAAAGGAGCCATCCCTGG	7
	rs12153798	GTAGAAACTCTGGAGTCAGG	TAAGTAGCTTGCTCAGGGTC	ACATCCCAGCTCTGACTC	10
	rs12188950	GCTGCCATCATTTTACATTTC	CTGAGAGCAAGCAGCAAATA	ACATCATTTTACATTTCTCCTGTTA	3
	rs12189147	GCCTTTCTATGGGCTTCTTC	TGGGTCGCTTCTACTTTGTG	GTGAGAGGCTTGCTCCC	2
	rs12515974	AGTAGGTGGAATCAAGAGTG	TCTCCTTCCCTCTTTTCCAC	TATCCACAGATCCTAACAA	1
	rs12518928	TTGTACATTTCCTGGCTGAC	GACAGAACTCATGTTTTTCC	AATAAAGTCAATTCTATTTTGTTAGTA	4
	rs12518928	ACGTTGGATGTTGTACATTTCCTGGCTGAC	ACGTTGGATGGACAGAACTCATGTTTTTCC	ATAAAGTCAATTCTATTTTGTTAGTA	17
	rs12522161	GGGACCACTAGGAAAAAGAG	GTGATTCCTTGAAGATATTGC	GTGAAGATATTGCTTTGTATCTAC	5
	rs12523473	ACTGATAATCTAGGGTACTC	AGAGATGTTACACCAACGGG	GGATCTAGGGTACTCATTTACA	5
	rs12658881	GCCTGGTTGGAAACTAAGAC	GGTTGTCTCAGGTGTTAAGG	GTTAAGGAACAAAGAAAAAGAG	2
	rs13166292	TGATTGGACAGTGGGTGCAG	GAAAAGGGAAATCCCCCAAC	CAGATGCCCCACCCTGCGCTGGCTCCC	9
	rs13179619	TCCATTGGTCTATATGTCTG	GCTAGAGACATCACATTTCC	TTTTGATCTCTGTGGTTTTG	8
	rs1396476	GGGCATATATCAGCTCTTAG	AGGCCTCAGCTTTGACTTAG	GGGCCACAGTGTGTATGTGTGAAC	5
	rs1423246	ACGTTGGATGCCTAAATTATCAAGTTTCCC	ACGTTGGATGATGTTGTAATCTCAGCGCCC	CTCCAGTTTCCCTGAACAGC	17
	rs1435077	CGATTCTACACTGGACTTAG	GTTGACTAATCTCTTCATCTG	AGAAAGACAGAGAGAAAATTCATACA	1
	rs1435077	ACGTTGGATGCGATTCTACACTGGACTTAG	ACGTTGGATGGTTGACTAATCTCTTCATCTG	AAGACAGAGAGAAAATTCATACA	17
	rs152341	ATGCTGGTGATTATGTGGAG	ACTGTTTTCCATAGTCGTTG	TCTACATTCCCAACAATGC	7
	rs1529842	TGTGAGGGTTATGATACTGG	TCTGTCCATGGACACAGAAC	CGAGATGCAAGAGGAAGTA	1
	rs1533019	GCTATAGATCTTCAGGCTAC	GAGTCCTACTTTCATGCACC	AGGCTACTGACCATGA	4
	rs159608	CTACCTATGCTTTCACAATC	GGAAATTTGCAGTGTTGTGG	AGTGGATAGATCACTTGTGTCAATA	4
	rs159616	CCTCTATCCGAAACTCTTAT	GGTTGGATTTGAATTATTCC	CCGAAACTCTTATTCTAATTTCTCATA	5

rs168883	TGTCTCAGCCTCTTGAGTAG	TGGCGAAACCCTGTCTCTAC	GCTGGGATTACAGGCA	7
rs168883	TGTCTCAGCCTCTTGAGTAG	TGGCGAAACCCTGTCTCTAC	TCTGTATCCAGGCATGGTGACAC	10
rs168883	ACGTTGGATGTGGCGAAACCCTGTCTCTAC	ACGTTGGATGTGTCTCAGCCTCTTGAGTAG	AAAAGTATCCAGGCATGGTGACAC	17
rs16890455	TCTAGACCCCCTTTTCCTTC	GGGTTTCATAATCCTTAAGT	TTCCTTCCATCAAGTTAGG	4
rs16890455	ACGTTGGATGTCTAGACCCCCTTTTCCTTC	ACGTTGGATGGGGTTTCATAATCCTTAAGT	ACCATTTTCCTTCCATCAAGTTAGG	17
rs16890459	TTGGTACTAGTGTGCTGCTG	AGAAAAGCAAGCCACAGCTC	TGTGTAGAAAGACCAGTG	10
rs17315957	CAGACCATCTATGTCCAGAG	CAACAGCATGCTTATGGTGG	CCAACTCTTGTCTACTCACTACC	9
rs2136203	AGCCTTTCTAGTGACTAAAC	ATGCTATGAAGCTTGATGAC	ATGCATAATTATTCCCATTCAG	1
rs2409741	GCTGTTATCATGAATTTTAG	GAATGTGTTACTTGAGAAAa	GAGTGTCACTAAATTCATAGTAG	5
rs2409741	ACGTTGGATGGCTGTTATCATGAATTTTAG	ACGTTGGATGGAATGTGTTACTTGAGAAA	GGGAAGTGTCACTAAATTCATAGTAG	18
rs26955	GATTACAGTGTGAGTCATGG	AAGTAGAAAGAGGATAGGAC	GGTGAGTCATGGGGCACACCAGCCAC	5
rs26955	ACGTTGGATGGATTACAGTGTGAGTCATGG	ACGTTGGATGAAGTAGAAAGAGGATAGGAC	TGGGCACACCAGCCAC	17
rs26956	GTGCCTCAAAAGCAATTTGG	GTCTCTTTTAGTTGCTGGTG	CATTTGGCCATGGAGATA	4
rs27565	GCTGTTACCAGACACTAAGG	GGACAGGATCCCACAAAAAG	GGGAAGTCAAAGGAAGAAAGGTTA	7
rs2898269	GGGCATAGTATTCATCTTCTG	GGCTCTCTTTCTCATCTCTC	CATCTCTCTGCCAGCG	2
rs2910829	TGAGGAAGAATAATGGATGC	CTCTAACCAAGTGCTTGCTG	GGATGCATCTAGTTTGGGAAAT	4
rs2910831	GGGTAAGAACTTACTACTGC	CCACAAAGAAGGGCAGTAAG	GGGGTTACTACTGCTATTTTGTTCA	7
rs2938784	CTGAGACCTCTGCAAGTTAC	ACTCAATCCCTACAACCACC	CTTAACTTCGCTCTTCCTCAGTGTTTT	3
rs2938787	ACATTGAGCAGAGAGCCCAG	ATGCCTGAGTGTTGCCATTC	CCAGTGAAGCTCACATG	9
rs2962964	AATCATTCACTCCAGGGCAC	CCTCAGTTCCTTGCCATGTG	GGGCGGACACTCACGCAGCA	10
rs2963820	AACCTAATGGTTCACAACTC	TCATGTAGGAGAGGCTTGAG	CTGGTTCACAACTCCCTTAC	3
rs2963820	ACGTTGGATGAACCTAATGGTTCACAACTC	ACGTTGGATGTCATGTAGGAGAGGCTTGAG	TGGTTCACAACTCCCTTAC	18
rs2963821	CCCTTCTATACCTGGAAGAC	AATGCCCAGAAAATAGCCCC	GATAGCCCCCTGCATAGGA	5
rs35382	CAGAATCCTAATAGATCATC	TTTTTGTGTCCCTAGTGCCC	CAAGTATTTTCCTCCCTAGAC	1
rs35382	ACGTTGGATGTTTTTGTGTCCCTAGTGCCC	ACGTTGGATGCAGAATCCTAATAGATCATC	CCTAGTGCCCAACATAT	17
rs35383	ACTCTCTTTCCTTGAGCTGG	GGAGAATCTGAAGTCCCAAG	GGGCATCTGAAGTCCCAAGAACTAG	10
rs35384	GATGCATTTTTCCATGCATA	CCTCACTGTATCTCTCTTC	GAAAATATATCCATGTGATATAGTG	4
rs35384	ACGTTGGATGCCTCACTGTATCTCTCTTC	ACGTTGGATGGATGCATTTTTCCATGCATA	TGTATCTCTCTTCATGAGATATTTT	17
rs35385	GCTAAGCTTGTTGTTAGAGAC	ATTTAGTGAACATGGGAGTC	CATGGGAGTCAACAGC	1
rs35386	AAAGGAAAAGAAACACCTC	GAATAACATGTTAACAGAGG	cACAGAGGTTATCTTTTGCTAAT	3
rs35386	ACGTTGGATGGAATAACATGTTAACAGAGG	ACGTTGGATGAAAGGAAAAGAAACACCTC	AACTGAACAGAGGTTATCTTTTGCTAAT	15
rs35387	GGACTTCCCGTAAGTAAATC	TCAAATTGTCTAAAAGACTA	CAATATTGCAACCATATTTTGTCCCA	8
rs364917	AACCCATGGTACTTTCTCCC	AGTAGGAAAGAGCTTGGCAC	CCCATGGTACTTTCTCCCTCCTTCAC	3
rs371424	ACGTTGGATGTGGAGAACTTTCCTTGTTAC	ACGTTGGATGGCCTGATGGACTGTAAACTC	TTGTTACTTTAGAATACAAAAAAAAT	18
rs37684	TTAGTGGTGGGTATCCTCTC	GCAGTGGAAAAGTCTACCAG	CCCGATTGGGAAACAGAGAGAA	7
rs37702	GAAGCCCATCAGACAAACAG	GAATGTCGAATATTGGCCCC	CCAAACAGCAGCTCTCTC	4
rs37702	GAAGCCCATCAGACAAACAG	GAATGTCGAATATTGGCCCC	CTGTCTTATAAGGTTTCTGC	8
rs37702	ACGTTGGATGGAATGTCGAATATTGGCCCC	ACGTTGGATGGAAGCCCATCAGACAAACAG	ACTATTTTCTGTCTTATAAGGTTTCTGC	18
rs37707	GAGAAAATGGATACTGCAGG	AGCTGTGTTAGTGGCTGTAG	TATGGGAGGAATGCTGAC	3
rs37707	ACGTTGGATGAGCTGTGTTAGTGGCTGTAG	ACGTTGGATGGAGAAAATGGATACTGCAGG	CATGGGAGGAATGCTGAC	18
rs3857232	CTCCTCTCCTTCAGTTTTCC	GTCCAACCTTGAGAATCAGC	CTTCAGTTTTCCTGTTAGCCC	3
rs3887175	TTCTTTGTGGCCCACAAAGG	CTTTAATCTGCTACTCTGGG	GTTGATGGGAGGGAGG	7
rs3887175	TTCTTTGTGGCCCACAAAGG	CTTTAATCTGCTACTCTGGG	GTTGGGGCCCACAAAGGGGCCTA	10
rs40512	TCCCTTCATCCACCTTTGAC	AGAAGCTGTGGGCATGCAAG	GCAGAAGTGTAGCCAAGAAGCAG	1
rs456009_A	GCTATGATTATAAGGCAGGAC	CTGTAGTGACTGGTTGAATG	TCAGCCTCAGTTGCCT	9
rs456009_G	GCTATGATTATAAGGCAGGAC	CTGTAGTGACTGGTTGAATG	TCAGCCTCAGTTGCCC	8
rs4580739	TCTTATCCATCTACCCCCAC	AAGGAGCATGTCATGTTGGG	CACCCCCACAAAACACAC	9
rs4604150	AGACACAGCTTCCTTGTCTC	GACCTTGGAATGGAAAGGAC	CCCTTGTCTCTTCTGTTCTGATGATTT	3
rs4700371	GTTTGCACACAGATGTTCCC	CCTAATTTACTTAACCATTC	AAAGATGTTCCCTAATAATTTGTATG	9
rs6449467	CTGTTGTCCAGATAGTGAAC	GGACTCCGAAAACAGAGAGG	CGTCCAGATAGTGAACATAGTATCCAA	1
rs6879326	AGACTGCACTTGGTGATGAG	CATTGCCAGACTACTCTTTG	TAGACTACTCTTTGATAAACCTGTCA	5
rs6887545	GAAGTAGGTACCATTACCCC	TTAGCTGCAAAACTGACCTG	GCTTTTTTAGTTCACTATTCTCATA	3
rs6889641	TGATCTCTGAGGAGATCCTT	ACTCAAAGCTTGTTTGGGAC	TTGTTTGGGACTTTGTAATATGG	7
rs6889641	ACGTTGGATGACTCAAAGCTTGTTTGGGAC	ACGTTGGATGTGATCTCTGAGGAGATCCTT	AGGTTTTGGGACTTTGTAATATGG	18
rs6889660	GCATGTATTCTCACTGCTTC	TTGTGTGTCAGGGAATAGGG	GAAAAGAGAGAAATTGGGGAATG	5
rs702553	GGCAGTCTAGCTTCAGAGTA	TGCCACATTCTTGCTCAAAG	CAGCTTCAGAGTACATGTTAA	3
rs7442640	AGGACTTAACTGAGCTCCTC	GCTGTCACATAACAAGTCTG	GCACCTCCCCAGGTAGGCTCTCCAGC	1
rs7710463	CTCTGTCAGTTTTTTCATCG	TGTGCTGAGCATTTTACATC	ATTATCTCATATAATCTTTACAGTAAC	1
rs7727343	TTATGCTTCTGCCTACCACG	AAGCTTCTGTGGGCTGAAAC	TAGAGTACAGCAAGGAAAATG	7

	rs7732249	GTTAAATTTTGTCAAACCC	CGTGAAATTCCACATTAAC	TGTCAAACCCTTTTTCTG	3, 10
	rs7732249	ACGTTGGATGGTTAAATTTTGTCAAACCC	ACGTTGGATGCGTGAAATTCCACATTAAC	TGTCAAACCCTTTTTCTG	18
	rs7733705	CACTACTCAGAACAGAATGC	CGGTTACCTGCAGTCAAC	CTGCAGTCAACCGTGTCC	4
	rs789395	CTTGTGTCTAGTACTGAGCC	TACTCAGGTAATGGCCTCTC	CTCCTAGTACTGAGCCTGGCATATA	9
	rs789396	CTTCTGGAAAGGAAGATTGG	TGCAAGTATATCCGTAGGAC	ATATCCGTAGGACAAAATCTTAAGG	1
	rs789396	CTTCTGGAAAGGAAGATTGG	TGCAAGTATATCCGTAGGAC	CGTAACTGCTTTTATCTAATAATTCTA	9
	rs789396	ACGTTGGATGTGCAAGTATATCCGTAGGAC	ACGTTGGATGCTTCTGGAAAGGAAGATTGG	TCCGTAGGACAAAATCTTAAGG	17
	rs966221	ATTGGAAGGATCTGCTGCTG	GGTATGAGTCCTGCTATTAA	GCTGCTGGATAAACCAC	5
ALOX5AP	rs10507391	TCCAGATGTATGTCCAAGCC	CTCTGTAAGGTAGGTCTATG	CTGGTTGCAACATTGAGGTTAA	3
	rs12429692	GTTGCAACCATAGACCTACC	CAGGAAATGTTCAGAATGGC	GAAAACCCTGCCTTTCTTCTTCCTCAT	2
	rs17074975	CACGAGACTGTGAAACTGAC	AAGGTACTGTGCTGCTTTGG	TGCAACATCTATAAACAATGCAA	9
	rs17216473	CAGACTGTCTTGAACTCCTG	GGTGGCTCATGCCTATAATC	CACAAAACTGTGGGAGGC	1
	rs17216473	CAGACTGTCTTGAACTCCTG	GGTGGCTCATGCCTATAATC	GTGATCTGCCTGCCTC	10
	rs17216473	ACGTTGGATGGGTGGCTCATGCCTATAATC	ACGTTGGATGCAGACTGTCTTGAACTCCTG	AAAACTGTGGGAGGC	18
	rs17222814	ATCAGCTAGTCTCTTTCCCC	TGTGTCCATACATAGCCCTG	TTCCCCAGCCACTGTT	5
	rs17222842	AGGAGTTTTCCTGGGATGTG	TGCACCCCACAAATACCTAC	CTTTTCCTGGGATGTGGTCCTTTC	4
	rs17612031	AGCTGCTGAGTACGTTTCTG	AATGCCTCCTTGACTTGACC	CCCTAAATGTCCTCCATC	5
	rs3885907	CTGTGGTTAACTAAAATCTC	TCACTCCACTGCTCAGTGAA	CACTGCTCAGTGAAGGAAAC	4
	rs3935644	AGCACCTTTCAGATTCCACC	CAGACACGTGTGTAATGCAG	CTTTCAGATTCCACCGGTTTTATC	4
	rs4076128	CACAGATCAGTGCAGCATAT	AAATATCCCACAGCGACAGG	GGCATATGCATAATTACGGAG	4
	rs4238139	GAGCTAGAGGGTTCACAAAG	TCCAGTTAGATTGAACGGGC	AGGGTTCACAAAGACTTAACCCGAC	1
	rs4445746	GGAGTGAAATTGCAACCAGG	ACCCACACTTCTGCAGAATC	TTCTGCCTGCACTCCAC	3
	rs4491352	GCTTCGTCTTGATGAAATGC	AAGAGTAGTAAACATGGGGg	TGAAAAATAAAGTAGCCATTATTTTTG	3
	rs4769055	TCCAGAATGGTAAGGAAAGC	AGCCTGACTTCCAAACAACC	CATCTCCCCTTCTGTTCTTCCCTGA	2
	rs4769062	ATGTGAAAGAGGCGCTTTCG	AAGCAGCTCTGGACACAAAC	CACAGCTCTCAACACTGTAATCTA	1
	rs4769063	TTTGTGTCCAGAGCTGCTTG	AAAGGCAGCCTCCTTGAAAC	GAGACCAGTGTTTATCCC	9
	rs4769870	TTTTGGAAGGGTCAAGGTTC	TCAATGGGGAAAGGAAAGTC	GGAAAGGAAAGTCTTTTTAACAAGG	5
	rs4769870	TTTTGGAAGGGTCAAGGTTC	TCAATGGGGAAAGGAAAGTC	CGAATGTACAGTTGTTCTAGCA	10
	rs4769874	TTTCAGGCATGCTCTGCACC	CACCAGGGAGCAAGCATTAG	TTAGCAATGCATTATCACA	1
	rs4769874	ACGTTGGATGCACCAGGGAGCAAGCATTAG	ACGTTGGATGTTTCAGGCATGCTCTGCACC	GCATTAGCAATGCATTATCACA	18
	rs9315050	CTGTCTCCAAATACAGTCCC	GTTGTGGGCTGAATTAGGTC	CAAAATTCATATGTTGAATCTTTAC	5
	rs9551963	AGTTCTTGACCTCACCAACC	GGGTTCAAGAGAGAAATTCC	AACCGAGGAGGAATTGCT	2
	rs9551964_A	GCAATGACTGCCAACTCTTG	ATGAGGAGTGAGTGAGCAAC	GAGGCCAATGAACTCAGGTT	7
	rs9578200	CGCACTTGCTTAAAATCCCC	GCCAAGGTGTGTGGGAATTA	AATCCCCCAGATAATTCT	2
	rs9579645	ACGTTGGATGTCCAGATCTAGGTATCCTGC	ACGTTGGATGGGAAAATTGTCCTGTTCAGC	AGGTATCCTGCTTGTTT	18
	rs9579648	CAGGATACTTTGACTCAGGG	CACTCATTCTGCCTGTGTTC	CAGGGAGATAAAGTAACTTG	3, 10
	rs9579653	CTACCCAATGCAGTAGTTCG	TACAGCACGTCTCAGTTCAG	GTGGCTCCTGAGCCCT	4
	rs9671124	AGATTTCTGGGCCCCAACAC	GAGAACATGTTAGAAATGC	CATGTTAGAAATGCAAATGTTCAC	3
	rs9671182	ACACAACTGCTGACTGATAC	CACACTGGCTAGTGTATCTG	GACTGCTGACTGATACAGATAGCTCAA	5
MPL	rs12731981	ACGTTGGATGCTACACTGTCCACAAAGAGG	ACGTTGGATGAAGTGCGTCTCTTCTTTCCG	GGGCGTCCGAGTCTGGTTTAGGAACA	16
	rs710252	ACGTTGGATGGTACCAAAAGGCAACTCAGC	ACGTTGGATGGCAGCATAGTAAGTGCTCAG	GCCTCTGTATCTGTGCTAA	15
ELOVL7	rs10440618	ACGTTGGATGCCCCTTATTGTTCATGAGGC	ACGTTGGATGATAGCTCTATCTGGGCTGTG	ATCTCCCCTTTAAAGTCTC	16
	rs13159372	ACGTTGGATGCCCTCCCCTTCTATTATTCG	ACGTTGGATGGGAGTTTAAGGATGGCTCAC	CCTAGTTCTCCCAATGAGTTGGACCA	16
	rs1563905	ACGTTGGATGCAAGCGAAAATCTCTGTGGG	ACGTTGGATGCTCCTACTATATGCCAACTG	GTTGGGATTGGTATTATTTGACTT	14
	rs16878412	ACGTTGGATGCTGACTCTTGCTACAGATCC	ACGTTGGATGTGCTGGGATTACAGGTGTAG	AGATCCATCATATCGATCAC	12
	rs17331746	ACGTTGGATGAGTGGAGTTGAAGGAAAGGG	ACGTTGGATGTCCTTAAACTCCACTGACCC	TTGGAGGGACTTTTAGGAGAAATATCAT	17
	rs17332108	ACGTTGGATGGAAGAGCATAGGCTCATTGG	ACGTTGGATGTGGCAGTTTGCATATGGTTC	ACTGTTAGATGTCAACGGTATTG	15
	rs1807017	ACGTTGGATGGTTCCATTTATCACAACTGC	ACGTTGGATGCCCAGAAAGACTTGAAATGC	ATCACAACTGCATTACCTAGATT	15
	rs1870682	ACGTTGGATGGCTCTTGAGTTAGCATATTG	ACGTTGGATGGAACAGATCTCTTGGTATCC	GTTAGCATATTGATTATTTGAGTC	17
	rs6449493	ACGTTGGATGAAGTTGGGTAATGGGTTCAC	ACGTTGGATGAAACTGGACAATTTCAAAC	GAAGAGTTTATTATACTCTTCTGTCTAC	12
	rs6869332	ACGTTGGATGGCCTCCTTTTCTTGTCTCAC	ACGTTGGATGGTGTAAATAACTTATTTGGG	GTTGCCCCACTCCCCAACTTCTGGTACT	12
	rs6872863	ACGTTGGATGCAGTTCCAACAAGGTAAGGG	ACGTTGGATGAGAAGAGAAGGAACATCCAC	TCAGGTTTTTCTGTTGTG	16
	rs702564	ACGTTGGATGGTGGGATGAAATCCCAAAGG	ACGTTGGATGGGATATGGAATTTGTATGGGC	TTTCTGAAATCCCAAAGGAACTCTATG	15
	rs7715147	ACGTTGGATGCACTGTGAATTAGGATACTGG	ACGTTGGATGTGATGGCATATCATGTAGGG	AACAATAGGGGTAGAGC	16
	rs7730827	ACGTTGGATGGGTTGAAGATAGAAGCTGGG	ACGTTGGATGGGATATTCCACCATCACAGG	AGCGGGGCAAGATGAGTTCCTTTTGG	12
	rs7731760	ACGTTGGATGTGTGCCCAGCCTGAGAATAC	ACGTTGGATGAGGTTGCCTTTGTCTCTCTC	AGGGGCCAGCCTGAGAATACTAAAAT	15
	rs898622	ACGTTGGATGGACTAAGTCTGAAAAGTCATC	ACGTTGGATGGGTAGCATTTTGGATAATGG	GAGGGAAAGTCATCATTTCTAATCAAAC	13
CDC14B	rs10121511	ACGTTGGATGATACAAAACAGGCAAAGGAC	ACGTTGGATGGATCACTGGTGAAGTTGAGC	GGGGAGGCAAAGGACATGAAT	16
	rs10978403	ACGTTGGATGCCCCATGCAACTTAACCATC	ACGTTGGATGTTGACCAGCCTCATTCATGC	CCCTAAACTTAACCATCCATCCA	13

	rs1411885	ACGTTGGATGCAACTCTTATGTAAATACTGC	ACGTTGGATGCATATGCCATCAAGTGCTCC	AGCGTATGTAAATACTGCTATTCCAAAG	13
	rs6477496	ACGTTGGATGCTGTCTTGTCGTGCCTTGTC	ACGTTGGATGCCTTCCCTAATAGTAGCCAG	ACAGTCAGGCAAGGA	13
	rs7852498	ACGTTGGATGCATGATAAGCCAAATCAGCC	ACGTTGGATGTAGAGCCTCTACTGAGCATC	TATGTCCAAATTGATTAGCAGTACTC	13
HEMGN	rs1059003	ACGTTGGATGCCACAACTTTATGGTTGAGC	ACGTTGGATGCCAGCAATTCTGAATGAGAG	CCCAAAACAAAACATAACTATAGACATC	14
	rs10760017	ACGTTGGATGATATCCTGGATTGGACCCTG	ACGTTGGATGCACTACCATTAACTACAGAC	CTGCTTGGACCCTGAAGCAGACAAAG	12
	rs10984462	ACGTTGGATGCCATGGTTTTGTTATCAGTG	ACGTTGGATGACAAACTCATAATGGGTAGC	AAATGAAATGAGATACTGCATATAA	13
	rs4743146	ACGTTGGATGGCCCATGGTTATGTAAGACG	ACGTTGGATGCAGAAAAGTTGTCATGATAG	TGGGCGAAAGGTATATAGTA	13
LCN2	rs10760543	ACGTTGGATGCCACCCTTCTCCTATTTTAC	ACGTTGGATGCAGTGCAAGACTCCATGTCA	CCCTTCTCCTATTTTACATATACC	16
	rs10987900	ACGTTGGATGTAGCATGAGAAGCTGGGAAG	ACGTTGGATGAGAAATGAGGGTCCACAGAG	GGGGCGGGAGCAGCTTCCCTGA	12
	rs6478823	ACGTTGGATGCCCCCTCAGCTGAAATAAAC	ACGTTGGATGTCTATGCCCTCCACTTCATC	CCTCTGAAATAAACACGTGCTGCCAAG	13
	rs878400	ACGTTGGATGTAGAAGCCTAGCCAGTCACC	ACGTTGGATGTGTGAGTAGTGTTCAGCGTG	CTTCCCAGCCTTCAAACAC	12
GOLGA2	rs2416981	ACGTTGGATGGTGAGTCGTCCAAGTGATTG	ACGTTGGATGCTGCTTTGGCTACCTAGAAG	GTGATAAGATCAAGCAAGTCAC	13
	rs7023913	ACGTTGGATGCCATAGGTACCATGGTAGTC	ACGTTGGATGCCAGGCTTGTTACCTGTATG	GCCCCGTGTAATCCCC	14
	rs7871866	ACGTTGGATGACTTTGACAGGCCTTTAGGG	ACGTTGGATGGAATACTGCAGGCGGCTGG	GCCCCATCTCAACCT	13
GFI1B	rs10751499	ACGTTGGATGAGGAATGCAGTCCTGCTAAC	ACGTTGGATGGAGGCCAGAAACTCAAAATG	CTTGATGTTAGCCCCA	13
	rs15906	ACGTTGGATGTAACTGCCAGTTTCTCCAGC	ACGTTGGATGACTTCCGACCAGCACTCAG	GGAGCTAGAATCCCCAG	17
	rs1888204	ACGTTGGATGAGGAACTCACCTGAAGTCAC	ACGTTGGATGGGTAAAAGGGACTTTGCAGG	GATCCCCTGAAGTCACACAGCTAGA	15
	rs1964196	ACGTTGGATGGTCCAAAGAATTCCTTCTGC	ACGTTGGATGTCTTAGTAGGCAACAGGCAG	GCATTCCTTCTGCAATACTTC	16
	rs2073577	ACGTTGGATGGCGGTTAGGAAGGACAAGAG	ACGTTGGATGTCAACCCAGTATAGCTGGTG	ATTCGGAAGGACAAGAGAACAAT	13
	rs2773818	ACGTTGGATGAGGGTCATGGTTTGCCATC	ACGTTGGATGTTGGCGAGTACCTTTGGTTC	ATTGCCATCAAAATATCAAAAGAT	12
	rs2773822	ACGTTGGATGTTGCAGGAGCACGTGGATTG	ACGTTGGATGGTTCAGTGTCTATCACTGGG	GTTGTATTGGCAATGTCCCCT	12
	rs2905069	ACGTTGGATGTGTTCGCCTCTTCTCCTGTG	ACGTTGGATGGACTCAAGACCTGAGAGGAC	GCTCCTCTTCTCCTGTGTCCATGAA	16
	rs3011266	ACGTTGGATGAAGCTGTCACCACGAACAGG	ACGTTGGATGAACTAGCCAGCCACATTAGG	TTTTATCTGAAAATTAATTCAGATGTT	16
	rs3827664	ACGTTGGATGTCATGGACACAGGAGAAGAG	ACGTTGGATGTTGGTTTCTTCTCCCACTCC	GGACCGGAGAAGAGGCGAACAG	12
	rs606141	ACGTTGGATGGGCAAGAAGAGGCCAGGAG	ACGTTGGATGTGGGTTGCTCAGAAAAACAG	TTATAGGCCAGGAGGAGACCC	13
	rs621940	ACGTTGGATGGGGAAGCCAGACAATGTTC	ACGTTGGATGAGCCCGACAAGCTTAATCTC	TATGTCGACTGAGGTTTGG	16
	rs633153	ACGTTGGATGGAGATAGGACCATCAGCATC	ACGTTGGATGTCCTCCTAGGAAGGGCGTG	CCTCAGCATCAATAGCCAC	13
	rs649651	ACGTTGGATGGAGATTCTAACCCACGAACC	ACGTTGGATGTGAGATTCACCTTCAGAGGC	AAGCTCAGACCACCC	13
	rs667805	ACGTTGGATGTTTCTGTGCTCTGCTAGAGG	ACGTTGGATGTGGGTCTTTCTCAGCAGAAG	AGGCTAGAGGAAGGCAC	18
	rs678946	ACGTTGGATGCTGATTACACTGAGCTCAGG	ACGTTGGATGGGGCCATTATTCAGCCAATC	TCCCATCTCAAGATCCTGAA	14
	rs686652	ACGTTGGATGCTATACCTGGTTTCGGTTTG	ACGTTGGATGAAGATTCCCACCCACCCTAC	GGAATCACTCAAAACCAG	13
	rs7036106	ACGTTGGATGTTTCCCTTTACCTTTAGGAC	ACGTTGGATGTTGAAAGGGTGCCGACATTG	GGGGTTACCTTTAGGACAAAAATCT	12
	rs8192999	ACGTTGGATGGGTGATTCTCATGCTTCTTC	ACGTTGGATGACTGTCACTTGCCAAGCTCA	TTTCTTGCCTCCCCC	14
	rs8193002	ACGTTGGATGAAGAGGCCAGAAGCAGCAG	ACGTTGGATGTTTTCCAGCCCAGAGACACC	TTAATAAGCAGCAGCTGGGCCGGTCC	14
	rs8193004	ACGTTGGATGGAATGTGAGTTGGCCATGAG	ACGTTGGATGTGCCCCCGGGTTCATTTTC	TGGCCATGAGAGAAAACA	12
	rs8193007	ACGTTGGATGAGAGAGGGGTTTGAAGTCAG	ACGTTGGATGATCGCCTAGAGAACTGAGCC	GGGGTCAGCCCTGAA	15
FAM69B	rs10117822	ACGTTGGATGGGACAAGAGCTCGTTGTAAG	ACGTTGGATGTTCACTGACTCTCTGACACG	GCTAATTTTGCTTTTTGAAAACAAG	16
	rs11793385	ACGTTGGATGACAACCTGAAGCCCCAAGTG	ACGTTGGATGTGGCCTCAGCTCCCCTTCT	AGTTGCCAGGTTCCATAC	14
	rs2275160	ACGTTGGATGGAGTTAGAAAGCCCCACTCAC	ACGTTGGATGCTGTTTTGTGGACAGGCATC	TTCTCCCAGATACCAGC	15
	rs2275161	ACGTTGGATGTGTGGTGCCAACAGGTCCAG	ACGTTGGATGAGGTAAGAGGTGAGGCGGAG	CCCCGTCCAGGACAGCATC	17
	rs3739940	ACGTTGGATGAGACCAGGGCTTGTGGAAAG	ACGTTGGATGTTCCCTGAGGATGAGAATCG	TCCAGTGCGTGCCCTC	16
	rs6874	ACGTTGGATGCATCTTGAAGTCGTAGGTGG	ACGTTGGATGGCTCTTACGGGACTTTCTAC	TTAGTTGGCTGTGTAGCCCAC	16
	rs945381	ACGTTGGATGGTTTCGTGGCCACTGTGTTG	ACGTTGGATGATCCGACCCATGTTACTCTC	GCCACTGTGTTGTAACTC	14
TMTC4	rs1051518	ACGTTGGATGGTGCACAATGTGCAGGTTAG	ACGTTGGATGAAAATGTGACCTTCCCCCTC	CTTTCTAACACAAGCATTACTC	14
	rs12427435	ACGTTGGATGAAGGAATGCGAAGGTGGATG	ACGTTGGATGTGCTGAGAAATAAAGACGCC	GATGGAGTGAACAGCG	14
	rs1283198	ACGTTGGATGCGACCTTGAACTGCTTAGTC	ACGTTGGATGCCTTCTGAAGAAAGTGGTGG	CTCACTGGCACAACA	12
	rs1572641	ACGTTGGATGGGGCAGAAGTAACACTTACG	ACGTTGGATGTCCCAGAAATACAAGGTGCG	CAACTAACTGCAAGAACTCA	15
	rs17578868	ACGTTGGATGACTCAAGGTGTGCCAGCTAC	ACGTTGGATGTCAGAAGCATGAAGTGTGGG	GGTTCCAGCTACACACCATCTTACCAT	17
	rs17579126	ACGTTGGATGAGGCCTCATAATTTCTTTCC	ACGTTGGATGCTGTAAGTACCAACATGACC	TACAAATGCTAAAAGAATGTTAC	18
	rs1765733	ACGTTGGATGTTTCCTGTGACCCTGTTCTC	ACGTTGGATGCCCAACCCTTTAGAGAAGAG	CTCTGTGGTCCCCTT	14
	rs2791686	ACGTTGGATGGCAAGTAGCATCCACATCAC	ACGTTGGATGGGAAGACGTGTTTTAAGGAG	AATTCACTCTCCTTTCTCGTGTTA	13
	rs7995648	ACGTTGGATGTAGGCTTAGTTCAAGTCCCC	ACGTTGGATGAGCCAACGTCTGGTAGTTTC	CCATTTGTCATTTATTCTAAGCCCATGT	15
	rs8002073	ACGTTGGATGGGGATGAACATGTTCGTGAG	ACGTTGGATGACCTGGGCGACTTGCATAAC	GGAGCCACCTGCCTCCCTGCAGT	16
	rs946845	ACGTTGGATGCAAACCACACACATTCTTGC	ACGTTGGATGTAAGGTATCGTCCTGCTGTG	CCCCCCTCGTGCGCTACTGAATTATAAA	16
	rs9513770	ACGTTGGATGATGGTTTTGGAGATGGTCCC	ACGTTGGATGAAGACCACCAGTTCCTTTGC	TCACTTTCTCAGGGGTACCAC	15
	rs9513777	ACGTTGGATGACCAGGCTTTAAGAGAGATG	ACGTTGGATGGGCACACTGGGTATTACTTC	GGAGGGCTTTAAGAGAGATGTAAATA	15
	rs9513779	ACGTTGGATGGCTCAGCATCTTGAAACTTC	ACGTTGGATGTCCCTCACAGTAGACTTCTC	AATTGGTTAAATGTAGATATTTCAATC	13
	rs9513782	ACGTTGGATGCACGACCTTGAACTGCTTAG	ACGTTGGATGTCTGAAGAAAGTGGTGGTGG	GTGCACTGCTTAGTCTTTCTAAGAAT	18
	rs9518104	ACGTTGGATGCTCTGGGTCTCAGTTTGTTC	ACGTTGGATGGTGCATTTTTTTGTATCACC	AGATCGTTATCTGCCCACCCTC	14

