EXOME SEQUENCING FOR RARE MUTATIONS IN YOUNG STROKE

EXOME SEQUENCING TO CHARACTERIZE THE ROLES OF MENDELIAN STROKE GENES AND NOVEL GENES IN YOUNG STROKE

By MICHAEL CHONG, B.ASc.

A Thesis Submitted to the School of Graduate Studies in Partial Fulfilment of the Requirements for the Degree Master of Science

McMaster University

 © Copyright by Michael Chong, August 2014

Master of Science (2014)

Medical Sciences

McMaster University

Hamilton, Ontario

TITLE: Exome sequencing to characterize the roles of Mendelian stroke genes and novel genes in young stroke.

AUTHOR: Michael Chong, B.ASc. (McMaster University)

SUPERVISOR: Dr. Guillaume Paré

NUMBER OF PAGES: xii, 72

**Abstract**

*Background*:Rare genetic mutations cause familial early-onset stroke disorders, known as “Mendelian strokes”. The broader relevance of rare mutations in unrelated young stroke patients is uncertain. We hypothesize that rare mutations in known and novel genes are important risk factors for stroke.

*Methods*:Exome sequencing was used to characterize rare disruptive protein-altering mutations in 185 young cases and 185 matched controls from INTERSTROKE, a large and globally representative stroke study. The major objectives were: 1) to precisely define the role of known Mendelian stroke genes and 2) to discover novel gene and pathway associations.

*Results*:A focused assessment of known Mendelian stroke genes revealed a significant contribution from *NOTCH3*, the causal gene for Cerebral Autosomal Dominant Arteriopathies with Subcortical Infarcts and Leucoencephalopathies (CADASIL). CADASIL mutations were identified in six cases and no controls (P=0.03). The clinical presentation of CADASIL mutation carriers deviated from known symptomatology, consisting of small-vessel ischemic strokes (SVIS) accompanied by secondary features including migraine and depression. A novel role for non-CADASIL *NOTCH3* mutations in ICH was also elucidated (OR=2.86; 95% CI, 1.13 to 7.93, P=0.02). Suchmutations were present in 22% of ICH cases and 8% of matching controls. An agnostic evaluation of all genes did not reveal any genome-wide significant associations. However, *NOTCH3* was among the top ICH genes out of 13,706 tested, and many others were also biologically relevant, notably, *AARS2* and *NBEAL2*. A protective association was identified for the renin angiotensin system (P=8.1x10-4), whereas type II diabetes mellitus was associated with increased risk (P=1.9x10-2).

*Conclusion*: Rare mutations influence risk of early-onset stroke. CADASIL mutations play an important role in unrelated stroke patients. Beyond CADASIL, a novel role was uncovered for other *NOTCH3* mutations as common and significant risk factors for ICH. Novel biologically relevant genes and pathways may also affect stroke susceptibility.

**Acknowledgements**

I would like to recognize the entire genetic and molecular epidemiology laboratory (GMEL) for their friendship throughout the past two years. It has been a pleasure to work with each and every member as I have never encountered such a talented group of individuals. Particularly, I would like to thank Reina Ditta and Amanda Hodge who organized, conducted, and troubleshooted the extensive lab work involved in exome sequencing. This thesis was truly a collaborative effort and I would like to also acknowledge Kripa Raman, Matt D’Mello, Jenny Sjaarda, Randa Stringer, and Stephanie Ross for editing my thesis.

Also, I would like to thank my family and friends for their emotional support. I could not have endured the late nights and stressful times without my best friends: Adrienne Yang, Michael Yoon, and Jason Binder.

I would like to express my gratitude to my committee members, Dr. Hart and Dr. Meyre, who have provided insightful feedback and who have exhibited incredible patience in booking of our meetings! I would also like to thank Dr. Samaan for serving as the external examiner.

Lastly, I would like to thank Dr. Paré, an incredible mentor and friend. I am extremely grateful that he provided me with an incredible opportunity to learn from the best and work on a cutting-edge dataset. There is no better role model for an aspiring researcher than Dr. Paré, who is the most driven, patient, and encouraging supervisor I know.

**Table of Contents**

1. **Chapter One: Introduction** **1**
	1. Stroke Biology and Epidemiology.............................................................2
	2. A Genetic Basis for Stroke........................................................................2
	3. Common Variants: A Lesson in Phenotypic Heterogeneity......................4
	4. Rare Variants: A Brave New world...........................................................6
	5. Objectives...................................................................................................8
	6. Hypotheses.................................................................................................8
	7. References..................................................................................................9
	8. Figures......................................................................................................14
2. **Chapter Two: Evaluation of Mendelian Stroke Genes in Young Stroke 15**
	1. Introduction..............................................................................................16
	2. Methods....................................................................................................17
	3. Results.......................................................................................................22
	4. Discussion….............................................................................................25
	5. Conclusion................................................................................................28
	6. References................................................................................................30
	7. Tables.......................................................................................................35
	8. Supplementary Material...........................................................................38
3. **Chapter Three: Systematic Exploration of Rare Variants influencing risk of Small-Vessel Stroke**  **42**
	1. Introduction..............................................................................................43
	2. Methods....................................................................................................44
	3. Results .....................................................................................................48
	4. Discussion….............................................................................................50
	5. Conclusion................................................................................................54
	6. References................................................................................................55
	7. Tables.......................................................................................................59
	8. Supplementary Material...........................................................................63
4. **Chapter Four: Conclusion**  **69**
	1. Summary of Findings...............................................................................70
	2. Research Implications...............................................................................71
	3. Future Directions .....................................................................................73
	4. Concluding Remarks …...........................................................................74
	5. References................................................................................................75

**List of Abbreviations and Symbols**

APOB/APOA1 – Apolipoprotein B/Apolipoprotein A1 Ratio

BMI – Body Mass Index

CAA – Cerebral Amyloid Angiopathy

CADASIL – Cerebral Autosomal Dominant Arteriopathies with Subcortical Infarcts and Leucoencephalopathy

CARASIL – Cerebral Autosomal Recessive Arteriopathies with Subcortical Infarcts and Leucoencephalopathy

CES – Cardioembolic Stroke

CHARGE – Cohorts for Heart and Aging in Genomic Epidemiology

CI – Confidence Interval

CVCD – Common Variant Common Disease

CVRD – Common Variant Rare Disease

CT – Computed Tomography

dbSNP137 – The Single Nucleotide Polymorphism Database 137

DNA – Deoxyribonucleic Acid

EDTA – Ethylenediaminetetraacetic acid

EGFR – Epidermal Growth Factor-like Repeat

ESP – National Institute of Health Heart, Lung and Blood Institute Grand Opportunity Exome Sequencing Project

GATK – Genome Analysis Tool Kit

GCTA – Genome Complex Trait Analysis

GEOS – Genetics of Early-Onset Stroke

GWAS – Genome-Wide Association Studies

Het/Hom – Heterozygous/Homozygous Ratio

HWE – Hardy Weinberg Equilibrium

ICH – Intracerebral Hemorrhage

INDEL – Insertion Deletion

KEGG – Kyoto Encyclopedia of Genes and Genomes

RefGene – NCBI Reference Sequence Database

RefSeq – NCBI Reference Sequence Database

LVIS – Large Vessel Ischemic Stroke

MAF – Minor Allele Frequency

MELAS – Mitochondrial Encephalomyopathy with Lactic Acidosis and Stroke-like episodes

MRI – Magnetic Resonance Imaging

NHLBI GO – National Institute of Health Lung Blood Institute Grand Opportunity

OCSP – Oxfordshire Community Stroke Project

OMIM – Online Mendelian Inheritance in Man

OR – Odds ratio

PAH – Pulmonary Arterial Hypertension

Polyphen-2 – Polymorphism Phenotyping v2

QC – Quality Control

RAS – Renin Angiotensin System

RVCD – Rare Variant Common Disease Hypothesis

SAH – Subarachnoid Hemorrhage

SD – Standard Deviation

SIFT – Sorting Intolerant from Tolerant

SKAT – Sequence Kernel Association Test

SKAT-O – Optimal Sequence Kernel Association Test

SMC – Smooth Muscle Cell

SNV – Single Nucleotide Variant

SVIS – Small Vessel Ischemic Stroke

Ti/Tv – Transition / Transversion Ratio

TOAST – Trial of ORG 10172 in Acute Stroke Treatment

T2DM – Type II Diabetes Mellitus

UNIPROT – Unified Protein Resource

1KG – 1000 Genomes

**List of Tables**

Table 2.1 –Characteristics of study subjects 35

Table 2.2 –Mutation carrier counts for Mendelian stroke genes 36

Table 2.3 –Clinical Features of putative CADASIL mutation carriers 36

Table 2.4 – Comparison of secondary CADASIL features among cases 37

Table 2.5 – Summary ofknown disease-causing mutation carrier counts 37

Table 3.1 –Characteristics of study subjects 59

Table 3.2 –Gene-based association results for all stroke 60

Table 3.3 –Gene-based association results for ICH 60

Table 3.4 –Gene-based association results for SVIS 61

Table 3.5 –Pathway-based association results 62

**List of Supplementary Tables**

Supplementary Table 2.1 – Candidate Mendelian stroke genes 38

Supplementary Table 2.2 –Non-CADASIL *NOTCH3* mutation carrier counts 39

Supplementary Table 2.3 –Coverage metrics for candidate genes 40

Supplementary Table 2.4 – Allele counts for candidate genes 41

Supplementary Table 3.1 –Quality control metrics for alignment and variant calling 63

Supplementary Table 3.2 –Variant counts by functional class 63

Supplementary Table 3.3 – Biologically relevant genes among top 50 genes 64

Supplementary Table 3.4 – Gene-based association results using SKAT-O 64

Supplementary Table 3.5 – Pathway-based association results for all stroke 65

Supplementary Table 3.6 – Pathway-based association results for ICH 66

Supplementary Table 3.7 –Pathway-based association results for SVIS 66

Supplementary Table 3.8– RAS pathway mutation carrier counts 67

Supplementary Table 3.9 –T2DM pathway mutation carrier counts 68

**List of Figures**

Figure 1.1 – Physiological comparison of the major stroke subtypes 14

**Chapter One: Introduction**

1.1 Stroke Biology and Epidemiology

Stroke imposes an enormous burden on society with more than 30 million people affected worldwide1. Defined as an acute neurological deficit, stroke is the result of abnormal blood flow to the brain2. The major subtypes are ischemic and hemorrhagic strokes (Figure 1.1). Ischemic stroke is characterized by thrombotic occlusion and can be further classified into small-vessel ischemic stroke (SVIS), large-vessel ischemic stroke (LVIS), or cardioembolic stroke (CES)3. In contrast, hemorrhagic strokes are characterized by ruptured vessels bleeding into the space surrounding the brain (subarachnoid hemorrhage (SAH)) or the brain itself (intracerebral hemorrhage (ICH)). ICH can be further classified as lobar or non-lobar (deep) ICH.

The composition of stroke subtypes is estimated to be roughly 73% ischemic strokes, 19% hemorrhagic strokes, and 8% of undetermined etiology4. The most common ischemic stroke subtype is SVIS, which accounts for ~50% of all ischemic strokes and ~35% of all strokes5, whereas ICH accounts for ~80% of all hemorrhagic strokes and ~15% of all strokes4.

1.2 A Genetic Basis for Stroke

 INTERSTROKE, a large international study of stroke across 22 different countries, demonstrated that 10 conventional risk factors (hypertension, diabetes, smoking, alcohol intake, cardiac causes, waist-to-hip ratio, APOB/APOA1 ratio, physical activity, stress, and diet) account for approximately 90% of stroke risk5. Other emerging risk factors, such as genetics, may explain the remaining fraction of risk. Age is by far the most important risk factor for stroke6, and while generally regarded as a disease of the old (mean age: 70 years1), INTERSTROKE estimates that 14% of all strokes occur in those below 45 years5. Combined with the fact that conventional risk factors are less prevalent among younger patients5, early stroke may be disproportionately the result of genetic predisposition, much like other early forms of disease (e.g. breast and colon cancer7,8).