	rs9518128	ACGTTGGATGAAGTCACATGGGACACATGC	ACGTTGGATGTGGCATCAGAAACCAAAGGG	CGGCAGATTCCTGGGTC	15
	rs9554712	ACGTTGGATGGATTATAGGGAGAAGACAA	ACGTTGGATGCGAAGATCTGATTGCTCTTG	GGAGGATAGGGAGAAGACAACCATTC	13
	rs9582406	ACGTTGGATGAGTAGATTGGCAGCTGGGAG	ACGTTGGATGAAAGAGAGAATGAGGAAAGG	AGCCTACTTCACCCC	12
TTC7B	rs10132961	ACGTTGGATGCAGGAATCAGGAATGGCATC	ACGTTGGATGTGATGGTGTCTCATAGTGGG	TGGCATCAGGGAGAA	17
	rs10139373	ACGTTGGATGTACAAATGGCCAATAAGCAC	ACGTTGGATGCTGGTAGGTGTGATGTGATA	GGGAGCAGGGAAAAGAAAATCAATCC	13
	rs10150862	ACGTTGGATGCAATGTGGTGAAACCCCGTC	ACGTTGGATGAATAGCTGGGACTGCAGGTG	CCCCCCCCGTCGCTACAAA	13
	rs1076958	ACGTTGGATGTGTGGTACTGTACAGCTGTC	ACGTTGGATGGGATTTAGCAATTACCTCCC	GCACCCAGTTTGTGTAA	14
	rs11620738	ACGTTGGATGATTAGAACTGCCACAAGCTG	ACGTTGGATGCATTGCAAAGGCAGAAGCAG	TTACCTGCTGCCTGATATATTA	13
	rs11629065	ACGTTGGATGTCTTCTGCTTTACTTTGTG	ACGTTGGATGCTTGTAAGAAGCAGTTGAGG	ATATCTAGTCTAATGTGCATTTAATT	14
	rs12147413	ACGTTGGATGCATTTCCGGCCAATTCCTTC	ACGTTGGATGCCAACAGTTTTGACCCGTTC	GCGTCCCTCTTGTTTGCTC	15
	rs12432719	ACGTTGGATGAGTGTTAGAAAATGCAGCTC	ACGTTGGATGGAGAAAGGAATTCAAGACAG	CCCCCGCAGCTCAATAAAATGCTTGCCA	13
	rs1286305	ACGTTGGATGTTGCCTTTGCCCTCCGTGTC	ACGTTGGATGTGCAACTCCGAGAGGAAAAG	TGCGCACCTGCTTCTCCGCTTA	12
	rs1286322	ACGTTGGATGTTTCTAGGAGAGGTGTGTGC	ACGTTGGATGAGAGAGCCGGGTGTAACAG	CCCTTCTCTTCTAACCCC	14
	rs1286459	ACGTTGGATGAATGGGCAACGGGAAGAATG	ACGTTGGATGGGTCCCAAGACTTTAAAGCC	GGACCTCATTTATTTTCTGACACTA	18
	rs1286464	ACGTTGGATGCTTGAAGAACCAGAATGTCC	ACGTTGGATGCCTCTTCCTCTCCCTTCAAG	GGCGAAGGAGTTAGTG	13
	rs1286470	ACGTTGGATGTAGATGTCTGCTCTGTGCCA	ACGTTGGATGGCACCAAAATGTTACTACTAC	GAGTCACTTGCGGTT	14
	rs1286477	ACGTTGGATGTTTGTCTTCAAGTGCCGTCC	ACGTTGGATGTGGGCTGTGCTCAATTTGGG	GTCCTACCCTGAGCTTC	16
	rs1286496	ACGTTGGATGGGCAGCTTCTGCTTCACTCA	ACGTTGGATGCCCTTGCCAGTTCAGTCTTG	TTTTTTGCTTCACTCAGAAGGTC	17
	rs12881399	ACGTTGGATGCTATCGCTTCTCTTCTTGGG	ACGTTGGATGTTCTGGCTGCCTGTTACTTC	CCTATTGGGAGTGCCTGCT	16
	rs12883490	ACGTTGGATGAAAAGGCTGTGACACACACC	ACGTTGGATGATCAGTGGCAGATGAGATCG	GGTAAATGGCCTTGGCACA	12
	rs12886545	ACGTTGGATGAACTTTACAGAACTGGACCG	ACGTTGGATGAATAGTGCCTGGCTACAGAG	ACTGGACCGGAATGTA	12
	rs12886812	ACGTTGGATGAGTCCAAAACACAAGGGCAG	ACGTTGGATGAGGCTTCTCTTGCGTCTTTC	AGGTGGCTTAAGGTTATAAA	12
	rs12889650	ACGTTGGATGCATTCTGCCGACTGTGTCTC	ACGTTGGATGCCACACCATTAATCCCAAGG	TGTCTCCAGCCCAGC	15
	rs12893100	ACGTTGGATGAGGCCATCAATCTTAGGTCA	ACGTTGGATGAGCAAACGATTACTCTGCAC	TCAAATGTTTCATTTAACCAAAGAT	13
	rs12894343	ACGTTGGATGGAAGATGCAGATGCCTTCCC	ACGTTGGATGGAAGAACCTGTGTGTTGTGC	GCCATTTATTGAGCATGGACT	12
	rs1294555	ACGTTGGATGAAACTGTGTCTTGGGCACTG	ACGTTGGATGTTCTGGAGTAGGCAGGATGG	GGGTTGCTGCTCCCATCCCTCCCCTAGC	14
	rs1294559	ACGTTGGATGACCATGGTGCTTAGACCAAG	ACGTTGGATGATGTGTGAGCAACCCTCTTG	AATTTCGTGGAGGAAGTGGC	14
	rs1294582	ACGTTGGATGAGCAGAATTCAGGGAGGTGG	ACGTTGGATGAGAAAGCTTCCAGAACAGCG	GGGATGTCTGAATCCGCCAT	15
	rs13379124	ACGTTGGATGCCTATATTTGACCCGAGTCC	ACGTTGGATGCCTGGAACACCTTTACCTAC	AAAATGACCTAGGAATCTTTGA	14
	rs1535321	ACGTTGGATGAACTCAGGCCCTATCTGATG	ACGTTGGATGCCAGTTTGACTGCAGGCTTC	TATGCCAGTGCTAAGG	13
	rs1742084	ACGTTGGATGCAGGTCTATGGCAGACATTG	ACGTTGGATGTAGGATACACTAGAGCCTTC	TTTTCTACAAATATAAAGGTGGGTACTT	12
	rs1742098	ACGTTGGATGGGGTCTGTAACCATGACTCT	ACGTTGGATGTAAGGATTGTCCTTCACCCG	CCCCGACTCTATTTTGTCTCTTTACATTC	15
	rs1742100	ACGTTGGATGTTCTCTAGCCTTGGTTTTCC	ACGTTGGATGTATTGCCATCTGTCCCTCTG	GGTTATTTTGGAGCTCTGGC	18
	rs1749715	ACGTTGGATGTTTTAAAAGCAGCTTTGTG	ACGTTGGATGACCATGCCCCGCCTATATAC	ACTGTTAAAAGCAGCTTTGTGTATAATT	14
	rs17721032	ACGTTGGATGTTCCGAATAGAAACTGCACC	ACGTTGGATGCCACGCTCTTCATATGAGTC	CCCACCTGCACCATCAGTTCAACTATTAC	13
	rs17793829	ACGTTGGATGTTGGAAAAGAGCACGAAGGC	ACGTTGGATGTGGACACAAGTGTGACCAAG	GCGACGGCAAGCTGAGAGCTGCCT	14
	rs17799388	ACGTTGGATGTTTGAATGGGCGTACCTTCG	ACGTTGGATGTGAGCTTAGCAGTGCAGATG	CCTTGGCTGTAGGTGA	16
	rs17799418	ACGTTGGATGCACAGAGGACTCAAGAGAAC	ACGTTGGATGGTGCCTGATGTATTTCTGTG	TCCAGAGAACACCGTCAATTTTC	14
	rs2147829	ACGTTGGATGTCCAAGAGTTCAGGTCTATG	ACGTTGGATGCTGTCATCACTGGGACACAC	GGGGCAGGTCTATGTTTTGCAT	14
	rs2180774	ACGTTGGATGGCTGTTGAACCCATTTCTCC	ACGTTGGATGAAAATTCACAGAGCTGGTCC	CGACTATCCACATGTCCAAAGCCTAT	14
	rs2277510	ACGTTGGATGAATCTGCTTCCCTTGAATGC	ACGTTGGATGATGGTGCATCATCCTGTGTG	TGTTGCCAAACGGCTCCTAC	12
	rs2343	ACGTTGGATGGCCCCCAGATTTGGAAAATG	ACGTTGGATGCGTATGTACTTAGCAGATGG	TGGGAAAATGAATTGAACAAGTGCCT	14
	rs4900055	ACGTTGGATGAGCTGGCCTTTCTCTTATGC	ACGTTGGATGATGATGTGGTGCACACAGAG	TCTCTCCCATCTCAAGAC	14
	rs4900057	ACGTTGGATGTGTCCTTGTGCTTTCCTTCC	ACGTTGGATGAAGCAGGAAGACATGCTGTG	AACCCTGTCCCCAACCAGCATTAA	12
	rs4904708	ACGTTGGATGGAAGACAAAGCATGAGACAG	ACGTTGGATGGGCGAGGTGTTTGTCATTTC	AACTGCAGACCGAGA	12
	rs4904725	ACGTTGGATGGGCCTTTCTTTGATTCACTG	ACGTTGGATGCTGGACTCTGACCACTAATG	GGGAACTGGCTTGGCAGGT	15
	rs6575138	ACGTTGGATGTAACAACCAGAGCCCTCCAG	ACGTTGGATGGGTGCTTCTCATTGGCCAAG	GGGTTTGGGCATGTG	16
	rs7145137	ACGTTGGATGGAATTTTGAACCTCACCCCC	ACGTTGGATGGTAACTGAACGCTGTTTCCC	ATGCAAATGGCACAGGATCCACGTTC	17
	rs7152676	ACGTTGGATGGATGATAAAGAGCAAGACAGG	ACGTTGGATGTAGAAACCATCCTGCAGGTG	GAATAGCAAGACAGGAGGAAT	17
	rs7154098	ACGTTGGATGTCTAAGGCCGAATAGGACAG	ACGTTGGATGTGTATCTCCCGCTTTACCAC	CGGCTTCATGGAACTTAC	12
	rs753310	ACGTTGGATGCAAAGCAGTGTGTGCAGATG	ACGTTGGATGCAACCTAGGCCAATCCTAAG	TCCCTTGTGCAGATGCCCCCACCACA	13
	rs8003019	ACGTTGGATGCTCCCAACACTGTTGCAATG	ACGTTGGATGGCTATGGTTTGAATGTGTCC	GGAGATGGGGATTAAGTCTTGA	15
	rs8005454	ACGTTGGATGGATACAAAGCTACTCGAGGG	ACGTTGGATGACAGGATGATGCACCATGCC	TCCAAGCTCTGGTCAGTTGT	17
	rs8016414	ACGTTGGATGCTAGCCAGTACCAGTCTGTG	ACGTTGGATGCGACGCCATTTTTAGAAAGC	TCTGACCCCTGCTATACAC	12
	rs8020106	ACGTTGGATGGATGAGGAGATACCTCGTTC	ACGTTGGATGTAGCGCTGCCACTCACCACA	TTCTGCAGAACTTAGCAGCCCTTATCT	16
	rs8022840	ACGTTGGATGGCTAGAATCCCATCCATCAG	ACGTTGGATGTGCAGCAGCTCACAAGACAC	GGGGCCTCTAAGCACATCCAAAAA	13
	rs881218	ACGTTGGATGCCTTAGGCTGTTTGCTCATC	ACGTTGGATGCTCTGTCCCATACACATACG	TTCCCATCCCAGTTATGTTTGCCTA	13
	rs9323852	ACGTTGGATGTGAGATGTGGAGGCAGAATC	ACGTTGGATGCCTAGAGCTCAGAAACAGAC	CGTTCTATGCCTCCGAAGAATC	17
	rs942738	ACGTTGGATGAATTATGAGCCTTGCAGTAG	ACGTTGGATGAGTCATTCATTCAGTCAGTC	GGGTCTGAGCCTTGCAGTAGTGCTTAG	16

	rs942746	ACGTTGGATGTTCCTTCAAGGCAGCTATCC	ACGTTGGATGATTTGCAGGGATGACCAGAG	TCACGGCAGCTATCCCTTAAAGACATC	12
	rs942748	ACGTTGGATGTTTCCCCTTTCTGTCCTAGC	ACGTTGGATGTCAAGAGATTGGGATGGCTG	GTCCTAGCTTGAGTTTAAG	14
	rs9944033	ACGTTGGATGAGAGGATGGGTGGGACACAG	ACGTTGGATGTGCAGACTGTGGAGGATATG	GGGGCGGGACACAGCAGGACA	17
	rs998338	ACGTTGGATGTAAAGATAAATCCCCTTAC	ACGTTGGATGGGAATAAGTTGGTAGTTCCT	GATACATCCCCTTACCCCACA	12
C14orf64	rs1038039	ACGTTGGATGTTCCAATCTTTGGCTCCCTC	ACGTTGGATGGGTTAAAGCGAAAGTCTTGC	CGTTGGCTCCCTCAAGTTTAG	14
	rs11628375	ACGTTGGATGTTCACACTCTGCTGTGAGAC	ACGTTGGATGGGTACCAATTTTAGGGCCTG	CTTTGAGACTAGAATTTGTATGTGATGC	15
	rs1466439	ACGTTGGATGAATTGTCCAGGTGACAAGGC	ACGTTGGATGATTCGGGAGATCCATGTATG	GGGCCGGTGACAAGGCAGTTATGTCCCT	16
	rs1551540	ACGTTGGATGGGCCAGAGCACAAATACATC	ACGTTGGATGCACCCTCTCTCTCTGACTTC	CTTTCTATAAGTGCTGTTTGGA	16
	rs2604989	ACGTTGGATGTTTTGTGGTGACAGGAGTGG	ACGTTGGATGTCCTCCTGCAAATCTCTGTC	GCACAGAGAGGAGATG	15
	rs2607049	ACGTTGGATGCAAAGAGCCAATGCCTAGAG	ACGTTGGATGCTTTTGGGGTTGTTTGTTTC	ACTGCCTAGAGGATGACAA	12
	rs2809121	ACGTTGGATGCCTACACAAAGGACTGAAAG	ACGTTGGATGCTGCAGCATATTGGGCTTAC	ACAGATGAAGAAGCTAAGG	17
	rs3818667	ACGTTGGATGGCTTCCAGAGACAGACTTTG	ACGTTGGATGCAGGCCATAGAACACTATTC	GGTCCCATCCCATCTTCCCACACTA	15
	rs754604	ACGTTGGATGCTCAGTCCCAGGCAATTCAG	ACGTTGGATGCCAACTCTCTAAGTTGAAGG	AGGGACAATTCGTCGTCA	13
	rs899117	ACGTTGGATGAGACGGACTTTGCAGATGTG	ACGTTGGATGCACCCTCTTCTTTCTTTCCC	ATGATTTTGAAATGGGGAGATTATC	16
PPP2R5C	rs10132483	ACGTTGGATGATCAGTCATGTCCACACTGG	ACGTTGGATGAGCTACTGACGACTTCTCAC	CTGCCTGATTCTTCCCA	12
	rs11624542	ACGTTGGATGTCCAAAACAAGGCTCCCTTC	ACGTTGGATGCATTAGACATGGAAGCTCCG	CTCCCTTCCTGCTGG	18
	rs12589350	ACGTTGGATGCAACTTGGGGCAAATTTGGG	ACGTTGGATGGCTAAGTTTCCTCATCTGGG	TGGGGCAAATTTGGGATTCAT	15
	rs1677990	ACGTTGGATGTGCAGTTTTGATCCAGGCTC	ACGTTGGATGCCATGGTAACTTAGCTCCTC	CGGCGCTCTCTGAGGACCCACAC	12
	rs1677999	ACGTTGGATGTCTCAAGTATGACTGGGTGG	ACGTTGGATGTCCCTTACCTCTCTTCACTG	TGGGAAGTCCTTGGTTTAAT	18
	rs1678002	ACGTTGGATGGCTGGTAAGTCACTTGATAC	ACGTTGGATGACAGTGCTTAGAAAAATCCC	CAACATTTATTGAGGACATGCTA	12
	rs1678019	ACGTTGGATGATAGATACTCCTCCCTTGTC	ACGTTGGATGCTGCTTCTGTGTTCAGAATG	CTCCTCCCTTGTCAACTAC	14
	rs1678032	ACGTTGGATGGGAATCTGTAAGGTGGAATC	ACGTTGGATGAGTACAAAGTAGACCACTTG	GGTAAGGTGGAATCTGAAAC	14
	rs1746587	ACGTTGGATGGACTTGTGCCTGAAGAAGAG	ACGTTGGATGCCCCAAGTGTCATTAGTCTG	CCCCAACTGAGCAGAGAATATATACCCT	12
	rs1746596	ACGTTGGATGAACAACCAAATGCCCATCAG	ACGTTGGATGCATTCATCTCTATCGCTTTGG	TTTAAATGCAAATGCAGTCTAT	15
	rs2256537	ACGTTGGATGTTTCCACTGGGCAACCTTTC	ACGTTGGATGTGCATGGATTCACATTCAG	GTTTAGTTTCCAAATGTCTTACC	13
	rs2281772	ACGTTGGATGACCATGTGGAGCTGGATGTC	ACGTTGGATGTTTTCCATGCATGGACCCTG	TGGAGCTGGATGTCTGATTG	16
	rs3405	ACGTTGGATGTCCAGCACAGCATTATTTCC	ACGTTGGATGTCCCCAAGGCTTTCTAGAAC	ACAGGCCATATTTCCCA	12
	rs3783370	ACGTTGGATGGTGACTGCTAGAAGATTCGG	ACGTTGGATGAGCCCTTTCCTGTGCAACTC	TGGATTAGCACATGGCTGA	14
	rs3993391	ACGTTGGATGCCGAAAGAAGAAACCTCCAG	ACGTTGGATGGCCTTGTTTCCAGTGCTATC	CCCATTTGCTAGAGAATCACC	15
	rs7143539	ACGTTGGATGCAGAAACTCTGCGTACGAAC	ACGTTGGATGACTGCCCAGGGAAGGTATTG	TCCCCTCTGCGTACGAACAACTTGG	12
ANKRD9	rs1007343	ACGTTGGATGATGTCTCCGAGCACCTTTAC	ACGTTGGATGGACATGGGCAGTTCTTCTTG	TGCCCAACATGCTGC	18
	rs2273905	ACGTTGGATGGGTCTCCAGGTTGTTTTCAC	ACGTTGGATGCACCAGCCCTCTATCCAAAC	GCTACAGGTTGTTTTCACCCTCCT	14
	rs3742440	ACGTTGGATGCCTTCACTATCCTCCCTAAC	ACGTTGGATGGCCCCGAATCCCAACCTTAT	TGTGCAGAGATGCTGA	12
	rs942024	ACGTTGGATGACTCCCCAGCGTCCTCACA	ACGTTGGATGAGGGAGCCACAGGCAGTAG	GGGGTCCTCACACACTTTCC	14
SDC4	rs1008953	ACGTTGGATGTGGTATTACAGGCATGAGCG	ACGTTGGATGGGTGGCTGTGGCAATTTTTC	TGAGCGACTATACCCAA	15
	rs1981430	ACGTTGGATGGGGCCAGACATTGCTCTAAT	ACGTTGGATGAGCCTCAGGACCAGTTTAAG	CGGAGCTTAATTCCAACAACAATCT	12
	rs2072786	ACGTTGGATGTCTTTTCGAGGACAGAGGAG	ACGTTGGATGCTTCTTGCAGAAAGCAGAGG	AAGCGGAGAGGAATGGTTTAGACT	12
	rs2251252	ACGTTGGATGAAGAACAAGGCTGGGAAGTG	ACGTTGGATGCACTCTGATCTCAGCAGTTC	CTTTTGGAGGGAGGAAGGAT	15
	rs2267867	ACGTTGGATGAACAGCAGCTCTGAGCAAAG	ACGTTGGATGCCAACTCTCCTGGCAAATAC	TGAGCAAAGTTCATTATTCTACTA	14
	rs2267871	ACGTTGGATGAGTGACAGTTCCCTTCCTTG	ACGTTGGATGAGAAAGCTTTCCAGGAGACC	ATGGGGTCCTATCTCTGAAGATTC	18
	rs2284278	ACGTTGGATGTGTGACCCCAGGATTTTCTG	ACGTTGGATGGCCAATGTTTCAATGCCCTC	TCTGATGTTTGTGAGGAC	15
	rs4599	ACGTTGGATGTCCAACAGATGGACATGCTC	ACGTTGGATGCTAACCACGTTTCTAGAGGC	GACAGTAGCCATGAACTACA	13
	rs6073708	ACGTTGGATGACAGTGTGGCAAAGTCTCAG	ACGTTGGATGAAACCCCCTCTGAACCAGTG	AGGCGAGATTTAAGCTGCCACGAGATTT	16
	rs6073714	ACGTTGGATGAACCCTTGCTCTGCAAAGTG	ACGTTGGATGGCCTGGCCTTGAAATTCTAC	AAACCAATAGCAACACC	13
	rs6104115	ACGTTGGATGCCCTCTGCAACACACATGTC	ACGTTGGATGCTCCTACCACTAAATCTGGG	ATCGGCAACACACATGTCACCTACA	14
TP53RK	rs11550540	ACGTTGGATGTTGGCTAAGTTGGAGAGACC	ACGTTGGATGAGGCTCAGTGACTGTTCGAG	AGTTTTTTCAGTCTCCATAG	13
	rs6012009	ACGTTGGATGATTTCCGAAAGGCCGGACAC	ACGTTGGATGTGAGGGACTCCCCTGCGCT	TTAAGGAAGGAGAAACTGAGGGAAATC	18
	rs971759	ACGTTGGATGAGTGCTTGTCACCTACAACC	ACGTTGGATGATCCAGCCATCTTTGCACAC	ACAACCCAGCAATATCACC	13
TUBB1	rs10485828	ACGTTGGATGGGGTGATGACTGAATGAAAG	ACGTTGGATGTACTACTGTTAGATCCCAGG	GGAAAAGACCATTTCAAAATAGC	16
	rs151337	ACGTTGGATGTAACCTGCTGGTTTCCAGAG	ACGTTGGATGCTGAATTCTCAGCACAGAGC	TCCCTAGGCACCCTCTTCCACTCAAT	17
	rs151348	ACGTTGGATGGAGCAGGGTACAAGAAAGTG	ACGTTGGATGGGCCATTATGGGTGACTCAG	GCTCAGCTCCTCTGG	16
	rs6070696	ACGTTGGATGTTTTCTCCCCACCTGAACTC	ACGTTGGATGGGCATGAATTTAATAGGAGG	CAAATGTCAAACACAAAGTACTAAAT	15
ROPN1-	rs1158012	ACGTTGGATGCCCCGAATTCTAGGTTTCTG	ACGTTGGATGAATACTCTGTTCTGAGGCCC	GGGTTCTGTGATATTGGAAAAAACA	17
KALRN	rs11708466	GGCACATCTTTCCCAGAATC	AGTGATTTGTGCCCCTGAAC	TTCTCTGAACTGGAATGGGAAAT	11
	rs11712619	GCAAAGCAATTGGTAACAGG	GGAAAATCCACAAGGCACAC	TGGTAACAGGTGATTAAGAA	11
	rs11929003	ATTCAAGTGAACACCCCAGC	TATGTGTACCCCTAAGCCTG	TGGTGGCAAAGGGCTGCTA	9
	rs12634530	CCTCTCTCTCTAGAGGAAAA	CAAGAACCACTAGGTAACCG	AGAGGTAACCGAGAGAAAACAACG	9
	rs12637456	CATTTCAGTCACAGCCACAG	TAGGAGACTGGAAGAAGGAG	AGGAGTTCGGAGTCCA	9
	rs12695434	CCCAAATGTACAAAGGACTC	CAGCTCCAATTTGAAAGACC	GCTTTTGTTACATTTATGATCCTAATTT	11

	rs13064819	ACGTTGGATGCCCGGAAGTGGTAAACTTTG	ACGTTGGATGATCCCCAGGAGATTTGTGTG	GTTGAGCCAAATTTGAAGGA	17
	rs13075202	CAGCTGTGTCCCTGGATATT	TCTGTCCACAAGGTGAAGAG	TCAGCTGTGTCCCTGGATATTCAGAAG	8
	rs1317671	AGCTGTTGTGGCCCTAGAAG	CCACCAGCACTTTTAAACGG	AAATCACTTTAGATCTCCCAGCCTC	11
	rs1373609	ATTCCCTGAGTAGGACGAAG	CAGGTGCAGTTGAATAAGGC	GGAAGGCGTAAGCCGAACAGC	11
	rs1444760	ACGTTGGATGACACACACAAAGCTCCAGTC	ACGTTGGATGCTTGACTCCATCTGGCTTAG	GGAAGTGCCAAGACA	17
	rs1444763	AAGGTGGTTCACTCATGTGG	CTTCTTTCCAGGTTCCTGTC	TTCCTGTCTGAGCAAAT	11
	rs1444766	ACGTTGGATGTGACCAGAGCTCAGTCATAC	ACGTTGGATGTCTCTTCTGAAGCACACGAC	GGTAGTCATACAACTGTTCCTA	18
	rs1444768	TGCAAAGTCTGCAATCCAGG	TTTGGACCAGTATGGCAAGG	CCTGAAATTCCAACAAAACAGC	10
	rs17221479	ACGTTGGATGCTTGTGTGCACATCAGCTAC	ACGTTGGATGCTGGCACATAGTAGGTGTTC	AGCTACCACCCATTCGG	14
	rs17286604	CTTTTCCTCATGTGGAAGGC	CTTTGACGTGATAACCACCC	CCCCTCATCTGCCTTGGAAAT	11
	rs17376453	GAAGAGATGGAAGTGTCCTG	TGAAGCCCCAAGCCTCACTA	ACTACCTCCAGCCAGAAAC	9
	rs17377867	TGAATTGGTACCAGGAGAGG	CCATATTTAGTAGGATATCAC	CCTCTTAGGATATCACTCCCACA	11
	rs17377867	ACGTTGGATGCCATATTTAGTAGGATATCAC	ACGTTGGATGTGAATTGGTACCAGGAGAGG	CATCTTTAGTAGGATATCACTCCCACA	17
	rs1950091	ACGTTGGATGTGCTCATGATCAGAAGCAGG	ACGTTGGATGAAAGCCTTTAAAAGCACCCC	AATCAATTAGAGGATGAGGAGGT	17
	rs2141664	AGGGATTTGGGTTATATCTG	TGACCCCCAGAAAAAAAATC	GAATGTGAGTGCATGAAACAATGA	11
	rs2141664	ACGTTGGATGAGGGATTTGGGTTATATCTG	ACGTTGGATGTGACCCCCAGAAAAAAAATC	GAAAATTTGAGTGCATGAAACAATGA	14
	rs2280422	CTCCAGATCCAGGTAGTATG	CAGGCAATGGTTAGCCCATC	GGCTTGACCCTGAAA	11
	rs2332719	CAAAGCACTAGCAGCTAGAG	CCCATGAAATTCCTCCCTTG	ATCCTCCCTTGTGTTTAAC	11
	rs3821525	AAGAGTCCTCTTCTGACTGG	ACATCAGTAGTGCCCATTTC	GTGTTGCCTCTGATAACTCAA	11
	rs4234218	CAACAACACTTTGAGAGGG	ACTGTGTGCCAGTAGGCAAG	AGTGCCAGTAGGCAAGTACTTCATCT	8
	rs4608634	GAAGAGAAGGCTAGGCAATG	GCTAGTTGTTACTTTTACTG	GGCGTTGTTACTTTTACTGTCTTTAA	11
	rs4608635	ATCTAACAATGGTGGGCAGG	GCAGCAACCTTCACACTTAG	TCGTTTCCAAAGAGAAGCTTCAGGAAAG	11
	rs4678085	CTCCCAGAAGTTCAGAAACC	GGGACAGGCAGGATAAATTC	ATCTCTCTCTGGGGC	11
	rs4678086	GGATCTAGGGAAGTTTTAGC	TTGCAGTCTCCTGCAGAAAC	GGAGCTAGGGAAGTTTTAGCAACTTA	11
	rs6774735	GTGCTCAACCATATGTCCTG	AGCTGCCTGGTGAATGAATG	TCCACCATATGTCCTGGGAATT	11
	rs6774735	ACGTTGGATGGTGCTCAACCATATGTCCTG	ACGTTGGATGAGCTGCCTGGTGAATGAATG	ACCATATGTCCTGGGAATT	18
	rs6781700	CTCTGACTATGAAGGAGTGC	AACAAGTGTCCTGTGTATCC	AGGGTCCACAGAATTAAATCTGAACCA	11
	rs6784664	ACGTTGGATGGGGAGTTATGTTTGTCCTTG	ACGTTGGATGGCTGGTAGGAAATGGGAAAG	AAGAGGAATGGATATTTTGGC	18
	rs6810298	GTCATGTTCACTGCACAAGG	TGCTCTTGTTGCATTGGTGG	ACAGTCACCTTCCCTCCTTTTATC	8
	rs7434266	ACCAGCCCCATCTGATGAAC	ACTAGGCTCAACTTGCTCC	GCATCTGATGAACAAGACC	9
	rs7613868	GGTCCTTTCTTGTCACTTGC	TTCTGAAGTCATCCCACGAG	AGGCATCCTGAAGGAC	10
	rs7633408	GGACTAGCCCTTTGAGCTTG	GACTGGATAGACAAGCAGAG	GGTCCGGAGGTTTAGT	11
	rs7633408	ACGTTGGATGGGACTAGCCCTTTGAGCTTG	ACGTTGGATGGACTGGATAGACAAGCAGAG	TTGGTCCGGAGGTTTAGT	17
	rs9289231	TTAGCAGAGGGCATGCATAG	AGTTGCCCTGGCTCAAAAAC	GGACTTGGTTATACAATTTAATTT	9
	rs9838361	GCACATGAGAAGTCTTGAAC	TACCATGCTGTCCCCAAATG	CACAAATATGAAGTGTCGATAATAA	11
	rs9880957	ACGTTGGATGCTCTTGCATTCAAAGGCCTG	ACGTTGGATGAGTCAGAGAAGCAAAAGAGC	CCGCCTTCAAAGGCCTGAACCAGAGTTA	17
CFH	rs1061170	GTTATGGTCCTTAGGAAAATG	ACGTCTATAGATTTACCCTG	GACCCTGTACAAACTTTCTTCCAT	2
EPO	rs1617640	TCTAAGGTGTGAGAGACCAG	TATGGCTTCTGGAAACCCTG	TTCTCTGGGAATCTCACTC	5
	rs4729607	AGACACTGGAATCTAAGCTG	CCTTCACAATGCCACAGTTT	ATCTAAGCTGAAGGCTAA	7
	rs564449	CTTTGGAGGCGATTTACCTG	AGAAGTCACAGCTTGCCACC	AGGGACAGGATGACCTGGA	7
HO2	rs2160567	GCCAGGGAAGTTAAAGAGTG	TTCCTTCTCTTCCCTTCCTC	TTTCCCTTCCTCCTCTGTCCTAATGG	9
	rs3761680	AGTGACCTTACCCCTCCCTG	GTGAGCCAAACCCCAGACT	GGCGTTCTTCAGGGCTAGGCGCCCTGG	7
	rs3761680	ACGTTGGATGAGTGACCTTACCCCTCCCTG	ACGTTGGATGAGGTGAGCCAAACCCCAGAC	GTGAAGGGCTAGGCGCCCTGG	18
	rs7665	AAGTGGGAGCGGGAAAGTGA	TGCTAACGGAGCTGGGACT	CCAAAGTGACCACTGAGC	7
	rs7702	TTAGTCTTCTGCCTGTCAGC	TTGAGCACAGCATCATCCAG	CTTTGGAGGCGCACCCTTAAGCAA	7
KLK1	rs2659058	TGGAGAAGTGGTGAGAACAG	CCCAAGTCAGTATGTGATTC	GTGCAGGACCCACTTGA	2
	rs266116	ATATGGCCAGTGGAGTACAG	ACCATGCATTTCACCTCCTG	GGGAGTACAGTTGGCTGATTGGCTTA	4
	rs266117	AGGGCTAAAAGGAGGAGATG	CCAACTTAGACTTCCAGCAG	CCCCCATCTCGGAAACAGGTA	7
	rs2739454	TGTGGGTATCAGCAGAGATG	TATGGATGAGAACTGGGGAG	CATCAGCAGAGATGACTTTG	1
	rs3212810	GTGTTGAATGAATGGAGCCC	CACTGTTCCTTTTCACCCAC	CCTAGGTGAGGAGAAGC	8
	rs3212820	ACGTTGGATGTGGGTGGGAAGGGACATTTG	ACGTTGGATGGAAAACCCTGCTTCCCAAAC	GTTTGCAGACCCCCCC	17

Table B.2: PCR (F and R) and extension (E) primer sequences designed to validate the imputed results of the SNPs rs7620580, rs6438833 and rs11712039 of the *KALRN*, using Sequenom's iPlex assays. The plex number where they were included is indicated.

Gene	SNP	F Primer	R Primer	E Primer	# Plex
KALRN	rs6438833	ACGTTGGATGAGTGCTTTGCTGAGTGTACC	ACGTTGGATGCATTGACTGTTAATGATGTGG	ATATTATTACCAATTTCAAGTGTTG	22
	rs11712039	ACGTTGGATGGTTGCCCCAAATTTCAGTAG	ACGTTGGATGCGTGAGGACCTCAAATCTTC	GCTGAATAGACATATACAATGA	22
	rs7620580	ACGTTGGATGTTTTGCTTGGTAGTGGTTGC	ACGTTGGATGCACCCCTTTGATCTGTGTTC	ATTCCACACATGCACACA	22

APPENDIX C – ASSOCIATION RESULTS

We genotyped in our Portuguese case-control biobank an appropriate number of SNPs per studied gene: the *PDE4D* and *ALOX5AP* that have been controversially implicated in IS; the genes prioritized based on the convergence of published linkage results and our gene profiling data; and our selected biological candidate genes. In this Appendix we present all association results with the IS risk obtained for these genes in the allelic and genotypic (unadjusted and adjusted for covariates) tests performed (Tables C.1, C.3, C.7 and C.9). Genotypic tests were made using the log-additive genetic model. For each studied SNP it is also presented its HWE in the control population as well as the ratio count and the frequencies of the alleles in the cases and controls. Results of haplotype association with the IS risk are presented for each gene too (Tables C.2, C.4, C.8 and C.10). For the genes prioritized following our GC procedures, only the ones that had significant associations have their obtained results presented here.

In this Appendix are accessible the detailed association results for the most significant imputed SNPs in the KALRN gene region (Table C.5), as well as the association testing results in our case-control biobank for the three imputed associated SNPs in *KALRN*, allowing the validation of the imputed results by directly genotyping (Table C.6).

There are also presented all association results with IS risk obtained for our GC genes in the Spanish population in the allelic and genotypic (unadjusted and adjusted for covariates) tests performed (Table C.11). The positive results of association with atherothrombotic, cardioembolic, and lacunar subtypes of stroke are also available (Table C.12). Genotypic tests were made using the log-additive genetic model, and for each studied SNP is presented its HWE in the control population as well as the ratio count and the frequencies of the alleles in the cases and controls. Positive results of haplotype association with the risk of IS are presented too (Table C.13).

Finally, the results of SNP and haplotype association testing with IS risk for TTC7B gene in the joint analysis of the Portuguese and Spanish datasets are presented in the Table C.14 and Table C.15, respectively.

Table C.1: Results of SNP association testing with IS risk for *PDE4D* and *ALOX5AP*. Significant p-values results are highlighted in bold. Multivariate logistic regression (log-additive model) was performed to adjust the analyses of association with risk for hypertension, diabetes and ever smoking.

SNP N	SNP ID	Gene	HWE p-value Controls	Assoc Allele	Case, Control Ratio Counts	Case, Control Frequencies	Unadjusted test Allel p-value	Unadjusted test Geno p-value	Adjusted test p-value*
1	rs2963821	*PDE4D*	0.374	C	522:600, 466:556	0.465, 0.456	0.667	0.657	0.746
2	rs2963820	*PDE4D*	0.707	G	248:806, 233:767	0.235, 0.233	0.902	0.883	0.507
3	rs2938784	*PDE4D*	0.455	T	430:686, 392:628	0.385, 0.384	0.963	0.962	0.822
4	rs2938787	*PDE4D*	0.294	T	887:239, 811:221	0.788, 0.786	0.915	0.914	0.422
5	rs11951422	*PDE4D*	0.445	T	258:864, 227:803	0.230, 0.220	0.596	0.653	0.251
6	rs7710463	*PDE4D*	0.808	C	121:981, 104:908	0.110, 0.103	0.600	0.600	0.334
7	rs6879326	*PDE4D*	0.537	T	557:565, 507:521	0.496, 0.493	0.881	0.876	0.745
8	rs1435077	*PDE4D*	0.281	C	388:706, 352:672	0.355, 0.344	0.599	0.576	0.876
9	rs1529842	*PDE4D*	1.000	C	947:163, 857:155	0.853, 0.847	0.684	0.686	0.673
10	rs10066510	*PDE4D*	1.000	C	981:143, 890:142	0.873, 0.862	0.478	0.474	0.099
11	rs6889660	*PDE4D*	1.000	C	634:488, 546:480	0.565, 0.532	0.126	0.123	0.375
12	rs17315957	*PDE4D*	0.458	T	301:823, 237:795	0.268, 0.230	**0.041**	**0.037**	**0.034**
13	rs12518928	*PDE4D*	0.700	A	265:783, 222:774	0.253, 0.223	0.112	0.105	0.108
14	rs1533019	*PDE4D*	0.481	T	786:344, 698:334	0.696, 0.676	0.336	0.329	0.350
15	rs16890459	*PDE4D*	0.729	C	172:954, 154:874	0.153, 0.150	0.849	0.849	0.780
16	rs7732249	*PDE4D*	0.414	T	956:134, 891:125	0.877, 0.877	0.995	0.901	0.686
17	rs2136203	*PDE4D*	0.525	T	798:314, 712:302	0.718, 0.702	0.433	0.425	0.321
18	rs1396476	*PDE4D*	1.000	G	164:958, 149:883	0.146, 0.144	0.906	0.906	0.777
19	rs2910831	*PDE4D*	1.000	T	720:378, 625:369	0.656, 0.629	0.199	0.189	0.062
20	rs2910829	*PDE4D*	0.182	A	600:480, 534:482	0.556, 0.526	0.169	0.175	0.234
21	rs966221	*PDE4D*	0.929	G	675:445, 582:446	0.603, 0.566	0.086	0.091	0.079
22	rs2898269	*PDE4D*	1.000	A	1027:99, 914:110	0.912, 0.893	0.127	0.133	0.239
23	rs2962964	*PDE4D*	0.533	G	809:321, 718:312	0.716, 0.697	0.337	0.331	0.570
24	rs6889641	*PDE4D*	0.524	G	994:134, 904:122	0.881, 0.881	0.994	0.993	0.686
25	rs7442640	*PDE4D*	0.738	G	206:906, 158:858	0.185, 0.156	0.069	0.068	**0.006**
26	rs4604150	*PDE4D*	0.089	C	408:694, 336:664	0.370, 0.336	0.101	0.107	0.088
27	rs11739760	*PDE4D*	0.242	T	901:225, 799:223	0.800, 0.782	0.295	0.305	0.335
28	rs12658881	*PDE4D*	0.258	C	883:241, 793:231	0.786, 0.774	0.532	0.476	0.419
29	rs12523473	*PDE4D*	0.149	A	499:533, 452:552	0.484, 0.450	0.132	0.143	0.316
30	rs12515974	*PDE4D*	0.293	T	801:313, 706:308	0.719, 0.696	0.248	0.257	0.239
31	rs10471476	*PDE4D*	0.791	A	613:509, 541:487	0.546, 0.526	0.351	0.360	0.339
32	rs6449467	*PDE4D*	0.527	G	261:853, 229:785	0.234, 0.226	0.644	0.638	0.891
33	rs12189147	*PDE4D*	0.506	C	330:796, 282:748	0.293, 0.274	0.321	0.311	0.864
34	rs6887545	*PDE4D*	0.424	A	197:923, 170:854	0.176, 0.166	0.544	0.550	0.967
35	rs37707	*PDE4D*	0.736	T	306:758, 275:735	0.288, 0.272	0.438	0.434	0.911
36	rs37684	*PDE4D*	0.714	T	674:448, 615:415	0.601, 0.597	0.864	0.864	0.942
37	rs789395	*PDE4D*	0.918	T	369:743, 319:701	0.332, 0.313	0.346	0.357	0.836
38	rs364917	*PDE4D*	0.912	C	328:798, 284:744	0.291, 0.276	0.440	0.444	0.959
39	rs371424	*PDE4D*	1.000	G	358:728, 319:701	0.330, 0.313	0.407	0.401	0.974
40	rs702553	*PDE4D*	1.000	T	370:736, 317:701	0.335, 0.311	0.255	0.263	0.700
41	rs11746901	*PDE4D*	0.133	G	158:968, 139:883	0.140, 0.136	0.773	0.864	0.958
42	rs12153798	*PDE4D*	0.134	C	164:954, 141:879	0.147, 0.138	0.577	0.573	0.748
43	rs26956	*PDE4D*	0.339	C	812:304, 726:302	0.728, 0.706	0.272	0.275	0.294
44	rs3857232	*PDE4D*	1.000	T	975:147, 886:140	0.869, 0.864	0.712	0.714	0.929
45	rs35386	*PDE4D*	0.297	T	417:669, 378:634	0.384, 0.374	0.622	0.672	0.912
46	rs40512	*PDE4D*	0.114	C	446:668, 399:615	0.400, 0.393	0.746	0.748	0.687
47	rs35384	*PDE4D*	0.080	G	738:320, 687:319	0.698, 0.683	0.472	0.507	0.354
48	rs35383	*PDE4D*	0.189	C	776:352, 699:331	0.688, 0.679	0.643	0.651	0.552
49	rs35382	*PDE4D*	0.183	G	604:478, 539:479	0.558, 0.529	0.186	0.219	0.141
50	rs7727343	*PDE4D*	1.000	C	162:962, 140:884	0.144, 0.137	0.622	0.625	0.525

51	rs10051720	*PDE4D*	0.088	G	760:354, 683:333	0.682, 0.672	0.623	0.631	0.359
52	rs27565	*PDE4D*	0.656	T	601:509, 545:469	0.541, 0.537	0.855	0.857	0.597
53	rs10939851	*PDE4D*	0.066	A	873:163, 840:158	0.843, 0.842	0.952	0.953	0.924
54	rs152341	*PDE4D*	0.930	A	563:559, 503:529	0.502, 0.487	0.505	0.503	0.541
1	rs17222814		1.000	A	84:1036, 76:954	0.075, 0.074	0.915	0.915	0.947
2	rs17216473		0.823	A	147:945, 113:875	0.135, 0.114	0.163	0.160	0.149
3	rs4076128		0.471	G	383:743, 319:711	0.340, 0.310	0.132	0.129	0.106
4	rs4769055	*ALOX5AP*	0.633	A	426:696, 369:651	0.380, 0.362	0.391	0.375	0.548
5	rs9579645	*ALOX5AP*	0.222	C	201:873, 180:836	0.187, 0.177	0.555	0.550	0.429
6	rs12429692	*ALOX5AP*	0.422	T	310:814, 266:758	0.276, 0.260	0.402	0.407	0.798
7	rs17612031	*ALOX5AP*	0.152	T	1040:82, 946:86	0.927, 0.917	0.376	0.389	0.200
8	rs9671182	*ALOX5AP*	0.594	C	519:601, 459:573	0.463, 0.445	0.386	0.394	0.169
9	rs9671124	*ALOX5AP*	0.329	C	623:487, 551:471	0.561, 0.539	0.305	0.310	0.209
10	rs9579648	*ALOX5AP*	0.770	G	936:190, 837:189	0.831, 0.816	0.347	0.359	0.104
11	rs17074975	*ALOX5AP*	0.219	T	66:1056, 58:956	0.059, 0.057	0.873	0.876	0.464
12	rs9551963	*ALOX5AP*	0.657	A	547:577, 473:551	0.487, 0.462	0.251	0.260	0.165
13	rs9551964	*ALOX5AP*	1.000	A	794:324, 729:301	0.710, 0.708	0.901	0.903	0.747
14	rs9315050	*ALOX5AP*	0.790	G	104:1018, 95:937	0.093, 0.092	0.959	0.960	0.627
15	rs3935644	*ALOX5AP*	0.903	A	277:853, 242:784	0.245, 0.236	0.615	0.621	0.182
16	rs4445746		0.611	A	270:850, 227:795	0.241, 0.222	0.299	0.312	0.472
17	rs9578200		0.459	T	208:918, 187:837	0.185, 0.183	0.900	0.898	0.984
18	rs9579653		0.090	C	241:883, 200:830	0.214, 0.194	0.245	0.231	0.186
19	rs4491352		0.514	C	488:626, 395:625	0.438, 0.387	**0.017**	**0.020**	0.062
20	rs4769062		0.849	A	178:932, 137:875	0.160, 0.135	0.106	0.112	0.267
21	rs4769063		0.848	T	176:938, 135:881	0.158, 0.133	0.101	0.107	0.295
22	rs4238139		0.563	G	316:798, 262:752	0.284, 0.258	0.190	0.190	0.320

*Multivariate logistic regression (log-additive model) with backward elimination of hypertension, diabetes and ever smoking

Table C.2: Results of haplotype association testing with IS risk for *PDE4D* and *ALOX5AP*. Significant p-values results are highlighted in bold. Haplotypes were estimated in Haploview using confidence intervals algorithm.

Gene	Haplotype	Freq.	Case, Control Ratio Counts	Case, Control Frequencies	p-value
PDE4D	Block 1: rs2963821, rs2963820, rs2938784, rs2938787, rs11951422				
	AATTC	0.377	426.0 : 700.0, 388.5 : 645.5	0.378, 0.376	0.901
	CACTT	0.218	248.0 : 878.0, 222.9 : 811.1	0.220, 0.216	0.795
	CGCCC	0.207	231.6 : 894.4, 214.7 : 819.3	0.206, 0.208	0.908
	AACTC	0.154	168.6 : 957.4, 163.3 : 870.7	0.150, 0.158	0.597
	CGCTC	0.023	22.4 : 1103.6, 27.5 : 1006.5	0.020, 0.027	0.298
	Block 2: rs7710463, rs6879326				
	TT	0.495	560.3 : 569.7, 510.3 : 523.7	0.496, 0.494	0.914
	TC	0.399	445.3 : 684.7, 418.0 : 616.0	0.394, 0.404	0.629
	CC	0.106	124.4 : 1005.6, 105.7 : 928.3	0.110, 0.102	0.554
	Block 3: rs1435077, rs1529842				
	TC	0.502	562.7 : 563.3, 520.7 : 513.3	0.500, 0.504	0.857
	CC	0.349	398.3 : 727.7, 355.3 : 678.7	0.354, 0.344	0.622
	TT	0.150	165.0 : 961.0, 158.0 : 876.0	0.147, 0.153	0.683
	Block 4: rs10066510, rs6889660, rs17315957, rs12518928, rs1533019, rs16890459, rs7732249, rs2136203				
	CTCGTTTC	0.279	299.6 : 812.4, 295.2 : 724.8	0.269, 0.289	0.305
	CCTATTTT	0.245	290.9 : 821.1, 232.0 : 788.0	0.262, 0.227	0.067
	CCCGTCTT	0.139	154.9 : 957.1, 141.3 : 878.7	0.139, 0.139	0.962
	ACCGCTTT	0.122	131.2 : 980.8, 129.4 : 890.6	0.118, 0.127	0.530
	CTCGCTCT	0.118	130.0 : 982.0, 121.0 : 899.0	0.117, 0.119	0.900
	CTCGCTTT	0.046	43.0 : 1069.0, 54.2 : 965.8	0.039, 0.053	0.111
	CCCGCTTT	0.022	26.6 : 1085.4, 21.2 : 998.8	0.024, 0.021	0.621

	Block 5: rs2910831, rs2910829				
	TA	0.540	626.2 : 503.8, 543.3 : 490.7	0.554, 0.525	0.181
	CG	0.360	394.5 : 735.5, 384.7 : 649.3	0.349, 0.372	0.267
	TG	0.100	109.4 : 1020.6, 106.0 : 928.0	0.097, 0.103	0.654
	Block 6: rs2898269, rs2962964, rs6889641, rs7442640, rs4604150, rs11739760, rs12658881, rs12523473				
	ACGATTCG	0.292	315.0 : 799.0, 309.0 : 713.0	0.283, 0.302	0.320
	AGGACTCA	0.170	193.7 : 920.3, 170.2 : 851.8	0.174, 0.166	0.648
	AGGGCTCA	0.165	196.5 : 917.5, 155.0 : 867.0	0.176, 0.152	0.124
	AGAATTCA	0.114	123.8 : 990.2, 120.6 : 901.4	0.111, 0.118	0.617
	AGGATGTG	0.112	126.2 : 987.8, 114.1 : 907.9	0.113, 0.112	0.906
	TGGATGTG	0.096	95.5 : 1018.5, 110.4 : 911.6	0.086, 0.108	0.082
	AGGACTCG	0.015	14.1 : 1099.9, 17.5 : 1004.5	0.013, 0.017	0.392
	AGGATTTG	0.011	16.1 : 1097.9, 7.4 : 1014.6	0.014, 0.007	0.109
	Block 7: rs37707, rs37684, rs789395				
	AGC	0.400	449.2 : 678.8, 414.7 : 619.3	0.398, 0.401	0.891
	TTT	0.280	327.5 : 800.5, 277.9 : 756.1	0.290, 0.269	0.265
	ATC	0.274	299.6 : 828.4, 293.4 : 740.6	0.266, 0.284	0.345
	ATT	0.042	48.0 : 1080.0, 43.2 : 990.8	0.043, 0.042	0.931
	Block 8: rs364917, rs371424, rs702553				
	TAA	0.676	751.0 : 377.0, 709.0 : 323.0	0.666, 0.687	0.292
	CGT	0.282	325.8 : 802.2, 283.6 : 748.4	0.289, 0.275	0.471
	TGT	0.040	49.1 : 1078.9, 37.4 : 994.6	0.044, 0.036	0.385
	Block 9: rs11746901, rs12153798				
	CT	0.857	962.9 : 165.1, 888.9 : 145.1	0.854, 0.860	0.689
	GC	0.138	157.0 : 971.0, 140.9 : 893.1	0.139, 0.136	0.847
	Block 10: rs26956, rs3857232, rs35386				
	CTT	0.379	432.5 : 695.5, 385.9 : 646.1	0.383, 0.374	0.647
	ATC	0.283	308.2 : 819.8, 303.3 : 728.7	0.273, 0.294	0.287
	CTC	0.204	239.1 : 888.9, 202.7 : 829.3	0.212, 0.196	0.371
	CAC	0.133	148.2 : 979.8, 140.2 : 891.8	0.131, 0.136	0.762
	Block 11: rs40512, rs35384, rs35383, rs35382, rs7727343				
	CGCGT	0.374	424.1 : 695.9, 379.9 : 650.1	0.379, 0.369	0.640
	TATAT	0.294	324.6 : 795.4, 307.8 : 722.2	0.290, 0.299	0.648
	TGCAT	0.137	149.2 : 970.8, 146.3 : 883.7	0.133, 0.142	0.551
	TGCGC	0.135	154.3 : 965.7, 137.0 : 893.0	0.138, 0.133	0.747
	TGCGT	0.030	34.1 : 1085.9, 31.3 : 998.7	0.030, 0.030	0.995
	CATAT	0.017	17.0 : 1103.0, 19.9 : 1010.1	0.015, 0.019	0.459
ALOX5AP	Block 1: rs17216473, rs4076128				
	GA	0.674	743.8 : 386.2, 713.1 : 318.9	0.658, 0.691	0.105
	GG	0.200	232.3 : 897.7, 199.8 : 832.2	0.206, 0.194	0.488
	AG	0.126	154.0 : 976.0, 119.1 : 912.9	0.136, 0.115	0.146
	Block 2: rs4769055, rs9579645				
	CA	0.625	696.8 : 433.2, 656.1 : 377.9	0.617, 0.634	0.392
	AA	0.192	221.9 : 908.1, 193.5 : 840.5	0.196, 0.187	0.589
	AC	0.180	207.5 : 922.5, 181.7 : 852.3	0.184, 0.176	0.633
	Block 3: rs9671182, rs9671124				
	CC	0.455	524.0 : 604.0, 460.7 : 573.3	0.465, 0.446	0.376
	GT	0.449	495.2 : 632.8, 475.1 : 558.9	0.439, 0.459	0.340
	GC	0.096	108.8 : 1019.2, 98.2 : 935.8	0.096, 0.095	0.908
	Block 4: rs17074975, rs9551963, rs9551964, rs9315050, rs3935644				
	ACTAG	0.291	328.0 : 802.0, 302.0 : 730.0	0.290, 0.293	0.904
	AAAAA	0.241	277.0 : 853.0, 244.0 : 788.0	0.245, 0.236	0.638
	AAAAG	0.234	272.6 : 857.4, 234.2 : 797.8	0.241, 0.227	0.433
	ACAAG	0.141	148.0 : 982.0, 156.5 : 875.5	0.131, 0.152	0.166
	TCAGG	0.060	67.9 : 1062.1, 61.2 : 970.8	0.060, 0.059	0.939
	ACAGG	0.033	36.4 : 1093.6, 34.0 : 998.0	0.032, 0.033	0.932
	Block 5: rs9578200, rs9579653, rs4491352, rs4769062, rs4769063, rs4238139				
	CTAGCA	0.395	415.6 : 706.4, 434.4 : 593.6	0.370, 0.423	**0.013**
	CCCGCG	0.202	236.3 : 885.7, 197.5 : 830.5	0.211, 0.192	0.288
	TTAGCA	0.181	204.3 : 917.7, 185.4 : 842.6	0.182, 0.180	0.921
	CTCATA	0.147	177.7 : 944.3, 137.8 : 890.2	0.158, 0.134	0.111
	CTCGCG	0.063	71.6 : 1050.4, 63.4 : 964.6	0.064, 0.062	0.837

Table C.3: Results of SNP association testing with IS risk for *KALRN*. See legend in table C.1.

SNP N	SNP ID	Gene	HWE p-value Controls	Assoc Allele	Case, Control Ratio Counts	Case, Control Frequencies	Unadjusted test Allel p-value	Unadjusted test Geno p-value	Adjusted test p-value*
1	rs7633408		0.110	G	146:928, 129:861	0.136, 0.130	0.707	0.692	0.723
2	rs6810298	*ROPN1*	0.846	G	400:716, 355:659	0.358, 0.350	0.688	0.694	0.792
3	rs17376453	*ROPN1*	0.111	G	932:150, 854:150	0.861, 0.851	0.484	0.496	0.546
4	rs2280422	*ROPN1*	0.147	A	517:599, 428:598	0.463, 0.417	**0.036**	**0.042**	0.085
5	rs7434266		0.084	A	230:846, 173:819	0.214, 0.174	**0.024**	**0.026**	**0.028**
6	rs2332719		0.162	G	421:707, 339:691	0.373, 0.329	**0.032**	**0.035**	0.150
7	rs7613868		0.320	T	422:708, 340:690	0.373, 0.330	**0.033**	**0.036**	0.185
8	rs12634530		1.000	T	281:819, 219:797	0.255, 0.216	**0.031**	**0.030**	0.207
9	rs12637456		0.365	A	399:665, 333:671	0.375, 0.332	**0.040**	**0.043**	0.215
10	rs1317671	*AK123068*	0.096	C	384:742, 314:718	0.341, 0.304	0.068	0.075	0.298
11	rs9289231	*AK123068*	0.808	T	1004:106, 902:104	0.905, 0.897	0.545	0.552	0.764
12	rs13075202	*KALRN*	**0.036****	G	376:744, 309:715	0.336, 0.302	0.092	0.101	0.334
13	rs4678085	*KALRN*	0.535	A	272:858, 238:794	0.241, 0.231	0.564	0.571	0.868
14	rs1950091	*KALRN*	0.297	G	952:160, 873:153	0.856, 0.851	0.732	0.731	0.838
15	rs1444760	*KALRN*	0.180	A	436:676, 370:654	0.392, 0.361	0.143	0.156	0.368
16	rs1444768	*KALRN*	0.270	G	491:625, 411:607	0.440, 0.404	0.091	0.096	0.169
17	rs1444766	*KALRN*	0.259	G	328:762, 274:746	0.301, 0.269	0.101	0.105	0.104
18	rs1444754	*KALRN*	0.232	C	474:610, 426:558	0.437, 0.433	0.842	0.847	0.903
19	rs1373609	*KALRN*	0.537	T	599:525, 529:495	0.533, 0.517	0.449	0.449	0.471
20	rs17286604	*KALRN*	0.129	C	774:354, 658:376	0.686, 0.636	**0.014**	**0.014**	**0.049**
21	rs1158012	*KALRN*	0.818	A	542:558, 470:546	0.493, 0.463	0.166	0.167	0.148
22	rs4608634	*KALRN*	0.526	C	447:583, 437:579	0.434, 0.430	0.860	0.857	0.650
23	rs4234218	*KALRN*	0.693	G	564:278, 606:348	0.670, 0.635	0.125	0.117	0.132
24	rs4608635	*KALRN*	0.617	C	380:730, 337:669	0.342, 0.335	0.721	0.725	0.603
25	rs6781700	*KALRN*	0.349	T	1037:87, 944:80	0.923, 0.922	0.950	0.951	0.932
26	rs1444763	*KALRN*	0.743	G	359:771, 288:744	0.318, 0.279	0.050	0.050	0.196
27	rs11929003	*KALRN*	1.000	G	542:586, 473:557	0.480, 0.459	0.323	0.314	0.509
28	rs13064819	*KALRN*	0.620	G	883:163, 824:160	0.844, 0.837	0.677	0.675	0.288
29	rs11712619	*KALRN*	0.127	C	775:343, 646:374	0.693, 0.633	**0.003**	**0.003**	**0.014**
30	rs17377867	*KALRN*	0.366	G	1022:88, 910:112	0.921, 0.890	**0.017**	**0.003**	0.056
31	rs9838361	*KALRN*	0.211	T	424:698, 366:666	0.378, 0.355	0.263	0.246	0.312
32	rs6784664	*KALRN*	0.899	A	534:506, 468:534	0.513, 0.467	**0.036**	**0.036**	0.054
33	rs3821525	*KALRN*	1.000	A	817:291, 746:278	0.737, 0.729	0.644	0.648	0.735
34	rs11708466	*KALRN*	0.754	T	811:319, 722:310	0.718, 0.700	0.355	0.360	0.187

*Multivariate logistic regression (log-additive model) with backward elimination of hypertension, diabetes and ever smoking

**Out of HWE in the control dataset (P<0.05), but it was genotyped consistently in two assays

Table C.4: Results of haplotype association testing with IS risk for *KALRN*. See legend in table C.2.

Gene	Haplotype	Freq.	Case, Control Ratio Counts	Case, Control Frequencies	p-value
ROPN1-KLRN	Block 1: rs6810298, rs17376453				
	AG	0.647	725.8 : 404.2, 673.4 : 360.6	0.642, 0.651	0.665
	GG	0.208	244.7 : 885.3, 205.3 : 828.7	0.217, 0.199	0.302
	GC	0.145	159.5 : 970.5, 155.4 : 878.6	0.141, 0.150	0.548
	Block 2: rs2280422, rs4499545				
	GG	0.552	595.5 : 532.5, 598.5 : 435.5	0.528, 0.579	**0.017**
	AG	0.249	287.1 : 840.9, 251.0 : 783.0	0.255, 0.243	0.528
	AA	0.193	237.6 : 890.4, 180.1 : 853.9	0.211, 0.174	**0.032**
	Block 3: rs2332719, rs7613868, rs12634530, rs12637456, rs1317671, rs9289231, rs13075202				
	ACCTTTA	0,634	684.6 : 435.4, 679.7 : 352.3	0.611, 0.659	**0,023**
	GTTACTG	0,210	252.1 : 867.9, 198.8 : 833.2	0.225, 0.193	0,064
	GTCACGG	0,093	104.5 : 1015.5, 95.9 : 936.1	0.093, 0.093	0,977
	GTTATTA	0,021	29.4 : 1090.6, 16.2 : 1015.8	0.026, 0.016	0,088
	GTCATTA	0,012	12.8 : 1107.2, 14.0 : 1018.0	0.011, 0.014	0,648
	Block 4: rs1950091, rs1444760				
	GG	0,622	684.2 : 439.8, 657.6 : 374.4	0.609, 0.637	0,172
	GA	0,230	277.7 : 846.3, 219.0 : 813.0	0.247, 0.212	0,055
	AA	0,147	162.1 : 961.9, 155.3 : 876.7	0.144, 0.151	0,681
	Block 5: rs1444768, rs1444766				
	AA	0,575	629.1 : 500.9, 615.4 : 418.6	0.557, 0.595	0,071
	GG	0,285	338.5 : 791.5, 278.2 : 755.8	0.300, 0.269	0,116
	GA	0,138	160.0 : 970.0, 139.1 : 894.9	0.142, 0.135	0,636
	Block 6: rs1373609, rs17286604				
	TC	0,525	603.0 : 527.0, 532.2 : 501.8	0.534, 0.515	0,379
	CT	0,338	355.4 : 774.6, 376.0 : 658.0	0.315, 0.364	**0,016**
	CC	0,137	171.6 : 958.4, 125.8 : 908.2	0.152, 0.122	**0,042**
	Block 7: rs4608634, rs4234218, rs4608635				
	GCG	0,343	368.1 : 761.9, 373.7 : 658.3	0.326, 0.362	0,075
	CGC	0,330	374.7 : 755.3, 337.8 : 694.2	0.332, 0.327	0,831
	GGG	0,212	253.0 : 877.0, 205.9 : 826.1	0.224, 0.199	0,166
	CGG	0,106	122.8 : 1007.2, 106.3 : 925.7	0.109, 0.103	0,671
	Block 8: rs6781700, rs1444763				
	TA	0,698	765.3 : 364.7, 743.2 : 288.8	0.677, 0.720	**0,030**
	TG	0,225	277.4 : 852.6, 208.0 : 824.0	0.245, 0.202	**0,014**
	CG	0,075	81.6 : 1048.4, 80.0 : 952.0	0.072, 0.078	0,639
	Block 9: rs13064819, rs11712619, rs17377867, rs9838361, rs6784664				
	GCGTA	0,350	394.7 : 715.7, 351.4 : 670.2	0.355, 0.344	0,579
	GTGGC	0,226	237.4 : 873.0, 245.3 : 776.3	0.214, 0.240	0,147
	CCGGC	0,164	180.0 : 930.5, 170.9 : 850.7	0.162, 0.167	0,745
	GCGGA	0,117	150.1 : 960.3, 99.0 : 922.6	0.135, 0.097	**0,006**
	GTAGC	0,090	85.7 : 1024.7, 106.6 : 915.0	0.077, 0.104	**0,029**
	GTGGA	0,021	23.6 : 1086.8, 20.4 : 1001.2	0.021, 0.020	0,831
	GCGTC	0,015	20.3 : 1090.1, 11.3 : 1010.3	0.018, 0.011	0,167
	GCGGC	0,011	14.3 : 1096.1, 9.2 : 1012.4	0.013, 0.009	0,399
	Block 10: rs3821525, rs11708466				
	AT	0,442	513.6 : 616.4, 441.7 : 590.3	0.455, 0.428	0,216
	AC	0,291	319.0 : 811.0, 310.0 : 722.0	0.282, 0.300	0,355
	GT	0,267	297.4 : 832.6, 280.3 : 751.7	0.263, 0.272	0,660

Table C.5: Detailed association results for the 32 imputed SNPs with p-value < 0.01 in the *KALRN* region. The presented values are from allelic association tests.

rs number	Position in chromosome 3 (bp)	Allele 1:2	NPRX[a]	SNP INFO[b]	Allele 2 frequency in controls	Allele 2 frequency in cases	OR	*P*
rs6790975	125033480	C:T	7	0.598	0.079	0.052	0.63	0.0006
rs1920616	125302273	G:A	10	0.936	0.061	0.093	1.59	0.0032
rs6794106	125303287	G:C	10	0.936	0.061	0.093	1.59	0.0032
rs1920625	125309976	C:T	10	0.933	0.062	0.096	1.61	0.0022
rs9811597	125315387	A:G	10	0.933	0.062	0.096	1.61	0.0022
rs6438833	125408285	A:T	10	0.866	0.084	0.124	1.54	0.0010
rs1373612	125409482	G:C	10	0.866	0.084	0.124	1.54	0.0010
rs11720960	125440632	C:T	10	0.737	0.146	0.110	0.72	0.0033
rs6779809	125447175	A:T	5	0.919	0.366	0.310	0.78	0.0042
rs11712039	125456276	T:C	5	0.919	0.366	0.310	0.78	0.0042
rs11719349	125460712	G:A	5	0.919	0.366	0.310	0.78	0.0042
rs11719623	125461508	C:A	5	0.883	0.382	0.326	0.78	0.0039
rs4289301	125465301	C:T	5	0.878	0.387	0.327	0.77	0.0021
rs9847323	125468374	C:T	5	0.823	0.406	0.348	0.78	0.0021
rs920894	125469906	G:A	5	0.878	0.387	0.327	0.77	0.0021
rs4420813	125472922	G:A	5	0.878	0.387	0.327	0.77	0.0021
rs4678105	125482955	G:A	5	0.885	0.378	0.320	0.77	0.0023
rs11706831	125503942	G:A	5	0.886	0.378	0.319	0.77	0.0023
rs4678111	125527213	T:C	10	0.930	0.136	0.189	1.48	0.0005
rs2289778	125527639	C:T	10	0.907	0.122	0.168	1.46	0.0013
rs7620580	125527993	G:A	10	0.930	0.136	0.189	1.48	0.0005
rs9820396	125528087	A:G	10	0.930	0.136	0.189	1.48	0.0005
rs1444770	125529086	G:A	10	0.930	0.136	0.189	1.48	0.0005
rs6438838	125530099	T:C	10	0.897	0.107	0.150	1.48	0.0015
rs2276740	125531454	T:C	10	0.868	0.106	0.145	1.43	0.0032
rs4371456	125532758	T:A	4	0.809	0.316	0.266	0.79	0.0049
rs1838473	125534520	C:T	4	0.812	0.316	0.267	0.79	0.0050
rs6438839	125536191	A:G	10	0.897	0.107	0.150	1.48	0.0015
rs1025957	125543770	C:T	4	0.812	0.316	0.267	0.79	0.0050
rs17289013	125602730	T:C	4	0.694	0.323	0.279	0.81	0.0085
rs11717452	125603289	C:T	4	0.728	0.312	0.268	0.81	0.0080
rs16835717	125813105	T:A	10	0.622	0.369	0.324	0.82	0.0050

[a]NPRX: number of proxy SNPs used in the imputation.

[b]SNP INFO: information content metric for the imputation.

Table C.6: Validation of SNP association testing with IS risk for three imputed associated SNPs in *KALRN*. See legend in table C.1.

SNP N	SNP ID	Gene	HWE p-value Controls	Assoc Allele	Case, Control Ratio Counts	Case, Control Frequencies	Unadjusted test Allel p-value	Unadjusted test Geno p-value	Adjusted test p-value*
35	rs7620580	*KALRN*	0.267	G	135:697, 112:850	0.162, 0.116	**0.005**	**0.006**	**0.007**
36	rs6438833	*KALRN*	0.199	A	102:738, 75:873	0.121, 0.079	**0.003**	**0.004**	**0.012**
37	rs11712039	*KALRN*	0.377	C	586:262, 607:353	0.691, 0.632	**0.009**	**0.007**	**0.027**

*Multivariate logistic regression (log-additive model) with backward elimination of hypertension, diabetes and ever smoking

Table C.7: Results of SNP association testing with IS risk for *EPO, HO2* and *KLK1*. See legend in table C.1.

SNP N	SNP ID	Gene	HWE p-value Controls	Assoc Allele	Case, Control Ratio Counts	Case, Control Frequencies	Unadjusted test Allel p-value	Unadjusted test Geno p-value	Adjusted test p-value*
1	rs1617640		0.917	C	352:760, 308:722	0.317, 0.299	0.380	0.405	0.858
2	rs564449	*EPO*	0.670	T	71:1047, 58:974	0.064, 0.056	0.476	0.496	0.952
3	rs4729607		0.604	A	181:839, 147:779	0.177, 0.159	0.271	0.276	0.610
1	rs3761680		0.699	A	715:365, 660:360	0.662, 0.647	0.471	0.459	0.227
2	rs2160567	*HO2*	0.924	T	742:378, 663:367	0.662, 0.644	0.360	0.356	0.177
3	rs7702		0.344	C	809:313, 721:307	0.721, 0.701	0.315	0.313	0.321
4	rs7665		0.518	G	815:305, 732:296	0.728, 0.712	0.421	0.421	0.380
1	rs266116		0.109	T	627:495, 556:470	0.559, 0.542	0.431	0.438	0.250
2	rs266117		0.319	A	657:465, 599:429	0.586, 0.583	0.893	0.893	0.628
3	rs2739454		0.467	G	649:461, 589:425	0.585, 0.581	0.859	0.859	0.681
4	rs3212820	*KLK1*	0.276	A	199:913, 173:855	0.179, 0.168	0.515	0.549	0.643
5	rs3212810		0.437	T	204:918, 174:854	0.182, 0.169	0.445	0.450	0.560

*Multivariate logistic regression (log-additive model) with backward elimination of hypertension, diabetes and ever smoking

Table C.8: Results of haplotype association testing with IS risk for *EPO, HO2* and *KLK1*. See legend in table C.2.

Gene	Haplotype	Freq.	Case, Control Ratio Counts	Case, Control Frequencies	p-value
EPO	Block 1: rs1617640, rs564449, rs4729607				
	AGT	0.692	762.5 : 351.5, 721.8 : 308.2	0.684, 0.701	0.414
	CGA	0.177	205.8 : 908.2, 173.3 : 856.7	0.185, 0.168	0.320
	CGT	0.072	75.7 : 1038.3, 77.9 : 952.1	0.068, 0.076	0.494
	CTT	0.059	70.0 : 1044.0, 57.0 : 973.0	0.063, 0.055	0.463
HO2	Block 1: rs3761680, rs2160567, rs7702, rs7665				
	ATCG	0.650	735.8 : 382.2, 660.9 : 371.1	0.658, 0.640	0.389
	CCGA	0.276	299.4 : 818.6, 294.2 : 737.8	0.268, 0.285	0.371
	CCCG	0.059	65.3 : 1052.7, 60.8 : 971.2	0.058, 0.059	0.963
KLK1	Block 1: rs266117, rs2739454, rs3212820, rs3212810				
	TACC	0.416	470.0 : 658.0, 430.0 : 604.0	0.417, 0.416	0.970
	AGCC	0.405	447.9 : 680.1, 426.9 : 607.1	0.397, 0.413	0.454
	AGAT	0.171	197.8 : 930.2, 171.9 : 862.1	0.175, 0.166	0.574

Table C.9: Results of SNP association testing with IS risk for GC genes. See legend in table C.1.