A genetic basis for stroke is supported by various lines of research. Firstly, stroke concordance is 65% higher between monozygotic twins than dizygotic twins9, who presumably share similar environments. Secondly, family history is a strong predictor of stroke. Independent of conventional risk factors, parental history of ischemic stroke is associated with two-fold higher risk of ischemic stroke10, whereas having a first-degree relative with ICH is associated with six-fold higher risk of ICH11. Furthermore, familial aggregation is more pronounced in younger patients12. Thirdly, the genetic component (heritability) for both stroke and its intermediate phenotypes (intima-media thickness13, intracranial aneurysm14,15, and white matter hyperintensities16,17) is substantial. The heritability of ischemic stroke and ICH is estimated to be 37.9% and 44%, respectively18,19. The Genetics of Early-Onset Stroke (GEOS) study also found that the heritability of ischemic stroke and its subtypes was slightly higher (non-significant) for those under 50 years20. Fourthly, genome-wide association studies (GWAS) have identified common genetic variants associated with stroke risk18. Lastly, rare protein-altering mutations are known to cause early stroke disorders, “Mendelian strokes”21–24.

1.3 Common Variant Studies: A Lesson in Phenotypic Heterogeneity

 The Common Disease-Common Variant (CDCV) hypothesis asserts that frequent mutations (MAF>5%) of modest effect (OR < 1.5) underlie a substantial fraction of diseased cases in the general population25. Common variants have been assessed through two approaches: candidate gene studies which evaluate certain biologically relevant genes and genome-wide association studies (GWAS) which systematically scan all genetic loci.

Bevan *et al*. (2012) performed a meta-analysis of GWAS data and revealed vast heterogeneity across ischemic stroke subtypes. While the total heritability of ischemic stroke was estimated to be 37.9%, the heritability of LVIS, SVIS, and CES were 40.3%, 16.1%, and 32.6%, respectively18. The largest GWAS meta-analysis including 25,736 ischemic stroke cases revealed subtype specificity for established stroke loci 26,27. *PITX2* and *ZFHX3* variants were only associated with risk of CES, whereas locus 9p21 and *HDAC9* variants were specific to LVIS. *PITX2* and *ZFHX3* mutations influence risk of atrial fibrillation28,29, a major risk factor for CES30, whereas *HDAC9* promotes carotid atherosclerosis31.

Similarly, there is also evidence for heterogeneity across ICH subtypes. Devan *et al.* (2013) estimated the total heritability of ICH be 41%19. Common *APOE* variants explain more than 30% of this heritability. *APOE* variants also exhibit heterogeneity across lobar and deep ICH subtypes19, accounting for 73% of the variation in lobar ICH risk, but only 34% of the variation in deep ICH risk. Deep ICH is primarily attributed to hypertension, whereas lobar ICH is characterized by amyloid accumulation in cortical vessel walls (cerebral amyloid angiopathies (CAA)). *APOE* variants are known to influence amyloid deposition for CAA and Alzheimer’s disease32. Thus, the stronger association with lobar ICH may reflect APOE’s role in amyloid pathology. Additionally, a polygenic risk score consisting of blood pressure-related loci was associated with deep ICH but not lobar ICH19, which is congruent with the hypertensive origins of deep ICH.

There is also evidence for a shared genetic basis across subtypes. The EuroCLOT study discovered that the *ABO* gene was associated with LVIS and CES33. This is consistent with the observation that people with non-O blood types are more susceptible to developing thromboembolism (pulmonary embolism34 and venous thrombosis35). While it is sensible that mutations in coagulation genes should also influence risk of thromboembolism, genetic studies can reveal more complex and unexpected relationships. For instance, Anderson *et al*. (2013) discovered that common variants within oxidative phosphorylation genes were associated with both deep ICH and SVIS, but not LVIS nor lobar ICH36.

In summary, findings from common variant studies underscore the importance of proper stroke subtyping. Heterogeneity exists not only between ischemic and hemorrhagic strokes but also within their subtypes. Although variants may influence risk of multiple stroke subtypes, most known associations are specific to one subtype. Consequently, subtypes must be analyzed as distinct phenotypes to properly decipher the genetic architecture of stroke.

1.4 Rare Variant Studies: A Brave New World

The role of rare variants in disease is only beginning to be elucidated due to previous limitations in technology. One of the first large-scale sequencing initiatives, the 1000 Genomes (1KG) project, estimated that every person carries approximately 20 rare disease-associated mutations37. Consequently, rare mutations may have a broader role in the general population than previously believed. The Common Disease – Rare Variant (CDRV) hypothesis asserts that the aggregate impact of individually rare mutations (MAF < 5%) with large effects (OR > 2)25, accounts for a substantial fraction of diseased cases.

In the context of stroke, the most compelling evidence supporting the CDRV hypothesis is the existence of “Mendelian strokes”, which are severe familial stroke disorders caused by rare protein-altering mutations38. Cerebral Autosomal Dominant Arteriopathies with Subcortical Infarcts and Leucoencephalopathies (CADASIL) and Fabry’s disease are the most extensively studied Mendelian stroke disorders. In the general population, the prevalence of CADASIL is estimated to be 1-2 per 100,000 individuals39,40, whereas the prevalence of Fabry’s disease is 14-50 per 100,000 indiviudals41,42. Conversely, among stroke patients, the prevalence is estimated to be higher at 500-6000 per 100,000 individuals for CADASIL43,44 and 500-3900 per 100,000 individuals for Fabry’s disease22,23. CADASIL is caused by rare mutations in *NOTCH3*, an important regulator of cerebral artery development45, and Fabry’s disease is caused by rare mutations in *GLA*, a metabolic enzyme which processes glycosphingolipids22. Both disorders are characterized by extremely high life-time risk of small-vessel strokes (up to 71%), early onset (before 50 years), and debilitating secondary complications46,47.

In the past, large-scale studies assessing rare variants were not possible; however, recent advances have led to an effective approach: exome sequencing. Exome sequencing enables the assessment of all types of genetic variation within the coding regions of the genome48. One major advantage over other conventional genotyping platforms is that exome sequencing can detect rare and even novel mutations49. The “exome” specifically refers to the 1-2% of the genome containing all ~20,000 protein-coding genes50. Just as genotyping arrays facilitated the transition from candidate gene studies to genome-wide scans for common variants, exome sequencing permits an agnostic exploration of rare mutations across all genes.

1.5 Objectives

*Section 1*:

* To determine which Mendelian stroke genes, if any, should be screened in young stroke patients
* To define the clinical features associated with rare mutations in Mendelian stroke genes

*Section 2:*

* To systematically identify novel gene associations for early stroke
* To systematically identify novel pathway associations for early stroke

1.6 Hypotheses

*Section 1*:

* Previously reported disease-causing mutations within known Mendelian stroke genes increase risk of early stroke
* Rare disruptive mutations within Mendelian stroke genes increase risk of early stroke

*Section 2:*

* Rare disruptive mutations within genes alter risk of early stroke
* Rare disruptive mutations within pathways alter risk of early stroke

1.7 References

1. Feigin, V. L. *et al.* Global and regional burden of stroke during 1990-2010: findings from the Global Burden of Disease Study 2010. *Lancet* **383,** 245–54 (2014).

2. Easton, J. D. *et al.* Definition and evaluation of transient ischemic attack: a scientific statement for healthcare professionals from the American Heart Association/American Stroke Association Stroke Council; Council on Cardiovascular Surgery and Anesthesia; Council on Cardio. *Stroke.* **40,** 2276–93 (2009).

3. Adams, H. P. *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).

4. Thrift, A. G., Dewey, H. M., Macdonell, R. A., McNeil, J. J. & Donnan, G. A. Incidence of the major stroke subtypes: initial findings from the North East Melbourne stroke incidence study (NEMESIS). *Stroke.* **32,** 1732–8 (2001).

5. O’Donnell, M. J. *et al.* Risk factors for ischaemic and intracerebral haemorrhagic stroke in 22 countries (the INTERSTROKE study): a case-control study. *Lancet* **376,** 112–23 (2010).

6. Brown, R. D., Whisnant, J. P., Sicks, J. D., O’Fallon, W. M. & Wiebers, D. O. Stroke incidence, prevalence, and survival: secular trends in Rochester, Minnesota, through 1989. *Stroke.* **27,** 373–80 (1996).

7. Giardiello, F. M. *et al.* The use and interpretation of commercial APC gene testing for familial adenomatous polyposis. *N. Engl. J. Med.* **336,** 823–7 (1997).

8. Lee, E.-H. *et al.* Effect of BRCA1/2 mutation on short-term and long-term breast cancer survival: a systematic review and meta-analysis. *Breast Cancer Res. Treat.* **122,** 11–25 (2010).

9. Flossmann, E., Schulz, U. G. R. & Rothwell, P. M. Systematic review of methods and results of studies of the genetic epidemiology of ischemic stroke. *Stroke.* **35,** 212–27 (2004).

10. Seshadri, S. *et al.* Parental occurrence of stroke and risk of stroke in their children: the Framingham study. *Circulation* **121,** 1304–12 (2010).

11. Woo, D. *et al.* Genetic and environmental risk factors for intracerebral hemorrhage: preliminary results of a population-based study. *Stroke.* **33,** 1190–5 (2002).

12. Cheng, Y.-C., Cole, J. W., Kittner, S. J. & Mitchell, B. D. Genetics of ischemic stroke in young adults. *Circ. Cardiovasc. Genet.* **7,** 383–92 (2014).

13. Zhao, J. *et al.* Heritability of carotid intima-media thickness: a twin study. *Atherosclerosis* **197,** 814–20 (2008).

14. Mackey, J. *et al.* Familial intracranial aneurysms: is anatomic vulnerability heritable? *Stroke.* **44,** 38–42 (2013).

15. Schievink, W. I., Schaid, D. J., Rogers, H. M., Piepgras, D. G. & Michels, V. V. On the inheritance of intracranial aneurysms. *Stroke* **25,** 2028–2037 (1994).

16. Carmelli, D. *et al.* Evidence For Genetic Variance in White Matter Hyperintensity Volume in Normal Elderly Male Twins. *Stroke* **29,** 1177–1181 (1998).

17. Atwood, L. D. *et al.* Genetic variation in white matter hyperintensity volume in the Framingham Study. *Stroke.* **35,** 1609–13 (2004).

18. Bevan, S. *et al.* Genetic heritability of ischemic stroke and the contribution of previously reported candidate gene and genomewide associations. *Stroke.* **43,** 3161–7 (2012).

19. Devan, W. J. *et al.* Heritability Estimates Identify a Substantial Genetic Contribution to Risk and Outcome of Intracerebral Hemorrhage. *Stroke.* (2013). doi:10.1161/STROKEAHA.111.000089

20. Jiang, B. *et al.* Prothrombin G20210A mutation is associated with young-onset stroke: the genetics of early-onset stroke study and meta-analysis. *Stroke.* **45,** 961–7 (2014).

21. Arboleda-Velasquez, J. F. *et al.* CADASIL mutations impair Notch3 glycosylation by Fringe. *Hum. Mol. Genet.* **14,** 1631–9 (2005).

22. Rolfs, A. *et al.* Acute cerebrovascular disease in the young: the stroke in young fabry patients study. *Stroke.* **44,** 340–9 (2013).

23. Rolfs, A. *et al.* Prevalence of Fabry disease in patients with cryptogenic stroke: a prospective study. *Lancet* **366,** 1794–6 (2005).

24. Meschia, J. F. New information on the genetics of stroke. *Curr. Neurol. Neurosci. Rep.* **11,** 35–41 (2011).

25. Bodmer, W. & Bonilla, C. Common and rare variants in multifactorial susceptibility to common diseases. *Nat. Genet.* **40,** 695–701 (2008).

26. Bellenguez, C. *et al.* Genome-wide association study identifies a variant in HDAC9 associated with large vessel ischemic stroke. *Nat. Genet.* **44,** 328–33 (2012).