SNP N	SNP ID	Gene	HWE p-value Controls	Assoc Allele	Case, Control Ratio Counts	Case, Control Frequencies	Unadjusted test Allel p-value	Unadjusted test Geno p-value	Adjusted test p-value*
1	rs10760017		0.067	G	780:302, 698:326	0.721, 0.682	**0.049**	0.051	0.272
2	rs4743146		1.000	A	980:86, 922:86	0.919, 0.915	0.702	0.705	0.888
3	rs1059003	*HEMGN*	1.000	G	139:971, 127:895	0.125, 0.124	0.947	0.895	0.764
4	rs10984462		1.000	G	291:723, 262:712	0.287, 0.269	0.371	0.366	0.647
1	rs1964196	*GFI1B*	0.724	A	613:499, 524:500	0.551, 0.512	0.067	0.065	0.055
2	rs686652	*GFI1B*	0.866	T	176:874, 157:837	0.168, 0.158	0.554	0.563	0.406
3	rs8192999	*GFI1B*	0.356	G	765:343, 706:322	0.690, 0.687	0.855	0.880	0.734
4	rs3827664	*GFI1B*	0.926	G	606:382, 582:404	0.613, 0.590	0.295	0.307	0.749
5	rs2905069	*GFI1B*	0.757	A	811:305, 712:318	0.727, 0.691	0.071	0.068	0.148
6	rs649651	*GFI1B*	0.420	A	141:953, 128:886	0.129, 0.126	0.855	0.844	0.764
7	rs2073577	*GFI1B*	0.114	T	638:384, 575:407	0.624, 0.586	0.076	0.089	0.251
8	rs606141	*GFI1B*	0.566	G	909:199, 816:194	0.820, 0.808	0.461	0.457	0.551
9	rs633153	*GFI1B*	0.710	G	472:596, 400:602	0.442, 0.399	**0.049**	0.056	0.159
10	rs8193002	*GFI1B*	0.615	A	282:804, 230:778	0.260, 0.228	0.094	0.109	0.316
11	rs8193004	*GFI1B*	0.187	T	677:395, 634:384	0.632, 0.623	0.680	0.675	0.712
12	rs667805	*GFI1B*	0.489	G	122:992, 109:923	0.110, 0.106	0.771	0.783	0.802
13	rs15906	*GFI1B*	0.833	A	132:976, 123:911	0.119, 0.119	0.990	0.998	0.817
14	rs10751499		0.545	G	158:944, 127:893	0.143, 0.125	0.203	0.200	0.169
15	rs678946		0.203	C	938:184, 836:200	0.836, 0.807	0.078	0.080	0.514
16	rs1888204		0.644	C	200:902, 178:850	0.181, 0.173	0.615	0.632	1.000
17	rs3011266		0.538	T	293:827, 239:789	0.262, 0.232	0.118	0.118	0.591
18	rs8193007		0.790	T	112:1004, 95:937	0.100, 0.092	0.515	0.530	0.810
1	rs8002073		0.225	T	249:867, 211:817	0.223, 0.205	0.314	0.328	0.238
2	rs9513770	*TMTC4*	0.652	T	822:280, 750:274	0.746, 0.732	0.479	0.497	0.242
3	rs9518104	*TMTC4*	0.780	G	453:657, 392:638	0.408, 0.381	0.193	0.178	0.428
4	rs9582406	*TMTC4*	0.081	G	808:290, 714:318	0.736, 0.692	**0.025**	**0.022**	**0.050**
5	rs7995648	*TMTC4*	0.086	T	662:426, 581:439	0.608, 0.570	0.070	0.093	0.099
6	rs17578868	*TMTC4*	0.099	A	935:155, 850:162	0.858, 0.840	0.253	0.251	0.233
7	rs1765733	*TMTC4*	0.147	G	274:814, 246:766	0.252, 0.243	0.642	0.647	0.987
8	rs9513779	*TMTC4*	0.102	T	366:656, 324:652	0.358, 0.332	0.219	0.219	0.188
9	rs9518128	*TMTC4*	0.077	T	568:534, 494:534	0.515, 0.481	0.108	0.112	0.066
10	rs9513782	*TMTC4*	0.199	A	236:870, 195:827	0.213, 0.191	0.195	0.204	0.326
11	rs946845	*TMTC4*	0.712	A	102:1016, 65:965	0.091, 0.063	**0.015**	**0.013**	**0.047**
1	rs1294555		0.669	G	783:317, 727:297	0.712, 0.710	0.925	0.869	0.780
2	rs1294559		0.574	C	311:799, 273:759	0.280, 0.265	0.417	0.470	0.938
3	rs1294582	*TTC7B*	0.373	C	490:608, 459:575	0.446, 0.444	0.913	0.911	0.510
4	rs12889650	*TTC7B*	0.331	T	518:600, 478:554	0.463, 0.463	0.995	0.970	0.319
5	rs17793829	*TTC7B*	0.064	C	885:205, 798:216	0.812, 0.787	0.153	0.158	0.108
6	rs11620738	*TTC7B*	0.640	G	663:413, 619:391	0.616, 0.613	0.877	0.877	0.665
7	rs1286496	*TTC7B*	0.294	G	722:390, 658:370	0.649, 0.640	0.657	0.668	0.145
8	rs1286477	*TTC7B*	0.858	C	970:148, 882:148	0.868, 0.856	0.447	0.450	0.253
9	rs1286463	*TTC7B*	0.740	C	92:1006, 74:952	0.084, 0.072	0.317	0.324	0.138
10	rs12432719	*TTC7B*	0.877	A	194:868, 176:826	0.183, 0.176	0.678	0.663	0.907
11	rs1286459	*TTC7B*	1.000	C	693:393, 638:380	0.638, 0.627	0.588	0.632	0.745
12	rs7145137	*TTC7B*	1.000	T	166:910, 144:872	0.154, 0.142	0.420	0.373	0.144
13	rs4904708	*TTC7B*	0.652	G	919:185, 844:184	0.832, 0.821	0.486	0.485	0.271
14	rs8020106	*TTC7B*	0.907	T	867:251, 770:262	0.775, 0.746	0.111	0.116	0.169
15	rs2343	*TTC7B*	0.095	C	714:386, 612:398	0.649, 0.606	**0.041**	**0.039**	0.105
16	rs942746	*TTC7B*	0.405	T	679:407, 622:400	0.625, 0.609	0.433	0.391	0.566
17	rs942748	*TTC7B*	0.713	G	467:643, 412:606	0.421, 0.405	0.454	0.453	0.786
18	rs8022840	*TTC7B*	0.873	C	195:873, 168:828	0.183, 0.169	0.407	0.413	0.703
19	rs2277510	*TTC7B*	0.472	A	186:890, 147:867	0.173, 0.145	0.082	0.084	0.148
20	rs10132961	*TTC7B*	0.772	C	718:374, 662:362	0.658, 0.646	0.595	0.595	0.706

21	rs12881399	*TTC7B*	0.857	C	502:612, 438:594	0.451, 0.424	0.221	0.225	0.831
22	rs17799388	*TTC7B*	0.422	G	956:158, 864:168	0.858, 0.837	0.177	0.182	0.476
23	rs12886545	*TTC7B*	0.462	T	673:411, 606:412	0.621, 0.595	0.230	0.232	0.656
24	rs881218	*TTC7B*	1.000	C	391:663, 358:642	0.371, 0.358	0.542	0.550	0.672
25	rs1076958	*TTC7B*	0.085	G	807:307, 741:295	0.724, 0.715	0.636	0.644	0.196
26	rs4900055	*TTC7B*	0.400	G	905:215, 828:202	0.808, 0.804	0.808	0.830	0.585
27	rs4900057	*TTC7B*	1.000	A	622:474, 576:450	0.568, 0.561	0.777	0.778	0.380
28	rs17799418	*TTC7B*	0.599	T	966:158, 883:153	0.859, 0.852	0.638	0.652	0.626
29	rs7152676	*TTC7B*	0.299	G	653:403, 622:384	0.618, 0.618	0.997	0.971	0.845
30	rs12147413	*TTC7B*	0.321	C	193:895, 158:824	0.177, 0.161	0.318	0.395	0.790
31	rs11629065	*TTC7B*	0.321	A	173:915, 131:885	0.159, 0.129	**0.050**	0.053	**0.026**
32	rs942738	*TTC7B*	0.583	A	664:442, 585:423	0.600, 0.580	0.350	0.357	0.935
33	rs12893100	*TTC7B*	0.803	T	267:751, 233:747	0.262, 0.238	0.206	0.217	0.531
34	rs1742100	*TTC7B*	0.645	A	671:411, 616:406	0.620, 0.603	0.413	0.452	0.350
35	rs1742098	*TTC7B*	0.648	C	469:617, 422:600	0.432, 0.413	0.379	0.406	0.785
36	rs1535321	*TTC7B*	0.369	G	244:844, 182:836	0.224, 0.179	**0.009**	**0.009**	0.092
37	rs13379124	*TTC7B*	1.000	C	111:1003, 99:933	0.100, 0.096	0.773	0.774	0.974
38	rs7154098	*TTC7B*	0.065	C	969:139, 887:141	0.875, 0.863	0.423	0.380	0.832
39	rs17721032	*TTC7B*	0.716	G	929:131, 863:143	0.876, 0.858	0.214	0.215	0.166
40	rs12883490	*TTC7B*	1.000	T	818:286, 740:286	0.741, 0.721	0.306	0.303	0.829
41	rs9944033	*TTC7B*	0.781	T	671:421, 610:402	0.614, 0.603	0.583	0.594	0.302
42	rs12894343	*TTC7B*	0.183	A	660:400, 621:379	0.623, 0.621	0.939	0.942	0.536
43	rs1286322	*TTC7B*	0.892	C	230:886, 208:826	0.206, 0.201	0.777	0.776	0.824
44	rs1286305	*TTC7B*	0.723	A	553:555, 480:546	0.499, 0.468	0.149	0.146	0.350
45	rs8016414		0.644	G	756:238, 713:259	0.761, 0.734	0.168	0.153	0.598
46	rs2180774		0.928	C	504:614, 444:582	0.451, 0.433	0.400	0.377	0.818
1	rs6104115		0.759	G	934:176, 854:178	0.841, 0.828	0.386	0.370	0.423
2	rs6073708		0.261	G	446:656, 391:623	0.405, 0.386	0.369	0.357	0.265
3	rs4599	*SDC4*	1.000	A	865:203, 789:217	0.810, 0.784	0.147	0.154	0.363
4	rs6073714	*SDC4*	0.357	G	798:272, 747:265	0.746, 0.738	0.690	0.700	0.226
5	rs2267867	*SDC4*	0.895	A	896:208, 799:215	0.812, 0.788	0.174	0.195	0.226
6	rs2251252	*SDC4*	0.535	G	643:469, 565:467	0.578, 0.547	0.151	0.153	**0.037**
7	rs2267871	*SDC4*	0.147	A	900:208, 811:219	0.812, 0.787	0.150	0.155	0.102
8	rs2284278	*SDC4*	0.432	A	853:247, 751:277	0.775, 0.731	**0.016**	**0.020**	**0.015**
9	rs1981430	*SDC4*	0.093	A	566:510, 512:510	0.526, 0.501	0.251	0.283	0.105
10	rs2072786	*SDC4*	0.068	G	618:370, 608:370	0.626, 0.622	0.861	0.874	0.540
11	rs1008953		0.203	T	252:838, 228:806	0.231, 0.221	0.556	0.581	0.331
1	rs151348		0.479	T	564:556, 486:544	0.504, 0.472	0.142	0.140	**0.021**
2	rs6070696	*TUBB1*	0.362	A	913:181, 843:181	0.835, 0.823	0.490	0.497	0.458
3	rs10485828	*TUBB1*	0.644	C	209:907, 178:850	0.187, 0.173	0.396	0.404	0.397
4	rs151337		0.538	C	786:330, 704:322	0.704, 0.686	0.362	0.359	0.103

*Multivariate logistic regression (log-additive model) with backward elimination of hypertension, diabetes and ever smoking

Table C.10: Results of haplotype association testing with IS risk for GC genes. See legend in table C.2.

Gene	Haplotype	Freq.	Case, Control Ratio Counts	Case, Control Frequencies	p-value
HEMGN	Block 1: rs10760017, rs4743146				
	GA	0.619	717.8 : 404.2, 619.7 : 418.3	0.640, 0.597	**0.041**
	CA	0.298	313.0 : 809.0, 330.5 : 707.5	0.279, 0.318	**0.045**
	GT	0.083	91.2 : 1030.8, 87.8 : 950.2	0.081, 0.085	0.782
	Block 2: rs1059003, rs10984462				
	AT	0.598	662.7 : 461.3, 629.8 : 406.2	0.590, 0.608	0.386
	AG	0.277	320.7 : 803.3, 277.5 : 758.5	0.285, 0.268	0.364
	GT	0.125	140.6 : 983.4, 128.7 : 907.3	0.125, 0.124	0.955
GFI1B	Block 1: rs686652, rs8192999				
	CG	0.526	587.7 : 534.3, 548.0 : 488.0	0.524, 0.529	0.812
	CC	0.311	346.9 : 775.1, 324.2 : 711.8	0.309, 0.313	0.853
	TG	0.162	185.7 : 936.3, 163.3 : 872.7	0.166, 0.158	0.620
	Block 2: rs649651, rs2073577, rs606141, rs633153, rs8193002				
	GGGTG	0.382	407.4 : 694.6, 400.6 : 615.4	0.370, 0.394	0.243
	GTGCA	0.240	278.8 : 823.2, 228.6 : 787.4	0.253, 0.225	0.131
	GTGCG	0.171	197.5 : 904.5, 163.6 : 852.4	0.179, 0.161	0.265
	ATATG	0.126	140.6 : 961.4, 126.1 : 889.9	0.128, 0.124	0.810
	GTATG	0.058	54.4 : 1047.6, 68.9 : 947.1	0.049, 0.068	0.070
	Block 3: rs8193004, rs667805				
	TA	0.625	704.3 : 417.7, 644.5 : 391.5	0.628, 0.622	0.787
	CA	0.267	295.0 : 827.0, 281.7 : 754.3	0.263, 0.272	0.638
	CG	0.106	120.0 : 1002.0, 108.3 : 927.7	0.107, 0.105	0.854
TMTC4	Block 1: rs8002073, rs9513770				
	GT	0.526	590.4 : 535.6, 546.2 : 489.8	0.524, 0.527	0.892
	GC	0.259	283.7 : 842.3, 277.0 : 759.0	0.252, 0.267	0.413
	TT	0.214	250.3 : 875.7, 212.1 : 823.9	0.222, 0.205	0.321
	Block 2: rs9582406, rs7995648				
	GT	0.590	683.5 : 440.5, 589.1 : 444.9	0.608, 0.570	0.070
	CA	0.286	299.1 : 824.9, 317.4 : 716.6	0.266, 0.307	**0.036**
	GA	0.125	141.5 : 982.5, 127.5 : 906.5	0.126, 0.123	0.858
	Block 3: rs17578868, rs1765733, rs9513779, rs9518128, rs9513782				
	AACCC	0.260	265.2 : 844.8, 291.0 : 741.0	0.239, 0.282	**0.023**
	AATTA	0.197	229.2 : 880.8, 192.6 : 839.4	0.206, 0.187	0.248
	AACTC	0.153	170.2 : 939.8, 157.0 : 875.0	0.153, 0.152	0.942
	GGCCC	0.152	161.6 : 948.4, 165.0 : 867.0	0.146, 0.160	0.359
	AATTC	0.143	164.7 : 945.3, 141.7 : 890.3	0.148, 0.137	0.466
	AGCCC	0.086	106.6 : 1003.4, 76.7 : 955.3	0.096, 0.074	0.073
TTC7B	Block 1: rs1294582, rs12889650, rs17793829, rs11620738, rs1286496				
	ATCTG	0.381	419.3 : 686.7, 395.6 : 636.4	0.379, 0.383	0.839
	CCTGA	0.197	202.3 : 903.7, 218.2 : 813.8	0.183, 0.211	0.097
	CCCGA	0.156	180.7 : 925.3, 152.1 : 879.9	0.163, 0.147	0.307
	ACCGG	0.092	104.4 : 1001.6, 91.8 : 940.2	0.094, 0.089	0.662
	CCCGG	0.088	102.6 : 1003.4, 85.6 : 946.4	0.093, 0.083	0.424
	ATCGG	0.079	88.5 : 1017.5, 80.7 : 951.3	0.080, 0.078	0.875
	Block 2: rs1286459, rs7145137				
	CC	0.628	696.1 : 401.9, 638.5 : 387.5	0.634, 0.622	0.581
	TC	0.223	232.7 : 865.3, 240.7 : 785.3	0.212, 0.235	0.210
	TT	0.145	164.2 : 933.8, 142.8 : 883.2	0.150, 0.139	0.500
	Block 3: rs8020106, rs2343				
	TC	0.625	725.1 : 400.9, 627.6 : 410.4	0.644, 0.605	0.059
	CT	0.239	251.3 : 874.7, 265.8 : 772.2	0.223, 0.256	0.073
	TT	0.136	149.6 : 976.4, 144.6 : 893.4	0.133, 0.139	0.663

	Block 4: rs8022840, rs2277510, rs10132961, rs12881399, rs17799388				
	TGTGG	0.346	381.4 : 734.6, 361.5 : 672.5	0.342, 0.350	0.705
	CGCCG	0.175	197.0 : 919.0, 178.5 : 855.5	0.176, 0.173	0.815
	TACCG	0.164	198.2 : 917.8, 154.0 : 880.0	0.178, 0.149	0.073
	TGCGA	0.152	157.5 : 958.5, 168.4 : 865.6	0.141, 0.163	0.160
	TGCCG	0.099	105.8 : 1010.2, 106.8 : 927.2	0.095, 0.103	0.508
	TGCGG	0.063	72.9 : 1043.1, 63.6 : 970.4	0.065, 0.061	0.718
	Block 5: rs881218, rs1076958, rs4900055, rs4900057				
	CGGA	0.363	416.2 : 707.8, 368.6 : 667.4	0.370, 0.356	0.486
	TAGG	0.278	308.1 : 815.9, 291.6 : 744.4	0.274, 0.281	0.704
	TGAA	0.193	213.4 : 910.6, 203.8 : 832.2	0.190, 0.197	0.684
	TGGG	0.155	174.1 : 949.9, 160.2 : 875.8	0.155, 0.155	0.991
	Block 6: rs17799418, rs7152676				
	TG	0.474	538.0 : 590.0, 488.5 : 549.5	0.477, 0.471	0.769
	TA	0.382	431.1 : 696.9, 396.0 : 642.0	0.382, 0.382	0.975
	CG	0.144	158.9 : 969.1, 153.5 : 884.5	0.141, 0.148	0.646
	Block 7: rs12147413, rs11629065, rs942738				
	AGA	0.421	476.7 : 651.3, 435.4 : 602.6	0.423, 0.419	0.882
	AGG	0.262	270.1 : 857.9, 296.5 : 741.5	0.239, 0.286	**0.015**
	CGA	0.170	200.2 : 927.8, 168.3 : 869.7	0.178, 0.162	0.340
	AAG	0.146	180.3 : 947.7, 135.3 : 902.7	0.160, 0.130	0.052
	Block 8: rs12893100, rs1742100, rs1742098, rs1535321, rs13379124, rs7154098				
	CGTATC	0.383	417.2 : 696.8, 405.8 : 626.2	0.374, 0.393	0.373
	TACGTC	0.200	246.4 : 867.6, 183.4 : 848.6	0.221, 0.178	**0.012**
	CATATG	0.129	138.1 : 975.9, 139.5 : 892.5	0.124, 0.135	0.442
	CACACC	0.097	109.7 : 1004.3, 99.5 : 932.5	0.098, 0.096	0.873
	CATATC	0.063	73.7 : 1040.3, 61.1 : 970.9	0.066, 0.059	0.506
	CACATC	0.063	63.4 : 1050.6, 71.1 : 960.9	0.057, 0.069	0.253
	TACATC	0.054	53.3 : 1060.7, 63.7 : 968.3	0.048, 0.062	0.158
	Block 9: rs17721032, rs12883490				
	GT	0.598	693.3 : 432.7, 601.8 : 436.2	0.616, 0.580	0.089
	GC	0.268	291.9 : 834.1, 288.8 : 749.2	0.259, 0.278	0.319
	AT	0.133	140.8 : 985.2, 147.4 : 890.6	0.125, 0.142	0.248
	Block 10: rs12894343, rs1286322				
	AT	0.621	698.8 : 427.2, 643.3 : 392.7	0.621, 0.621	0.986
	GC	0.202	229.3 : 896.7, 208.2 : 827.8	0.204, 0.201	0.877
	GT	0.175	194.5 : 931.5, 184.1 : 851.9	0.173, 0.178	0.760
	Block 11: rs1286305, rs8016414				
	AG	0.481	550.4 : 563.6, 479.9 : 550.1	0.494, 0.466	0.192
	GG	0.263	291.9 : 822.1, 272.3 : 757.7	0.262, 0.264	0.901
	GT	0.253	266.5 : 847.5, 275.7 : 754.3	0.239, 0.268	0.130
SDC4	Block 1: rs6104115, rs6073708, rs4599				
	GAA	0.603	664.6 : 455.4, 636.1 : 399.9	0.593, 0.614	0.328
	GGA	0.191	236.7 : 883.3, 174.3 : 861.7	0.211, 0.168	**0.011**
	AGG	0.161	171.3 : 948.7, 176.5 : 859.5	0.153, 0.170	0.271
	GGG	0.040	40.1 : 1079.9, 46.8 : 989.2	0.036, 0.045	0.270
	Block 2: rs6073714, rs2267867				
	GA	0.542	624.4 : 495.6, 544.0 : 492.0	0.558, 0.525	0.131
	TA	0.258	284.5 : 835.5, 272.4 : 763.6	0.254, 0.263	0.637
	GG	0.200	211.0 : 909.0, 219.6 : 816.4	0.188, 0.212	0.172
	Block 3: rs1981430, rs2072786				
	AG	0.511	572.4 : 523.6, 514.0 : 514.0	0.522, 0.500	0.305
	CC	0.377	407.1 : 688.9, 394.4 : 633.6	0.371, 0.384	0.561
	CG	0.109	113.0 : 983.0, 118.3 : 909.7	0.103, 0.115	0.375
TUBB1	Block 1: rs6070696, rs10485828, rs151337				
	AGC	0.346	398.4 : 727.6, 350.0 : 688.0	0.354, 0.337	0.418
	AGT	0.304	332.7 : 793.3, 325.5 : 712.5	0.296, 0.314	0.360
	ACC	0.180	209.7 : 916.3, 179.0 : 859.0	0.186, 0.172	0.404
	GGC	0.170	185.2 : 940.8, 183.4 : 854.6	0.164, 0.177	0.449

Table C.11: Results of SNP association testing with IS risk for GC genes in the Spanish population. Significant p-values results are highlighted in bold. Multivariate logistic regression (log-additive model) was performed to adjust the analyses of association with risk for hypertension, diabetes, dyslipidemic status and cigarette smoking.

SNP N	SNP ID	Gene	HWE p-value Controls	Assoc Allele	Case, Control Ratio Counts	Case, Control Frequencies	Unadjusted test Allel p-value	Unadjusted test Geno p-value	Adjusted test p-value
1	rs10760017	*HEMGN*	0.209	C	342:756, 217:543	0.311, 0.286	0.231	0.231	0.194
9	rs633153	*GFI1B*	0.351	G	491:649, 330:450	0.431, 0.423	0.740	0.731	0.282
4	rs9582406	*TMTC4*	0.890	C	328:802, 187:587	0.290, 0.242	**0.019**	**0.019**	0.076
11	rs946845		1.000	A	1044:86, 717:63	0.924, 0.919	0.709	0.317	0.608
15	rs2343	*TTC7B*	0.584	T	442:696, 282:490	0.388, 0.365	0.307	0.322	0.371
30	rs12147413		0.879	A	844:200, 577:161	0.808, 0.782	0.169	0.171	0.110
31	rs11629065		0.375	G	974:138, 652:102	0.876, 0.865	0.479	0.483	0.325
32	rs942738		0.200	A	627:443, 429:307	0.586, 0.583	0.896	0.897	0.638
33	rs12893100		0.886	T	271:797, 176:562	0.254, 0.238	0.460	0.464	0.116
34	rs1742100		0.307	A	616:508, 420:354	0.548, 0.543	0.816	0.818	0.468
35	rs1742098		0.827	C	459:679, 283:491	0.403, 0.366	0.097	0.102	**0.031**
37	rs13379124		1.000	T	1066:66, 729:49	0.942, 0.937	0.673	0.669	0.370
38	rs7154098		0.068	C	1024:112, 677:101	0.901, 0.870	**0.033**	**0.038**	0.095
1	rs6104115	*SDC4*	0.637	G	921:217, 622:156	0.809, 0.799	0.594	0.599	0.420
2	rs6073708		0.538	A	693:445, 436:344	0.609, 0.559	**0.029**	**0.027**	**0.028**
3	rs4599		0.416	A	881:255, 583:193	0.776, 0.751	0.219	0.225	0.199
6	rs2251252		0.102	G	650:480, 414:360	0.575, 0.535	0.082	0.090	0.117
8	rs2284278		0.299	A	842:288, 565:207	0.745, 0.732	0.517	0.524	0.738

Table C.12: Positive results of SNP association testing with atherothrombotic, cardioembolic and lacunar stroke risk for GC genes in the Spanish population. See legend in table C.11.

SNP N	SNP ID	Gene	HWE p-value Controls	Assoc Allele	Case, Control Ratio Counts	Case, Control Frequencies	Unadjusted test Allel p-value	Unadjusted test Geno p-value	Adjusted test p-value
Atherothrombotic									
4	rs9582406	*TMTC4*	0.890	C	103:237, 187:587	0.303, 0.242	**0.032**	**0.036**	0.086
15	rs2343	*TTC7B*	0.584	A	148:194, 282:490	0.433, 0.365	**0.033**	**0.041**	0.057
6	rs2251252	*SDC4*	0.102	G	208:132, 414:360	0.612, 0.535	**0.017**	**0.022**	**0.025**
Cardioembolic									
33	rs12893100	*TTC7B*	0.886	T	114:306, 176:562	0.271, 0.238	0.214	0.216	**0.028**
38	rs7154098	*TTC7B*	0.068	C	386:44, 677:101	0.898, 0.870	0.159	**0.029**	0.113
2	rs6073708	*SDC4*	0.538	A	275:157, 436:344	0.637, 0.559	**0.009**	**0.008**	**0.016**
Lacunar									
4	rs9582406	*TMTC4*	0.890	C	108:236, 187:587	0.314, 0.242	**0.011**	**0.012**	**0.049**

Table C.13: Positive results of haplotype association testing with IS risk for GC genes in the Spanish population. See legend in table C.2.

Gene	Haplotype	Freq.	Case, Control Ratio Counts	Case, Control Frequencies	p-value
TTC7B	Block 8*: rs12893100, rs1742100, 1742098, 13379124, rs7154098				
	CGTTC	0.453	512.5 : 625.5, 353.7 : 422.3	0.450, 0.456	0.814
	TACTC	0.254	297.5 : 840.5, 188.1 : 587.9	0.261, 0.242	0.346
	CATTG	0.112	112.7 : 1025.3, 101.0 : 675.0	0.099, 0.130	**0.034**
	CACTC	0.073	93.8 : 1044.2, 46.8 : 729.2	0.082, 0.060	0.068
	CACCC	0.061	67.2 : 1070.8, 49.0 : 727.0	0.059, 0.063	0.715
	CATTC	0.048	54.2 : 1083.8, 37.4 : 738.6	0.048, 0.048	0.951
SDC4	Block 1: rs6104115, rs6073708, rs4599				
	GAA	0.588	690.8 : 445.2, 436.0 : 344.0	0.608, 0.559	**0.032**
	AGG	0.193	215.8 : 920.2, 154.9 : 625.1	0.190, 0.199	0.638
	GGA	0.176	189.2 : 946.8, 147.8 : 632.2	0.167, 0.190	0.195
	GGG	0.041	40.2 : 1095.8, 39.2 : 740.8	0.035, 0.050	0.108

Table C.14: Results of SNP association testing with IS risk for TTC7B gene in the joint analysis of the Portuguese and Spanish datasets. Significant p-values results are highlighted in bold. Multivariate logistic regression (recessive model) was performed to adjust the analyses of association with risk for hypertension, diabetes, ever smoking, and sample origin.

SNP N	SNP ID	Gene	HWE p-value Controls	Assoc Allele	Case, Control Ratio Counts	Case, Control Frequencies	Unadjusted test Allel p-value	Unadjusted test Geno p-value	Adjusted test p-value
15	rs2343	*TTC7B*	0.103	C	1408:828, 1102:678	0.630, 0.619	0.491	0.506	0.705
30	rs12147413		0.368	A	1739:393, 1401:319	0.816, 0.815	0.928	0.967	0.782
31	rs11629065		0.184	A	311:1889, 233:1535	0.141, 0.132	0.383	0.858	0.962
32	rs942738		0.210	A	1291:885, 1014:730	0.593, 0.581	0.453	0.350	0.894
33	rs12893100		0.707	T	538:1548, 409:1309	0.258, 0.238	0.159	**0.013**	**0.003**
34	rs1742100		0.682	A	1287:919, 1036:758	0.583, 0.577	0.706	0.953	0.655
35	rs1742098		0.889	C	926:1296, 705:1091	0.417, 0.393	0.120	**0.020**	**0.007**
37	rs13379124		0.825	T	2067:177, 1662:148	0.921, 0.918	0.736	0.389	0.153

Table C.15: Positive results of haplotype association testing with IS risk for for TTC7B gene in the joint analysis of the Portuguese and Spanish datasets. See legend in table C.2.

Gene	Haplotype	Freq.	Case, Control Ratio Counts	Case, Control Frequencies	p-value
TTC7B	Block 8**: rs1742100, rs1742098, rs13379124				
	GTT	0.418	939.8 : 1322.2, 763.9 : 1052.1	0.415, 0.421	0.738
	ACT	0.325	763.7 : 1498.3, 559.8 : 1256.2	0.338, 0.308	**0.046**
	ATT	0.176	377.9 : 1884.1, 340.1 : 1475.9	0.167, 0.187	0.092
	ACC	0.08	176.8 : 2085.2, 148.6 : 1667.4	0.078, 0.082	0.666

APPENDIX D – DIFFERENTIALLY EXPRESSED GENES AMONG CASES AND CONTROLS

To identify genes whose expression pattern suggests their involvement in the pathogenesis of IS we compare and contrast the genetic profiles in PBMCs from affected individuals and controls. We found 709 probe sets differentially expressed among IS cases and controls (331 probe sets were down-regulated in cases vs. controls) with a threshold of 1.2 fold-change and an uncorrected p-value < 0.05 (Table D.1). In the Tables D.2 and D.3 are presented the significant differentially expressed groups of genes according their molecular and cellular functions obtained using different software, and in the Table D.4 are presented the significant differentially expressed groups of genes associated with genetic disorders, injuries or malformations. The canonical pathways they significantly affect, according to the IPA software, are presented in the Table D.5.

Table D.1: Differentially expressed genes among cases and controls. Probe sets differentially expressed among IS cases and controls with a threshold of 1.2 fold-change and an uncorrected p-value <0.05, performing the ANOVA of the normalized expression results including as factors the type, sex, age and the combinations among them (type*age, type*age*sex, type*sex), the geographic origin of the participants and the scan-date of the microarrays. Results were obtained using the Partek software.

Probeset ID	Gene Symbol	Gene Title	p-value	Fold-change (cases vs. controls)
230760_at	ZFY	Zinc finger protein, Y-linked	$2.58x10^{-5}$	2.19
209170_s_at	GPM6B	glycoprotein M6B	$2.92x10^{-5}$	-1.54
223128_at	FOXRED1	FAD-dependent oxidoreductase domain containing 1	$3.33x10^{-5}$	-1.25
205001_s_at	DDX3Y	DEAD (Asp-Glu-Ala-Asp) box polypeptide 3, Y-linked	$3.43x10^{-5}$	2.16
205000_at	DDX3Y	DEAD (Asp-Glu-Ala-Asp) box polypeptide 3, Y-linked	$5.96x10^{-5}$	3.12
1567009_at	---	Full length insert cDNA clone YU51G05	$9.80x10^{-5}$	1.23
209167_at	GPM6B	glycoprotein M6B	$1.01x10^{-4}$	-1.40
1559893_at	CCDC75	coiled-coil domain containing 75	$1.04x10^{-4}$	-1.29
241133_at	PRSS1	Protease, serine, 1 (trypsin 1)	$1.77x10^{-4}$	3.31
202820_at	AHR	aryl hydrocarbon receptor	$1.97x10^{-4}$	1.46
235139_at	GNGT2	guanine nucleotide binding protein (G protein), gamma transducing activity polypeptide 2	$2.03x10^{-4}$	-1.36
208965_s_at	IFI16	interferon, gamma-inducible protein 16	$2.77x10^{-4}$	-1.26
230787_at	---	---	$3.18x10^{-4}$	1.36
214983_at	TTTY15	testis-specific transcript, Y-linked 15	$3.30x10^{-4}$	1.52
212113_at	LOC552889	hypothetical LOC552889	$3.43x10^{-4}$	-1.34
209349_at	RAD50	RAD50 homolog (S. cerevisiae)	$3.54x10^{-4}$	1.70
208924_at	RNF11	ring finger protein 11	$4.54x10^{-4}$	1.42
213180_s_at	GOSR2	golgi SNAP receptor complex member 2	$4.65x10^{-4}$	-1.21
221286_s_at	PACAP	proapoptotic caspase adaptor protein	$5.06x10^{-4}$	-1.48
207652_s_at	CMKLR1	chemokine-like receptor 1	$5.25x10^{-4}$	-1.34
244482_at	EIF1AY	Eukaryotic translation initiation factor 1A, Y-linked	$5.67x10^{-4}$	1.56

229307_at	ANKRD28	ankyrin repeat domain 28	$5.88x10^{-4}$	1.58
201909_at	RPS4Y1	ribosomal protein S4, Y-linked 1	$6.19x10^{-4}$	3.09
210322_x_at	UTY	ubiquitously transcribed tetratricopeptide repeat gene, Y-linked	$6.21x10^{-4}$	1.30
1569898_a_at	---	CDNA clone IMAGE:5259766	$7.02x10^{-4}$	-1.21
205463_s_at	PDGFA	platelet-derived growth factor alpha polypeptide	$7.11x10^{-4}$	1.44
204482_at	CLDN5	claudin 5 (transmembrane protein deleted in velocardiofacial syndrome)	$8.01x10^{-4}$	1.46
235643_at	SAMD9L	sterile alpha motif domain containing 9-like	$8.01x10^{-4}$	-1.38
203966_s_at	PPM1A	protein phosphatase 1A (formerly 2C), magnesium-dependent, alpha isoform	$8.18x10^{-4}$	1.25
218327_s_at	SNAP29	synaptosomal-associated protein, 29kDa	$8.43x10^{-4}$	-1.25
204555_s_at	PPP1R3D	protein phosphatase 1, regulatory subunit 3D	$8.69x10^{-4}$	-1.21
201905_s_at	CTDSPL	CTD (carboxy-terminal domain, RNA polymerase II, polypeptide A) small phosphatase-like	$9.00x10^{-4}$	1.39
226273_at	LOC158563	hypothetical protein LOC158563	$9.20x10^{-4}$	-1.24
201980_s_at	RSU1	Ras suppressor protein 1	$9.40x10^{-4}$	1.29
235173_at	LOC401093	hypothetical LOC401093	$9.80x10^{-4}$	1.33
205390_s_at	ANK1	ankyrin 1, erythrocytic / ankyrin 1, erythrocytic	$1.00x10^{-3}$	1.31
205088_at	CXorf6	chromosome X open reading frame 6	$1.06x10^{-3}$	1.27
225666_at	TMTC4	transmembrane and tetratricopeptide repeat containing 4	$1.07x10^{-3}$	-1.23
1558647_at	SH3D19	SH3 domain protein D19	$1.08x10^{-3}$	1.20
214108_at	MAX	MYC associated factor X	$1.10x10^{-3}$	1.31
336_at	TBXA2R	thromboxane A2 receptor	$1.11x10^{-3}$	1.40
209933_s_at	CD300A	CD300a molecule	$1.13x10^{-3}$	-1.34
1555338_s_at	AQP10	aquaporin 10	$1.18x10^{-3}$	1.60
209839_at	DNM3	dynamin 3	$1.18x10^{-3}$	1.73
203305_at	F13A1	coagulation factor XIII, A1 polypeptide	$1.21x10^{-3}$	1.78
233033_at	ZFHX1B	Zinc finger homeobox 1b	$1.26x10^{-3}$	1.23
241250_at	RFFL	Ring finger and FYVE-like domain containing 1	$1.27x10^{-3}$	1.57
242776_at	ZCCHC6	zinc finger, CCHC domain containing 6	$1.36x10^{-3}$	-1.30
1560762_at	LOC285972	hypothetical protein LOC285972	$1.40x10^{-3}$	-1.27
219608_s_at	FBXO38	F-box protein 38	$1.43x10^{-3}$	-1.26
205733_at	BLM	Bloom syndrome	$1.45x10^{-3}$	-1.24
208753_s_at	NAP1L1	nucleosome assembly protein 1-like 1	$1.47x10^{-3}$	1.28
203946_s_at	ARG2	arginase, type II	$1.52x10^{-3}$	1.48
220215_at	ZNF669	zinc finger protein 669	$1.54x10^{-3}$	-1.24
241888_at	---	Transcribed locus	$1.56x10^{-3}$	1.20
240300_at	TK2	Thymidine kinase 2, mitochondrial	$1.56x10^{-3}$	1.20
206641_at	TNFRSF17	tumor necrosis factor receptor superfamily, member 17	$1.58x10^{-3}$	-2.42
222532_at	SRPRB	signal recognition particle receptor, B subunit	$1.62x10^{-3}$	-1.20
219861_at	DNAJC17	DnaJ (Hsp40) homolog, subfamily C, member 17	$1.63x10^{-3}$	-1.22
213568_at	OSR2	odd-skipped related 2 (Drosophila)	$1.68x10^{-3}$	1.58
211149_at	UTY	ubiquitously transcribed tetratricopeptide repeat gene, Y-linked	$1.71x10^{-3}$	1.43
201904_s_at	CTDSPL	CTD (carboxy-terminal domain, RNA polymerase II, polypeptide A) small phosphatase-like	$1.74x10^{-3}$	1.65
217959_s_at	TRAPPC4	trafficking protein particle complex 4	$1.83x10^{-3}$	-1.25
205345_at	BARD1	BRCA1 associated RING domain 1	$1.86x10^{-3}$	-1.23
224494_x_at	DHRS10	dehydrogenase/reductase (SDR family) member 10	$1.92x10^{-3}$	1.20