27. Traylor, M. *et al.* Genetic risk factors for ischaemic stroke and its subtypes (the METASTROKE collaboration): a meta-analysis of genome-wide association studies. *Lancet Neurol.* **11,** 951–62 (2012).

28. Olesen, M. S. *et al.* Genetic loci on chromosomes 4q25, 7p31, and 12p12 are associated with onset of lone atrial fibrillation before the age of 40 years. *Can. J. Cardiol.* **28,** 191–5 (2012).

29. Gudbjartsson, D. F. *et al.* A sequence variant in ZFHX3 on 16q22 associates with atrial fibrillation and ischemic stroke. *Nat. Genet.* **41,** 876–8 (2009).

30. Flegel, K. M., Shipley, M. J. & Rose, G. RISK OF STROKE IN NON-RHEUMATIC ATRIAL FIBRILLATION. *Lancet* **329,** 526–529 (1987).

31. Markus, H. S. *et al.* Evidence HDAC9 genetic variant associated with ischemic stroke increases risk via promoting carotid atherosclerosis. *Stroke.* **44,** 1220–5 (2013).

32. Ramanan, V. K. *et al.* APOE and BCHE as modulators of cerebral amyloid deposition: a florbetapir PET genome-wide association study. *Mol. Psychiatry* 1–7 (2013). doi:10.1038/mp.2013.19

33. Williams, F. M. K. *et al.* Ischemic stroke is associated with the ABO locus: The EuroCLOT study. *Ann. Neurol.* **73,** 16–31 (2013).

34. Goldhaber, S. Z., Jr, C. A. C. & Fuchs, C. S. embolism in two large cohort studies. **104,** 962–971 (2011).

35. Sode, B., Allin, K., Dahl, M., Gyntelberg, F. & Nordestgaard, B. G. Risk of venous thromboembolism and myocardial mutations and blood type. *CMAJ* **185,** 229–237 (2013).

36. Anderson, C. D. *et al.* Common variants within oxidative phosphorylation genes influence risk of ischemic stroke and intracerebral hemorrhage. *Stroke.* **44,** 612–9 (2013).

37. Abecasis, G. R. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).

38. Oyston, J. Online Mendelian Inheritance in Man. *Anesthesiology* **89,** 811–2 (1998).

39. Razvi, S. S. M., Davidson, R., Bone, I. & Muir, K. W. The prevalence of cerebral autosomal dominant arteriopathy with subcortical infarcts and leucoencephalopathy (CADASIL) in the west of Scotland. *J. Neurol. Neurosurg. Psychiatry* **76,** 739–41 (2005).

40. Narayan, S. K., Gorman, G., Kalaria, R. N., Ford, G. A. & Chinnery, P. F. The minimum prevalence of CADASIL in northeast England. *Neurology* **78,** 1025–7 (2012).

41. Inoue, T. *et al.* Newborn screening for Fabry disease in Japan: prevalence and genotypes of Fabry disease in a pilot study. *J. Hum. Genet.* **58,** 548–52 (2013).

42. Hwu, W.-L. *et al.* Newborn screening for Fabry disease in Taiwan reveals a high incidence of the later-onset GLA mutation c.936+919G>A (IVS4+919G>A). *Hum. Mutat.* **30,** 1397–405 (2009).

43. Choi, J. C. *et al.* Screening for NOTCH3 gene mutations among 151 consecutive Korean patients with acute ischemic stroke. *J. Stroke Cerebrovasc. Dis.* **22,** 608–14 (2013).

44. Dong, Y. *et al.* Yield of Screening for CADASIL Mutations in Lacunar Stroke and Leukoaraiosis. *Stroke* **34,** 203–206 (2002).

45. Zaucker, A., Mercurio, S., Sternheim, N., Talbot, W. S. & Marlow, F. L. notch3 is essential for oligodendrocyte development and vascular integrity in zebrafish. *Dis. Model. Mech.* **6,** 1246–59 (2013).

46. Adib-Samii, P., Brice, G., Martin, R. J. & Markus, H. S. Clinical spectrum of CADASIL and the effect of cardiovascular risk factors on phenotype: study in 200 consecutively recruited individuals. *Stroke.* **41,** 630–4 (2010).

47. De Brabander, I. *et al.* Phenotypical characterization of α-galactosidase A gene mutations identified in a large Fabry disease screening program in stroke in the young. *Clin. Neurol. Neurosurg.* **115,** 1088–93 (2013).

48. Biesecker, L. G., Shianna, K. V & Mullikin, J. C. Exome sequencing: the expert view. *Genome Biol.* **12,** 128 (2011).

49. Do, R., Kathiresan, S. & Abecasis, G. R. Exome sequencing and complex disease: practical aspects of rare variant association studies. *Hum. Mol. Genet.* **21,** R1–9 (2012).

50. Pruitt, K. D. *et al.* RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res.* **42,** D756–63 (2014).

1.8. Figures



Figure 1.1 Physiological comparison of the major stroke subtypes51

**Chapter Two: An Evaluation of Mendelian Stroke Genes in Young Stroke Patients**

2.1 Introduction

Suffering from stroke at a young age is a distressing event for patients and their families given the current lack of information on both etiology and prognosis. “Mendelian strokes” are genetic disorders characterized by extreme risk of stroke early on in life. Specifically, they are caused by rare protein-altering mutations in single genes1. We hypothesize that rare disruptive protein-altering mutations within such genes are important risk factors for stroke in young patients. To evaluate this hypothesis, 10 known genes were screened in a subset of participants (185 early-onset cases and 185 matched controls) from INTERSTROKE, a large international study of stroke2.

The most extensively studied Mendelian strokes are Cerebral Autosomal Dominant Arteriopathies with Subcortical Infarcts and Leukoencephalopathies (CADASIL)3 and Fabry’s disease4. Both are small-vessel pathologies leading to enormous risk of early-onset stroke. Recurrent ischemic episodes are reported to be present in 71% of CADASIL patients (median onset: 46.1 years)5, and 18.8% of Fabry’s disease patients (mean onset: 38.3 years)6. In the general population, the prevalence of CADASIL is estimated to be 1-2 per 100,000 individuals7,8 and for Fabry’s disease, 14-50 per 100,000 indiviudals9,10. Conversely, in stroke patients, the estimated prevalence is higher but ranges widely from 500-6000 per 100,000 individuals for CADASIL11,12 and 500-3900 per 100,000 individuals for Fabry’s disease13,14.

While rare genetic disorders are considered significant determinants of stroke in isolated families carrying Mendelian mutations, their relevance in a group of unrelated individuals is unclear. Previous studies investigating this issue lacked controls, were not comprehensive in genetic testing, and only included participants from a single ethnicity. In the present study, both cases and controls were screened to properly verify the pathogenicity of Mendelian stroke genes as one would expect a higher frequency in cases. Additionally, this marks the first exploration of multiple Mendelian genes within a single cohort, thus permitting an estimate of the combined prevalence of Mendelian stroke disorders as a whole. Newly identified Mendelian stroke disorders caused by mutations in *SAMHD1* and *CECR1*15,16 were also investigated in the current study. Ultimately, our primary objectives were to characterize the risk, prevalence, and clinical features associated with Mendelian mutations in a diverse population of unrelated patients and to determine which known genes, if any, should be considered for genetic screening. As a subproject of INTERSTROKE, which includes participants from 6 continents, 35 countries, and 18 ethnicities, this study provides globally relevant insights into the genetic basis of young stroke.

2.2 Methods

*Sample Selection, Collection, & Exome Sequencing*

Mendelian strokes tend to affect the small cerebral vessels17,18 and occur in the absence of conventional risk factors19. To ensure maximal representation of Mendelian strokes, the youngest INTERSTROKE cases with small-vessel pathology (ICH or small-vessel ischemic stroke (SVIS)) without risk factors were prioritized. Specifically, we selected the youngest stroke cases that could be matched to an appropriate control (sex, ethnicity, recruitment center). When two or more cases were the same age, the case with the least risk factors (hypertension, diabetes) was chosen. Conversely, when multiple controls could be selected, the oldest control with the most risk factors was chosen. The sample selection process is illustrated in Figure 2.1.

ICH was defined by clinical evaluation and neuroimaging (MRI or CT scans), while SVIS was determined according to Trial of Org 10172 in Acute Ischemic Stroke Treatment (TOAST) guidelines20. Cardiovascular risk factors were defined in the same manner as described for the main INTERSTROKE study2. Hypertension was defined by either self-report or having a blood pressure greater than 160/90 mmHg (mean of 3 measurements for cases). Diabetes, migraine, and depression were self-reported. Specifically, migraine was assessed by asking, “do you have migraine?”, whereas depression was assessed by asking “during the past twelve months, was there ever a time when you felt sad, blue, or depression for two weeks or more in a row”. Blood samples for all study participants were collected in EDTA Whole Blood DNA tubes. DNA was extracted using the QIAGEN QIAsymphony DSP DNA Midi kit. Exomic sequences were captured with the Illumina TrueSeq Exome Enrichment Kit and subsequently sequenced on HiSeq with paired-end reads (2 x 100 bp).

*Variant Calling & Quality Control*

Sequence reads were mapped to the hg19 reference genome with the Burrows Wheeler Aligner21 and processed according to the Genome Analysis Tool Kit (GATK22) protocol (Unified Genotyper) to generate single nucleotide variant (SNV) and insertion/deletion (INDEL) calls. Variant quality control (QC) filters were applied in KGGSEQ23 using default settings, except for the read depth filter which was increased to require eight reads per call. Variants deviating from Hardy Weinberg Equilibrium (HWE) within at least one ethnicity were excluded in all samples. Variants within regions that cannot be confidently called for most next-generation sequencing platforms were excluded based upon Zook *et al.* (2014)24. Variant QC metrics were generated in PicardTools25, BedTools26, GATK22, and SnpSift27. Sample QC included checks for ethnicity, sex, cryptic relatedness, and genotypic concordance with exome chip data. 146 exome sequenced samples were also genotyped on the Illumina HumanExome Chip V 1-1. Using exome chip genotypes as the benchmark, samples with more than 10% discrepant genotypes were excluded. Samples failing any quality control check or left without a matching sample were removed. All QC checks were performed using PLINK28, GCTA29, GATK22, and Variant Tools30. After QC, 185 case-control pairs remained (Table 2.1).

*Candidate Gene Search*

A list of Mendelian stroke genes was compiled from searching the Online Mendelian Inheritance of Man (OMIM) database using the key terms, “stroke” OR “intracerebral hemorrhage” OR “ischemic stroke” OR “cerebral aneurysm” OR “arterial thrombosis”. Search results were limited to those with “phenotype description, molecular basis known”, “Mendelian phenotype or locus, molecular basis unknown”, or “other, mainly phenotypes with suspected Mendelian basis”. Genes were excluded if (1) stroke was not the primary feature, (2) cardieombolic stroke, thoracic aortic aneurysm, large-vessel strokes, or venous thrombosis was the primary subtype, (3) disease onset was not adult (neonatal or pediatric), (4) they did not show clear evidence of Mendelian inheritance, or (5) they were encoded in the mitochondrial genome. More recently discovered stroke genes not yet recognized by the OMIM search, *CECR1* 15and *SAMHD1*16, were also added resulting in 10 stroke genes in total (S. Table 2.1). All candidate genes were reviewed by stroke neurologist, Dr. Michael Sharma, to ensure clinical relevance.

*Gene Association Testing*

All rare disruptive protein-altering mutations within the same gene were treated as a single unit for association testing. Rare mutations were defined as those having a minor allele frequency (MAF) less than 5% in all 185 INTERSTROKE controls and participants from both the NHLBI GO Exome Sequencing Project (ESP) and 1000 Genomes (1KG). This MAF threshold was applied within each ethnic subdivision for ESP and 1KG. If a variant was common (MAF>0.05) to even a single ethnicity within ESP or 1KG, then it was excluded. Disruptiveness was determined using two criteria: either 1) manual curation of known disease-causing variants (OMIM1, UNIPROT31, or PubMed) or 2) bioinformatic prediction of “disease-causing” or “probably-disease causing” variants in either Polyphen-II (HDIV or HVAR)32 or SIFT33.

To assess whether individual genes carried an excess burden of rare disruptive mutations in cases, a two-sided Fisher’s exact test was used in R34. Three separate analyses for all stroke, ICH, and SVIS status were conducted. P-values were corrected for multiple hypothesis testing by experiment-wise permutation of phenotypes. Case-control status was permutated within pairs to maintain appropriate matching by ethnicity, sex, and recruitment center. From every permutation, the most significant p-value out of all genes was extracted. 1000 permutations were performed thus generating a null distribution of 1000 permutated p-values for each analysis (all stroke, ICH, SVIS). Adjusted p-values were calculated as the proportion of permutated p-values that were more significant than the initial p-value.

*Characterization of CADASIL mutations*

Variants within the EGFR domains of *NOTCH3* were considered CADASIL-causing if 1) they involved the gain or loss of a cysteine35, 2) they resulted in an INDEL36, or 3) they affected the *NOTCH3* R75 residue37,38. These criteria are based upon features of known, disease-causing CADASIL mutations.

*Other Statistical Analyses*

Characteristics between cases and controls were compared using the two-sided Student’s t-test for quantitative variables and two-sided Fisher’s test for count data in R34. Using Variant Tools30, HWE was evaluated within the controls of individual ethnicities (P<0.05/N).

*Statistical Power*

 Power calculations were performed using the “statmod” package in R assuming an allele frequency of 5%, and accordingly, a mutation carrier frequency of 9.75%. Sample sizes were sufficiently powered to detect associations of OR=2.32 for all stroke, OR=2.98 for ICH, and OR=3.11 for SVIS.

2.3 Results

*Study Subject Characteristics*

Overall, men comprised 57.6% of all study participants (Table 2.1). On average, cases (46.7 years) were younger than controls (74.1 years) (P<2.2x10-16). ICH and SVIS accounted for 54% and 46% of all stroke cases, respectively. Among all participants, there were 120 (32.4%) Europeans, 86 (23.2%) Latin Americans, 70 (18.9%) South East Asians, 48 (13%) Africans, 30 (8.1%) Persians, 8 (2.2%) South Asians, and 8 (2.2%) others. As expected because of the selection criteria, the prevalence of hypertension and diabetes did not vary significantly between cases and controls as 51.3% of cases and 57.3% of controls were hypertensive (P=0.30) and 18.9% of cases and 14.6% of controls were diabetic (P=0.33). However, other risk factors such as BMI (27.3 vs. 26.1; P=0.02), current smoking status (28.1% vs. 9.2%; P=3.94x10-6), and APOB/APOA1 ratio (0.88 vs. 0.80; P=0.01), were elevated in cases. Parental history of stroke was similar between cases (16.2%) and controls (13.5%) (P=0.56).

*Mendelian Stroke Gene Analysis*

The presence of rare disruptive mutations within any of the 10 Mendelian genes was significantly associated with stroke status (OR=1.62; 95% CI, 1.02-2.62; P=0.04) (Table 2.2). Overall, 66 (35.7%) cases and 47 (25.4%) controls were mutation carriers. Stratifying by stroke subtype, the presence of mutations was associated with ICH status (OR=2.82; 95% CI, 1.46-5.59; P=0.001) but not SVIS status (OR=0.84; 95% CI, 0.41-1.72; P=0.74). 43 (43%) ICH cases and 21 (21%) controls were mutation carriers. Upon closer examination of individual genes, there was only one significant gene association. *NOTCH3* (P=0.009)was significantly associated with ICH status and remained significant after adjustment for multiple hypothesis testing (P=0.01). *COL4A1* exhibited a trend towards significance (P=0.06) but not after adjustment (P=0.12). 22 (22%) ICH cases and 8 (8%) controls were *NOTCH3* mutation carriers (OR=3.23; 95% CI: 1.29 to 8.87; P=0.009). Notably, one ICH case had two rare disruptive *NOTCH3* mutations. To evaluate whether this association was solely accounted for by *NOTCH3* mutations, we tested whether significance persisted after excluding *NOTCH3.* The presence of mutations within any of the other nine genes was not associated with ICH status, nor was there a trend towards association (OR=1.77; 95% CI: 0.79-4.13; P=0.19).

*NOTCH3 Analysis*

Rare *NOTCH3* CADASIL-causingmutations follow a highly stereotyped pattern. We evaluated whether such mutations contributed to the *NOTCH3* signal. Overall, there were six (3.2%) stroke cases, of which three (3%) were ICH and three (3.5%) were SVIS, carrying putative CADASIL mutations (Table 2.3). No CADASIL mutation was observed in controls. The presence of CADASIL mutations was significantly associated with stroke status (P=0.03). The ethnic distribution was Persian (3), African (2), and European (1). Migraine and depression are common secondary symptoms of CADASIL39. However, migraine was absent, and depression was only present in a single carrier. Additionally, five of the six mutation carriers had at least one cardiovascular risk factor.

Excluding CADASIL mutations, the presence of other *NOTCH3* mutations *per se* conferred significantly higher risk to ICH (OR=2.86; 95% CI, 1.13 to 7.93; P=0.02). 19 (86.4%) of the 22 total *NOTCH3* mutation carriers with ICH possessed mutations of the non-CADASIL variety. Non-CADASIL *NOTCH3* mutations were observed among multiple ethnic groups; there were nine European, six African, two Latin American, one South East Asian, and one individual of unknown ethnicity who carried non-CADASIL *NOTCH3* mutations. Stratifying by ethnicity, the presence of non-CADASIL *NOTCH3* mutations was not associated with ICH status in any single ethnicity though we were underpowered to detect associations within individual ethnic groups. Among all stroke cases, the prevalence of migraine and depression was similar between non-CADASIL mutation carriers and non-carriers (Table 2.4). Specifically, migraine was present in one (4.2%) non-CADASIL mutation carrier as compared to 20 (13.0%) non-carriers, whereas depression was present in seven (29.2%) non-CADASIL mutation carriers as compared to 44 (28.6%) non-carriers.

*Previously Reported Disease-Causing Mutations*

Known Mendelian stroke mutations were found across *CECR1, COL4A2, GLA,* and *NOTCH3* genes (Table 2.5). In total, there were 11 (2.97%) mutation carriers, of which six (3.2%) were cases and five (2.7%) were controls. The presence of these mutations was not significantly associated with stroke (OR=1.20; 95% CI, 0.30-5.06; P=1), ICH (OR=0.74; 95% CI, 0.11-4.52; P=1), or SVIS status (OR=3.05; 95% CI, 0.24-163.08; P=0.62).

2.4 Discussion

*Summary of Findings*

 The presence of rare disruptive mutations within Mendelian stroke genes significantly increases the risk of stroke in young patients (OR=1.62; 95% CI, 1.02-2.62; P=0.04). Specifically, *NOTCH3* mutations were associated with ICH susceptibility (OR=3.23; 95% CI: 1.29 to 8.87; P=0.009), and there was also a trend towards significant for *COL4A1* mutations and ICH (P=0.06). Our findings also confirm the pathogenicity of CADASIL mutations, which were exclusively found in cases, but not Mendelian mutations in other genes that have been previously reported as disease-causing. Finally, a novel role for non-CADASIL *NOTCH3* mutations was uncovered. While individually rare, collectively, these mutations were common risk factors for ICH, being present in 19% of ICH cases and 8% of controls (OR=2.86; 95% CI, 1.13 to 7.93; P=0.02).

*NOTCH3 mutations are important risk factors for young stroke*

Based on our findings, *NOTCH3* mutations can be categorized into two classes with respect to stroke risk: 1) CADASIL mutations which are very rare and lead to an extreme risk of stroke, and 2) non-CADASIL mutations that are more common and lead to a more moderate, albeit still substantial, risk of stroke. Non-CADASIL *NOTCH3* mutationshave not been studied in the context of ICH; however, Schmidt *et al.* (2011) also identified rare disruptive mutations in elderly individuals with severe white matter lesions, a hallmark of small-vessel disease 40. Moreover, Fouillade *et al.* (2008) pinpointed another non-CADASIL *NOTCH3*  variant, L1515P, as the cause of a hereditary small-vessel disease that was distinct from CADASIL pathology41. In conjunction with these studies, there may be a more pervasive role for *NOTCH3* mutations in stroke than formerly recognized.

*Clinical Implications*

CADASIL is characterized by SVIS, migraine, and depression, among other debilitating symptoms5. In contrast, putative CADASIL mutation carriers deviated from this classic presentation. Firstly, ICH was the primary manifestation in 50% of carriers, suggesting that ICH is a major feature of CADASIL. While the number of CADASIL patients identified in our study was small, this notion is consistent with larger studies reporting ICH in ~25% of South East Asian CADASIL patients42–46. Secondly, migraine was not reported by any mutation carriers, which was unexpected given that it is more prevalent than stroke in some Caucasian studies19,39,47. Conversely, Wang *et al.*(2009) demonstrated that migraine was very rare (5%) in Chinese CADASIL patients48. Even the occurrence of MRI features once believed to be pathognomonic of CADASIL is less common in non-Caucasians42,47–49.

Altogether, our results are consistent with emerging CADASIL literature suggesting that classic CADASIL symptoms may be deceiving due to variability across ethnic groups. Mutation carriers generally resembled other young stroke patients without *NOTCH3* mutations. As such, *NOTCH3* screening might be warranted in all young stroke patients regardless of clinical presentation. While the risk of stroke is extremely high in CADASIL mutation carriers, it is not absolute. Conventional risk factors, especially hypertension and smoking, still modulate symptom severity and onset in CADASIL patients19,39. As an example, five of the six CADASIL mutation carriers in our study still had at least one cardiovascular risk factor. Accordingly, therapy should focus on aggressively minimizing the presence of these conventional risk factors. “Cascade testing” could be implemented to identify entire families at high genetic risk for stroke. Our research also has specific implications for *NOTCH3* screening. Current diagnostic testing for CADASIL only encompasses the EGFR domains and leaves approximately 30% of potentially functional coding regions unexamined. 9 (47.3%) *NOTCH3* mutation carriers with ICH had mutations beyond EGFR domains (S. Table 2.2). To comprehensively detect disease-causing *NOTCH3* mutations, the spectrum of mutations should be expanded to all coding regions of the gene.

*Research Implications*

Our findings corroborate the rare-variant common disease hypothesis50, which asserts that many individually rare variants of large-effect account for a substantial proportion of diseased cases in the population. Additionally, using diverse cohorts appears to be an effective way to study rare genetic disorders. For example, had participants been limited to solely Caucasians, the sample size would need to be more than seven times as large to be equally powered to detect the significant case-enrichment in CADASIL mutations. Results also highlight the importance of controls since a similar frequency between cases and controls was observed for mutations presumed to be validated for clinical pathogenicity. This also suggests that some known, disease-causing mutations either confer minor risk to stroke or are false-positives. We identified one such mutation, *GLA* D313Y, which was originally assumed to be pathogenic in a Fabry’s disease patient with two different *GLA* variants. However, after Yasuda *et al.* (2003) demonstrated its presence in 4 (0.5%) out of 800 control chromosomes51, it became clear that this mutation was a benign polymorphism. In our study, the same variant was identified in a single control (0.54%), which is consistent with non-pathogenicity. Overall, our data suggest that researchers should be cautious to label variants as disease-causing on the basis of case-only observations.

*Limitations*

There were several limitations in our study: 1) potentially important mitochondrial stroke disorders52–56 were not evaluated, 2) the pathogenicity of putative CADASIL mutations were not verified by MRI features or skin biopsy testing for granular osmiophilic material57, 3) the ESP database used for filtering variants based on frequency did not make stroke status publically available, 4) we did not have perfect coverage of candidate genes denoting that causal variants could have been missed (e.g. *HTRA1* and *CST3*) (S. Table 2.3), 5) implications of research are restricted to small-vessel strokes (ICH or SVIS), and 6) we evaluated the role of 10 known Mendelian stroke genes, but many more novel genes yet to be identified may also influence stroke susceptibility.