244740_at	MGC9913	hypothetical protein MGC9913	$1.93x10^{-3}$	1.30
209767_s_at	GP1BB / SEPT5	glycoprotein Ib (platelet), beta polypeptide / septin 5	$1.96x10^{-3}$	1.47
234897_s_at	C6orf21	chromosome 6 open reading frame 21	$2.00x10^{-3}$	1.57
222793_at	DDX58	DEAD (Asp-Glu-Ala-Asp) box polypeptide 58	$2.04x10^{-3}$	-1.43
239335_at	ZNF710	zinc finger protein 710	$2.14x10^{-3}$	-1.26
235294_at	---	---	$2.15x10^{-3}$	-1.21
219204_s_at	SRR	serine racemase	$2.15x10^{-3}$	-1.43
214307_at	LOC642252	similar to Homogentisate 1,2-dioxygenase (Homogentisicase) (Homogentisate oxygenase)	$2.15x10^{-3}$	1.49
206624_at	USP9Y	ubiquitin specific peptidase 9, Y-linked (fat facets-like, Drosophila)	$2.19x10^{-3}$	1.56
200700_s_at	KDELR2	KDEL (Lys-Asp-Glu-Leu) endoplasmic reticulum protein retention receptor 2	$2.22x10^{-3}$	-1.25
1553842_at	CXorf20	chromosome X open reading frame 20	$2.24x10^{-3}$	1.84
232122_s_at	VEPH1	ventricular zone expressed PH domain homolog 1 (zebrafish)	$2.35x10^{-3}$	1.36
232173_at	CLEC2L	C-type lectin domain family 2, member L	$2.37x10^{-3}$	1.22
228492_at	USP9Y	ubiquitin specific peptidase 9, Y-linked (fat facets-like, Drosophila)	$2.39x10^{-3}$	1.70
219944_at	RSNL2	restin-like 2	$2.43x10^{-3}$	-1.22
205401_at	AGPS	alkylglycerone phosphate synthase	$2.49x10^{-3}$	-1.24
211430_s_at	IGH@ / IGHG1 / IGHG2 / IGHG3 / IGHM	immunoglobulin heavy locus / immunoglobulin heavy constant gamma / heavy constant mu	$2.51x10^{-3}$	-2.53
242323_at	CASP6	Caspase 6, apoptosis-related cysteine peptidase	$2.52x10^{-3}$	1.42
213258_at	TFPI	tissue factor pathway inhibitor (lipoprotein-associated coagulation inhibitor)	$2.53x10^{-3}$	1.45
214974_x_at	CXCL5	chemokine (C-X-C motif) ligand 5	$2.57x10^{-3}$	2.55
204993_at	GNAZ	guanine nucleotide binding protein (G protein), alpha z polypeptide	$2.61x10^{-3}$	1.60
236694_at	CYorf15A	chromosome Y open reading frame 15A	$2.61x10^{-3}$	1.64
205731_s_at	NCOA2	nuclear receptor coactivator 2	$2.80x10^{-3}$	-1.22
48031_r_at	C5orf4	chromosome 5 open reading frame 4	$2.86x10^{-3}$	1.34
1563228_x_at	MGC15523	Hypothetical protein MGC15523	$2.89x10^{-3}$	1.25
223615_at	ABI3	ABI gene family, member 3	$2.92x10^{-3}$	-1.21
218986_s_at	FLJ20035	hypothetical protein FLJ20035	$2.96x10^{-3}$	-1.30
239489_at	UBLCP1	Ubiquitin-like domain containing CTD phosphatase 1	$2.96x10^{-3}$	-1.22
231963_at	---	Homo sapiens, clone IMAGE:3869276, mRNA	$2.97x10^{-3}$	1.22
217963_s_at	NGFRAP1	nerve growth factor receptor (TNFRSF16) associated protein 1	$3.04x10^{-3}$	1.39
228618_at	PEAR1	platelet endothelial aggregation receptor 1	$3.06x10^{-3}$	1.37
216640_s_at	PDIA6	protein disulfide isomerase family A, member 6	$3.06x10^{-3}$	-1.21
227941_at	LOC339803	hypothetical protein LOC339803	$3.07x10^{-3}$	-1.42
204409_s_at	EIF1AY	eukaryotic translation initiation factor 1A, Y-linked	$3.13x10^{-3}$	2.12
205170_at	STAT2	signal transducer and activator of transcription 2, 113kDa	$3.19x10^{-3}$	-1.34
1554478_a_at	HEATR3	HEAT repeat containing 3	$3.28x10^{-3}$	-1.46
202223_at	STT3A	STT3, subunit of the oligosaccharyltransferase complex, homolog A (*S. cerevisiae*)	$3.32x10^{-3}$	-1.22
209894_at	LEPR	leptin receptor	$3.36x10^{-3}$	1.39
223819_x_at	COMMD5	COMM domain containing 5	$3.40x10^{-3}$	-1.23
219431_at	ARHGAP10	Rho GTPase activating protein 10	$3.47x10^{-3}$	-1.22
212077_at	CALD1	caldesmon 1	$3.49x10^{-3}$	2.25
229002_at	FAM69B	family with sequence similarity 69, member B	$3.50x10^{-3}$	1.24
217078_s_at	CD300A	CD300a molecule	$3.53x10^{-3}$	-1.39
211048_s_at	PDIA4	protein disulfide isomerase family A, member 4	$3.55x10^{-3}$	-1.20

217761_at	ADI1	acireductone dioxygenase 1	$3.63x10^{-3}$	1.27
206700_s_at	SMCY	Smcy homolog, Y-linked (mouse)	$3.66x10^{-3}$	1.84
218599_at	REC8L1	REC8-like 1 (yeast)	$3.71x10^{-3}$	-1.23
215101_s_at	CXCL5	chemokine (C-X-C motif) ligand 5	$3.71x10^{-3}$	2.32
206883_x_at	GP9	glycoprotein IX (platelet)	$3.75x10^{-3}$	1.30
1568592_at	RNF36	ring finger protein 36	$3.90x10^{-3}$	-1.26
203095_at	MTIF2	mitochondrial translational initiation factor 2	$3.99x10^{-3}$	-1.22
243952_at	---	Homo sapiens, clone IMAGE:4685786, mRNA	$4.09x10^{-3}$	1.28
212651_at	RHOBTB1	Rho-related BTB domain containing 1	$4.09x10^{-3}$	1.67
1552470_a_at	ABHD11	abhydrolase domain containing 11	$4.32x10^{-3}$	-1.29
235446_at	XIST	X (inactive)-specific transcript	$4.33x10^{-3}$	-1.37
206279_at	PRKY	protein kinase, Y-linked	$4.34x10^{-3}$	1.26
205485_at	RYR1	ryanodine receptor 1 (skeletal)	$4.35x10^{-3}$	-1.37
210493_s_at	MFAP3L	microfibrillar-associated protein 3-like	$4.41x10^{-3}$	1.32
215240_at	ITGB3	integrin, beta 3 (platelet glycoprotein IIIa, antigen CD61)	$4.45x10^{-3}$	1.52
232553_at	PCYT1B	phosphate cytidylyltransferase 1, choline, beta	$4.46x10^{-3}$	1.33
201906_s_at	CTDSPL	CTD (carboxy-terminal domain, RNA polymerase II, polypeptide A) small phosphatase-like	$4.57x10^{-3}$	1.64
219785_s_at	FBXO31	F-box protein 31	$4.73x10^{-3}$	-1.29
217629_at	GNGT2	Guanine nucleotide binding protein (G protein), gamma transducing activity polypeptide 2	$4.77x10^{-3}$	-1.27
1561651_s_at	TAL1	T-cell acute lymphocytic leukemia 1	$4.84x10^{-3}$	1.25
206513_at	AIM2	absent in melanoma 2	$4.93x10^{-3}$	-1.47
202555_s_at	MYLK	myosin, light polypeptide kinase	$4.95x10^{-3}$	1.65
244555_at	---	---	$4.96x10^{-3}$	1.22
227088_at	PDE5A	phosphodiesterase 5A, cGMP-specific	$4.99x10^{-3}$	1.73
205730_s_at	ABLIM3	actin binding LIM protein family, member 3	$5.00x10^{-3}$	1.53
210922_at	SIPA1L3	Signal-induced proliferation-associated 1 like 3	$5.00x10^{-3}$	1.21
238669_at	PTGS1	prostaglandin-endoperoxide synthase 1	$5.00x10^{-3}$	1.66
225775_at	TSPAN33	tetraspanin 33	$5.03x10^{-3}$	1.46
227092_at	---	---	$5.04x10^{-3}$	1.21
240679_at	STK32B	Serine/threonine kinase 32B	$5.04x10^{-3}$	1.21
205425_at	HIP1	huntingtin interacting protein 1	$5.04x10^{-3}$	-1.39
216100_s_at	TOR1AIP1	torsin A interacting protein 1	$5.05x10^{-3}$	-1.23
225166_at	ARHGAP18	Rho GTPase activating protein 18	$5.15x10^{-3}$	1.28
214545_s_at	PROSC	proline synthetase co-transcribed homolog (bacterial)	$5.18x10^{-3}$	-1.28
236125_at	---	CDNA FLJ31332 fis, clone MAMGL1000096	$5.23x10^{-3}$	-1.20
207813_s_at	FDXR	ferredoxin reductase	$5.26x10^{-3}$	-1.26
227044_at	TBC1D22A	TBC1 domain family, member 22A	$5.32x10^{-3}$	1.63
1554465_s_at	ZNF673 / ZNF674	zinc finger protein 673 / zinc finger protein 674	$5.32x10^{-3}$	-1.25
227671_at	XIST	X (inactive)-specific transcript	$5.33x10^{-3}$	-2.49
217998_at	PHLDA1	pleckstrin homology-like domain, family A, member 1	$5.36x10^{-3}$	1.42
223165_s_at	IHPK2	inositol hexaphosphate kinase 2	$5.39x10^{-3}$	-1.36
224588_at	XIST	X (inactive)-specific transcript	$5.39x10^{-3}$	-2.69
217445_s_at	GART	phosphoribosylglycinamide formyltransferase, phosphoribosylglycinamide synthetase	$5.39x10^{-3}$	-1.30
223565_at	PACAP	proapoptotic caspase adaptor protein	$5.39x10^{-3}$	-1.50
224795_x_at	IGKC / IGKV1-5	immunoglobulin kappa constant / immunoglobulin kappa variable 1-5	$5.39x10^{-3}$	-1.65
218597_s_at	C10orf70	chromosome 10 open reading frame 70	$5.51x10^{-3}$	-1.20
221253_s_at	TXNDC5	thioredoxin domain containing 5	$5.61x10^{-3}$	-1.40
201738_at	EIF1B	eukaryotic translation initiation factor 1B	$5.63x10^{-3}$	1.31

220117_at	ZNF659	zinc finger protein 659	$5.65x10^{-3}$	1.24
223646_s_at	CYorf15B	chromosome Y open reading frame 15B	$5.69x10^{-3}$	1.55
242943_at	ST8SIA4	ST8 alpha-N-acetyl-neuraminide alpha-2,8-sialyltransferase 4	$5.86x10^{-3}$	-1.35
206049_at	SELP	selectin P (granule membrane protein 140kDa, antigen CD62)	$5.87x10^{-3}$	1.64
230222_at	---	---	$5.90x10^{-3}$	1.22
210365_at	RUNX1	runt-related transcription factor 1 (acute myeloid leukemia 1; aml1 oncogene)	$5.95x10^{-3}$	1.26
63825_at	ABHD2	abhydrolase domain containing 2	$6.02x10^{-3}$	-1.20
209213_at	CBR1	carbonyl reductase 1	$6.07x10^{-3}$	-1.33
225737_s_at	FBXO22	F-box protein 22	$6.10x10^{-3}$	-1.23
204384_at	GOLGA2	golgi autoantigen, golgin subfamily a, 2	$6.15x10^{-3}$	-1.24
209922_at	BRAP	BRCA1 associated protein	$6.15x10^{-3}$	-1.27
220751_s_at	C5orf4	chromosome 5 open reading frame 4	$6.26x10^{-3}$	1.70
218723_s_at	RGC32	response gene to complement 32	$6.32x10^{-3}$	1.71
1555905_a_at	C3orf23	chromosome 3 open reading frame 23	$6.36x10^{-3}$	-1.36
1558045_a_at	---	Transcribed locus	$6.36x10^{-3}$	-1.30
202734_at	TRIP10	thyroid hormone receptor interactor 10	$6.51x10^{-3}$	1.28
219014_at	PLAC8	placenta-specific 8	$6.52x10^{-3}$	-1.30
203882_at	ISGF3G	interferon-stimulated transcription factor 3, gamma 48kDa	$6.54x10^{-3}$	-1.23
242680_at	---	Transcribed locus	$6.57x10^{-3}$	1.23
230574_at	---	---	$6.73x10^{-3}$	1.50
1555789_s_at	PHF23	PHD finger protein 23	$6.74x10^{-3}$	-1.24
206465_at	ACSBG1	acyl-CoA synthetase bubblegum family member 1	$6.84x10^{-3}$	1.48
206268_at	LEFTY1	left-right determination factor 1	$6.94x10^{-3}$	1.26
229744_at	SSFA2	Sperm specific antigen 2	$6.94x10^{-3}$	1.26
227600_at	---	Full-length cDNA clone CS0DK012YA15 of HeLa cells Cot 25-normalized of *Homo sapiens*	$6.95x10^{-3}$	-1.24
230535_s_at	---	Transcribed locus, strongly similar to NP_110400.1 beta tubulin 1, class VI	$7.04x10^{-3}$	1.41
222343_at	BCL2L11	BCL2-like 11 (apoptosis facilitator)	$7.11x10^{-3}$	-1.21
1552552_s_at	CLEC4C	C-type lectin domain family 4, member C	$7.19x10^{-3}$	-1.54
1554217_a_at	FLJ20097	hypothetical protein LOC55610, isoform b	$7.21x10^{-3}$	-1.26
206857_s_at	FKBP1B	FK506 binding protein 1B, 12.6 kDa	$7.29x10^{-3}$	1.33
237289_at	FAM119A	Family with sequence similarity 119, member A	$7.34x10^{-3}$	-1.23
214218_s_at	XIST	X (inactive)-specific transcript	$7.39x10^{-3}$	-1.97
226018_at	Ells1	hypothetical protein Ells1	$7.40x10^{-3}$	1.59
221651_x_at	IGKC / IGKV1-5	immunoglobulin kappa constant / immunoglobulin kappa variable 1-5	$7.40x10^{-3}$	-1.62
1556209_at	CLEC2B	C-type lectin domain family 2, member B	$7.51x10^{-3}$	-1.63
214131_at	CYorf15B	chromosome Y open reading frame 15B	$7.54x10^{-3}$	1.52
1555841_at	---	---	$7.58x10^{-3}$	1.27
216574_s_at	RPE	ribulose-5-phosphate-3-epimerase / rcRPE	$7.58x10^{-3}$	-1.25
212592_at	IGJ	Immunoglobulin J polypeptide, linker protein for immunoglobulin alpha and mu polypeptide	$7.63x10^{-3}$	-1.73
212667_at	SPARC	secreted protein, acidic, cysteine-rich (osteonectin)	$7.66x10^{-3}$	1.59
1558111_at	MBNL1	Muscleblind-like (Drosophila)	$7.77x10^{-3}$	1.35
214463_x_at	HIST1H4K / HIST1H4J	histone 1, H4k / histone 1, H4j	$7.78x10^{-3}$	-1.22
210251_s_at	RUFY3	RUN and FYVE domain containing 3	$7.82x10^{-3}$	-1.22
239942_at	---	Transcribed locus	$7.84x10^{-3}$	1.36

214836_x_at	IGKC / IGKV1-5	immunoglobulin kappa constant / immunoglobulin kappa variable 1-5	$7.89x10^{-3}$	-1.55
239474_at	SLC6A6	Solute carrier family 6 (neurotransmitter transporter, taurine), member 6	$7.94x10^{-3}$	1.23
219677_at	SPSB1	splA/ryanodine receptor domain and SOCS box containing 1	$8.14x10^{-3}$	1.24
221671_x_at	IGKC / IGKV1-5	immunoglobulin kappa constant / immunoglobulin kappa variable 1-5	$8.15x10^{-3}$	-1.60
213327_s_at	USP12	ubiquitin specific peptidase 12	$8.16x10^{-3}$	1.24
205037_at	RABL4	RAB, member of RAS oncogene family-like 4	$8.33x10^{-3}$	-1.21
219503_s_at	TMEM40	transmembrane protein 40	$8.42x10^{-3}$	1.25
216521_s_at	BRCC3	BRCA1/BRCA2-containing complex, subunit 3	$8.52x10^{-3}$	-1.30
218978_s_at	SLC25A37	solute carrier family 25, member 37	$8.69x10^{-3}$	1.38
229830_at	---	Transcribed locus, strongly similar to NP_148983.1 platelet-derived growth factor	$8.85x10^{-3}$	1.46
213876_x_at	U2AF1L2	U2 small nuclear RNA auxiliary factor 1-like 2	$8.85x10^{-3}$	-1.20
205483_s_at	ISG15	ISG15 ubiquitin-like modifier	$8.90x10^{-3}$	-1.50
203029_s_at	PTPRN2	protein tyrosine phosphatase, receptor type, N polypeptide 2	$8.94x10^{-3}$	-1.23
242903_at	IFNGR1	Interferon gamma receptor 1	$9.07x10^{-3}$	-1.35
1552277_a_at	C9orf30	chromosome 9 open reading frame 30	$9.07x10^{-3}$	1.23
227703_s_at	SYTL4	synaptotagmin-like 4 (granuphilin-a)	$9.29x10^{-3}$	1.43
241955_at	HECTD1	HECT domain containing 1	$9.32x10^{-3}$	-1.22
215537_x_at	DDAH2	dimethylarginine dimethylaminohydrolase 2	$9.38x10^{-3}$	-1.20
1556829_at	TIPARP	TCDD-inducible poly(ADP-ribose) polymerase	$9.47x10^{-3}$	1.21
232902_s_at	RARSL	arginyl-tRNA synthetase-like	$9.47x10^{-3}$	-1.21
228003_at	RAB30	RAB30, member RAS oncogene family	$9.54x10^{-3}$	1.21
222803_at	PRTFDC1	phosphoribosyl transferase domain containing 1	$9.63x10^{-3}$	1.31
237403_at	GFI1B	growth factor independent 1B (potential regulator of CDKN1A, translocated in CML)	$9.64x10^{-3}$	1.46
234764_x_at	LOC96610 / IGL@	Hypothetical protein similar to KIAA0187 gene product / Immunoglobulin lambda locus	$9.65x10^{-3}$	-1.72
218340_s_at	UBE1L2	ubiquitin-activating enzyme E1-like 2	$9.70x10^{-3}$	-1.75
201059_at	CTTN	cortactin	$9.87x10^{-3}$	1.77
209651_at	TGFB1I1	transforming growth factor beta 1 induced transcript 1	$9.89x10^{-3}$	1.72
1557261_at	WHDC1L1 / WHDC1L2	WAS protein homology region 2 domain containing 1-like 1 / - like 2	$9.93x10^{-3}$	1.50
227230_s_at	KIAA1211	KIAA1211 protein	$9.95x10^{-3}$	1.30
223645_s_at	CYorf15B	chromosome Y open reading frame 15B	$1.00x10^{-2}$	1.47
223063_at	C1orf198	chromosome 1 open reading frame 198	$1.01x10^{-2}$	1.26
201438_at	COL6A3	collagen, type VI, alpha 3	$1.01x10^{-2}$	1.29
214669_x_at	IGKC	Immunoglobulin kappa constant	$1.03x10^{-2}$	-1.65
202071_at	SDC4	syndecan 4 (amphiglycan, ryudocan)	$1.03x10^{-2}$	1.55
231341_at	SLC35D3	solute carrier family 35, member D3	$1.03x10^{-2}$	1.26
221556_at	CDC14B	CDC14 cell division cycle 14 homolog B (*S. cerevisiae*)	$1.03x10^{-2}$	1.47
205552_s_at	OAS1	2',5'-oligoadenylate synthetase 1, 40/46kDa	$1.04x10^{-2}$	-1.52
239105_at	---	Transcribed locus	$1.05x10^{-2}$	1.21
202729_s_at	LTBP1	latent transforming growth factor beta binding protein 1	$1.06x10^{-2}$	1.55
203680_at	PRKAR2B	protein kinase, cAMP-dependent, regulatory, type II, beta	$1.06x10^{-2}$	1.81
217144_at	UBB / LOC648390	ubiquitin B / similar to ubiquitin B precursor	$1.07x10^{-2}$	1.21
1554999_at	RASGEF1B	RasGEF domain family, member 1B	$1.08x10^{-2}$	1.59
230261_at	ST8SIA4	ST8 alpha-N-acetyl-neuraminide alpha-2,8-sialyltransferase 4	$1.09x10^{-2}$	-1.32
217022_s_at	IGHA1 / IGHA2	immunoglobulin heavy constant alpha 1 / immunoglobulin heavy constant alpha 2	$1.09x10^{-2}$	-2.38

37966_at	PARVB	parvin, beta	$1.10x10^{-2}$	1.40
208637_x_at	ACTN1	actinin, alpha 1	$1.11x10^{-2}$	1.23
221728_x_at	XIST	X (inactive)-specific transcript	$1.11x10^{-2}$	-1.80
215017_s_at	FNBP1L	formin binding protein 1-like	$1.11x10^{-2}$	1.25
221160_s_at	CABP3 / CABP5	calcium binding protein 3 / calcium binding protein 5	$1.13x10^{-2}$	1.57
231699_at	NFKBIA	Nuclear factor of kappa light polypeptide gene enhancer in B-cells inhibitor, alpha	$1.14x10^{-2}$	1.33
214073_at	CTTN	cortactin	$1.15x10^{-2}$	1.72
214777_at	---	Immunoglobulin light chain variable region complementarity determining region	$1.15x10^{-2}$	-1.72
210313_at	LILRA4	leukocyte immunoglobulin-like receptor, subfamily A (with TM domain), member 4	$1.16x10^{-2}$	-1.38
243683_at	MORF4L2	Mortality factor 4 like 2	$1.16x10^{-2}$	-1.40
206493_at	ITGA2B	integrin, alpha 2b (platelet glycoprotein IIb of IIb/IIIa complex, antigen CD41)	$1.17x10^{-2}$	1.83
205914_s_at	GRIN1	glutamate receptor, ionotropic, N-methyl D-aspartate 1	$1.17x10^{-2}$	1.20
221540_x_at	GTF2H2	general transcription factor IIH, polypeptide 2, 44kDa	$1.17x10^{-2}$	-1.23
209339_at	SIAH2	seven in absentia homolog 2 (Drosophila)	$1.18x10^{-2}$	1.32
222632_s_at	LZTFL1	leucine zipper transcription factor-like 1	$1.18x10^{-2}$	-1.26
225354_s_at	SH3BGRL2	SH3 domain binding glutamic acid-rich protein like 2	$1.19x10^{-2}$	1.85
206925_at	ST8SIA4	ST8 alpha-N-acetyl-neuraminide alpha-2,8-sialyltransferase 4	$1.20x10^{-2}$	-1.29
1570360_s_at	DDX3Y	DEAD (Asp-Glu-Ala-Asp) box polypeptide 3, Y-linked	$1.20x10^{-2}$	1.42
223754_at	MGC13057	hypothetical protein MGC13057	$1.20x10^{-2}$	1.52
240456_at	FLJ11795	FLJ11795 protein	$1.21x10^{-2}$	1.47
202886_s_at	PPP2R1B	protein phosphatase 2 (formerly 2A), regulatory subunit A (PR 65), beta isoform	$1.21x10^{-2}$	-1.24
230009_at	FAM118B	family with sequence similarity 118, member B	$1.21x10^{-2}$	-1.27
214074_s_at	CTTN	cortactin	$1.23x10^{-2}$	1.44
220586_at	CHD9	chromodomain helicase DNA binding protein 9	$1.23x10^{-2}$	-1.21
207540_s_at	SYK	spleen tyrosine kinase	$1.23x10^{-2}$	-1.21
242961_x_at	DDX58	DEAD (Asp-Glu-Ala-Asp) box polypeptide 58	$1.24x10^{-2}$	-1.78
212528_at	---	CDNA clone IMAGE:3878236	$1.26x10^{-2}$	-1.23
220840_s_at	C1orf112	chromosome 1 open reading frame 112	$1.26x10^{-2}$	-1.21
222892_s_at	TMEM40	transmembrane protein 40	$1.28x10^{-2}$	1.49
202221_s_at	EP300	E1A binding protein p300	$1.28x10^{-2}$	1.21
215946_x_at	CTA-246H3.1	similar to omega protein	$1.29x10^{-2}$	-1.48
229121_at	---	CDNA FLJ44441 fis, clone UTERU2020242	$1.29x10^{-2}$	-1.37
230127_at	---	Transcribed locus	$1.29x10^{-2}$	1.50
242094_at	ITGB3	Integrin, beta 3 (platelet glycoprotein IIIa, antigen CD61)	$1.30x10^{-2}$	1.39
1563357_at	---	MRNA; cDNA DKFZp564C203 (from clone DKFZp564C203)	$1.31x10^{-2}$	1.44
209853_s_at	PSME3	proteasome (prosome, macropain) activator subunit 3 (PA28 gamma; Ki)	$1.31x10^{-2}$	-1.22
229330_at	SSU72	SSU72 RNA polymerase II CTD phosphatase homolog (*S. cerevisiae*)	$1.32x10^{-2}$	1.20
223851_s_at	TNFRSF18	tumor necrosis factor receptor superfamily, member 18	$1.33x10^{-2}$	-1.25
210005_at	GART	phosphoribosylglycinamide formyltransferase, phosphoribosylglycinamide synthetase	$1.34x10^{-2}$	-1.21
218894_s_at	FLJ10292	mago-nashi homolog	$1.34x10^{-2}$	-1.23
205128_x_at	PTGS1	prostaglandin-endoperoxide synthase 1	$1.34x10^{-2}$	1.46
208814_at	HSPA4	Heat shock 70kDa protein 4	$1.34x10^{-2}$	-1.28
206204_at	GRB14	growth factor receptor-bound protein 14	$1.36x10^{-2}$	1.41
216542_x_at	IGHA1 / IGHG1	immunoglobulin heavy constant alpha 1 / immunoglobulin heavy constant gamma 1	$1.36x10^{-2}$	-1.29

204614_at	SERPINB2	serpin peptidase inhibitor, clade B (ovalbumin), member 2	$1.37x10^{-2}$	1.69
202619_s_at	PLOD2	procollagen-lysine, 2-oxoglutarate 5-dioxygenase 2	$1.37x10^{-2}$	1.22
219412_at	RAB38	RAB38, member RAS oncogene family	$1.37x10^{-2}$	-1.35
209417_s_at	IFI35	interferon-induced protein 35	$1.37x10^{-2}$	-1.31
223939_at	SUCNR1	succinate receptor 1	$1.38x10^{-2}$	1.24
207098_s_at	MFN1	mitofusin 1	$1.39x10^{-2}$	-1.22
202728_s_at	LTBP1	latent transforming growth factor beta binding protein 1	$1.39x10^{-2}$	1.33
203264_s_at	ARHGEF9	Cdc42 guanine nucleotide exchange factor (GEF) 9	$1.40x10^{-2}$	-1.22
228195_at	MGC13057	Hypothetical protein MGC13057	$1.40x10^{-2}$	1.68
237181_at	PPP2R5C	Protein phosphatase 2, regulatory subunit B (B56), gamma isoform	$1.41x10^{-2}$	-1.27
224590_at	XIST	X (inactive)-specific transcript	$1.41x10^{-2}$	-2.09
202607_at	NDST1	N-deacetylase/N-sulfotransferase (heparan glucosaminyl) 1	$1.42x10^{-2}$	1.23
230778_at	---	Transcribed locus	$1.42x10^{-2}$	1.32
203920_at	NR1H3	nuclear receptor subfamily 1, group H, member 3	$1.47x10^{-2}$	-1.25
211343_s_at	COL13A1	collagen, type XIII, alpha 1	$1.50x10^{-2}$	-1.60
244105_at	WHDC1 / WHDC1L1 / WHDC1L2	WAS protein homology region 2 domain containing 1 / -like 1 / -like 2	$1.52x10^{-2}$	1.20
201125_s_at	ITGB5	integrin, beta 5	$1.52x10^{-2}$	1.49
226778_at	C8orf42	Chromosome 8 open reading frame 42	$1.52x10^{-2}$	1.21
226668_at	WDSUB1	WD repeat, sterile alpha motif and U-box domain containing 1	$1.53x10^{-2}$	-1.23
217179_x_at	LOC96610	Hypothetical protein similar to KIAA0187 gene product	$1.57x10^{-2}$	-1.48
227386_s_at	TTMB	TTMB protein	$1.57x10^{-2}$	1.25
225066_at	---	---	$1.57x10^{-2}$	1.39
225369_at	ESAM	endothelial cell adhesion molecule	$1.58x10^{-2}$	1.36
211644_x_at	IGKC	immunoglobulin kappa constant	$1.59x10^{-2}$	-1.76
230026_at	MRPL43	mitochondrial ribosomal protein L43	$1.59x10^{-2}$	1.46
242784_at	ETS2	V-ets erythroblastosis virus E26 oncogene homolog 2 (avian)	$1.59x10^{-2}$	1.67
222693_at	FNDC3B	fibronectin type III domain containing 3B	$1.60x10^{-2}$	-1.21
224496_s_at	TMEM107	transmembrane protein 107	$1.60x10^{-2}$	1.61
205514_at	ZNF415	zinc finger protein 415	$1.60x10^{-2}$	-1.20
1557450_s_at	WHDC1L1	WAS protein homology region 2 domain containing 1-like 1	$1.60x10^{-2}$	1.32
226733_at	PFKFB2	6-phosphofructo-2-kinase/fructose-2,6-biphosphatase 2	$1.61x10^{-2}$	-1.31
222746_s_at	BSPRY	B-box and SPRY domain containing	$1.61x10^{-2}$	-1.22
1557008_at	LOC340107	hypothetical protein LOC340107	$1.61x10^{-2}$	1.22
238079_at	TPM3	Tropomyosin 3	$1.62x10^{-2}$	-1.25
209729_at	GAS2L1	growth arrest-specific 2 like 1	$1.62x10^{-2}$	1.27
209274_s_at	HBLD2	HESB like domain containing 2	$1.62x10^{-2}$	1.22
219211_at	USP18	ubiquitin specific peptidase 18	$1.63x10^{-2}$	-1.53
228607_at	OAS2	2'-5'-oligoadenylate synthetase 2, 69/71kDa	$1.64x10^{-2}$	-1.31
205127_at	PTGS1	prostaglandin-endoperoxide synthase 1	$1.64x10^{-2}$	1.52
228974_at	ZNF677	Zinc finger protein 677	$1.67x10^{-2}$	-1.39
235593_at	ZFHX1B	zinc finger homeobox 1b	$1.68x10^{-2}$	1.31
220519_s_at	LIM2	lens intrinsic membrane protein 2, 19kDa	$1.69x10^{-2}$	-1.21
213913_s_at	KIAA0984	KIAA0984 protein	$1.69x10^{-2}$	-1.25
205409_at	FOSL2	FOS-like antigen 2	$1.70x10^{-2}$	-1.36
202364_at	MXI1	MAX interactor 1	$1.71x10^{-2}$	1.31
210169_at	SEC14L5	SEC14-like 5 (*S. cerevisiae*)	$1.71x10^{-2}$	1.35
1555938_x_at	VIM	vimentin	$1.72x10^{-2}$	1.60

242577_at	LOC642398	hypothetical protein LOC642398	$1.72x10^{-2}$	-1.21
1555964_at	ARL17P1	ADP-ribosylation factor-like 17 pseudogene 1	$1.72x10^{-2}$	-1.25
236700_at	LOC653352	similar to eukaryotic translation initiation factor 3, subunit 8	$1.74x10^{-2}$	-1.27
223340_at	SPG3A	spastic paraplegia 3A (autosomal dominant)	$1.75x10^{-2}$	1.30
230972_at	ANKRD9	ankyrin repeat domain 9	$1.75x10^{-2}$	1.31
206494_s_at	ITGA2B	integrin, alpha 2b (platelet glycoprotein IIb of IIb/IIIa complex, antigen CD41)	$1.75x10^{-2}$	2.03
243999_at	SLFN5	schlafen family member 5	$1.78x10^{-2}$	-1.32
224891_at	FOXO3A	forkhead box O3A	$1.78x10^{-2}$	1.25
212086_x_at	LMNA	lamin A/C	$1.79x10^{-2}$	1.22
212923_s_at	C6orf145	chromosome 6 open reading frame 145	$1.80x10^{-2}$	1.27
214020_x_at	ITGB5	Integrin, beta 5	$1.80x10^{-2}$	1.36
224175_s_at	TRIM34 / TRIM6-TRIM34	tripartite motif-containing 34 / tripartite motif-containing 6 and tripartite	$1.80x10^{-2}$	-1.29
225331_at	CCDC50	coiled-coil domain containing 50	$1.80x10^{-2}$	-1.23
222503_s_at	WDR41	WD repeat domain 41	$1.81x10^{-2}$	-1.21
212151_at	PBX1	Pre-B-cell leukemia transcription factor 1	$1.81x10^{-2}$	1.34
242370_at	MTHFD2L	methylenetetrahydrofolate dehydrogenase (NADP+ dependent) 2-like	$1.84x10^{-2}$	1.21
230833_at	ACRBP	acrosin binding protein	$1.85x10^{-2}$	1.58
211300_s_at	TP53	tumor protein p53 (Li-Fraumeni syndrome)	$1.85x10^{-2}$	-1.33
205013_s_at	ADORA2A	adenosine A2a receptor	$1.87x10^{-2}$	1.24
215936_s_at	KIAA1033	KIAA1033	$1.88x10^{-2}$	-1.21
205442_at	MFAP3L	microfibrillar-associated protein 3-like	$1.88x10^{-2}$	1.84
225102_at	MGLL	monoglyceride lipase	$1.88x10^{-2}$	1.32
226020_s_at	DAB1 / OMA1	disabled homolog 1 (Drosophila) / OMA1 homolog, zinc metallopeptidase (*S. cerevisiae*)	$1.89x10^{-2}$	-1.22
221219_s_at	KLHDC4	kelch domain containing 4	$1.89x10^{-2}$	-1.27
215492_x_at	PTCRA	pre T-cell antigen receptor alpha	$1.89x10^{-2}$	1.20
214505_s_at	FHL1	four and a half LIM domains 1	$1.91x10^{-2}$	1.33
220416_at	ATP8B4	ATPase, Class I, type 8B, member 4	$1.91x10^{-2}$	-1.35
205187_at	SMAD5	SMAD, mothers against DPP homolog 5 (Drosophila)	$1.91x10^{-2}$	-1.20
241832_at	FAM98A	family with sequence similarity 98, member A	$1.91x10^{-2}$	-1.24
227974_at	---	Transcribed locus, moderately similar to NP_775735.1 l(3)mbt-like 4	$1.91x10^{-2}$	-1.26
226603_at	SAMD9L	sterile alpha motif domain containing 9-like	$1.93x10^{-2}$	-1.28
1553055_a_at	SLFN5	schlafen family member 5	$1.93x10^{-2}$	-1.44
239012_at	IBRDC2	IBR domain containing 2	$1.93x10^{-2}$	1.39
206176_at	BMP6	bone morphogenetic protein 6	$1.94x10^{-2}$	1.42
213908_at	WHDC1L1 / WHDC1L2	WAS protein homology region 2 domain containing 1-like 1 / · like 2	$1.94x10^{-2}$	1.33
201540_at	FHL1	four and a half LIM domains 1	$1.94x10^{-2}$	1.46
204115_at	GNG11	guanine nucleotide binding protein (G protein), gamma 11	$1.96x10^{-2}$	1.80
205930_at	GTF2E1	general transcription factor IIE, polypeptide 1, alpha 56kDa	$1.96x10^{-2}$	-1.26
218817_at	SPCS3	signal peptidase complex subunit 3 homolog (*S. cerevisiae*)	$1.97x10^{-2}$	-1.23
212958_x_at	PAM	peptidylglycine alpha-amidating monooxygenase	$1.98x10^{-2}$	-1.20
219599_at	PRO1843	hypothetical protein PRO1843	$1.99x10^{-2}$	1.32
214973_x_at	IGHD	immunoglobulin heavy constant delta	$2.00x10^{-2}$	-1.45
207808_s_at	PROS1	protein S (alpha)	$2.02x10^{-2}$	1.76
204410_at	EIF1AY	eukaryotic translation initiation factor 1A, Y-linked	$2.03x10^{-2}$	1.60

221555_x_at	CDC14B	CDC14 cell division cycle 14 homolog B (*S. cerevisiae*)	$2.03x10^{-2}$	1.27
226152_at	TTC7B	tetratricopeptide repeat domain 7B	$2.03x10^{-2}$	1.54
210659_at	CMKLR1	chemokine-like receptor 1	$2.04x10^{-2}$	-1.35
225402_at	TP53RK	TP53 regulating kinase	$2.05x10^{-2}$	-1.21
238504_at	C6orf57	chromosome 6 open reading frame 57	$2.06x10^{-2}$	-1.21
213891_s_at	---	CDNA FLJ37747 fis, clone BRHIP2022986	$2.06x10^{-2}$	-1.21
1563118_at	---	CDNA: FLJ20923 fis, clone ADSE00893	$2.08x10^{-2}$	1.23
231894_at	SARS	Seryl-tRNA synthetase	$2.08x10^{-2}$	-1.26
204037_at	EDG2	endothelial differentiation, lysophosphatidic acid G-protein-coupled receptor, 2	$2.09x10^{-2}$	-1.21
219860_at	LY6G5C	lymphocyte antigen 6 complex, locus G5C	$2.10x10^{-2}$	1.27
205126_at	VRK2	vaccinia related kinase 2	$2.11x10^{-2}$	-1.21
205072_s_at	XRCC4	X-ray repair complementing defective repair in Chinese hamster cells 4	$2.12x10^{-2}$	-1.22
206302_s_at	NUDT4 / NUDT4P1	nudix (nucleoside diphosphate linked moiety X)-type motif 4 / -pseudogene 1	$2.12x10^{-2}$	1.38
1560869_a_at	---	Full length insert cDNA clone YQ50C11	$2.13x10^{-2}$	1.27
1555659_a_at	TREML1	triggering receptor expressed on myeloid cells-like 1	$2.15x10^{-2}$	1.72
216693_x_at	HDGFRP3	hepatoma-derived growth factor, related protein 3	$2.15x10^{-2}$	1.39
205391_x_at	ANK1	ankyrin 1, erythrocytic	$2.15x10^{-2}$	1.24
212148_at	PBX1	Pre-B-cell leukemia transcription factor 1	$2.17x10^{-2}$	1.49
214881_s_at	UBTF	upstream binding transcription factor, RNA polymerase I	$2.17x10^{-2}$	-1.25
227180_at	ELOVL7	ELOVL family member 7, elongation of long chain fatty acids (yeast)	$2.18x10^{-2}$	1.78
205336_at	PVALB	parvalbumin	$2.19x10^{-2}$	1.65
210357_s_at	SMOX	spermine oxidase	$2.19x10^{-2}$	1.27
225833_at	LOC221955	KCCR13L	$2.20x10^{-2}$	-1.21
233937_at	ZNF403	zinc finger protein 403	$2.21x10^{-2}$	-1.44
227473_at	CTTN	Cortactin	$2.22x10^{-2}$	1.70
213096_at	TMCC2	transmembrane and coiled-coil domain family 2	$2.22x10^{-2}$	1.30
219994_at	APBB1IP	amyloid beta (A4) precursor protein-binding, family B, member 1 interacting protein	$2.22x10^{-2}$	-1.26
215379_x_at	IGL@ / IGLC1 / IGLC2 / IGLV3-25 / IGLV2-14 / IGLJ3	immunoglobulin lambda locus	$2.23x10^{-2}$	-1.57
1555611_s_at	MBD1	methyl-CpG binding domain protein 1	$2.24x10^{-2}$	-1.26
214677_x_at	IGL@ / IGLC1 / IGLC2 / IGLV3-25 / IGLV2-14 / IGLJ3	immunoglobulin lambda locus	$2.25x10^{-2}$	-1.68
206655_s_at	GP1BB / SEPT5	glycoprotein Ib (platelet), beta polypeptide / septin 5	$2.25x10^{-2}$	1.73
233031_at	ZFHX1B	zinc finger homeobox 1b	$2.26x10^{-2}$	1.34
214455_at	HIST1H2BC	histone 1, H2bc	$2.26x10^{-2}$	1.42
237224_at	---	Transcribed locus	$2.26x10^{-2}$	1.21
1565579_at	---	CDNA clone IMAGE:3689276	$2.27x10^{-2}$	1.51
242727_at	ARL5B	ADP-ribosylation factor-like 5B	$2.28x10^{-2}$	1.30
221942_s_at	GUCY1A3	guanylate cyclase 1, soluble, alpha 3	$2.29x10^{-2}$	1.32
222644_s_at	GLT25D1	glycosyltransferase 25 domain containing 1	$2.31x10^{-2}$	-1.22
216956_s_at	ITGA2B	integrin, alpha 2b (platelet glycoprotein IIb of IIb/IIIa complex, antigen CD41)	$2.32x10^{-2}$	1.74
215850_s_at	NDUFA5	NADH dehydrogenase (ubiquinone) 1 alpha subcomplex, 5, 13kDa	$2.32x10^{-2}$	-1.35
228723_at	NPTN	Neuroplastin	$2.33x10^{-2}$	-1.28
204042_at	WASF3	WAS protein family, member 3	$2.33x10^{-2}$	1.59