2.5 Conclusion

To the best of our knowledge, we conducted the first case-control study of Mendelian stroke. The presence of rare disruptive mutations within Mendelian genes was associated with stroke status, particularly ICH. *NOTCH3* was the major gene driving the association. In contrast to putative CADASIL mutations which were the most rare, other non-CADASIL *NOTCH3* mutations were pervasive and potentially represent a novel class of disease-causing *NOTCH3* mutations. Classic CADASIL symptomatology may be deceiving as migraine and depression were uncommon among mutation carriers, whereas ICH was actually common. Consequently, young stroke patients should be tested for all rare disruptive *NOTCH3* mutations even in the absence of CADASIL signs. Ultimately, our results strongly support a role for rare mutations in early-onset stroke. Future studies should focus on more precisely defining the clinical features of *NOTCH3* mutation carriers and verifying the pathogenicity of specific mutations in larger cohorts.

2.6 References

1. Oyston, J. Online Mendelian Inheritance in Man. *Anesthesiology* **89,** 811–2 (1998).

2. O’Donnell, M. J. *et al.* Risk factors for ischaemic and intracerebral haemorrhagic stroke in 22 countries (the INTERSTROKE study): a case-control study. *Lancet* **376,** 112–23 (2010).

3. NOTCH3 mutations in CADASIL cause stroke and dementia (1998) Joutel et al.pdf.

4. Garman, S. C. & Garboczi, D. N. The molecular defect leading to Fabry disease: structure of human alpha-galactosidase. *J. Mol. Biol.* **337,** 319–35 (2004).

5. Dichgans, M. *et al.* The phenotypic spectrum of CADASIL: clinical findings in 102 cases. *Ann. Neurol.* **44,** 731–9 (1998).

6. Mehta, a *et al.* Fabry disease: a review of current management strategies. *QJM* **103,** 641–59 (2010).

7. Razvi, S. S. M., Davidson, R., Bone, I. & Muir, K. W. The prevalence of cerebral autosomal dominant arteriopathy with subcortical infarcts and leucoencephalopathy (CADASIL) in the west of Scotland. *J. Neurol. Neurosurg. Psychiatry* **76,** 739–41 (2005).

8. Narayan, S. K., Gorman, G., Kalaria, R. N., Ford, G. A. & Chinnery, P. F. The minimum prevalence of CADASIL in northeast England. *Neurology* **78,** 1025–7 (2012).

9. Inoue, T. *et al.* Newborn screening for Fabry disease in Japan: prevalence and genotypes of Fabry disease in a pilot study. *J. Hum. Genet.* **58,** 548–52 (2013).

10. Hwu, W.-L. *et al.* Newborn screening for Fabry disease in Taiwan reveals a high incidence of the later-onset GLA mutation c.936+919G>A (IVS4+919G>A). *Hum. Mutat.* **30,** 1397–405 (2009).

11. Choi, J. C. *et al.* Screening for NOTCH3 gene mutations among 151 consecutive Korean patients with acute ischemic stroke. *J. Stroke Cerebrovasc. Dis.* **22,** 608–14 (2013).

12. Dong, Y. *et al.* Yield of Screening for CADASIL Mutations in Lacunar Stroke and Leukoaraiosis. *Stroke* **34,** 203–206 (2002).

13. Rolfs, A. *et al.* Prevalence of Fabry disease in patients with cryptogenic stroke: a prospective study. *Lancet* **366,** 1794–6 (2005).

14. Rolfs, A. *et al.* Acute cerebrovascular disease in the young: the stroke in young fabry patients study. *Stroke.* **44,** 340–9 (2013).

15. Zhou, Q. *et al.* Early-Onset Stroke and Vasculopathy Associated with Mutations in ADA2. *N. Engl. J. Med.* 140219140049003 (2014). doi:10.1056/NEJMoa1307361

16. Xin, B. *et al.* Homozygous mutation in SAMHD1 gene causes cerebral vasculopathy and early onset stroke. (2011). doi:10.1073/pnas.1014265108/-/DCSupplemental.www.pnas.org/cgi/doi/10.1073/pnas.1014265108

17. Volonghi, I. *et al.* Role of COL4A1 in basement-membrane integrity and cerebral small-vessel disease. The COL4A1 stroke syndrome. *Curr. Med. Chem.* **17,** 1317–24 (2010).

18. Meschia, J. F. New information on the genetics of stroke. *Curr. Neurol. Neurosci. Rep.* **11,** 35–41 (2011).

19. Ciolli, L. *et al.* Influence of vascular risk factors and neuropsychological profile on functional performances in CADASIL: results from the MIcrovascular LEukoencephalopathy Study (MILES). *Eur. J. Neurol.* **21,** 65–71 (2014).

20. Adams, H. P. *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).

21. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–60 (2009).

22. DePristo, M. a *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43,** 491–8 (2011).

23. Li, M.-X. *et al.* Predicting mendelian disease-causing non-synonymous single nucleotide variants in exome sequencing studies. *PLoS Genet.* **9,** e1003143 (2013).

24. Zook, J. M. *et al.* Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nat. Biotechnol.* **32,** 246–51 (2014).

25. Picard. No Title. at < http://picard.sourceforge.net>

26. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–2 (2010).

27. Cingolani, P. *et al.* Using Drosophila melanogaster as a Model for Genotoxic Chemical Mutational Studies with a New Program, SnpSift. *Front. Genet.* **3,** 35 (2012).

28. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81,** 559–75 (2007).

29. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88,** 76–82 (2011).

30. San Lucas, F. A., Wang, G., Scheet, P. & Peng, B. Integrated annotation and analysis of genetic variants from next-generation sequencing studies with variant tools. *Bioinformatics* **28,** 421–2 (2012).

31. Leinonen, R. *et al.* UniProt archive. *Bioinformatics* **20,** 3236–7 (2004).

32. Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet.* **Chapter 7,** Unit7.20 (2013).

33. Teer, J. K., Green, E. D., Mullikin, J. C. & Biesecker, L. G. VarSifter: visualizing and analyzing exome-scale sequence variation data on a desktop computer. *Bioinformatics* **28,** 599–600 (2012).

34. R Development Core Team. R: A language and environment for statistical computing. (2008). at <http://www.r-project.org>

35. Peters, N. *et al.* Spectrum of mutations in biopsy-proven CADASIL: implications for diagnostic strategies. *Arch. Neurol.* **62,** 1091–4 (2005).

36. Mazzei, R. *et al.* A novel Notch3 gene mutation not involving a cysteine residue in an Italian family with CADASIL. *Neurology* **63,** 561–564 (2004).

37. Mizuno, T. *et al.* Two Japanese CADASIL Families Exhibiting Notch3 Mutation R75P Not Involving Cysteine Residue. *Intern. Med.* **47,** 2067–2072 (2008).

38. Kim, Y.-E. *et al.* Spectrum of NOTCH3 mutations in Korean patients with clinically suspicious cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy. *Neurobiol. Aging* **35,** 726.e1–6 (2014).

39. Adib-Samii, P., Brice, G., Martin, R. J. & Markus, H. S. Clinical spectrum of CADASIL and the effect of cardiovascular risk factors on phenotype: study in 200 consecutively recruited individuals. *Stroke.* **41,** 630–4 (2010).

40. Schmidt, H. *et al.* Genetic variants of the NOTCH3 gene in the elderly and magnetic resonance imaging correlates of age-related cerebral small vessel disease. *Brain* **134,** 3384–97 (2011).

41. Fouillade, C. *et al.* Activating NOTCH3 mutation in a patient with small-vessel-disease of the brain. *Hum. Mutat.* **29,** 452 (2008).

42. Lee, Y.-C. *et al.* Population-specific spectrum of NOTCH3 mutations, MRI features and founder effect of CADASIL in Chinese. *J. Neurol.* **256,** 249–55 (2009).

43. Lian, L. *et al.* Spontaneous intracerebral hemorrhage in CADASIL. *J. Headache Pain* **14,** 98 (2013).

44. Choi, J. C., Kang, S.-Y., Kang, J.-H. & Park, J.-K. Intracerebral hemorrhages in CADASIL. *Neurology* **67,** 2042–4 (2006).

45. Rinnoci, V. *et al.* Cerebral hemorrhages in CADASIL: report of four cases and a brief review. *J. Neurol. Sci.* **330,** 45–51 (2013).

46. Yagi, T., Konoeda, F., Mizuta, I., Mizuno, T. & Suzuki, N. Increasing microbleeds in CADASIL. *Eur. Neurol.* **69,** 352–3 (2013).

47. Pantoni, L. *et al.* Comparison of clinical, familial, and MRI features of CADASIL and NOTCH3-negative patients. *Neurology* **74,** 57–63 (2010).

48. Wang, Z. *et al.* NOTCH3 mutations and clinical features in 33 mainland Chinese families with CADASIL. *J. Neurol. Neurosurg. Psychiatry* **82,** 534–9 (2011).

49. Van den Boom, R., Lesnik Oberstein, S. A. J., Ferrari, M. D., Haan, J. & van Buchem, M. A. Cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy: MR imaging findings at different ages--3rd-6th decades. *Radiology* **229,** 683–90 (2003).

50. Bodmer, W. & Bonilla, C. Common and rare variants in multifactorial susceptibility to common diseases. *Nat. Genet.* **40,** 695–701 (2008).

51. Yasuda, M. *et al.* Fabry disease: characterization of alpha-galactosidase A double mutations and the D313Y plasma enzyme pseudodeficiency allele. *Hum. Mutat.* **22,** 486–92 (2003).

52. Sasarman, F., Antonicka, H. & Shoubridge, E. a. The A3243G tRNALeu(UUR) MELAS mutation causes amino acid misincorporation and a combined respiratory chain assembly defect partially suppressed by overexpression of EFTu and EFG2. *Hum. Mol. Genet.* **17,** 3697–707 (2008).

53. Goto, M. *et al.* MELAS phenotype associated with m.3302A>G mutation in mitochondrial tRNA(Leu(UUR)) gene. *Brain Dev.* 10–12 (2013). doi:10.1016/j.braindev.2013.03.001

54. Goodfellow, J. A. *et al.* Mitochondrial myopathy, encephalopathy, lactic acidosis and stroke-like episodes: an important cause of stroke in young people. *Postgrad. Med. J.* **88,** 326–34 (2012).

55. Conforto, A. B. *et al.* Screening for MELAS mutations in young patients with stroke of undetermined origin. *Arq. Neuropsiquiatr.* **65,** 371–6 (2007).

56. Blakely, E. L. *et al.* Pathogenic mitochondrial tRNA point mutations: nine novel mutations affirm their importance as a cause of mitochondrial disease. *Hum. Mutat.* 1–30 (2013). doi:10.1002/humu.22358

57. Morroni, M. *et al.* Role of electron microscopy in the diagnosis of cadasil syndrome: a study of 32 patients. *PLoS One* **8,** e65482 (2013).

2.7 Tables

Table 2.1.Characteristics of study subjects.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | SVIS | ICH | AllCase | AllControl | All StrokeP-value | SVIS P-value | ICH P-value |
| N | 85 | 100 | 185 | 185 | - | - | - |
| Men, n (%) | 49(57.6%) | 53(53%) | 102(55.1%) | 102(55.1%) | 1 | 1 | 1 |
| Age(SD) | 50.1(9.79) | 45.6(12.55) | 47.6(11.55) | 74.1(8.36) | <2.2x10-16 | <2.2x10-16 | <2.2x10-16 |
| Hypertension(%) | 46(54.12%) | 49(49%) | 95(51.35%) | 106(57.30%) | 0.30 | 1 | 0.12 |
| Diabetes(%) | 23(27.06%) | 12(12%) | 35(18.92%) | 27(14.59%) | 0.33 | 0.20 | 1 |
| Current Smoker(%) | 31(36.47%) | 21(21%) | 52(28.11%) | 17(9.19%) | 3.94x10-6 | 4.08 x10-5 | 0.03 |
| BMI(SD) | 26.9(4.88) | 27.7(5.51) | 27.3(5.23) | 26.1(4.83) | 0.02 | 0.14 | 0.08 |
| APOB/APOA1(SD) | 0.90(0.29) | 0.87(0.28) | 0.88(0.28) | 0.80(0.31) | 0.01 | 0.05 | 0.12 |
| Parental History(%) | 23(27.1%) | 7(7%) | 30(16.2%) | 25(13.5%) | 0.56 | 0.47 | 1 |

Table 2.2Mutation carrier counts for Mendelian stroke genes.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Gene | SVIS Case | SVIS Control | SVIS P-value | ICH Case | ICH Control | ICH P-value | All StrokeCase | All Stroke Control | All StrokeP-value |
| *APP* | 2 | 2 | 1 | 1 | 1 | 1 | 3 | 3 | 1 |
| *CECR1* | 0 | 0 | 1 | 2 | 1 | 1 | 2 | 1 | 1 |
| *COL4A1* | 2 | 3 | 1 | 5 | 0 | 0.06 | 7 | 3 | 0.33 |
| *COL4A2* | 12 | 12 | 1 | 11 | 9 | 0.81 | 23 | 21 | 0.87 |
| *CST3* | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 |
| *GLA* | 0 | 0 | 1 | 0 | 2 | 0.50 | 0 | 2 | 0.50 |
| *HTRA1* | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 1 |
| *NOTCH3* | 7 | 8 | 1 | 22 | 8 | 0.009\*\* | 29 | 16 | 0.06 |
| *SAMHD1* | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 |
| *TREX1* | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 |
| Total | **23** | **26** | **0.74** | **43** | **21** | **0.001\*\*** | **66** | **47** | **0.04\*** |

Table 2.3.Clinical features of putative CADASIL mutation carriers.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Variant | ESP MAF | Stroke Type | Ethnicity | Sex | Risk Factors | Age | Migraine | Depression | Parental Stroke |
| R1285C | Novel  | SVIS | Persian | F | Diabetes, APOB/APOA1 | 45 | No | No | Yes |
| R1231C | 7.7x10-5 | ICH | Persian | F | None | 52 | No | No | No |
| R1231C | 7.7x10-5 | SVIS | Persian | F | Diabetes, APOB/APOA1 | 45 | No | Yes | No |
| R1100C | Novel | ICH | African | F | Obesity | 50 | No | No | No |
| R75W | Novel | ICH | African | M | Hypertension, Obesity | 60 | No | No | No |
| F1273- | Novel | SVIS | European | M | Hypertension | 54 | No | No | No |

Table 2.4. Comparison of secondary CADASIL features among cases.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | CADASILMutation Carrier(CAD) | NON-CADASIL Mutation Carrier(NCAD) | NON-CARRIER(NCAR) | CAD/NCARP-value | NCAD/NCARP-value |
| N | 6 | 24 | 155 | – | – |
| ICH/SVIS | 3/3 | 19/5 | 78/77 | – | – |
| Migraine | 0 (0%) | 1(4.2%) | 20 (13.0%) | 1 | 0.29 |
| Depression | 1 (16.7%) | 7 (29.2%) | 44 (28.6%) | 1 | 1 |
| Parental Stroke | 1 (16.7%) | 3 (12.5%) | 26 (16.9%) | 1 | 0.77 |

Table 2.5. Summary ofknown disease-causing mutation carrier counts.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Gene | Disease | Variant | ESP MAF | SVIS Case | SVIS Control | ICH Case | ICH Control | All StrokeCase | All Stroke Control |
| *CECR1* | Stroke and Vasculopathy | G47W | . | 0 | 0 | 0 | 1 | 0 | 1 |
| *COL4A2* | ICH | E1123G | 9.78x10-3 | 2 | 1 | 0 | 2 | 2 | 3 |
| *COL4A2* | ICH | Q1150K | 6.67x10-3 | 0 | 0 | 2 | 0 | 2 | 0 |
| *GLA* | Fabry's Disease | D313Y | 2.94x10-3 | 0 | 0 | 0 | 1 | 0 | 1 |
| *NOTCH3* | CADASIL | R1231C | 7.70x10-5 | 1 | 0 | 1 | 0 | 2 | 0 |
| Total | **–** | **–** | **–** | **3** | **1** | **3** | **4** | **6** | **5** |
| P-value | **–** | **–** | **–** | **0.62** | **1** | **1** |

2.8 Supplementary Material

Supplementary Table 2.1. Candidate Mendelian stroke genes

|  |  |  |
| --- | --- | --- |
| Gene | Disease | Mode of Inheritance |
| *APP*  | Amyloid Angiopathies  | Dominant  |
| *CST3*  | Amyloid Angiopathies  | Dominant  |
| *CECR1*  | Early-onset stroke and cerebral vasculopathy  | Recessive  |
| *COL4A1*  | Porencephaly / ICH  | Dominant  |
| *COL4A2*  | Porencephaly / ICH  | Dominant  |
| *GLA*  | Fabry’s Disease  | Recessive  |
| *HTRA1*  | CARASIL  | Recessive  |
| *NOTCH3*  | CADASIL  | Dominant  |
| *SAMHD1*  | Early-onset stroke and cerebral vasculopathy  | Recessive  |
| *TREX1*  | Retinal vasculopathy with cerebral leukodystrophy | Dominant  |

Supplementary Table 2.2.Non-CADASIL *NOTCH3* mutation carrier counts.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Variant | ESP MAF | In EGFR domain? | SVIS Case | SVISControl | SVIS P-value | ICH Case | ICH Control | ICH P-value | All Stroke Case | All Stroke Control | All Stroke P-value |
| A64D | 0.0006 | Yes | 0 | 0 | N/A | 1 | 0 | 1 | 1 | 0 | 1 |
| V159M | Novel | Yes | 0 | 0 | N/A | 0 | 1 | 1 | 0 | 1 | 1 |
| H170R | 0.0011 | Yes | 1 | 0 | 1 | 0 | 0 | N/A | 1 | 0 | 1 |
| V237M | Novel | Yes | 0 | 0 | N/A | 1 | 0 | 1 | 1 | 0 | 1 |
| D367N | Novel | Yes | 0 | 0 | N/A | 0 | 1 | 1 | 0 | 1 | 1 |
| P380L | Novel | Yes | 0 | 0 | N/A | 1 | 0 | 1 | 1 | 0 | 1 |
| P496L | 0.0155 | Yes | 1 | 1 | 1 | 3 | 1 | 0.62 | 4 | 2 | 0.68 |
| T575M | 0.0001 | Yes | 0 | 0 | N/A | 1 | 0 | 1 | 1 | 0 | 1 |
| A979T | Novel | Yes | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 1 |
| G1105A | 0.0013 | Yes | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 1 |
| H1133Q | 0.0131 | Yes | 0 | 0 | N/A | 4 | 2 | 0.68 | 4 | 2 | 0.68 |
| L1518M | 0.0014 | No | 2 | 3 | 1 | 2 | 1 | 1 | 4 | 4 | 1 |
| R1560P | 0.0035 | No | 0 | 1 | 1 | 2 | 0 | 0.50 | 2 | 1 | 1 |
| E1638Q | Novel | No | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 1 |
| R1834G | Novel | No | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 1 |
| R1857Q | 0.0001 | No | 0 | 1 | 1 | 0 | 0 | N/A | 0 | 1 | 1 |
| V1952M | 0.0078 | No | 1 | 1 | 1 | 3 | 1 | 0.62 | 4 | 2 | 0.68 |
| EGFR | **–** | **11** | **2** | **2** | **1** | **11** | **6** | **0.31** | **13** | **9** | **0.51** |
| Non-EGFR | **–** | **6** | **3** | **6** | **0.50** | **9** | **2** | **0.06** | **12** | **8** | **0.49** |
| Total | **–** | **17** | **5** | **8** | **0.60** | **20** | **8** | **0.02\*** | **25** | **16** | **0.18** |

Supplementary Table 2.3.Coverage metrics for candidate genes.

|  |  |  |  |
| --- | --- | --- | --- |
| Gene | Average Coverage (SD) | % Targets > 8x(SD) | % Targets > 20x (SD) |
| *APP*  | 33.00 (9.02) | 86.59(1.73) | 69.17(13.63) |
| *CECR1*  | 34.93 (9.59) | 99.48(1.82) | 81.13(16.03) |
| *COL4A1*  | 50.81 (11.89) | 94.26(3.29) | 79.81(8.81) |
| *COL4A2*  | 60.26 (14.08) | 96.33 (2.23) | 87.48(5.63) |
| *CST3*  | 14.19 (4.16) | 52.78(12.96) | 34.30(13.05) |
| *GLA*  | 20.20(8.40) | 90.10(13.49) | 43.53(29.76) |
| *HTRA1*  | 34.67(8.92) | 66.23(2.21) | 56.26(6.40) |
| *NOTCH3*  | 73.48(17.39) | 89.46(2.32) | 82.50(4.66) |
| *SAMHD1*  | 30.99(10.02) | 98.59(4.26) | 76.11(21.41) |
| *TREX1*  | 157.24(36.46) | 93.38(1.15) | 90.62(1.34) |

Supplementary Table 2.4. Allele counts for candidate genes.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Gene | SVIS Case | SVIS Control | SVIS P-value | ICH Case | ICH Control | ICH P-value | All StrokeCase | All Stroke Control | All StrokeP-value |
| *APP* | 2 | 2 | 1 | 1 | 1 | 1 | 3 | 3 | 1 |
| *CECR1* | 0 | 0 | 1 | 2 | 1 | 1 | 2 | 1 | 1 |
| *COL4A1* | 2 | 3 | 1 | 5 | 0 | 0.06 | 7 | 3 | 0.34 |
| *COL4A2* | 13 | 14 | 1 | 13 | 9 | 0.50 | 26 | 23 | 0.77 |
| *CST3* | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 |
| *GLA* | 0 | 0 | 1 | 0 | 2 | 0.50 | 0 | 2 | 0.50 |
| *HTRA1* | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 1 |
| *NOTCH3* | 8 | 8 | 1 | 23 | 8 | 0.008\*\* | 31 | 16 | 0.03\* |
| *SAMHD1* | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 |
| *TREX1* | 0 | 1 | 1 | 2 | 0 | 0.50 | 2 | 1 | 1 |
| Total | **25** | **28** | **0.74** | **47** | **21** | **0.0002\*\*\*** | **72** | **49** | **0.01\*** |

**Chapter Three: A Systematic Exploration of Rare Variants influencing risk of Small-Vessel Stroke**

3.1 Introduction

Stroke is the third most common cause of death worldwide and the leading cause of adult-acquired disability in Canada1,2. INTERSTROKE, a large international case-control study, demonstrated that approximately 90% of stroke risk can be attributed to 10 conventional risk factors3. Other undetermined risk factors, such as genetics, may explain the remaining portion of stroke risk, especially in young patients in which conventional risk factors are less prevalent3. We previously conducted exome sequencing in 185 early-onset case-control pairs and demonstrated that rare disruptive mutations within Mendelian stroke genes increased stroke risk, yet only a minority of cases had mutations. In the present investigation, we extend the previous analysis to an agnostic exploration of all genes. Thus, we hypothesize that rare disruptive protein-altering mutations in novel genes and pathways are important risk factors for early stroke.