243115_at	ITGB3	Integrin, beta 3 (platelet glycoprotein IIIa, antigen CD61)	$2.33x10^{-2}$	1.42
216207_x_at	IGKV1D-13 / LOC649876	immunoglobulin kappa variable 1D-13 / similar to Ig kappa chain V-I region HK1	$2.34x10^{-2}$	-1.45
207414_s_at	PCSK6	proprotein convertase subtilisin/kexin type 6	$2.36x10^{-2}$	1.61
210971_s_at	ARNTL	aryl hydrocarbon receptor nuclear translocator-like	$2.38x10^{-2}$	-1.20
1558750_a_at	---	CDNA FLJ34964 fis, clone NTONG2004095	$2.39x10^{-2}$	1.21
230546_at	VASH1	vasohibin 1	$2.39x10^{-2}$	1.21
202087_s_at	CTSL	cathepsin L	$2.39x10^{-2}$	-1.40
207426_s_at	TNFSF4	tumor necrosis factor (ligand) superfamily, member 4	$2.39x10^{-2}$	1.51
204081_at	NRGN	neurogranin (protein kinase C substrate, RC3)	$2.42x10^{-2}$	1.61
230690_at	TUBB1	tubulin, beta 1	$2.42x10^{-2}$	1.79
219352_at	HERC6	hect domain and RLD 6	$2.43x10^{-2}$	-1.21
206503_x_at	PML	promyelocytic leukemia	$2.43x10^{-2}$	-1.23
1568695_s_at	INTS6	integrator complex subunit 6	$2.43x10^{-2}$	1.21
223984_s_at	NUPL1	nucleoporin like 1	$2.46x10^{-2}$	-1.41
224764_at	ARHGAP21	Rho GTPase activating protein 21	$2.47x10^{-2}$	1.21
238010_at	C1orf174	chromosome 1 open reading frame 174	$2.47x10^{-2}$	-1.28
220334_at	RGS17	regulator of G-protein signalling 17	$2.48x10^{-2}$	-1.20
1563397_at	---	EST from clone 114659, full insert	$2.48x10^{-2}$	1.27
228250_at	KIAA1961	KIAA1961 gene	$2.49x10^{-2}$	-1.22
222142_at	CYLD	cylindromatosis (turban tumor syndrome)	$2.49x10^{-2}$	1.30
213502_x_at	LOC91316	similar to bK246H3.1 (immunoglobulin lambda-like polypeptide 1)	$2.50x10^{-2}$	-1.24
237076_at	---	---	$2.50x10^{-2}$	-1.20
1556153_s_at	NFKBIZ	Nuclear factor of kappa light polypeptide gene enhancer in B-cells inhibitor, zeta	$2.51x10^{-2}$	1.76
76897_s_at	KIAA0674	KIAA0674	$2.52x10^{-2}$	-1.52
241114_s_at	---	Transcribed locus	$2.52x10^{-2}$	-1.26
207594_s_at	SYNJ1	synaptojanin 1	$2.52x10^{-2}$	-1.23
209301_at	CA2	carbonic anhydrase II	$2.53x10^{-2}$	1.51
203711_s_at	HIBCH	3-hydroxyisobutyryl-Coenzyme A hydrolase	$2.53x10^{-2}$	-1.23
244257_at	TMEM104	Transmembrane protein 104	$2.54x10^{-2}$	-1.22
221477_s_at	MGC5618	hypothetical protein MGC5618	$2.54x10^{-2}$	1.29
203414_at	MMD	monocyte to macrophage differentiation-associated	$2.54x10^{-2}$	1.42
219863_at	HERC5	hect domain and RLD 5	$2.56x10^{-2}$	-1.35
229389_at	ATG16L2	ATG16 autophagy related 16-like 2 (*S. cerevisiae*)	$2.56x10^{-2}$	-1.21
207519_at	SLC6A4	solute carrier family 6 (neurotransmitter transporter, serotonin), member 4	$2.56x10^{-2}$	1.21
223631_s_at	C19orf33	chromosome 19 open reading frame 33	$2.57x10^{-2}$	1.62
217257_at	---	(clone B3B3E13) Huntington's disease candidate region mRNA fragment	$2.57x10^{-2}$	-1.32
222869_s_at	ELAC1	elaC homolog 1 (*E. coli*)	$2.57x10^{-2}$	-1.25
238551_at	FUT11	fucosyltransferase 11 (alpha (1,3) fucosyltransferase)	$2.58x10^{-2}$	-1.27
230942_at	CMTM5	CKLF-like MARVEL transmembrane domain containing 5	$2.59x10^{-2}$	1.50
216565_x_at	---	---	$2.60x10^{-2}$	-1.24
238356_at	DOCK11	dedicator of cytokinesis 11	$2.61x10^{-2}$	-1.27
206369_s_at	PIK3CG	phosphoinositide-3-kinase, catalytic, gamma polypeptide	$2.62x10^{-2}$	-1.21
218349_s_at	ZWILCH	Zwilch, kinetochore associated, homolog (Drosophila)	$2.63x10^{-2}$	-1.21
240103_at	---	Full-length cDNA clone CS0DI080YO16 of Placenta Cot 25-normalized of *Homo sapiens*	$2.63x10^{-2}$	2.00
211013_x_at	PML	promyelocytic leukemia	$2.64x10^{-2}$	-1.24
224760_at	SP1	Sp1 transcription factor	$2.64x10^{-2}$	-1.24

205692_s_at	CD38	CD38 molecule	$2.67x10^{-2}$	-1.45
202974_at	MPP1	membrane protein, palmitoylated 1, 55kDa	$2.68x10^{-2}$	1.29
204363_at	F3	coagulation factor III (thromboplastin, tissue factor)	$2.70x10^{-2}$	1.80
210662_at	KYNU	kynureninase (L-kynurenine hydrolase)	$2.71x10^{-2}$	1.28
243423_at	---	Transcribed locus	$2.71x10^{-2}$	1.55
243927_x_at	KIAA1429	KIAA1429	$2.71x10^{-2}$	-1.20
1554447_at	LOC554203	hypothetical LOC554203	$2.74x10^{-2}$	-1.22
217996_at	PHLDA1	pleckstrin homology-like domain, family A, member 1	$2.75x10^{-2}$	1.66
223168_at	RHOU	ras homolog gene family, member U	$2.76x10^{-2}$	-1.29
227110_at	HNRPC	heterogeneous nuclear ribonucleoprotein C (C1/C2)	$2.76x10^{-2}$	1.21
1556583_a_at	---	CDNA FLJ37694 fis, clone BRHIP2015224	$2.79x10^{-2}$	-1.30
208075_s_at	CCL7	chemokine (C-C motif) ligand 7	$2.80x10^{-2}$	1.21
209138_x_at	IGL@	Immunoglobulin lambda locus	$2.80x10^{-2}$	-1.63
222717_at	SDPR	serum deprivation response (phosphatidylserine binding protein)	$2.81x10^{-2}$	1.71
211643_x_at	LOC651961	Myosin-reactive immunoglobulin light chain variable region	$2.81x10^{-2}$	-1.54
216841_s_at	SOD2	superoxide dismutase 2, mitochondrial	$2.82x10^{-2}$	1.30
229054_at	FLJ39779	FLJ39779 protein	$2.82x10^{-2}$	1.25
210504_at	KLF1	Kruppel-like factor 1 (erythroid)	$2.82x10^{-2}$	1.38
228675_at	USP30	Ubiquitin specific peptidase 30	$2.83x10^{-2}$	-1.23
206283_s_at	TAL1	T-cell acute lymphocytic leukemia 1	$2.84x10^{-2}$	1.57
241819_at	---	---	$2.84x10^{-2}$	1.29
207550_at	MPL	myeloproliferative leukemia virus oncogene	$2.86x10^{-2}$	1.71
203595_s_at	IFIT5	interferon-induced protein with tetratricopeptide repeats 5	$2.86x10^{-2}$	-1.23
236016_at	---	CDNA FLJ38419 fis, clone FEBRA2009846	$2.87x10^{-2}$	-1.34
234388_at	TRA@	T cell receptor alpha locus	$2.87x10^{-2}$	1.22
215853_at	SDCCAG8	Serologically defined colon cancer antigen 8	$2.87x10^{-2}$	1.52
238729_x_at	LOC646561	similar to WW45 protein	$2.88x10^{-2}$	1.26
239208_s_at	C21orf57	Chromosome 21 open reading frame 57	$2.88x10^{-2}$	1.48
235068_at	ZDHHC21	zinc finger, DHHC-type containing 21	$2.89x10^{-2}$	-1.23
202947_s_at	GYPC	glycophorin C (Gerbich blood group)	$2.90x10^{-2}$	1.26
218783_at	INTS7	integrator complex subunit 7	$2.90x10^{-2}$	-1.25
215813_s_at	PTGS1	prostaglandin-endoperoxide synthase 1	$2.91x10^{-2}$	1.42
221491_x_at	HLA-DRB1 / HLA-DRB3 / HLA-DRB4	major histocompatibility complex, class II, DR beta	$2.92x10^{-2}$	2.63
221478_at	BNIP3L	BCL2/adenovirus E1B 19kDa interacting protein 3-like	$2.92x10^{-2}$	1.22
212813_at	JAM3	junctional adhesion molecule 3	$2.93x10^{-2}$	1.38
1564002_a_at	C6orf199	chromosome 6 open reading frame 199	$2.94x10^{-2}$	-1.20
211645_x_at	---	Immunoglobulin kappa light chain (IGKV) mRNA variable region, joining region	$2.97x10^{-2}$	-1.61
1554628_at	ZNF57	zinc finger protein 57	$2.97x10^{-2}$	-1.21
1560562_a_at	ZNF677	Zinc finger protein 677	$2.97x10^{-2}$	-1.28
205838_at	GYPA	glycophorin A (MNS blood group)	$2.99x10^{-2}$	1.28
239690_at	---	Transcribed locus	$2.99x10^{-2}$	1.24
201334_s_at	ARHGEF12	Rho guanine nucleotide exchange factor (GEF) 12	$3.00x10^{-2}$	1.26
202283_at	SERPINF1	serpin peptidase inhibitor, clade F (alpha-2 antiplasmin, pigment epithelium derived factor) 1	$3.02x10^{-2}$	-1.29
206116_s_at	TPM1	tropomyosin 1 (alpha)	$3.02x10^{-2}$	1.39
212437_at	CENPB	centromere protein B, 80kDa	$3.03x10^{-2}$	-1.25
234491_s_at	SAV1	salvador homolog 1 (Drosophila)	$3.03x10^{-2}$	1.26

205327_s_at	ACVR2A	activin A receptor, type IIA	$3.03x10^{-2}$	1.29
221425_s_at	HBLD2	HESB like domain containing 2	$3.07x10^{-2}$	1.27
206073_at	COLQ	collagen-like tail subunit (single strand of homotrimer) of asymmetric acetylcholinesterase	$3.08x10^{-2}$	-1.36
234311_s_at	DKFZP686A10121	hypothetical protein	$3.10x10^{-2}$	-1.20
215111_s_at	TSC22D1	TSC22 domain family, member 1	$3.12x10^{-2}$	1.31
201121_s_at	PGRMC1	progesterone receptor membrane component 1	$3.12x10^{-2}$	1.24
204187_at	GMPR	guanosine monophosphate reductase	$3.13x10^{-2}$	1.57
201058_s_at	MYL9	myosin, light polypeptide 9, regulatory	$3.15x10^{-2}$	2.26
206157_at	PTX3	pentraxin-related gene, rapidly induced by IL-1 beta	$3.15x10^{-2}$	1.65
215121_x_at	IGL@ / IGLC1 / IGLC2 / IGLV4-3 / IGLV3-25 / IGLV2-14	immunoglobulin lambda locus	$3.15x10^{-2}$	-1.56
244631_at	LOC389834 / LOC642398	hypothetical gene supported by AK123403 / hypothetical protein LOC642398	$3.17x10^{-2}$	-1.48
235965_at	BHLHB8	Basic helix-loop-helix domain containing, class B, 8	$3.18x10^{-2}$	-1.27
204625_s_at	ITGB3	integrin, beta 3 (platelet glycoprotein IIIa, antigen CD61)	$3.18x10^{-2}$	1.29
201560_at	CLIC4	chloride intracellular channel 4	$3.18x10^{-2}$	1.32
204972_at	OAS2	2'-5'-oligoadenylate synthetase 2, 69/71kDa	$3.19x10^{-2}$	-1.32
229748_x_at	LOC389830	similar to Tektin-3	$3.23x10^{-2}$	-1.42
232298_at	LOC401093	hypothetical LOC401093	$3.24x10^{-2}$	1.24
235962_at	---	Transcribed locus	$3.26x10^{-2}$	-1.25
201739_at	SGK	serum/glucocorticoid regulated kinase	$3.26x10^{-2}$	1.22
230903_s_at	C8orf42	Chromosome 8 open reading frame 42	$3.27x10^{-2}$	1.30
1569203_at	CXCL2	chemokine (C-X-C motif) ligand 2	$3.27x10^{-2}$	1.37
211810_s_at	GALC	galactosylceramidase	$3.27x10^{-2}$	-1.21
207433_at	IL10	interleukin 10	$3.27x10^{-2}$	-1.26
238018_at	LOC285016	hypothetical protein LOC285016	$3.28x10^{-2}$	-1.26
218435_at	DNAJC15	DnaJ (Hsp40) homolog, subfamily C, member 15	$3.29x10^{-2}$	-1.21
228693_at	CCDC50	coiled-coil domain containing 50	$3.32x10^{-2}$	-1.23
227394_at	NCAM1	Neural cell adhesion molecule 1	$3.33x10^{-2}$	-1.48
203148_s_at	TRIM14	tripartite motif-containing 14	$3.33x10^{-2}$	-1.21
207113_s_at	TNF	tumor necrosis factor (TNF superfamily, member 2)	$3.33x10^{-2}$	2.24
206210_s_at	CETP	cholesteryl ester transfer protein, plasma	$3.34x10^{-2}$	1.24
244523_at	MMD	monocyte to macrophage differentiation-associated	$3.35x10^{-2}$	1.28
215159_s_at	NADK	NAD kinase	$3.35x10^{-2}$	-1.27
226319_s_at	THOC4	THO complex 4	$3.35x10^{-2}$	-1.42
208406_s_at	GRAP2	GRB2-related adaptor protein 2	$3.36x10^{-2}$	1.31
213338_at	RIS1	Ras-induced senescence 1	$3.36x10^{-2}$	1.51
225782_at	MSRB3	methionine sulfoxide reductase B3	$3.39x10^{-2}$	1.42
218543_s_at	PARP12	poly (ADP-ribose) polymerase family, member 12	$3.39x10^{-2}$	-1.20
204627_s_at	ITGB3	integrin, beta 3 (platelet glycoprotein IIIa, antigen CD61)	$3.40x10^{-2}$	1.91
202145_at	LY6E	lymphocyte antigen 6 complex, locus E	$3.41x10^{-2}$	-1.39
213023_at	UTRN	utrophin (homologous to dystrophin)	$3.44x10^{-2}$	-1.21
208636_at	ACTN1	actinin, alpha 1	$3.44x10^{-2}$	1.21
236772_s_at	---	Transcribed locus	$3.44x10^{-2}$	-1.21
206553_at	OAS2	2'-5'-oligoadenylate synthetase 2, 69/71kDa	$3.46x10^{-2}$	-1.26
202975_s_at	RHOBTB3	Rho-related BTB domain containing 3	$3.47x10^{-2}$	-1.40
205896_at	SLC22A4	solute carrier family 22 (organic cation transporter), member 4	$3.48x10^{-2}$	1.32
201295_s_at	WSB1	WD repeat and SOCS box-containing 1	$3.48x10^{-2}$	-1.33

218276_s_at	SAV1	salvador homolog 1 (Drosophila)	$3.49x10^{-2}$	1.23
212573_at	ENDOD1	endonuclease domain containing 1	$3.52x10^{-2}$	1.24
229778_at	C12orf39	Chromosome 12 open reading frame 39	$3.53x10^{-2}$	1.89
219998_at	HSPC159	HSPC159 protein	$3.54x10^{-2}$	1.35
201925_s_at	CD55	CD55 molecule, decay accelerating factor for complement (Cromer blood group)	$3.54x10^{-2}$	1.21
224589_at	XIST	X (inactive)-specific transcript	$3.58x10^{-2}$	-1.64
233078_at	API5	apoptosis inhibitor 5	$3.59x10^{-2}$	-1.21
226047_at	MRVI1	Murine retrovirus integration site 1 homolog	$3.59x10^{-2}$	-1.31
222600_s_at	UBE1L2	ubiquitin-activating enzyme E1-like 2	$3.60x10^{-2}$	-1.22
229843_at	FAM82B	Family with sequence similarity 82, member B	$3.61x10^{-2}$	-1.42
211138_s_at	KMO	kynurenine 3-monooxygenase (kynurenine 3-hydroxylase)	$3.61x10^{-2}$	-1.32
243802_at	DNHD2	dynein heavy chain domain 2	$3.62x10^{-2}$	1.23
204786_s_at	IFNAR2	interferon (alpha, beta and omega) receptor 2	$3.62x10^{-2}$	-1.21
222573_s_at	SAV1	salvador homolog 1 (Drosophila)	$3.63x10^{-2}$	1.20
235417_at	SPOCD1	SPOC domain containing 1	$3.64x10^{-2}$	1.25
204505_s_at	EPB49	erythrocyte membrane protein band 4.9 (dematin)	$3.64x10^{-2}$	1.47
205904_at	MICA	MHC class I polypeptide-related sequence A	$3.64x10^{-2}$	-1.20
216925_s_at	TAL1	T-cell acute lymphocytic leukemia 1	$3.66x10^{-2}$	1.27
1560485_at	HIVEP1	human immunodeficiency virus type I enhancer binding protein 1	$3.66x10^{-2}$	1.28
203725_at	GADD45A	growth arrest and DNA-damage-inducible, alpha	$3.70x10^{-2}$	1.26
236242_at	---	---	$3.71x10^{-2}$	-1.23
219382_at	SERTAD3	SERTA domain containing 3	$3.72x10^{-2}$	-1.26
231532_at	NCAM1	Neural cell adhesion molecule 1	$3.72x10^{-2}$	-1.25
226188_at	HSPC159	HSPC159 protein	$3.72x10^{-2}$	1.74
233819_s_at	ZNF294	zinc finger protein 294	$3.73x10^{-2}$	-1.36
236213_at	HNRPA3	Heterogeneous nuclear ribonucleoprotein A3	$3.73x10^{-2}$	1.49
219269_at	HMBOX1	homeobox containing 1	$3.75x10^{-2}$	1.29
208601_s_at	TUBB1	tubulin, beta 1	$3.77x10^{-2}$	1.77
212531_at	LCN2	lipocalin 2 (oncogene 24p3)	$3.79x10^{-2}$	1.79
232155_at	KIAA1618	KIAA1618	$3.80x10^{-2}$	-1.26
212946_at	RP11-125A7.3	KIAA0564 protein	$3.81x10^{-2}$	1.20
215223_s_at	SOD2	superoxide dismutase 2, mitochondrial	$3.81x10^{-2}$	1.43
228084_at	CASP6	Caspase 6, apoptosis-related cysteine peptidase	$3.82x10^{-2}$	1.31
226364_at	HIP1	Huntingtin interacting protein 1	$3.83x10^{-2}$	-1.29
235490_at	TMEM107	transmembrane protein 107	$3.84x10^{-2}$	1.63
208653_s_at	CD164	CD164 molecule, sialomucin	$3.84x10^{-2}$	-1.22
236470_at	---	Transcribed locus	$3.85x10^{-2}$	-1.36
208352_x_at	ANK1	ankyrin 1, erythrocytic	$3.87x10^{-2}$	1.23
202254_at	SIPA1L1	Signal-induced proliferation-associated 1 like 1	$3.87x10^{-2}$	-1.21
200923_at	LGALS3BP	lectin, galactoside-binding, soluble, 3 binding protein	$3.88x10^{-2}$	-1.38
1552950_at	C15orf26	chromosome 15 open reading frame 26	$3.89x10^{-2}$	1.25
1552554_a_at	CARD12	caspase recruitment domain family, member 12	$3.89x10^{-2}$	-1.46
201539_s_at	FHL1	four and a half LIM domains 1	$3.92x10^{-2}$	1.34
1556332_at	---	CDNA FLJ38412 fis, clone FEBRA2009385	$3.93x10^{-2}$	1.30
219159_s_at	SLAMF7	SLAM family member 7	$3.93x10^{-2}$	-1.32
240646_at	GIMAP8	GTPase, IMAP family member 8	$3.94x10^{-2}$	-1.27
242206_at	---	---	$3.94x10^{-2}$	-1.21
226278_at	DKFZp313A2432	hypothetical protein DKFZp313A2432	$3.94x10^{-2}$	1.25
237625_s_at	---	Immunoglobulin light chain variable region complementarity determining region	$3.95x10^{-2}$	-1.36
219983_at	HRASLS	HRAS-like suppressor	$3.97x10^{-2}$	1.45

205978_at	KL	klotho	$3.98x10^{-2}$	-1.27
210455_at	C10orf28	chromosome 10 open reading frame 28	$3.98x10^{-2}$	-1.25
239186_at	MGC39372	hypothetical protein MGC39372	$3.99x10^{-2}$	1.56
242335_at	SLC25A37	solute carrier family 25, member 37	$4.00x10^{-2}$	1.57
215498_s_at	MAP2K3	mitogen-activated protein kinase kinase 3	$4.01x10^{-2}$	1.22
224277_at	MOP-1 / LOC649524	MOP-1 /// hypothetical protein LOC649524	$4.02x10^{-2}$	1.38
231996_at	N4BP2	Nedd4 binding protein 2	$4.02x10^{-2}$	-1.31
204838_s_at	MLH3	mutL homolog 3 (*E. coli*)	$4.03x10^{-2}$	1.49
221383_at	NMUR1	neuromedin U receptor 1	$4.03x10^{-2}$	-1.28
208436_s_at	IRF7	interferon regulatory factor 7	$4.04x10^{-2}$	-1.29
217157_x_at	---	Immunoglobulin kappa chain, V-region (SPK.3)	$4.05x10^{-2}$	-1.46
201315_x_at	IFITM2	interferon induced transmembrane protein 2 (1-8D)	$4.06x10^{-2}$	-1.21
203817_at	---	---	$4.08x10^{-2}$	1.42
225842_at	PHLDA1	pleckstrin homology-like domain, family A, member 1	$4.09x10^{-2}$	1.48
202984_s_at	BAG5	BCL2-associated athanogene 5	$4.10x10^{-2}$	-1.22
218711_s_at	SDPR	serum deprivation response (phosphatidylserine binding protein)	$4.11x10^{-2}$	1.66
223669_at	HEMGN	hemogen	$4.16x10^{-2}$	1.64
228376_at	GGTA1	Glycoprotein, alpha-galactosyltransferase 1	$4.16x10^{-2}$	1.59
222347_at	LOC644450	hypothetical protein LOC644450	$4.17x10^{-2}$	-1.42
221542_s_at	SPFH2	SPFH domain family, member 2	$4.17x10^{-2}$	-1.21
219787_s_at	ECT2	epithelial cell transforming sequence 2 oncogene	$4.20x10^{-2}$	-1.26
206034_at	SERPINB8	serpin peptidase inhibitor, clade B (ovalbumin), member 8	$4.22x10^{-2}$	1.21
201695_s_at	NP	nucleoside phosphorylase	$4.25x10^{-2}$	1.26
206271_at	TLR3	toll-like receptor 3	$4.26x10^{-2}$	-1.24
202869_at	OAS1	2',5'-oligoadenylate synthetase 1, 40/46kDa	$4.26x10^{-2}$	-1.46
201616_s_at	CALD1	caldesmon 1	$4.29x10^{-2}$	1.21
207206_s_at	ALOX12	arachidonate 12-lipoxygenase	$4.32x10^{-2}$	1.70
236898_at	---	Transcribed locus	$4.32x10^{-2}$	1.57
210613_s_at	SYNGR1	synaptogyrin 1	$4.32x10^{-2}$	-1.24
237315_at	---	Transcribed locus	$4.33x10^{-2}$	1.21
217999_s_at	PHLDA1	pleckstrin homology-like domain, family A, member 1	$4.33x10^{-2}$	1.50
216412_x_at	---	Clone ds1-1 immunoglobulin lambda chain VJ region, (IGL)	$4.34x10^{-2}$	-1.27
235940_at	C9orf64	chromosome 9 open reading frame 64	$4.35x10^{-2}$	-1.27
207165_at	HMMR	hyaluronan-mediated motility receptor (RHAMM)	$4.35x10^{-2}$	-1.29
217216_x_at	MLH3	mutL homolog 3 (*E. coli*)	$4.36x10^{-2}$	1.34
215739_s_at	TUBGCP3	tubulin, gamma complex associated protein 3	$4.37x10^{-2}$	-1.25
244443_at	LOC440309	Hypothetical LOC440309	$4.39x10^{-2}$	-1.30
205209_at	ACVR1B	activin A receptor, type IB	$4.39x10^{-2}$	-1.25
210933_s_at	FSCN1	fascin homolog 1, actin-bundling protein (*Strongylocentrotus purpuratus*)	$4.39x10^{-2}$	1.21
235683_at	SESN3	sestrin 3	$4.40x10^{-2}$	1.44
202388_at	RGS2	regulator of G-protein signalling 2, 24kDa	$4.40x10^{-2}$	-1.22
240182_at	HLA-A	Major histocompatibility complex, class I, A	$4.40x10^{-2}$	-1.24
227724_at	LOC642351	hypothetical protein LOC642351	$4.40x10^{-2}$	1.28
210255_at	RAD51L1	RAD51-like 1 (S. cerevisiae)	$4.41x10^{-2}$	-1.20
243124_at	---	---	$4.41x10^{-2}$	-1.26
241881_at	OR2W3	olfactory receptor, family 2, subfamily W, member 3	$4.42x10^{-2}$	1.95
229751_s_at	PUS7L	pseudouridylate synthase 7 homolog (*S. cerevisiae*)-like	$4.42x10^{-2}$	-1.24
211141_s_at	CNOT3	CCR4-NOT transcription complex, subunit 3	$4.43x10^{-2}$	-1.26
210299_s_at	FHL1	four and a half LIM domains 1	$4.43x10^{-2}$	1.44
212843_at	NCAM1	neural cell adhesion molecule 1	$4.44x10^{-2}$	-1.60

232530_at	LOC652226	hypothetical protein LOC652226	$4.45x10^{-2}$	1.55
1570362_at	---	Homo sapiens, Similar to LOC153081, clone IMAGE:5749586, mRNA	$4.45x10^{-2}$	1.29
205476_at	CCL20	chemokine (C-C motif) ligand 20	$4.46x10^{-2}$	2.32
1554152_a_at	OGDH	oxoglutarate (alpha-ketoglutarate) dehydrogenase (lipoamide)	$4.48x10^{-2}$	-1.31
208353_x_at	ANK1	ankyrin 1, erythrocytic	$4.48x10^{-2}$	1.24
201660_at	ACSL3	Acyl-CoA synthetase long-chain family member 3	$4.49x10^{-2}$	-1.24
205016_at	TGFA	transforming growth factor, alpha	$4.50x10^{-2}$	-1.26
222043_at	CLU	clusterin	$4.51x10^{-2}$	1.39
1558102_at	TM6SF1	Transmembrane 6 superfamily member 1	$4.52x10^{-2}$	1.30
223386_at	FAM118B	family with sequence similarity 118, member B	$4.54x10^{-2}$	-1.22
208180_s_at	HIST1H4H	histone 1, H4h	$4.56x10^{-2}$	1.31
216576_x_at	---	Rearranged Ig kappa light chain variable region (I.114)	$4.57x10^{-2}$	-1.55
240065_at	FAM81B	family with sequence similarity 81, member B	$4.57x10^{-2}$	1.27
202767_at	ACP2	acid phosphatase 2, lysosomal	$4.59x10^{-2}$	-1.23
203940_s_at	VASH1	vasohibin 1	$4.59x10^{-2}$	1.38
231630_at	FLJ16341	Hypothetical gene supported by AK122786	$4.59x10^{-2}$	1.52
223717_s_at	ACRBP	acrosin binding protein	$4.60x10^{-2}$	1.44
209347_s_at	MAF	v-maf musculoaponeurotic fibrosarcoma oncogene homolog (avian)	$4.61x10^{-2}$	-1.24
208022_s_at	CDC14B	CDC14 cell division cycle 14 homolog B (*S. cerevisiae*)	$4.62x10^{-2}$	1.36
218847_at	IGF2BP2	insulin-like growth factor 2 mRNA binding protein 2	$4.63x10^{-2}$	1.30
206390_x_at	PF4	platelet factor 4 (chemokine (C-X-C motif) ligand 4)	$4.63x10^{-2}$	1.52
1559097_at	C14orf64	chromosome 14 open reading frame 64	$4.64x10^{-2}$	1.24
33494_at	ETFDH	electron-transferring-flavoprotein dehydrogenase	$4.64x10^{-2}$	-1.21
1565717_s_at	FUS	fusion (involved in t(12;16) in malignant liposarcoma)	$4.67x10^{-2}$	-1.24
37028_at	PPP1R15A	protein phosphatase 1, regulatory (inhibitor) subunit 15A	$4.67x10^{-2}$	1.50
201160_s_at	CSDA	cold shock domain protein A	$4.67x10^{-2}$	1.21
218039_at	NUSAP1	nucleolar and spindle associated protein 1	$4.67x10^{-2}$	-1.26
37201_at	ITIH4	inter-alpha (globulin) inhibitor H4 (plasma Kallikrein-sensitive glycoprotein)	$4.67x10^{-2}$	-1.26
208792_s_at	CLU	clusterin	$4.70x10^{-2}$	1.57
242020_s_at	ZBP1	Z-DNA binding protein 1	$4.70x10^{-2}$	-1.24
230082_at	LRRFIP2	Leucine rich repeat (in FLII) interacting protein 2	$4.71x10^{-2}$	1.24
221185_s_at	IQCG	IQ motif containing G	$4.72x10^{-2}$	-1.24
228108_at	---	CDNA clone IMAGE:5263177	$4.72x10^{-2}$	-1.32
242051_at	---	Transcribed locus	$4.74x10^{-2}$	1.25
206254_at	EGF	epidermal growth factor (beta-urogastrone)	$4.76x10^{-2}$	1.35
200665_s_at	SPARC	secreted protein, acidic, cysteine-rich (osteonectin)	$4.77x10^{-2}$	1.62
209273_s_at	HBLD2	HESB like domain containing 2	$4.77x10^{-2}$	1.20
227678_at	XRCC6BP1	XRCC6 binding protein 1	$4.77x10^{-2}$	-1.22
206964_at	CML2	putative N-acetyltransferase Camello 2	$4.79x10^{-2}$	1.31
209524_at	HDGFRP3	hepatoma-derived growth factor, related protein 3	$4.79x10^{-2}$	1.38
220173_at	C14orf45	chromosome 14 open reading frame 45	$4.79x10^{-2}$	1.31
1556669_a_at	---	Full length insert cDNA clone YR71G12	$4.83x10^{-2}$	1.28
215118_s_at	IGHG1	Immunoglobulin heavy constant gamma 1 (G1m marker)	$4.83x10^{-2}$	-1.67
1559249_at	ATXN1	Ataxin 1	$4.84x10^{-2}$	1.44
1558234_at	FLJ36644	hypothetical gene supported by AK093963	$4.86x10^{-2}$	1.66
215698_at	JARID1A	Jumonji, AT rich interactive domain 1A (RBBP2-like)	$4.86x10^{-2}$	-1.24
31874_at	GAS2L1	growth arrest-specific 2 like 1	$4.94x10^{-2}$	1.26
235843_at	DGKG	Diacylglycerol kinase, gamma 90kDa	$4.99x10^{-2}$	1.29
242787_at	---	Transcribed locus	$4.99x10^{-2}$	-1.21

Table D.2: Significant differentially expressed groups of genes according their molecular and cellular functions. Results were obtained using the IPA software.