Only one other case-control study of stroke using exome sequencing has been published to date. The NHLBI GO Exome Sequencing Project (ESP) replicated a known association with *PON1*4,5 (P= 3.01x10-3) in 496 ischemic stroke cases. While this study was important in revealing a role for rare mutations within *PON1*, common variants within this gene have already been associated with ischemic stroke6. Alternatively, novel genes and pathways may play a role, where pathways are defined as a collection of genes involved in the same biological process. For instance, common mutations within genes related to oxidative phosphorylation7 and blood pressure8 are associated with increased stroke risk. Whether rare variants in these or other pathways alter stroke risk is uncertain. Secondly, the ESP study consisted primarily of Europeans, thus limiting the generalisability of its implications. Lastly, ischemic stroke subtypes were pooled together and treated as a single disease outcome despite recognized heterogeneity across subtypes9.

The current study is the first agnostic exploration of rare variants in the context of stroke. Exome sequencing was conducted in an ethnically diverse case-control study of small-vessel stroke (100 ICH and 85 SVIS). Not only does this study have globally relevant implications for stroke genetics, but it is also the first exome sequencing study of ICH. An extreme phenotype study design was adopted to enrich for a genetic etiology to stroke. Specifically, the youngest cases without major risk factors were matched to the oldest controls with risk factors. Accordingly, our main objective was to systematically identify genes and pathways associated with either protective or harmful effects.

3.2 Methods

*Sample Selection, Collection, & Exome Sequencing*

The youngest INTERSTROKE cases of small-vessel subtype (ICH or SVIS) were matched with the oldest controls by sex, ethnicity, and recruitment center. The rationale for selecting cases without conventional stroke risk factors and controls with risk factors was to enrich for a genetic basis for stroke in the phenotypic extremes. Specifically, we selected the youngest stroke cases that could be matched to an appropriate control (sex, ethnicity, recruitment center). When two or more cases were the same age, the case with the least risk factors (hypertension, diabetes) was chosen. Conversely, when multiple controls could be selected, the oldest control with the most risk factors was chosen. Age is the strongest risk factor for stroke10 and was prioritized above all other risk factors. Hypertension and diabetes were minimized in cases but maximized in controls when possible. Because the major stroke risk factors do not account for controls’ protection and cases’ susceptibility to stroke, genetic factors are more likely to influence risk in selected individuals.

ICH was defined by clinical evaluation and neuroimaging (MRI or CT scans), whereas SVIS was determined according to Trial of Org 10172 in Acute Ischemic Stroke Treatment (TOAST) guidelines 11. Cardiovascular risk factors including hypertension and diabetes were defined in the same manner as previously described for the entire INTERSTROKE study3. Blood samples were collected in EDTA Whole Blood DNA tubes. DNA was extracted using the QIAGEN QIAsymphony DSP DNA Midi kit. Exomic sequences were captured with the Illumina TrueSeq Exome Enrichment Kit and subsequently sequenced on HiSeq with paired-end reads (2 x 100 bp).

*Variant Calling & Quality Control*

Read pairs were mapped to the hg19 reference genome with the Burrows Wheeler Aligner12 and processed according to the Genome Analysis Tool Kit (GATK13) protocol (Unified Genotyper) to generate single nucleotide variant (SNV) and insertion/deletion (INDEL) calls. Default variant quality control (QC) filters were applied in KGGSEQ14, except for the read depth filter which was increased to require eight reads per call. Variants deviating from Hardy Weinberg Equilibrium (HWE) within at least one ethnicity were excluded in all samples. Regions that cannot be confidently called for next-generation sequencing platforms were excluded based upon Zook *et al.* (2014)15. Variant QC metrics were generated in PicardTools16, BedTools17, GATK13, and SnpSift18. Sample QC included checks for ethnicity, sex, and cryptic relatedness. Accuracy of genotyping calls was confirmed through secondary genotyping using the Illumina HumanExome Chip V 1-1 for a subset of 146 samples. Using exome chip genotypes as the benchmark, samples with more than 10% discrepant genotypes were excluded. Samples failing any of the aforementioned QC checks were removed along with their matching sample. After QC, 185 case-control pairs remained (S. Table 3.1). QC checks were performed using PLINK19, GCTA20, GATK13, and Variant Tools21.

Cases and controls did not differ in terms of sequencing depth, target coverage, or other alignment metrics (S. Table 3.1). On average, there were 9.1% duplicates, and 57.5% of unique reads were on target indicating efficient capture of exomic regions. Mean target coverage was 52x, and 94.4% of all targeted regions were covered by at least two reads. 73.6% of targeted regions were covered by more than 20 reads. Genotypic concordance was 99.8% with no significant differences between cases (99.8%) and controls (99.9%) (P=0.34). 17,917 coding mutations were detected per sample with similar numbers in cases (17,599) and controls (17,670) (P=0.65) (S. Table 3.1). Of all coding mutations, 17,635 (98.4%) were single nucleotide polymorphisms (SNPs) whereas 283 (1.6%) were insertions or deletions (INDELs). On average, 17,382 (98.57%) SNPs and 240 (85.23%) INDELs were recorded in dbSNP137. The number of mutations did not differ between cases and controls when stratifying by functional class (S. Table 3.2).

*Gene and Pathway-based Association Testing*

Rare disruptive protein-altering mutations within the same gene, as defined by RefGene22, were consolidated into a single unit for association testing. Mutations were also grouped by KEGG pathways23. Rare mutations were defined as those having a minor allele frequency (MAF) less than 5% in all 185 INTERSTROKE controls, NHLBI GO Exome Sequencing Project (ESP), and 1000 Genomes (1KG) participants. The MAF threshold was applied within each ethnic subdivision for ESP and 1KG. Variants that were common (MAF>0.05) to a single ethnicity in ESP or 1KG were excluded from the analyses. Variants designated as “disease-causing” or “probably-disease causing” according to either Polyphen-II (HDIV or HVAR)24 or SIFT25 algorithms were considered “disruptive”.

Under the Variant Tools framework21, the collapsing test was applied to genes and pathways. This method overcomes the bias incurred when rare variants cosegregate due to genomic proximity (i.e. linkage disequilibrium) and not genuine association with disease status. Three separate analyses were conducted for all stroke, ICH, and SVIS. Moreover, cases were coded as 1 to identify associations conferring risk, and then controls were coded as 1 to identify associations conferring protection. P-values were corrected for multiple hypothesis testing by experiment-wise permutation of phenotypes within matched pairs. The major advantage of this method is that it does not assume an identical error rate for each association test. Instead, the family-wise error rate is empirically derived without this assumption. Specifically, this entails extracting the most significant p-value among all genes from each of 1000 permutations to create a null distribution of 1000 p-values. Adjusted p-values were calculated as the proportion of permutated p-values that were more significant than the initial p-value. A sensitivity analysis was also conducted using an additional rare variant association test, SKAT-O26. Results were generally consistent among top genes using either method (S. Table 3.4).

*Biological Relevance*

A PubMed search for the top 50 genes and 10 pathways of each analysis was perfrmed to ascertain biological relevance. Genes or pathways related to stroke pathology (e.g. thrombosis, vessel wall integrity, etc.), major risk factors (e.g. diabetes, hypertension, dyslipidemia), or intermediate phenotypes (e.g. white-matter disease) were reported.

*Other Statistical Analyses*

Characteristics between cases and controls were compared using the two-sided Student’s t-test for quantitative variables and two-sided Fisher’s test for count data in R27. Using Variant Tools21, HWE was evaluated within the controls of individual ethnicities (P<0.05/N).

3.3 Results

*Study Subject Characteristics*

Overall, men comprised 57.6% of all study participants (Table 3.1). On average, cases (47.6 years) were younger than controls (74.1 years) (P<2.2x10-16). ICH and SVIS accounted for 54% and 46% of all stroke cases, respectively. Among all study participants, there were 120 (32.4%) Europeans, 86 (23.2%) Latin Americans, 70 (18.9%) South East Asians, 48 (13%) Africans, 30 (8.1%) Persians, 8 (2.2%) South Asians, and 8 (2.2%) others. As expected because of the selection criteria, the prevalence of hypertension and diabetes did not differ between cases and controls. 51.3% of cases and 57.3% of controls were hypertensive (P=0.30) and 18.9% of cases and 14.6% of controls were diabetic (P=0.33). However, other risk factors such as BMI (27.3 vs. 26.1; P=0.02), current smoking status (28.1% vs. 9.2%; P=3.94x10-6), and APOB/APOA1 ratio (0.88 vs. 0.80; P=0.01), were elevated in cases. Parental history of stroke was similar between cases (16.2%) and controls (13.5%) (P=0.56).

*Variant Summary*

In total, 174,464 protein-altering mutations were detected, of which 148,379 (85.05%) were rare and 86,687 (49.69%) were both rare and disruptive. At least one rare disruptive mutation was observed in 13,706 genes and 208 KEGG pathways. On average, each individual possessed 378 rare and disruptive protein-altering mutations.

*Gene-based Analysis*

The top five genes associated with all stroke status were *CDK5RAP2* (P=1.2x10-4), *EEF1D* (P=1.7x10-4), *GTSF1L* (P=2.2x10-4), *BCAN* (P=3.6x10-4)*,* and *POLG2* (P=5.4x10-4) (Table 3.2). Among the 50 most significant genes, *POLG2*, *VEGFC, DOK2, THBS4,* and *PEAR1* were biologically relevant. Stratifying by stroke subtype, the top genes for ICH were *CDK5RAP2* (P=1.4x10-4), *GGT5* (P=4.4x10-4)*, MYOM1* (P=5.6x10-4)*, CPAMD8* (P=1.0x10-3)*,* and *BCAN* (P=1.1x10-3) (Table 3.3), whereas the top genes for SVIS were *KCNH6* (P=1.0x10-3), *TRANK1* (1.0x10-3)*, DENND2C* (P=1.8x10-3)*, FOCAD* (P=1.9x10-3)*,* and *STXBP2* (P=1.9x10-3)(Table 3.4). Among the top genes in either subtype analysis, biologically relevant genes included *DOK2, PEAR1, AARS2, PODN, POLG2, NOTCH3, EPN1, ALDH1L1* and *MTOR* for ICH and *EIF2AK3, NBEAL2, NINJ2,* and *MPL* for SVIS*.* Biologically relevant genes are described in more detail in S. Table 3.3. No gene was significant after adjustment for multiple hypothesis testing.

*Pathway-based Analysis*

 Pathway results are summarized in Table 3.5. No pathways were significant after correction for multiple hypothesis testing. The most significant pathways for all stroke, ICH, and SVIS analyses are presented in detail in Supplementary Tables 3.5-3.7. Biologically relevant pathways included the renin-angiotensin system (RAS) which conferred protection against all stroke (P=8.1x10-4), ICH (P=1.3x10-2), and SVIS (P=1.2x10-2), and type II diabetes mellitus (T2DM) which increased risk of all stroke (P=1.9x10-2) and ICH (P=1.1x10-2), but not SVIS (P=0.27). The genes most strongly contributing to the pathway associations were *ACE*, *CMA1*, *CPA3*, and *MME* for the RAS pathway (S. Table 3.8); and *ANGPTL7*, *INSR*, *PIK3CG*, *PRKCE*, *PRKCZ* for the T2DM pathway (S. Table 3.9)*.* Mutations in the RAS pathway were not associated with hypertension status (OR=0.97; 95% CI, 0.62-1.52; P=0.91), nor were mutations in the T2DM pathway associated with diabetes status (OR=1.28; 95% CI, 0.71-2.32; P=0.41). Pathways did not remain significant after adjustment for age, sex, ethnicity, case status, and cardiovascular risk factors (APOB/APOA1, BMI, smoking status, hypertension, and diabetes).

3.4 Discussion

*Suggestive evidence for novel stroke genes*

 We conducted the first agnostic assessment of rare variants for small-vessel stroke. *NOTCH3* was the 17th most significant gene for ICH overall. *NOTCH3* is the causative gene for Cerebral Autosomal Dominant Arteriopathies with Subcortical Infarcts and Leucoencephalopathies (CADASIL), a Mendelian stroke disorder characterized by early ICH and SVIS28,29. Although some associations were nominally significant, results were encouraging considering that an established small-vessel stroke gene was among top genes out of more than 10,000 tested.