Category	p-value	Molecules
Cancer	$3.80x10^{-12}$ - $1.64x10^{-3}$	ACVR1B, ACVR2A, ADI1, ADORA2A, AHR, AIM2, ALOX12, ANKRD9, ARG2, ARHGAP18, ARHGEF12, ATXN1, BARD1, BCL2L11, BLM, BMP6, BNIP3L, C13ORF15, CA2, CALD1, CBR1, CD38, CD55, CLCN5, CLU, COL6A3, CREB1, CTDSPL, CTSL1, CTTN, CXCL2, CYLD, DNAJC15, ECT2, EGF, EIF4B, EP300, F3, FDXR, FOXO3, FSCN1, FUS, GADD45A, GART, GFI1B, GNAZ, GNG11, GRIN1, HIP1, HIST4H4, HLA-B, HMMR, HNRNPC, HSPA4, IFI16, IFITM2, IFNAR2, IFNGR1, IGHG1, IGHM, IGJ, IGKC, IGL@, IL10, IP6K2, ISG15, ITGB3, ITGB5, ITIH4, JAM3, JARID1A, KIAA1211, KL, LCN2, LEPR, LGALS3BP, LPAR1, LTBP1, LY6E, MAF, MAP2K3, MAX, MBD1, MBNL1, MFAP3L, MFN1, MGC13057, MGLL, MLH3, MPL, MRVI1, MXI1, MYL9, MYLK, NCAM1, NDST1, NFKBIZ, NLRC4, NP, PARVB, PBX1, PCSK6, PCYT1B, PDE5A, PDGFA, PDIA6, PF4, PHLDA1, PIK3CG, PLD1, PLOD2, PML, PPP1R15A, PPP1R3D, PPP2R1B, PRKAR2B, PROSC, PTCRA, PTGS1, PTPRN2, PTX3, PUS7L, RAD50, RAD51L1, RNF144B, RNF160, RPE, RSU1, RUNX1, SDPR, SELP, SERPINB2, SERPINF1, SGK1, SLC6A4, SMAD5, SMOX, SOD2, SP1, SPARC, SPSB1, ST8SIA4, STAT2, SYK, SYNGR1, TAL1, TBXA2R, TCF4, TFPI, TGFA, TGFB1I1, TLR3, TNF, TNFRSF17, TNFRSF18, TNFSF8, TP53, TP53RK, TPM1, TUBB1, UBTF, VASH1, VIM, VRK2, ZEB2
Cell Death	$4.02x10^{-12}$ - $8.08x10^{-4}$	ACVR1B, ACVR2A, ADI1, ADORA2A, AHR, AIM2, ALOX12, ARG2, ATXN1, BAG5, BARD1, BCL2L11, BLM, BMP6, BNIP3L, BRCC3, CA2, CCL7, CD300A, CD38, CD55, CLIC4, CLU, CREB1, CSDA, CXCL2, CYLD, DDX58, DNAJC15, EGF, EP300, F13A1, F3, FDXR, FKBP1B, FOSL2, FOXO3, FUS, GADD45A, GFI1B, GRAP2, GRIN1, HIP1, HLA-B, HMMR, HNRNPC, IFI16, IFNAR2, IFNGR1, IGHG1, IGHM, IGKC, IL10, IP6K2, IRF9, ISG15, ITGB3, ITGB5, KL, KLF1, LCN2, LEPR, LGALS3BP, LMNA, LPAR1, MAF, MAP2K3, MAX, MBD1, MFN1, MGC29506, MPL, MYLK, NCAM1, NCOA2, NDST1, NFKBIZ, NGFRAP1, NLRC4, NP, NR1H3, NRGN, OAS1, OGDH, PARVB, PCSK6, PDGFA, PF4, PFKFB2, PHLDA1, PIK3CG, PLAC8, PLD1, PML, PPM1A, PPP1R15A, PPP1R3D, PPP2R1B, PRKAR2B, PSME3, PTCRA, PTGS1, PTPRN2, RAD50, RNASE2, RUNX1, SDC4, SEPT5, SERPINB2, SERPINF1, SGK1, SIAH2, SLAMF7, SLC8A1, SMAD5, SMOX, SOD2, SP1, SPARC, ST8SIA4, STAT2, SYK, TAL1, TBXA2R, TCF4, TGFA, TGFB1I1, TLR3, TMEM158, TNF, TNFRSF17, TNFRSF18, TNFSF8, TP53, TPM1, TPM3, TRIP10, TXNDC5, UBTF, USP18, XRCC4, ZEB2

Cellular Growth and Proliferation	$1.46x10^{-10}$ - $9.10x10^{-4}$	ACTN1, ACVR1B, ACVR2A, ADORA2A, AGPS, AHR, AIM2, ALOX12, BARD1, BCL2L11, BLM, BMP6, BNIP3L, BRAP, CBR1, CCL20, CD164, CD38, CD55, CENPB, CLU, COL6A3, COMMD5, CREB1, CSDA, CTDSPL, CTSL1, CXCL2, CXCL5, CYLD, EGF, EIF1AY, EIF4B, EP300, ETFDH, F3, FDXR, FKBP1B, FOSL2, FOXO3, FSCN1, FUS, GADD45A, GFI1B, GOLGA2, GRAP2, GRB14, GUCY1A3, GUCY1B3, HIP1, HMMR, HNRNPC, HRASLS, IFI16, IFNAR2, IFNGR1, IGHG1, IGHM, IGKC, IL10, IP6K2, ISG15, ITGA2B, ITGB3, ITGB5, KLF1, LCN2, LEFTY1, LEPR, LMNA, LPAR1, LTBP1, LY6E, MAP2K3, MAX, MBD1, MPL, MXI1, MYL9, NAP1L1, NCAM1, NCOA2, NCSTN, NGFRAP1, NP, NR1H3, OSR2, PARVB, PBX1, PCYT1B, PDGFA, PF4, PHLDA1, PIK3CG, PLAC8, PLD1, PML, PPM1A, PPP1R15A, PRKAR2B, PTCRA, PTGS1, PTX3, PUS7L, RGS2, RUNX1, SERPINB2, SERPINF1, SERTAD3, SGK1, SH3BP2, SLAMF7, SLC6A4, SMAD5, SMOX, SOD2, SP1, SPARC, SYK, TAL1, TBXA2R, TCF4, TFPI, TGFA, TGFB1I1, TLR3, TNF, TNFRSF17, TNFRSF18, TNFSF4, TNFSF8, TP53, TPM1, TPM3, TSC22D1, UBC, UBTF, VASH1, VIM, XRCC4, ZEB2
Hematological Disease	$1.99x10^{-8}$ - $1.60x10^{-3}$	ADORA2A, AHR, ARHGEF12, BCL2L11, BNIP3L, CALD1, CCL7, CD55, CLU, CTSL1, CXCL2, EP300, EPB49, F13A1, F3, FOSL2, FOXO3, FUS, GADD45A, GFI1B, GP9, GRAP2, GYPA, GYPC, IFNAR2, IFNGR1, IGHM, IL10, ITGA2B, ITGB3, ITGB5, KLF1, LCN2, LEPR, LGALS3BP, MAP2K3, MAX, MFN1, MPL, MXI1, NCAM1, NLRC4, NP, PBX1, PDE5A, PDGFA, PDIA6, PF4, PIK3CG, PLD1, PML, PPP1R15A, PROS1, PSME3, PTCRA, PTGS1, PTX3, RAD50, RNASE2, RUNX1, SELP, SLC6A4, SLC8A1, SOD2, SYK, TAL1, TBXA2R, TFPI, TGFA, TNF, TNFRSF17, TNFRSF18, TNFSF4, TNFSF8, TP53, TUBB1
Cellular Movement	$3.30x10^{-8}$ - $1.07x10^{-3}$	ABHD2, ABI3, ADI1, ADORA2A, ALOX12, BMP6, CALD1, CCL20, CCL7, CD38, CD55, CLIC4, CLU, CMKLR1, CTSL1, CTTN, CXCL2, CXCL5, DAB1, EGF, ESAM, F3, FOSL2, FOXO3, FSCN1, GADD45A, GNAZ, GOLGA2, GP9, GRIN1, GUCY1A3, GUCY1B3, HMMR, IGHG1, IL10, ITGA2B, ITGB3, ITGB5, JAM3, LCN2, LEFTY1, LPAR1, MAP2K3, MAX, MGLL, MYLK, NCAM1, NDST1, NFKBIZ, PARVB, PCSK6, PDGFA, PF4, PIK3CG, PLD1, RNASE2, SDC4, SELP, SERPINB2, SOD2, SP1, SPARC, ST8SIA4, SYK, TBXA2R, TFPI, TGFA, TLR3, TNF, TNFRSF18, TNFSF4, TP53, TPM1, TPM3, VASH1, VIM, WASF3, ZEB2
Hematological System Development and Function	$4.23x10^{-8}$ - $1.36x10^{-3}$	ACVR1B, ACVR2A, ADORA2A, AHR, ALOX12, ANK1, ARNTL, BCL2L11, BLM, CCL20, CCL7, CD300A, CD38, CD55, CLEC4C, CLU, CMKLR1, CREB1, CXCL2, CXCL5, CYLD, EGF, EP300, F13A1, F3, FKBP1B, FOXO3, FUS, GADD45A, GFI1B, GNAZ, GP9, GRAP2, GUCY1A3, GUCY1B3, HIP1, HIST4H4, HSPA4, IFI16, IFNAR2, IFNGR1, IGHG1, IGHM, IGKC, IL10, IRF7, ISG15, ITGA2B, ITGB3, JAM3, JARID1D, KL, KLF1, LCN2, LEPR, LTBP1, MAF, MAP2K3, MAX, MPL, MYLK, NDST1, NFKBIZ, NP, PBX1, PF4, PIK3CG, PLAC8, PML, PROS1, PTCRA, PTX3, RGS2, RNASE2, RUNX1, SELP, SEPT5, SH3BP2, SLAMF7, SLC8A1, SMAD5, SOD2, SYK, TAL1, TBXA2R, TCF4, TFPI, TLR3, TNF, TNFRSF17, TNFRSF18, TNFSF4, TNFSF8, TP53, TRIP10, XRCC4

Genetic Disorder	$6.72x10^{-8}$ - $1.62x10^{-3}$	ABHD2, ACP2, ACSBG1, ACTN1, ACVR1B, ACVR2A, ADORA2A, AGPS, AHR, ALOX12, ANK1, ANKRD28, ANKRD9, ARHGAP10, ARHGAP18, ARHGEF12, ARHGEF9, ARNTL, ATL1, ATP8B4, ATXN1, BARD1, BCL2L11, BLM, BMP6, BRAP, BRCC3, C10ORF28, C13ORF15, C14ORF181, C14ORF64, C15ORF26, C6ORF199, C6ORF57, C8ORF42, CA2, CALD1, CCDC132, CCDC50, CCL20, CCL7, CD300A, CD38, CETP, CHD2, CLCN5, CLDN5, CLEC2B, CLIC4, CLIP4, CLU, COL13A1, COL6A3, COLQ, CREB1, CSDA, CTDSPL, CTSL1, CXCL2, CXCL5, CYLD, DAB1, DDX58, DNAH12, DNAJC15, DNM3, ECT2, EGF, EIF1B, EIF3C, EIF4B, ELOVL7, ENDOD1, EP300, ETFDH, F13A1, F3, FAM69B, FAM81B, FBXO22, FBXO38, FDXR, FHL1, FKBP15, FNDC3B, FOSL2, FOXO3, FSCN1, GADD45A, GALC, GART, GGNBP2, GNAZ, GOSR2, GP9, GRAP2, GRIN1, GTF2E1, GUCY1A3, GUCY1B3, GYPA, GYPC, HDGFRP3, HEATR3, HERC6, HGD, HIBCH, HIP1, HIVEP1, HLA-B, HMBOX1, HMMR, HNRNPC, IFITM2, IFNAR2, IFNGR1, IGF2BP2, IGHD, IGHG1, IGHM, IGJ, IGKC, IGL@, IL10, INTS7, ISG1, ITGA2B, ITGB3, ITGB5, ITIH4, JAM3, JARID1A, KIAA0564, KIAA1211, KIAA1618, KL, KLHDC4, KMO, KYNU, LCN2, LEPR, LMNA, LPAR1, LTBP1, LY6E, LY6G5C, LY6G6D, MAF, MAMLD1, MBD1, MBNL1, MFN1, MGC39372, MGC9913, MGLL, MLH3, MMD, MPL, MSRB3, MXI1, MYL9, MYLK, NCAM1, NCOA2, NCSTN, NDUFA5, NFKBIZ, NGFRAP1, NLRC4, NP, NR1H3, NRGN, OAS1, OAS2, OGDH, OSR2, PAM, PARP12, PARVB, PBX1, PCSK6, PDE5A, PDGFA, PGRMC1, PIK3CG, PLD1, PLOD2, PML, PPP2R1B, PRKAR2B, PROS1, PRTFDC1, PTGS1, PTPRN2, PVALB, RAB38, RAD51L1, RARS2, RASGEF1B, RGS17, RHOBTB3, RNASE2, RNF160, RUFY3, RUNX1, RYR1, SDC4, SELP, SERPINB8, SESN3, SETP5, SGK1, SH3BP2, SLC22A4, SLC25A37, SLC35D3, SLC6A4, SLC8A1, SLFN5, SMAD5, SNAP29, SOD2, SP1, SPARC, SRR, SSFA2, SYK, SYNGR1, SYNJ1, SYTL4, TBC1D30, TBXA2R, TCF4, TGFA, TIPARP, TK2, TLR3, TMCC2, TNF, TNFRSF17, TNFSF4, TNFSF8, TP53, TPM1, TPM3, TRIM14, TRIM69, TTC7B, TUBB1, TUBGCP3, UBA6, USP12, USP9Y, UTRN, VEPH1, VIM, VRK2, WASF3, WDR41, XIST, ZCCHC6, ZEB2, ZNF385D
Organismal Survival	$7.12x10^{-8}$ - $6.93x10^{-4}$	ADORA2A, AHR, ALOX12, BARD1, BCL2L11, BLM, BNIP3L, CREB1, CXCL2, DDX58, EP300, F13A1, F3, FKBP1B, FOXO3, GADD45A, GALC, GFI1B, GNAZ, GUCY1B3, HIP1, HNRNPC, IFI16, IFNGR1, IGHG1, IGHM, IL10, ITGA2B, ITGB3, KL, KLF1, LPAR1, LY6E, MAX, MFN1, MXI1, NCOA2, NCSTN, NP, PBX1, PCSK6, PDGFA, PIK3CG, PTGS1, PTX3, RAD50, SDC4, SMAD5, SOD2, SP1, ST8SIA4, STAT2, SYK, TAL1, TGFA, TLR3, TNF, TNFRSF18, TP53, TPM3, UBA6, USP18, UTRN, XRCC4
Infectious Disease	$7.92x10^{-8}$ - $1.40x10^{-3}$	ACTN1, ALOX12, ANKRD9, ARHGAP18, ARHGEF12, ATG16L2, BCL2L11, C1ORF174, CA2, CALD1, CCL20, CCL7, CD164, CD55, CLU, CMKLR1, CTSL1, CXCL5, DDX58, EGF, EIF1AY, EP300, F3, GOSR2, GRIN1, GTF2H2, GUCY1A3, GUCY1B3, GYPA, GYPC, HIBCH, IFNAR2, IFNGR1, IGHM, IL10, INTS6, INTS7, ITGA2B, ITGB3, JARID1D, LCN2, LEFTY1, MGLL, MPL, PARVB, PCSK6, PDE5A, PDIA6, PF4, PLOD2, PML, PROS1, PTGS1, PTPRN2, PTX3, RNASE2, RPS4Y1, SLC6A4, SPARC, SPCS3, STT3A, TAL1, TFPI, TLR3, TNF, TNFRSF18, TNFSF4, TUBB1, UTRN, UTY, WASF3, ZDHHC21
Tissue Morphology	$9.40x10^{-8}$ - $1.34x10^{-3}$	ACVR2A, ADORA2A, AHR, ARG2, ARNTL, BCL2L11, BLM, BMP6, CALD1, CCL7, CD38, CREB1, CXCL2, CYLD, EGF, EP300, FKBP1B, FOSL2, GRAP2, HRASLS, IFNGR1, IGHM, IGJ, IGKC, IL10, ISG15, ITGA2B, ITGB3, JAM3, KL, LEPR, LPAR1, MAP2K3, MPL, MYL9, MYLK, NP, PBX1, PCYT1B, PDGFA, PF4, PIK3CG, PML, PTCRA, PTGS1, RUNX1, RYR1, SDC4, SELP, SLC8A1, SMAD5, SOD2, SPARC, ST8SIA4, TAL1, TBXA2R, TCF4, TGFA, TLR3, TNF, TNFSF4, TP53, TPM1, UTRN

Cell-To-Cell Signaling and Interaction	$3.77x10^{-7}$ - $1.36x10^{-3}$	ADORA2A, AHR, ALOX12, BMP6, CALD1, CCL20, CCL7, CD164, CD300A, CD38, CD55, CLEC4C, CLU, CXCL2, CXCL5, DAB1, EGF, ESAM, F13A1, F3, FOXO3, GFI1B, GP9, GRIN1, HSPA4, IFI16, IFNAR2, IFNGR1, IGHA1, IGHG1, IGHM, IGKC, IGL@, IL10, ITGA2B, ITGB3, ITGB5, JAM3, LCN2, LEPR, LGALS3BP, LTBP1, MAF, MAP2K3, MXI1, MYLK, NCAM1, NDST1, PARVB, PDGFA, PF4, PHLDA1, PIK3CG, PLD1, PPP1R15A, PRKAR2B, PROS1, PTX3, RNASE2, RSU1, SDC4, SELP, SERPINB2, SERPINF1, SLC6A4, SOD2, SPARC, ST8SIA4, SYK, TBXA2R, TFPI, TGFA, TGFB1I1, TLR3, TNF, TNFRSF17, TNFSF4, TNFSF8, TP53, TRIP10, VIM, XRCC4
Cellular Function and Maintenance	$3.77x10^{-7}$ - $1.18x10^{-3}$	ADORA2A, BMP6, CCL20, CLU, EGF, F3, FOXO3, GRAP2, HIP1, IFI16, IGHA1, IGHG1, IGHM, IL10, ITGB3, ITGB5, LTBP1, MXI1, PF4, PIK3CG, PML, PPP1R15A, PRKAR2B, PROS1, PSME3, PTX3, RSU1, STAT2, SYK, TGFA, TLR3, TNF, TNFSF8, TP53
Tissue Development	$4.04x10^{-7}$ - $9.10x10^{-4}$	ACVR1B, ACVR2A, ADORA2A, AHR, ALOX12, BCL2L11, BMP6, CA2, CCL20, CCL7, CD164, CD38, CD55, CLU, COL6A3, CXCL2, DAB1, EGF, EP300, ESAM, F3, FOSL2, GNAZ, GP9, GUCY1A3, GUCY1B3, HIP1, HMMR, IFI16, IFNGR1, IGHA1, IGHM, IL10, ITGA2B, ITGB3, ITGB5, JAM3, KL, LCN2, LGALS3BP, LPAR1, LY6E, MAF, MAX, MBNL1, MPL, NCAM1, NCSTN, NDST1, PARVB, PBX1, PCSK6, PDGFA, PF4, PHLDA1, PIK3CG, PLD1, PML, PTX3, RUNX1, SELP, SEPT5, SERPINB2, SERPINF1, SH3BP2, SMAD5, SPARC, ST8SIA4, SYK, TAL1, TCF4, TFPI, TGFA, TGFB1I1, TLR3, TNF, TNFSF4, TP53, TPM1, UTRN, ZEB2
Cellular Development	$4.87x10^{-7}$ - $1.34x10^{-3}$	ACVR1B, ACVR2A, ADORA2A, AHR, ALOX12, ANK1, ARHGEF9, BCL2L11, BLM, BMP6, CA2, CCL7, CD164, CD300A, CD38, CD55, CLIC4, CLU, CMKLR1, CREB1, CTTN, CXCL2, ECT2, EGF, EP300, ESAM, FNDC3B, FOSL2, FOXO3, FUS, GADD45A, GFI1B, GNAZ, GRB14, HEMGN, HIP1, HIST4H4, IFI16, IFITM2, IGHG1, IGHM, IL10, IRF7, ITGA2B, ITGB3, ITGB5, JARID1D, KLF1, LCN2, LEPR, LMNA, LPAR1, LTBP1, MAF, MAP2K3, MAX, MBNL1, MPL, NCAM1, NCOA2, NDST1, NP, PARVB, PBX1, PCSK6, PDGFA, PF4, PIK3CG, PLAC8, PLD1, PML, PRKAR2B, PTCRA, REC8, RGS2, RHOU, RNASE2, RUNX1, SDC4, SERPINB2, SERPINF1, SIAH2, SMAD5, SOD2, SP1, SPARC, ST8SIA4, SYK, TAL1, TCF4, TGFA, TGFB1I1, TLR3, TNF, TNFSF4, TNFSF8, TP53, TPM1, VASH1, XRCC4
Gene Expression	$1.10x10^{-6}$ - $1.49x10^{-3}$	ABLIM3, ACVR1B, ACVR2A, ADORA2A, AHR, APBB1IP, ARNTL, ATXN1, BARD1, BLM, BMP6, CHD2, CLU, CREB1, CSDA, CTDSPL, CXCL2, DDX58, ECT2, EGF, EP300, FOSL2, FOXO3, FUS, GADD45A, GFI1B, GRIN1, GTF2E1, HIP1, HIVEP1, IFI16, IFNAR2, IGF2BP2, IL10, IP6K2, IRF7, IRF9, ITGB3, JARID1A, KLF1, LEPR, LMNA, LPAR1, MAF, MAP2K3, MAX, MBD1, MGC29506, MPL, MXI1, NAP1L1, NCAM1, NCOA2, NFKBIZ, NR1H3, PBX1, PDGFA, PIK3CG, PML, PPM1A, PRKAR2B, RNF11, RP5-1000E10.4, RUNX1, SERTAD3, SGK1, SH3BP2, SMAD5, SOD2, SP1, STAT2, SYK, TAL1, TCF4, TGFA, TGFB1I1, THOC4, TLR3, TNF, TNFRSF17, TP53, TSC22D1, UBTF, XIST, ZBP1
Organismal Development	$1.53x10^{-6}$ - $1.34x10^{-3}$	ACVR1B, ADORA2A, ALOX12, BARD1, BCL2L11, BLM, BMP6, BRCC3, CD164, CLU, CREB1, CSDA, DDX3Y, DDX58, EGF, EP300, F3, FOSL2, GMPR, GTF2H2, HIP1, HMMR, HNRNPC, HSPA4, IFNGR1, IGHG1, IL10, ITGB3, JARID1D, KLF1, LCN2, LEPR, LMNA, LY6E, MAX, NCOA2, NCSTN, NDST1, NGFRAP1, PBX1, PCSK6, PDGFA, PIK3CG, PML, PSME3, PTGS1, RPS4Y1, RUNX1, SELP, SERPINF1, SMAD5, SOD2, SP1, ST8SIA4, SYNJ1, TGFB1I1, TLR3, TNF, TNFRSF17, TP53, TPM1, TRIM14, USP9Y, UTY, XRCC4, ZEB2

Organismal Injury and Abnormalities	$1.93x10^{-6}$ - $1.53x10^{-3}$	ADORA2A, AHR, ATXN1, BCL2L11, BMP6, BNIP3L, CA2, CBR1, CLU, CREB1, DDX58, EGF, EP300, F13A1, F3, FHL1, GRIN1, GUCY1A3, GUCY1B3, HLA-B, HMMR, IFI35, IFNAR2, IFNGR1, IGHG1, IL10, IRF7, IRF9, ITGA2B, ITGB3, LMNA, MLH3, MPL, MXI1, NCOA2, NFKBIZ, OAS1, OAS2, PCYT1B, PDE5A, PF4, PIK3CG, PTGS1, SDC4, SELP, SLC6A4, SLC8A1, SMAD5, SOD2, SPARC, STAT2, TGFA, TLR3, TNF, TNIP1, TP53, TRIM69, TUBB1, USP18
Cell Cycle	$2.70x10^{-6}$ - $8.46x10^{-4}$	ACVR1B, AHR, BARD1, BCL2L11, BLM, BMP6, BRCC3, C13ORF15, CD38, CENPB, CLU, COMMD5, CREB1, CYLD, ECT2, EGF, EP300, FOSL2, FOXO3, GADD45A, GFI1B, GOLGA2, GYPA, IFI16, IGHG1, IGHM, IL10, IRF7, IRF9, ITGB3, LMNA, MAP2K3, MAX, MLH3, MPL, MXI1, NUSAP1, PLAC8, PML, PPM1A, PPP1R15A, PROS1, RAD50, RAD51L1, REC8, RHOU, RUNX1, SEPT5, SGK1, SH3BP2, SOD2, SP1, SPARC, SYK, TGFA, TNF, TP53, TPM1, TUBB1, XRCC4
Neurological Disease	$3.34x10^{-6}$ - $1.55x10^{-3}$	ACP2, ACVR2A, ADORA2A, ALOX12, ANK1, ARHGAP10, ARHGAP18, ARHGEF9, ARNTL, ATL1, ATP8B4, ATXN1, BAG5, BARD1, BCL2L11, BMP6, BRCC3, C14ORF64, CA2, CCDC132, CCDC50, CD38, CLDN5, CLIP4, CLU, COL13A1, COL6A3, COLQ, CREB1, CSDA, CXCL2, DAB1, DDX58, DNAJC15, DNM3, EGF, EIF3C, ENDOD1, EP300, F13A1, F3, FBXO38, FDXR, FHL1, FNDC3B, FOXO3, GADD45A, GALC, GART, GNAZ, GRAP2, GRIN1, GUCY1B3, GYPC, HDGFRP3, HERC6, HIP1, HIVEP1, HLA-B, HMBOX1, HNRNPC, IFNAR2, IFNGR1, IGHG1, IGHM, IGKC, IGL@, IL10, ITGA2B, ITGB3, ITGB5, ITIH4, JAM3, JARID1A, KIAA1211, KL, KMO, LCN2, LMNA, LPAR1, LTBP1, MGC29506, MGC39372, MGLL, MLH3, MMD, MYLK, NCAM1, NCSTN, NDUFA5, NGFRAP1, NLRC4, NRGN, PARVB, PBX1, PCSK6, PDE5A, PDGFA, PF4, PGRMC1, PLOD2, PML, PRKAR2B, PROS1, PTGS1, PTPRN2, PVALB, RAB38, RARS2, RASGEF1B, RGS17, RHOBTB3, RNASE2, RUFY3, RUNX1, RYR1, SDC4, SEPT5, SERPINF1, SESN3, SGK1, SLC22A4, SLC25A37, SLC35D3, SLC6A4, SMAD5, SNAP29, SOD2, SP1, SPARC, SRR, ST8SIA4, STAT2, SYK, SYNGR1, SYNJ1, SYTL4, TCF4, TGFA, TIPARP, TLR3, TMCC2, TNF, TNFSF4, TP53, TPM3, TRIP10, TTC7B, TUBB1, UBA6, USP18, VIM, XRCC4, ZEB2, ZNF385D
Reproductive System Disease	$3.44x10^{-6}$ - $1.50x10^{-3}$	ADI1, AHR, AIM2, ALOX12, BARD1, BCL2L11, BNIP3L, C13ORF15, CETP, CLCN5, CLU, COL6A3, CREB1, CTDSPL, CTSL1, CXCL5, CYLD, DNAJC15, ECT2, EGF, EIF4B, EP300, F3, FOSL2, FOXO3, FSCN1, GADD45A, GART, GNG11, GRIN1, GUCY1B3, HIP1, HLA-B, HMMR, IFI16, IFITM2, IFNAR2, IFNGR1, IGHM, IGJ, IGKC, IGL@, IL10, IP6K2, ISG15, ITGB3, ITGB5, KIAA1211, LCN2, LEPR, LPAR1, LTBP1, LY6E, MAF, MAMLD1, MAX, MFAP3L, MGC13057, MRVI1, MXI1, MYL9, PCSK6, PCYT1B, PDE5A, PDGFA, PML, PPP1R3D, PPP2R1B, PRKAR2B, PROSC, PTGS1, PTPRN2, RAD51L1, RHOU, RNF144B, RNF160, RPE, RUNX1, SDPR, SERPINF1, SGK1, SLC6A4, SMAD5, SOD2, SP1, SPARC, SPSB1, SRPRB, TAL1, TCF4, TGFA, TLR3, TNF, TNFRSF17, TP53, TPM1, TUBB1, USP9Y, VIM, XIST

Cardiovascular Disease	$4.42x10^{-6}$ - $1.38x10^{-3}$	ACSBG1, ADORA2A, AGPS, AHR, ANK1, ARG2, BARD1, BLM, BMP6, BNIP3L, BSPRY, CA2, CCDC132, CCDC50, CCL20, CCL7, CETP, CLU, CMKLR1, CREB1, CTTN, DAB1, EIF4B, EP300, F13A1, F3, FBXO22, FKBP1B, FOXO3, GOSR2, GRIN1, GUCY1A3, GUCY1B3, GYPC, HDGFRP3, HEATR3, HIBCH, HIP1, HMBOX1, IFNGR1, IGHG1, IL10, ITGA2B, ITGB3, ITIH4, KIAA0564, KL, KMO, LCN2, LEPR, LMNA, LTBP1, MBNL1, MGLL, MMD, MSRB3, MYLK, NR1H3, OGDH, PBX1, PDE5A, PDGFA, PIK3CG, PLD1, PPP1R15A, PROS1, PRTFDC1, PTGS1, PTPRN2, RAD50, RAD51L1, RGS2, RNASE2, RNF160, SELP, SERPINB8, SLC25A37, SLC6A4, SLC8A1, SOD2, SYK, TBXA2R, TFPI, TGFA, TNF, TNFSF4, TNIP1, TPM1, TRIM14, TUBB1, UBA6, UBTF, VEPH1, WASF3, ZNF385D, ZNF710
Cell-mediated Immune Response	$4.76x10^{-6}$ - $1.57x10^{-3}$	ABHD2, ABI3, ADORA2A, AHR, AIM2, ALOX12, ANK1, BCL2L11, BLM, BMP6, CALD1, CCL20, CCL7, CD164, CD300A, CD38, CD55, CLEC4C, CLU, CMKLR1, CREB1, CTSL1, CTTN, CXCL2, CXCL5, CYLD, DDX58, EGF, EP300, F3, FKBP1B, FOSL2, FOXO3, FSCN1, FUS, GADD45A, GFI1B, GNAZ, GP9, GRAP2, GRIN1, HIST4H4, HLA-B, HMMR, HSPA4, IFITM2, IFNAR2, IFNGR1, IGHG1, IGHM, IL10, IRF7, ITGA2B, ITGB3, ITGB5, JAM3, JARID1D, KLF1, LCN2, LEPR, LPAR1, LTBP1, MAF, MAP2K3, MGLL, MYLK, NCAM1, NDST1, NFKBIZ, NLRC4, NP, OAS1, PARVB, PBX1, PCSK6, PDGFA, PF4, PIK3CG, PLD1, PML, PROS1, PTCRA, PTGS1, PTX3, RGS2, RNASE2, RUNX1, SDC4, SELP, SERPINF1, SH3BP2, SOD2, SPARC, STAT2, SYK, TAL1, TCF4, TGFA, TGFB1I1, TLR3, TNF, TNFRSF17, TNFRSF18, TNFSF4, TNFSF8, TNIP1, TP53, TPM1, TPM3, VIM, XRCC4
Inflammatory Disease	$6.32x10^{-6}$ - $1.36x10^{-3}$	ACSBG1, ACVR1B, ADORA2A, AGPS, AHR, ALOX12, ANK1, ANKRD28, ARG2, ARHGAP18, ATXN1, BCL2L11, BNIP3L, C13ORF15, C14ORF181, C14ORF64, C6ORF199, CA2, CALD1, CCL7, CD55, CLEC2B, CLU, CMKLR1, COLQ, CTSL1, CXCL2, CXCL5, CYLD, DAB1, DDX58, DNAH12, DNM3, EGF, EIF1B, EP300, F13A1, F3, FBXO22, FOSL2, FOXO3, GADD45A, GART, GRIN1, GUCY1A3, HDGFRP3, HIP1, HIVEP1, HLA-B, HMBOX1, HMMR, IFNAR2, IFNGR1, IGHD, IGHG1, IGHM, IGJ, IGKC, IGL@, IL10, INTS7, ITGB3, KIAA0564, KIAA1211, KIAA1618, KL, KYNU, LEPR, LMNA, LTBP1, LY6G5C, LY6G6D, MBD1, MGC9913, MPL, MXI1, MYLK, NCOA2, NCSTN, NFKBIZ, OAS2, PAM, PBX1, PDE5A, PDGFA, PIK3CG, PLD1, PRKAR2B, PTGS1, PTPRN2, RAD51L1, RUNX1, SELP, SERPINB2, SLC22A4, SLC25A37, SLC6A4, SLC8A1, SLFN5, SOD2, SSFA2, TBC1D30, TCF4, TLR3, TNF, TNFRSF18, TNFSF4, TNFSF8, TP53, TRIM69, TUBB1, UTRN, VIM, WDR41, ZCCHC6, ZEB2
Renal and Urological Disease	$6.32x10^{-6}$ - $1.33x10^{-3}$	ADORA2A, AHR, CA2, CLCN5, CLU, CTSL1, EGF, EP300, GADD45A, GRIN1, IFNGR1, IL10, KL, MYL9, PDE5A, PTGS1, PUS7L, PVALB, SELP, SGK1, SLC6A4, SLC8A1, TGFA, TLR3, TNF, TP53, TRIM69
Infection Mechanism	$8.28x10^{-6}$ - $7.71x10^{-4}$	CD38, DDX58, EGF, EP300, IGKC, IL10, IRF9, KLF1, NCOA2, NR1H3, PLD1, PML, RP5-1000E10.4, SP1, STAT2, TGFB1I1, TNF, TNIP1, TP53

Inflammatory Response	$8.79x10^{-6}$ - $1.07x10^{-3}$	ABHD2, ADORA2A, AIM2, ALOX12, ARNTL, BCL2L11, BMP6, CCL20, CCL7, CD164, CD300A, CD38, CD55, CLEC4C, CMKLR1, CTSL1, CXCL2, CXCL5, DDX58, EGF, F3, FOXO3, GFI1B, GNAZ, GP9, GRAP2, GUCY1A3, GUCY1B3, HLA-B, HSPA4, IFITM2, IFNAR2, IFNGR1, IGHA1, IGHG1, IGHM, IL10, IRF7, ITGA2B, ITGB3, JAM3, LCN2, LEPR, MAF, MAP2K3, MGLL, MPL, MYLK, NDST1, NFKBIZ, NLRC4, NP, OAS1, PF4, PIK3CG, PML, PROS1, PTGS1, PTX3, RGS2, RNASE2, SELP, SEPT5, SERPINF1, SH3BP2, STAT2, SYK, TAL1, TLR3, TNF, TNFRSF17, TNFRSF18, TNFSF4, TNIP1, TP53
Skeletal and Muscular Disorders	$1.23x10^{-5}$ - $1.34x10^{-3}$	ACP2, ACSBG1, ACVR1B, ACVR2A, ADORA2A, ANK1, ANKRD28, ARHGAP10, ATXN1, BCL2L11, C13ORF15, C14ORF181, CA2, CALD1, CCDC132, CCL7, CD38, CLCN5, CLEC2B, CLIP4, CLU, COL13A1, COL6A3, COLQ, CREB1, CXCL2, CXCL5, DAB1, DNM3, EGF, EIF1B, F13A1, F3, FHL1, FOXO3, GADD45A, GART, GRIN1, GUCY1B3, GYPC, HDGFRP3, HIP1, HIVEP1, HLA-B, HMBOX1, HMMR, IFNAR2, IFNGR1, IGHD, IGHG1, IGHM, IGKC, IGL@, IL10, ITGA2B, ITGB3, ITIH4, KIAA1618, KL, LCN2, LEPR, LMNA, LTBP1, LY6G5C, LY6G6D, MAP2K3, MAX, MBD1, MGC39372, MGC9913, NCOA2, NDUFA5, NLRC4, NRGN, PAM, PBX1, PGRMC1, PIK3CG, PLAC8, PLD1, PLOD2, PRKAR2B, PTGS1, PTPRN2, PVALB, RAD51L1, RHOBTB3, RNASE2, RUFY3, RUNX1, RYR1, SDC4, SELP, SEPT5, SGK1, SLC22A4, SLC25A37, SLC35D3, SLC6A4, SLFN5, SMAD5, SNAP29, SOD2, SP1, SPARC, SYNJ1, SYTL4, TBC1D30, TCF4, TGFA, TLR3, TNF, TNFRSF18, TNFSF4, TP53, TPM3, TRIM69, TUBB1, UBA6, USP18, UTRN, VIM, ZCCHC6
Gastrointestinal Disease	$1.37x10^{-5}$ - $1.47x10^{-3}$	EGF, EP300, FOXO3, GADD45A, MXI1, TGFA, TNF, TP53, EGF, EP300, GADD45A, TNF, TP53, CD55, MYLK, SLC6A4, TNF, EGF, FOXO3, MXI1, PPP1R15A, ACVR2A, ADORA2A, AHR, AIM2, CA2, CLU, COL6A3, CREB1, FDXR, FSCN1, FUS, IFNAR2, IL10, ITGB3, ITGB5, JARID1A, LCN2, LEPR, PML, PTGS1, SLC6A4, TNF, TP53, TUBB1, VIM, VRK2, AHR, EP300, GADD45A, IL10, NR1H3, PML, TGFA, TNF, BCL2L11, CLU, EGF, EP300, FDXR, FOXO3, GADD45A, PPP1R15A, SGK1, SPARC, TNF, TP53
Antigen Presentation	$1.46x10^{-5}$ - $1.34x10^{-3}$	ABHD2, ADORA2A, AIM2, ARNTL, BCL2L11, BMP6, CCL20, CCL7, CD164, CD300A, CD38, CLEC4C, CMKLR1, CTSL1, CXCL2, CXCL5, DDX58, F3, FOXO3, GFI1B, GNAZ, GRAP2, HLA-B, HSPA4, IFITM2, IFNAR2, IFNGR1, IGHG1, IGHM, IL10, IRF7, LCN2, LEPR, MAF, MAP2K3, MGLL, MYLK, NFKBIZ, NLRC4, NP, OAS1, PF4, PIK3CG, PROS1, PTGS1, PTX3, RGS2, RNASE2, SELP, SERPINF1, STAT2, SYK, TLR3, TNF, TNFRSF17, TNFSF4, TNIP1, TP53
Humoral Immune Response	$1.46x10^{-5}$ - $1.33x10^{-3}$	ABHD2, ADORA2A, AIM2, BCL2L11, BMP6, CCL20, CCL7, CD164, CD300A, CD38, CLEC4C, CMKLR1, CTSL1, CXCL2, CXCL5, DDX58, FOXO3, GFI1B, GNAZ, HLA-B, HSPA4, IFITM2, IFNAR2, IFNGR1, IGHG1, IGHM, IL10, IRF7, LCN2, LEPR, MAF, MAP2K3, MGLL, MYLK, NFKBIZ, NLRC4, NP, OAS1, PF4, PIK3CG, PROS1, PTGS1, PTX3, RGS2, RNASE2, SELP, SERPINF1, SH3BP2, STAT2, SYK, TCF4, TLR3, TNF, TNFRSF17, TNFSF4, TNFSF8, TNIP1, TP53
Tumor Morphology	$1.46x10^{-5}$ - $1.62x10^{-3}$	AHR, BMP6, CBR1, CLU, CREB1, CTDSPL, EGF, EP300, FUS, GADD45A, HIP1, HSPA4, IFNAR2, IGHG1, IL10, ITGB3, ITGB5, JAM3, MXI1, NDST1, PBX1, PCYT1B, RUNX1, SERPINB2, TGFA, TNF, TNFRSF18, TP53, TPM1, VASH1
Connective Tissue Development and Function	$1.84x10^{-5}$ - $6.87x10^{-4}$	AHR, ARHGAP10, EGF, FOSL2, GADD45A, HMMR, ITGB3, ITGB5, LGALS3BP, LPAR1, NCSTN, PDGFA, PF4, PML, SLC6A4, SOD2, SPARC, TBXA2R, TGFA, TGFB1I1, TNF, TP53, TPM1, XRCC4

Hematopoiesis	$1.89x10^{-5}$ - $9.10x10^{-4}$	ACVR1B, ACVR2A, ANK1, BCL2L11, BLM, CD300A, CREB1, EGF, EP300, FOXO3, FUS, GFI1B, GRAP2, HIST4H4, IFI16, IFNGR1, IGHG1, IGHM, IL10, IRF7, ITGB3, JARID1D, KLF1, LCN2, LTBP1, MAF, MAX, NP, PBX1, PF4, PIK3CG, PML, PTCRA, RUNX1, SELP, SMAD5, SOD2, SYK, TAL1, TCF4, TLR3, TNF, TNFSF4, TNFSF8, TP53, XRCC4
Lymphoid Tissue Structure and Development	$1.89x10^{-5}$ - $4.70x10^{-5}$	ADORA2A, AHR, ANK1, BCL2L11, CREB1, ECT2, EGF, EP300, FOXO3, FUS, GFI1B, HIST4H4, IGHG1, IGHM, IL10, ITGB3, ITGB5, JAM3, JARID1D, KLF1, LCN2, LEPR, LTBP1, MAF, NP, PBX1, PDGFA, PF4, PIK3CG, PML, PTX3, RUNX1, SELP, SERPINF1, SMAD5, SOD2, SPARC, SYK, TAL1, TBXA2R, TCF4, TGFA, TLR3, TNF, TNFSF4, TNFSF8, TP53, VASH1, XRCC4
Respiratory Disease	$2.50x10^{-5}$ - $9.61x10^{-4}$	ACTN1, ADORA2A, ARG2, BCL2L11, CCL7, CD164, CLU, CREB1, EGF, EIF1AY, EP300, F3, FDXR, FOSL2, GADD45A, GRIN1, IFNGR1, IGJ, IL10, ITGA2B, ITGB3, JARID1D, KL, LCN2, NLRC4, NR1H3, PDE5A, PDGFA, PIK3CG, PML, PTGS1, PTX3, RNASE2, RPS4Y1, SLC6A4, SLC8A1, SPARC, TAL1, TNF, TP53
Cardiovascular System Development and Function	$2.85x10^{-5}$ - $1.05x10^{-3}$	ADORA2A, AHR, ALOX12, BMP6, ECT2, EGF, F3, GUCY1A3, GUCY1B3, HMMR, IGHG1, IL10, ITGB3, ITGB5, JAM3, LEPR, PDGFA, PF4, PIK3CG, PML, PTGS1, PTX3, RUNX1, SELP, SERPINF1, SMAD5, SPARC, TAL1, TBXA2R, TFPI, TGFA, TNF, TP53, VASH1
Nervous System Development and Function	$3.01x10^{-5}$ - $9.10x10^{-4}$	ADORA2A, AHR, CBR1, CREB1, EGF, GRIN1, HIP1, IL10, ITGB3, LMNA, NCAM1, SERPINF1, SGK1, SLC6A4, SNAP29, SP1, ST8SIA4, SYNGR1, TGFA, TNF
Immune Cell Trafficking	$3.11x10^{-5}$ - $1.20x10^{-3}$	ALOX12, CCL20, CCL7, CD300A, CD38, CD55, CLEC4C, CLU, CMKLR1, CXCL2, CXCL5, F13A1, F3, FOXO3, GFI1B, GNAZ, HSPA4, IGHG1, IGHM, IL10, ITGB3, JAM3, LCN2, MAF, MAP2K3, MYLK, NDST1, NFKBIZ, PF4, PIK3CG, PML, PROS1, PTX3, RNASE2, SELP, SH3BP2, SYK, TAL1, TLR3, TNF, TNFRSF17, TNFRSF18, TNFSF4, TP53, ADORA2A, AHR, CCL20, CCL7, CD55, CXCL5, FOXO3, IGHG1
Cell Morphology	$3.58x10^{-5}$ - $1.57x10^{-3}$	AHR, ALOX12, ARHGAP10, BCL2L11, BLM, CLIC4, CREB1, CTTN, ECT2, EGF, FDXR, FOXO3, GADD45A, IGHM, IL10, ITGA2B, ITGB3, ITGB5, KL, KLF1, LCN2, LEPR, LPAR1, MAP2K3, MPL, PAM, PARVB, PBX1, PCSK6, PDGFA, PHLDA1, PIK3CG, PLA2G12A, PLD1, PRKAR2B, PSME3, RGS2, SDC4, SIPA1L1, SOD2, SPARC, SYK, TBXA2R, TCF4, TGFA, TGFB1I1, TNF, TP53, TPM1, TPM3, TUBB1, VIM
Skeletal and Muscular System Development and Function	$3.80x10^{-5}$ - $6.87x10^{-4}$	AHR, ARG2, CALD1, EGF, FKBP1B, MAP2K3, MYL9, MYLK, PIK3CG, PTGS1, RYR1, SDC4, SLC8A1, SMAD5, TBXA2R, TNF, TPM1, UTRN
Cell Signaling	$6.35x10^{-5}$ - $1.68x10^{-3}$	ADORA2A, ARG2, ARHGEF12, CCL20, CCL7, CD300A, CD38, CLEC4C, CMKLR1, CXCL2, CXCL5, DDAH2, EGF, F3, FKBP1B, GP9, GRAP2, GRIN1, HIP1, IFNGR1, IGHM, IGKC, IL10, ITGA2B, ITGB3, KL, LPAR1, LTBP1, MAP2K3, MRVI1, NCAM1, NMUR1, NRGN, PIK3CG, PLD1, PTCRA, PTX3, RGS17, RGS2, SELP, SH3BP2, SLC8A1, SOD2, SYK, TBXA2R, TGFA, TLR3, TNF, TP53
Molecular Transport	$6.35x10^{-5}$ - $1.41x10^{-3}$	ADORA2A, ARHGEF12, ARNTL, CCL20, CCL7, CD300A, CD38, CLEC4C, CMKLR1, CXCL2, CXCL5, EGF, F3, FKBP1B, GP9, GRAP2, GRIN1, GUCY1A3, GUCY1B3, HIP1, IGHM, IGKC, IL10, ITGA2B, ITGB3, KL, LPAR1, LTBP1, MRVI1, NCAM1, NMUR1, NRGN, PDE5A, PIK3CG, PLD1, PML, PTCRA, RGS17, RGS2, SELP, SH3BP2, SLC8A1, SOD2, SYK, TBXA2R, TGFA, TLR3, TNF, TP53

Vitamin and Mineral Metabolism	$6.35x10^{-5}$ - $1.41x10^{-3}$	ADORA2A, ARHGEF12, CCL20, CCL7, CD300A, CD38, CLEC4C, CMKLR1, CXCL2, CXCL5, EGF, F3, FKBP1B, GP9, GRAP2, GRIN1, HIP1, IGHM, IGKC, IL10, ITGA2B, ITGB3, KL, LPAR1, LTBP1, MRVI1, NCAM1, NMUR1, NRGN, PIK3CG, PLD1, PTCRA, RGS17, RGS2, SELP, SH3BP2, SLC8A1, SYK, TBXA2R, TGFA, TLR3, TNF, TP53
Psychological Disorders	$6.42x10^{-5}$ - $1.36x10^{-3}$	ARNTL, ATXN1, CLDN5, EGF, GNAZ, GRIN1, IL10, LEPR, NCAM1, SLC6A4, SNAP29, SOD2, SRR, SYNGR1, SYNJ1, TCF4, TNF, TP53
Visual System Development and Function	$7.04x10^{-5}$ - $1.04x10^{-4}$	EGF, IGHG1, MAF, PDGFA, SERPINF1, TGFA, TNF, TP53
Embryonic Development	$7.58x10^{-5}$ - $1.34x10^{-3}$	ACVR1B, BLM, CREB1, EGF, IL10, KLF1, MAX, PBX1, RAD50, RUNX1, SMAD5, SP1, TGFA, TNF, TP53, TPM3, XRCC4
Dermatological Diseases and Conditions	$9.73x10^{-5}$ - $9.10x10^{-4}$	ALOX12, BLM, BNIP3L, CCL20, CCL7, CD55, CLIC4, CLU, COL13A1, COL6A3, EGF, FOXO3, GADD45A, GRIN1, HLA-B, IFNAR2, IGHD, IGHG1, IGHM, IGKC, IL10, ITGB3, ITGB5, LMNA, MGC29506, MXI1, NCOA2, NFKBIZ, NRGN, PDE5A, PHLDA1, PIK3CG, PPM1A, PPP1R15A, PROS1, PTGS1, SELP, SERPINF1, SGK1, SLC6A4, SNAP29, TCF4, TGFA, TGFB1I1, TNF, TNF, TNFSF8, TP53, TP53, TP53, TP53, TUBB1
Organ Development	$1.04x10^{-4}$ - $6.87x10^{-4}$	IGHG1, EGF, PDGFA, SERPINF1, TGFA, TNF, TP53
Lipid Metabolism	$1.72x10^{-4}$ - $1.62x10^{-3}$	EGF, IGHM, IL10, IP6K2, NCAM1, PIK3CG, PLD1, PTGS1, RGS2, SYNJ1, TGFA, TLR3, TNF, VIM
Small Molecule Biochemistry	$1.72x10^{-4}$ - $1.62x10^{-3}$	ABI3, ACVR1B, ARG2, DDAH2, EGF, EP300, GUCY1A3, GUCY1B3, IFNGR1, IGHM, IL10, IP6K2, LEPR, MYLK, NCAM1, PDE5A, PDGFA, PIK3CG, PLD1, PTGS1, PTX3, RGS2, SGK1, SOD2, SRR, SYK, SYNJ1, TGFA, TLR3, TNF, VIM
Hypersensitivity Response	$1.75x10^{-4}$	CCL7, IL10, PIK3CG, TLR3, TNF
Cellular Assembly and Organization	$1.92x10^{-4}$ - $9.10x10^{-4}$	ARHGEF12, CALD1, CLU, ECT2, EGF, FHL1, ITGA2B, ITGB3, LPAR1, MYLK, PAM, PLD1, SDC4, SERPINF1, SGK1, SIPA1L1, SOD2, SPARC, TGFA, TNF, TP53, TPM1, TPM3, TRIP10, TUBGCP3, VIM, WASF3
Drug Metabolism	$1.92x10^{-4}$	EGF, IGHM, PTGS1, TGFA, TNF, VIM
Antimicrobial Response	$2.07x10^{-4}$	DDX58, IFNAR2, IFNGR1, IRF7, LCN2, OAS1, TLR3, TNF
Cellular Compromise	$3.38x10^{-4}$ - $6.87x10^{-4}$	EGF, ITGA2B, ITGB3, MYLK, TNF
Hepatic System Development and Function	$3.38x10^{-4}$ - $5.80x10^{-4}$	EGF, TGFA, TNF
Ophthalmic Disease	$3.38x10^{-4}$	SERPINF1, TNF, TP53
Connective Tissue Disorders	$3.71x10^{-4}$ - $1.34x10^{-3}$	ACSBG1, ACVR1B, ANK1, ANKRD28, BCL2L11, C13ORF15, C14ORF181, CALD1, CCL7, CLEC2B, CLU, COL13A1, COL6A3, COLQ, CXCL2, CXCL5, DAB1, EGF, EIF1B, F13A1, FOXO3, GART, HIP1, HIVEP1, HLA-B, HMBOX1, HMMR, IFNAR2, IFNGR1, IGHD, IGHG1, IGHM, IGKC, IGL@, IL10, KIAA1618, KL, LEPR, LMNA, LTBP1, LY6G5C, LY6G6D, MBD1, MGC9913, NCOA2, PAM, PBX1, PIK3CG, PLD1, PLOD2, PRKAR2B, PTGS1, PTPRN2, RAD51L1, RUNX1, SELP, SERPINB2, SLC22A4, SLC6A4, SLFN5, SOD2, TBC1D30, TCF4, TGFA, TLR3, TNF, TNFRSF18, TNFSF4, TP53, TRIM69, TUBB1, VIM, ZCCHC6
Hepatic System Disease	$3.80x10^{-4}$ - $1.62x10^{-3}$	ADORA2A, AHR, BCL2L11, CA2, CLU, CREB1, EGF, EP300, GADD45A, GART, IFNAR2, IFNGR1, IL10, ISG15, ITIH4, MPL, NR1H3, PML, PTGS1, SLC6A4, TGFA, TNF, TP53, TUBB1

Endocrine System Disorders	$6.91x10^{-4}$ - $1.20x10^{-3}$	ABHD2, ACTN1, AHR, ALOX12, ARNTL, ATP8B4, ATXN1, BARD1, BLM, BRAP, C10ORF28, C14ORF181, C15ORF26, C6ORF57, C8ORF42, CA2, CCDC50, CD38, CETP, CHD2, CLEC2B, CLIC4, CLU, COL13A1, COLQ, CREB1, CSDA, CTSL1, DAB1, DNM3, ECT2, EGF, ELOVL7, ENDOD1, F13A1, FAM69B, FAM81B, FKBP15, FNDC3B, FOXO3, GGNBP2, GTF2E1, GYPC, HGD, HIP1, HLA-B, HMBOX1, IFNGR1, IGF2BP2, IGHG1, IL10, ITGB3, KIAA0564, KIAA1211, KIAA1618, KLHDC4, LCN2, LEPR, LY6G5C, LY6G6D, MBNL1, MGC29506, MGC9913, MSRB3, MXI1, MYLK, NCAM1, NCOA2, NR1H3, OAS1, PARP12, PARVB, PDE5A, PTGS1, PTPRN2, RAD51L1, RUNX1, SERPINF1, SLC22A4, SLC6A4, SLC8A1, SYK, TCF4, TGFA, TNF, TNFSF4, TP53, TRIM69, TTC7B, TUBGCP3, USP12, UTRN, VEPH1, WASF3, ZNF385D
DNA Replication, Recombination, and Repair	$8.46x10^{-4}$ - $1.61x10^{-3}$	BLM, CLU, EGF, EP300, FOXO3, GADD45A, IFI16, IFNGR1, MBD1, MLH3, N4BP2, NP, PML, PPP1R15A, RAD50, RAD51L1, REC8, RUNX1, SGK1, SOD2, TNF, TP53, UBTF, XRCC4, XRCC6BP1
Protein Trafficking	$1.01x10^{-3}$	ARNTL, PML, SOD2, TNF, TP53
Protein Synthesis	$1.07x10^{-3}$	AHR, CREB1, NCOA2, TP53
Nucleic Acid Metabolism	$1.18x10^{-3}$	EGF, GUCY1A3, GUCY1B3, PDE5A, TNF
Metabolic Disease	$1.20x10^{-3}$	ABHD2, ACTN1, ALOX12, ARNTL, ATP8B4, ATXN1, BARD1, BLM, BRAP, C10ORF28, C14ORF181, C15ORF26, C6ORF57, C8ORF42, CA2, CCDC50, CD38, CETP, CHD2, CLEC2B, CLIC4, CLU, COL13A1, COLQ, CSDA, CTSL1, DAB1, DNM3, EGF, ELOVL7, ENDOD1, F13A1, FAM69B, FAM81B, FKBP15, FNDC3B, FOXO3, GGNBP2, GTF2E1, GYPC, HGD, HIP1, HLA-B, HMBOX1, IFNGR1, IGF2BP2, IGHG1, IL10, KIAA0564, KIAA1211, KIAA1618, KLHDC4, LCN2, LEPR, LY6G5C, LY6G6D, MBNL1, MGC9913, MSRB3, MXI1, MYLK, NCAM1, NCOA2, OAS1, PARP12, PARVB, PDE5A, PTGS1, PTPRN2, RAD51L1, RUNX1, SERPINF1, SLC22A4, SLC6A4, SLC8A1, SYK, TCF4, TGFA, TNF, TNFSF4, TRIM69, TTC7B, TUBGCP3, USP12, UTRN, VEPH1, WASF3, ZNF385D
Developmental Disorder	$1.43x10^{-3}$ - $1.62x10^{-3}$	AHR, EGF, FOSL2, HIP1, IGHG1, OSR2, PCSK6, PDGFA, PSME3, TGFA, TNF, TP53, UTRN
Carbohydrate Metabolism	$1.62x10^{-3}$	IGHM, IP6K2, PIK3CG, SYNJ1

Table D.3: Significant differentially expressed groups of genes according to their cellular function. Results were obtained using the dChip 2009 software.

Function Annotation	p-value	Molecules
Antigen binding	$1.00x10^{-9}$	IGH@ / IGHG1 / IGHG2 / IGHG3 / IGHM, IGHA1 / IGHA2, IGHA1 / IGHG1, IGHD, IGHG1, IGJ, IGKC / IGKV1-5, IGKC, IGKV1D-13, IGL@ / IGLC1 / IGLC2 / IGLV3-25 / IGLV2-14 / IGLJ3, IGL@ / IGLC1 / IGLC2 / IGLV4-3 / IGLV3-25 / IGLV2-14 / IGLJ3, LOC96610, PPP2R1B
Immune response	$1.00x10^{-9}$	AIM2, BMP6, CCL20, CCL7, CD164, CD300A, CD55, CLEC4C, CLU, CMKLR1, CXCL2, CXCL5, DDX58, HLA-DRB1 / HLA-DRB3 / HLA-DRB4, IFI35, IFITM2, IGH@ / IGHG1 / IGHG2 / IGHG3 / IGHM, IGHA1 / IGHA2, IGHA1 / IGHG1, IGHD, IGHG1, IGJ, IGKC / IGKV1-5, IGKC, IGKV1D-13, IGL@ / IGLC1 / IGLC2 / IGLV3-25 / IGLV2-14 / IGLJ3, IGL@ / IGLC1 / IGLC2 / IGLV4-3 / IGLV3-25 / IGLV2-14 / IGLJ3, IL10, LILRA4, LOC96610, MICA, OAS1, OAS2, PF4, TLR3, TNF, TNFRSF17, TNFSF4
Platelet alpha granule membrane	$1.20x10^{-5}$	GP1BB / SEPT5, GP9, ITGA2B, ITGB3, SELP
Response to virus	$3.60x10^{-5}$	DDX58, IFI16, IFI35, IFNAR2, IFNGR1, IRF7, ISG15, ISGF3G, STAT2, TLR3, TNF
Oxidoreductase activity	$1.21x10^{-4}$	ADI1, ALOX12, JARID1A, LOC642252, PLOD2, PTGS1, SMCY, UTY
Response to DNA damage stimulus	$2.85x10^{-4}$	BARD1, BLM, BRCC3, FUS, GADD45A, GTF2H2, IRF7, MICA, MLH3, MORF4L2, PML, PPP1R15A, RAD50, RAD51L1, SGK, TP53, XRCC4
Inflammatory response	$8.82x10^{-4}$	ADORA2A, BMP6, CARD12, CCL20, CCL7, CXCL2, IL10, IRF7, MAP2K3, MGLL, NDST1, PTX3, SELP, TLR3, TNF, TNFSF4

Table D.4: Significant differentially expressed groups of genes associated with genetic disorders, injuries or malformations. Results were obtained using the IPA software.

Functional Annotation	p-value	Molecules
Genetic Disorder:		
Genetic disorder	$6.72x10^{-8}$	ABHD2, ACP2, ACSBG1, ACTN1, ACVR1B, ACVR2A, ADORA2A, AGPS, AHR, ALOX12, ANK1, ANKRD9, ANKRD28, ARHGAP10, ARHGAP18, ARHGEF9, ARHGEF12, ARNTL, ATL1, ATP8B4, ATXN1, BARD1, BCL2L11, BLM, BMP6, BRAP, BRCC3, C10ORF28, C13ORF15, C14ORF64, C14ORF181, C15ORF26, C6ORF57, C6ORF199, C8ORF42, CA2, CALD1, CCDC50, CCDC132, CCL7, CCL20, CD38, CD300A, CETP, CHD2, CLCN5, CLDN5, CLEC2B, CLIC4, CLIP4, CLU, COL13A1, COL6A3, COLQ, CREB1, CSDA, CTDSPL, CTSL1, CXCL2, CXCL5, CYLD, DAB1, DDX58, DNAH12, DNAJC15, DNM3, ECT2, EGF, EIF1B, EIF3C, EIF4B, ELOVL7, ENDOD1, EP300, ETFDH, F3, F13A1, FAM69B, FAM81B, FBXO22, FBXO38, FDXR, FHL1, FKBP15, FNDC3B, FOSL2, FOXO3, FSCN1, GADD45A, GALC, GART, GGNBP2, GNAZ, GOSR2, GP9, GRAP2, GRIN1, GTF2E1, GUCY1A3, GUCY1B3, GYPA, GYPC, HDGFRP3, HEATR3, HERC6, HGD, HIBCH, HIP1, HIVEP1, HLA-B, HMBOX1, HMMR, HNRNPC, IFITM2, IFNAR2, IFNGR1, IGF2BP2, IGHD, IGHG1, IGHM, IGJ, IGKC, IGL@, IL10, INTS7, ISG15, ITGA2B, ITGB3, ITGB5, ITIH4, JAM3, JARID1A, KIAA0564, KIAA1211, KIAA1618, KL, KLHDC4, KMO, KYNU, LCN2, LEPR, LMNA, LPAR1, LTBP1, LY6E, LY6G5C, LY6G6D, MAF, MAMLD1, MBD1, MBNL1, MFN1, MGC9913, MGC39372, MGLL, MLH3, MMD, MPL, MSRB3, MXI1, MYL9, MYLK, NCAM1, NCOA2, NCSTN, NDUFA5, NFKBIZ, NGFRAP1, NLRC4, NP, NR1H3, NRGN, OAS1, OAS2, OGDH, OSR2, PAM, PARP12, PARVB, PBX1, PCSK6, PDE5A, PDGFA, PGRMC1, PIK3CG, PLD1, PLOD2, PML, PPP2R1B, PRKAR2B, PROS1, PRTFDC1, PTGS1, PTPRN2, PVALB, RAB38, RAD51L1, RARS2, RASGEF1B, RGS17, RHOBTB3, RNASE2, RNF160, RUFY3, RUNX1, RYR1, SDC4, SELP, SEPT5, SERPINB8, SESN3, SGK1, SH3BP2, SLC22A4, SLC25A37, SLC35D3, SLC6A4, SLC8A1, SLFN5, SMAD5, SNAP29, SOD2, SP1, SPARC, SRR, SSFA2, SYK, SYNGR1, SYNJ1, SYTL4, TBC1D30, TBXA2R, TCF4, TGFA, TIPARP, TK2, TLR3, TMCC2, TNF, TNFRSF17, TNFSF4, TNFSF8, TP53, TPM1, TPM3, TRIM14, TRIM69, TTC7B, TUBB1, TUBGCP3, UBA6, USP12, USP9Y, UTRN, VEPH1, VIM, VRK2, WASF3, WDR41, XIST, ZCCHC6, ZEB2,
Prostate cancer	$3.44x10^{-6}$	CLU, CTDSPL, DNAJC15, EP300, GART, HIP1, IGHM, IGJ, IGKC, IGL@, IL10, ISG15, ITGB3, ITGB5, KIAA1211, LY6E, MXI1, MYL9, PDE5A, PML, PTGS1, PTPRN2, RUNX1, SGK1, TCF4, TNF, TP53, TUBB1
Sjogren's syndrome	$4.28x10^{-4}$	CCL20, IFITM2, PML, SYK, TNF, TNFRSF17
Chronic hepatitis C	$6.62x10^{-4}$	IFNAR2, IL10, SLC6A4, TNF
Psoriasis of humans	$9.10x10^{-4}$	IGHG1, IL10, TNF
Autoimmune disease	$1.21x10^{-3}$	ABHD2, ACSBG1, ACVR1B, ANK1, ANKRD28, ATXN1, BARD1, BCL2L11, BRAP, C13ORF15, C14ORF181, C6ORF57, CALD1, CCL7, CCL20, CHD2, CLEC2B, CLIC4, CLU, COLQ, CTSL1, CXCL2, CXCL5, DAB1, EGF, EIF1B, F13A1, FAM81B, FNDC3B, GADD45A, GART, GYPC, HGD, HIP1, HIVEP1, HLA-B, HMBOX1, IFITM2, IFNAR2, IFNGR1, IGHG1, IGL@, IL10, KIAA0564, KIAA1618, KLHDC4, LTBP1, LY6G5C, LY6G6D, MBD1, MGC9913, MXI1, MYLK, NCAM1, NCOA2, NP, OAS1, PAM, PARVB, PBX1, PLD1, PML, PRKAR2B, PTGS1, PTPRN2, RAD51L1, RUNX1, SELP, SLC22A4, SLC6A4, SLFN5, SYK, TBC1D30, TCF4, TGFA, TLR3, TNF, TNFRSF17, TNFSF4, TP53, TRIM69, VIM, ZCCHC6, ZNF385D
Schizophrenia of humans	$1.36x10^{-3}$	ATXN1, CLDN5, EGF, GRIN1, IL10, NCAM1, SLC6A4, SNAP29, SOD2, SRR, SYNGR1, TNF, TP53
Hepatitis C	$1.36x10^{-3}$	IFNAR2, IL10, MPL, SLC6A4, TNF
HIV infection	$1.40x10^{-3}$	ALOX12, GRIN1, IFNAR2, IFNGR1, MPL, PDE5A, PTGS1, SLC6A4, TLR3, TNF, TUBB1
Cleft palate syndrome of mice	$1.62x10^{-3}$	AHR, EGF, OSR2, TGFA

Genetic Disorder / Hematological Disease:		
Chronic b-cell leukemia	5.80×10^{-4}	MFN1, NP, TP53
Glanzmann's thrombasthenia	6.87×10^{-4}	ITGA2B, ITGB3
Genetic Disorder / Organismal Injury and Abnormalities:		
Migraines	5.88×10^{-4}	CA2, CLU, FHL1, ITGA2B, ITGB3, MLH3, PTGS1, SPARC, TNF
Hematological Disease:		
Hematological disorder	1.99×10^{-8}	ADORA2A, AHR, BNIP3L, CALD1, CCL7, CLU, CTSL1, EP300, EPB49, F3, F13A1, FOSL2, FUS, GADD45A, GP9, GYPA, GYPC, IFNGR1, IGHM, IL10, ITGA2B, ITGB3, ITGB5, KLF1, LCN2, MAX, MPL, MXI1, PBX1, PDE5A, PDGFA, PIK3CG, PLD1, PML, PPP1R15A, PROS1, PSME3, PTGS1, PTX3, RAD50, RNASE2, RUNX1, SELP, SLC8A1, SOD2, SYK, TBXA2R, TFPI, TNF, TNFSF4, TNFSF8, TP53, TUBB1
Hematologic cancer	1.03×10^{-6}	ADORA2A, ARHGEF12, BCL2L11, F3, IFNAR2, ITGB3, ITGB5, MFN1, MPL, NCAM1, NP, PBX1, PDE5A, PDIA6, PF4, PML, PTCRA, PTGS1, RUNX1, SLC6A4, SOD2, SYK, TAL1, TNF, TP53, TUBB1
Ischemia	4.42×10^{-6}	ADORA2A, CLU, ITGA2B, ITGB3, LCN2, PLD1, PPP1R15A, PTGS1, TNF
Ischemia of humans	6.87×10^{-4}	ITGA2B, ITGB3
Leukemia	4.75×10^{-5}	ADORA2A, ARHGEF12, BCL2L11, F3, IFNAR2, ITGB3, ITGB5, MFN1, MPL, NP, PBX1, PDE5A, PML, PTGS1, RUNX1, SLC6A4, SYK, TAL1, TNF, TP53, TUBB1
Thrombosis	6.06×10^{-5}	AHR, F3, PDGFA, PIK3CG, PROS1, PTGS1, TBXA2R, TFPI, TNF
Thrombosis of blood vessel	1.96×10^{-4}	AHR, F3, TFPI, TNF
Myeloproliferative syndrome	1.20×10^{-4}	IFNAR2, ITGB3, ITGB5, MPL, PML, PTGS1, RUNX1, TNF, TP53
B-cell leukemia	1.75×10^{-4}	MFN1, NP, PBX1, SYK, TP53
Bleeding	2.32×10^{-4}	F3, F13A1, IFNGR1, IL10, ITGB3, PTGS1, PTX3, SELP, SYK, TNF, TP53, TUBB1
Lymphocytic leukemia	2.78×10^{-4}	ADORA2A, BCL2L11, F3, IFNAR2, MFN1, NP, PBX1, PML, RUNX1, SYK, TAL1, TP53, TUBB1
Acute lymphocytic leukemia	2.94×10^{-4}	ADORA2A, BCL2L11, F3, NP, RUNX1, SYK, TAL1, TP53, TUBB1
Apoptosis of hematopoietic progenitor cells	3.42×10^{-4}	BCL2L11, MPL, PF4, PML, TNF
Apoptosis of leukocytes	6.28×10^{-4}	ADORA2A, BCL2L11, CXCL2, FOXO3, GRAP2, IFNGR1, IGHM, IL10, LCN2, LEPR, MAP2K3, NP, PF4, PIK3CG, SOD2, TBXA2R, TNF, TNFRSF17, TNFRSF18, TP53
Apoptosis of mucosal mast cells	6.87×10^{-4}	FOXO3, TP53
Apoptosis of leukemia cell lines	1.07×10^{-3}	BCL2L11, CD55, GFI1B, LGALS3BP, MAX, PML, RUNX1, SOD2, TAL1, TGFA, TNF, TNFRSF17, TNFRSF18, TP53
Anemia of mammalia	3.94×10^{-4}	EP300, EPB49, IL10, ITGB3, KLF1, MPL, PBX1, SOD2
Myeloid leukemia	6.08×10^{-4}	ADORA2A, ARHGEF12, F3, IFNAR2, ITGB3, ITGB5, MPL, PML, PTGS1, RUNX1, TNF, TP53
Leukemogenesis	6.62×10^{-4}	PTCRA, RUNX1, TAL1, TP53
Cell death of leukocytes	1.12×10^{-3}	ADORA2A, BCL2L11, CXCL2, FOXO3, GRAP2, IFNGR1, IGHM, IL10, LCN2, LEPR, MAP2K3, NLRC4, NP, PF4, PIK3CG, SOD2, TBXA2R, TNF, TNFRSF17, TNFRSF18, TP53
Transformation of hematopoietic progenitor cells	1.34×10^{-3}	FOXO3, PBX1, TNF
Multiple myeloma	1.60×10^{-3}	ADORA2A, IFNAR2, ITGB3, ITGB5, NCAM1, PDIA6, PF4, PML, SOD2, TNF

Hematological Disease / Organismal Injury and Abnormalities:		
Thrombocytosis of mice	5.80×10^{-4}	BNIP3L, EP300, MPL
Organismal Injury and Abnormalities:		
Degeneration of organ	1.93×10^{-6}	ATXN1, CLU, DDX58, MXI1, NCOA2, PCYT1B, TNF
Fibrosis	1.09×10^{-4}	AHR, EGF, HLA-B, IFI35, IFNAR2, IFNGR1, IL10, IRF7, IRF9, OAS1, OAS2, PDE5A, PF4, SLC6A4, STAT2, TGFA, TNF
Pathogenesis	1.78×10^{-4}	AHR, CLU, EGF, F13A1, IGHG1, IL10, PIK3CG, SLC8A1, SOD2, TNF, TNIP1, TP53, USP18
Headache	1.94×10^{-4}	CA2, CLU, FHL1, ITGA2B, ITGB3, MLH3, PTGS1, SLC6A4, SPARC, TNF
Quantity of nodule	1.96×10^{-4}	BMP6, HMMR, SPARC, TP53
Injury of tissue	2.21×10^{-4}	ADORA2A, CLU, F3, IL10, TNF
Injury of organ	2.51×10^{-4}	ADORA2A, AHR, BCL2L11, CBR1, F3, IL10, SDC4, SELP, TNF, TRIM69
Organismal abnormalities of mice	2.39×10^{-4}	AHR, BNIP3L, CREB1, EP300, F3, MPL, PF4, PIK3CG, SELP, SMAD5, TLR3
Thrombus	3.08×10^{-4}	F3, IL10, ITGB3, PF4, PIK3CG, SELP, TNF
Formation of skin lesion	3.38×10^{-4}	IL10, NFKBIZ, TNF
Necrosis	3.71×10^{-4}	AHR, CLU, EGF, F13A1, IGHG1, IL10, PIK3CG, SLC8A1, SOD2, TNF, TP53, USP18
Pain	8.81×10^{-4}	CA2, CLU, FHL1, GRIN1, GUCY1A3, GUCY1B3, ITGA2B, ITGB3, MLH3, PDE5A, PTGS1, SLC6A4, SPARC, TNF, TUBB1
Damage of tissue	1.53×10^{-3}	ADORA2A, CLU, F3, IL10, LMNA, SOD2, TNF

Table D.5: Significant canonical pathways affected by the differentially expressed genes among IS cases and controls. The genes that are differentially expressed in each pathway are presented. Results were obtained using the IPA software.

Canonical pathway	p-value	Molecules
Role of pattern recognition receptors in recognition of bacteria and viruses	$7.24x10^{-7}$	PTX3, OAS1, IRF7, OAS2, IL10, PIK3CG, SYK, DDX58, CREB1, TLR3, TNF, NLRC4
Activation of IRF by cytosolic pattern recognition receptors	$3.24x10^{-5}$	IRF7, IL10, DDX58, ZBP1, RP5-1000E10.4, STAT2, IRF9, TNF, ISG15
Interferon signaling	$4.68x10^{-5}$	OAS1, IFNGR1, IFI35, STAT2, IRF9, IFNAR2
Primary immunodeficiency signaling	$4.47x10^{-4}$	SP1, IGKC, IGL@, IGHM, IGHG1, EP300, IGHD
Androgen signaling	$2.28x10^{-3}$	HSPA4, TGFB1I1, GNG11, PRKAR2B, NCOA2, GTF2E1, GTF2H2, GNAZ, EP300
Huntington's disease signaling	$2.81x10^{-3}$	TP53, SGK1, DNM3, EGF, HIP1, EP300, HSPA4, GNG11, SP1, PIK3CG, CREB1, UBC, GOSR2
Dendritic cell maturation	$3.31x10^{-3}$	IL10, LEPR, PIK3CG, FSCN1, CREB1, HLA-B, STAT2, IGHG1, TLR3, TNF
Hepatic fibrosis / Hepatic stellate cell activation	$5.75x10^{-3}$	MYL9, IL10, PDGFA, LEPR, TGFA, EGF, IFNGR1, IFNAR2, TNF
Clathrin-mediated endocytosis	$5.75x10^{-3}$	SYNJ1, PDGFA, PIK3CG, DNM3, EGF, HIP1, UBC, CTTN, ITGB5, ITGB3
Allograft Rejection signaling	$6.61x10^{-3}$	IL10, HLA-B, IGHG1, TNF
LXR/RXR activation	$8.13x10^{-3}$	CCL7, NR1H3, ARG2, TLR3, TNF, CETP
Role of BRCA1 in DNA damage response	$8.91x10^{-3}$	TP53, GADD45A, BARD1, BLM, RAD50
Caveolar-mediated endocytosis	$1.00x10^{-2}$	ITGA2B, CD55, HLA-B, EGF, ITGB5, ITGB3
Coagulation system	$1.29x10^{-2}$	PROS1, F13A1, TFPI, F3
VDR/RXR activation	$1.48x10^{-2}$	COL13A1, SP1, NCOA2, GADD45A, PDGFA, EP300
Cell cycle: G2/M DNA damage checkpoint regulation	$1.55x10^{-2}$	TP53, GADD45A, UBC, EP300
IL-12 signaling and production in macrophages	$2.04x10^{-2}$	IL10, PIK3CG, MAF, ALOX12, IFNGR1, TNF, EP300
IL-10 signaling	$2.19x10^{-2}$	SP1, IL10, ARG2, MAP2K3, TNF
Role of PKR in interferon induction and antiviral response	$2.19x10^{-2}$	TP53, MAP2K3, TLR3, TNF
Thrombin signaling	$2.34x10^{-2}$	MYLK, MYL9, ARHGEF12, GNG11, PIK3CG, CREB1, RHOU, EGF, GNAZ, ARHGEF9
Macropinocytosis	$2.63x10^{-2}$	PDGFA, PIK3CG, EGF, ITGB5, ITGB3
Non-small cell lung cancer signaling	$2.75x10^{-2}$	TP53, PIK3CG, FOXO3, TGFA, EGF
Calcium signaling	$3.16x10^{-2}$	MYL9, TPM1, GRIN1, PRKAR2B, TPM3, CREB1, RYR1, SLC8A1, EP300
Circadian rhythm signaling	$3.31x10^{-2}$	GRIN1, ARNTL, CREB1
HMGB1 signaling	$3.39x10^{-2}$	SP1, PIK3CG, RHOU, IFNGR1, MAP2K3, TNF
Relaxin signaling	$3.89x10^{-2}$	GNG11, PRKAR2B, GUCY1A3, PIK3CG, CREB1, GNAZ, GUCY1B3
Fc epsilon RI signaling	$3.89x10^{-2}$	PLA2G12A, GRAP2, PIK3CG, SYK, MAP2K3, TNF
Insulin receptor signaling	$4.17x10^{-2}$	PPP1R3D, PRKAR2B, TRIP10, SGK1, PIK3CG, FOXO3, HLA-B
Nicotinate and nicotinamide metabolism	$4.27x10^{-2}$	SGK1, NP, CD38, MAP2K3, NADK, ACVR2A
Reelin signaling in neurons	$4.27x10^{-2}$	ARHGEF12, PIK3CG, DAB1, ARHGEF9, ITGB3
ATM signaling	$4.47x10^{-2}$	TP53, GADD45A, CREB1, RAD50
Phototransduction pathway	$4.47x10^{-2}$	PRKAR2B, GUCY1A3, GNGT2, GUCY1B3
Autoimmune thyroid disease signaling	$4.68x10^{-2}$	IL10, HLA-B, IGHG1
Tight junction signaling	$4.79x10^{-2}$	MYLK, MYL9, CLDN5, PRKAR2B, JAM3, CSDA, PPP2R1B, TNF