*CDK5RAP2* was the top association for all stroke (P=1.4x10-2) and ICH (P=1.2x10-4) but not SVIS (P>0.10). CDK5RAP2 is essential for chromosome segregation in neuronal progenitor cells and regulates the DNA damage response30,31. Conversely, *KCNH6* and *TRANK1* were tied as the top genes for SVIS status (P=1.0x10-3). *KCNH6* is a voltage-gated potassium ion channel whose function is unknown, while *TRANK1* is a risk locus for bipolar disorder32. Numerous top genes were also biologically relevant. In the present study, *AARS2* was the second most significant gene among those associated with increased risk of ICH. Rare *AARS2* mutations were recently found to cause early-onset leukodystrophy, an indicator for cerebral small-vessel dysfunction, though no strokes were reported33. Another biologically relevant gene, *NBEAL2,* was nominally associated with SVIS. Rare *NBEAL2* mutations cause gray platelet syndrome, a genetic bleeding disorder characterized by deficiency in platelet α-granules, which are the intracellular vesicles that store clotting factors34. Therefore, *AARS2* and *NBEAL2* may be intriguing candidates to cause novel Mendelian stroke disorders. Other pertinent top genes and their relation to stroke are described in S. Table 3. Additionally, the *PON1* gene, which was replicated bythe ESP as an ischemic stroke locus, was not associated with all stroke (P=0.28), ICH (P=0.49), or SVIS (P=0.16) in our study.

*Suggestive evidence for stroke pathways*

Diabetes and hypertension are important risk factors for small-vessel stroke. Although diabetic and hypertensive cases were minimized by design, results suggested that genes related to these risk factors may alter risk of stroke. RAS mutations were nominally protective against all stroke, ICH, and SVIS. Assuming that mutations resulted in loss-of-function, this protective effect is consistent with RAS inhibition. Moreover, this protective effect was independent of hypertension, which was unexpected given that RAS is the primary regulatory mechanism underlying blood pressure. However, Mollsten *et al.* (2008) also identified a blood-pressure-independent association with a common variant in *ACE*, the gene encoding angiotensin converting enzyme35. Similarly, Hajjar *et al.* (2010) discovered a blood-pressure-independent association between a common angiotensin gene polymorphism and vasoreactivity36. Consistent with Das *et al.* (2014) reporting an association between an *ACE* polymorphism and both ischemic and hemorrhagic subtypes37, RAS mutations influenced risk of both SVIS and ICH in the present study. Also, T2DM mutations conferred risk for all stroke and ICH but not SVIS, a surprising finding given that diabetes is a risk factor for ischemic but not hemorrhagic stroke3.

*Future Implications*

 Our results validate the utility of exome sequencing as a discovery tool for novel stroke genes and pathways. Although no genome-wide significant associations were identified, *NOTCH3* was among the top genes thus providing external validity to the study design. Many nominally significant genes were found to have protective effects just as rare variants have been reported to protect against other diseases, such as hypercholesterolemia38, coronary heart disease39,40, and type II diabetes41.Likewise, our findings prompt future studies to search for protective stroke genes to better understand the mechanisms underlying stroke resistance. Additionally, the notion that SVIS and ICH have distinct genetic etiologies is supported by the fact that none of the top genes within subtype analyses overlapped. Furthermore, most top genes associated with all stroke status were primarily driven by association with one stroke subtype. This is consistent with the fact that nearly all replicable stroke loci are subtype-specific42, thus emphasizing the importance of proper stroke subtyping.

*Limitations*

 The major limitation of our study is the small sample size. As such, results must be validated in additional, larger stroke cohorts. Secondly, results were dependent on the accuracy of prediction algorithms. Polyphen II and SIFT are widely used, but overall, specificity is very low (~15%) for both algorithms43. As a result, inclusion of non-disruptive mutations may mask real associations. Thirdly, coverage of protein-coding genes was incomplete. While 13,706 genes were tested, numerous genes were not assessed either due to low sequencing depth or because no rare and disruptive mutation was identified. Additionally, genes encoded in the mitochondrial genome could not be evaluated even though they are known causes of Mendelian stroke44–46. Lastly, it was assumed that genes were either associated with protective or disease-causing effects though rare mutations within the same gene could have opposing effects. However, results using SKAT-O, a rare variant test robust to effect direction, were generally concordant for top genes.

3.5 Conclusion

An agnostic, large-scale investigation of rare variants for early stroke was performed. *NOTCH3*, a classic stroke gene, was among the top genes out of 13,706 tested. *AARS2* and *NBEAL2,* which were nominally associated with risk of ICH and SVIS, respectively, are plausible candidates for novel Mendelian stroke disorders. Pathway analyses revealed nominally significant associations for RAS and T2DM, which are closely linked with major risk factors despite associations being independent of hypertension and diabetes, respectively. Future studies are required to replicate the suggestive findings reported in this study.

3.6 References

1. Warlow, C., Sudlow, C., Dennis, M., Wardlaw, J. & Sandercock, P. Stroke. *Lancet* **362,** 1211–24 (2003).

2. Whisnant, J. P. Modeling of risk factors for ischemic stroke. The Willis Lecture. *Stroke.* **28,** 1840–4 (1997).

3. O’Donnell, M. J. *et al.* Risk factors for ischaemic and intracerebral haemorrhagic stroke in 22 countries (the INTERSTROKE study): a case-control study. *Lancet* **376,** 112–23 (2010).

4. Bis, J. C. *et al.* Associations of NINJ2 Sequence Variants with Incident Ischemic Stroke in the Cohorts for Heart and Aging in Genomic Epidemiology (CHARGE) Consortium. *PLoS One* **9,** e99798 (2014).

5. Kim, D. S. *et al.* Rare coding variation in paraoxonase-1 is associated with ischemic stroke in the NHLBI Exome Sequencing Project. *J. Lipid Res.* (2014). doi:10.1194/jlr.P049247

6. Liu, M.-E. *et al.* A functional polymorphism of PON1 interferes with microRNA binding to increase the risk of ischemic stroke and carotid atherosclerosis. *Atherosclerosis* 1–7 (2013). doi:10.1016/j.atherosclerosis.2013.01.036

7. Anderson, C. D. *et al.* Common variants within oxidative phosphorylation genes influence risk of ischemic stroke and intracerebral hemorrhage. *Stroke.* **44,** 612–9 (2013).

8. Devan, W. J. *et al.* Heritability Estimates Identify a Substantial Genetic Contribution to Risk and Outcome of Intracerebral Hemorrhage. *Stroke.* (2013). doi:10.1161/STROKEAHA.111.000089

9. Hacke, W. & Grond-Ginsbach, C. Commentary on a GWAS: HDAC9 and the risk for ischaemic stroke. *BMC Med.* **10,** 70 (2012).

10. Brown, R. D., Whisnant, J. P., Sicks, J. D., O’Fallon, W. M. & Wiebers, D. O. Stroke incidence, prevalence, and survival: secular trends in Rochester, Minnesota, through 1989. *Stroke.* **27,** 373–80 (1996).

11. Adams, H. P. *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).

12. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–60 (2009).

13. DePristo, M. a *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43,** 491–8 (2011).

14. Li, M.-X. *et al.* Predicting mendelian disease-causing non-synonymous single nucleotide variants in exome sequencing studies. *PLoS Genet.* **9,** e1003143 (2013).

15. Zook, J. M. *et al.* Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nat. Biotechnol.* **32,** 246–51 (2014).

16. Picard. No Title. at < http://picard.sourceforge.net>

17. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–2 (2010).

18. Cingolani, P. *et al.* Using Drosophila melanogaster as a Model for Genotoxic Chemical Mutational Studies with a New Program, SnpSift. *Front. Genet.* **3,** 35 (2012).

19. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81,** 559–75 (2007).

20. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88,** 76–82 (2011).

21. San Lucas, F. A., Wang, G., Scheet, P. & Peng, B. Integrated annotation and analysis of genetic variants from next-generation sequencing studies with variant tools. *Bioinformatics* **28,** 421–2 (2012).

22. Pruitt, K. D. *et al.* RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res.* **42,** D756–63 (2014).

23. Kanehisa, M., Goto, S., Sato, Y., Furumichi, M. & Tanabe, M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.* **40,** D109–14 (2012).

24. Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet.* **Chapter 7,** Unit7.20 (2013).

25. Teer, J. K., Green, E. D., Mullikin, J. C. & Biesecker, L. G. VarSifter: visualizing and analyzing exome-scale sequence variation data on a desktop computer. *Bioinformatics* **28,** 599–600 (2012).

26. Lee, S. *et al.* Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am. J. Hum. Genet.* **91,** 224–37 (2012).

27. R Development Core Team. R: A language and environment for statistical computing. (2008). at <http://www.r-project.org>

28. Choi, J. C., Kang, S.-Y., Kang, J.-H. & Park, J.-K. Intracerebral hemorrhages in CADASIL. *Neurology* **67,** 2042–4 (2006).

29. Dichgans, M. *et al.* The phenotypic spectrum of CADASIL: clinical findings in 102 cases. *Ann. Neurol.* **44,** 731–9 (1998).

30. Barr, A. R., Kilmartin, J. V & Gergely, F. CDK5RAP2 functions in centrosome to spindle pole attachment and DNA damage response. *J. Cell Biol.* **189,** 23–39 (2010).

31. Buchman, J. J. *et al.* Cdk5rap2 interacts with pericentrin to maintain the neural progenitor pool in the developing neocortex. *Neuron* **66,** 386–402 (2010).

32. Mühleisen, T. W. *et al.* Genome-wide association study reveals two new risk loci for bipolar disorder. *Nat. Commun.* **5,** 3339 (2014).

33. Dallabona, C. *et al.* Novel (ovario) leukodystrophy related to AARS2 mutations. *Neurology* (2014). doi:10.1212/WNL.0000000000000497

34. Gunay-Aygun, M. *et al.* NBEAL2 is mutated in gray platelet syndrome and is required for biogenesis of platelet α-granules. *Nat. Genet.* **43,** 732–4 (2011).

35. Möllsten, A., Stegmayr, B. & Wiklund, P.-G. Genetic polymorphisms in the renin-angiotensin system confer increased risk of stroke independently of blood pressure: a nested case-control study. *J. Hypertens.* **26,** 1367–72 (2008).

36. Hajjar, I. *et al.* Renin angiotensin system gene polymorphisms and cerebral blood flow regulation: the MOBILIZE Boston study. *Stroke.* **41,** 635–40 (2010).

37. Das, S., Roy, S. & Sharma, V. Association of ACE gene I / D polymorphism and ACE levels with hemorrhagic stroke : comparison with ischemic stroke. (2014). doi:10.1007/s10072-014-1880-8

38. Scartezini, M. *et al.* The PCSK9 gene R46L variant is associated with lower plasma lipid levels and cardiovascular risk in healthy U.K. men. *Clin. Sci. (Lond).* **113,** 435–41 (2007).

39. Heart, N. Loss-of-Function Mutations in APOC3, Triglycerides, and Coronary Disease. *N. Engl. J. Med.* (2014). doi:10.1056/NEJMoa1307095

40. Jørgensen, A. B., Frikke-Schmidt, R., Nordestgaard, B. G. & Tybjærg-Hansen, A. Loss-of-Function Mutations in APOC3 and Risk of Ischemic Vascular Disease. *N. Engl. J. Med.* 32–41 (2014). doi:10.1056/NEJMoa1308027

41. Flannick, J. *et al.* Loss-of-function mutations in SLC30A8 protect against type 2 diabetes. *Nat. Genet.* **46,** 357–63 (2014).

42. Traylor, M. *et al.* Genetic risk factors for ischaemic stroke and its subtypes (the METASTROKE collaboration): a meta-analysis of genome-wide association studies. *Lancet Neurol.* **11,** 951–62 (2012).

43. Flanagan, S. E., Patch, A.-M. & Ellard, S. Using SIFT and PolyPhen to predict loss-of-function and gain-of-function mutations. *Genet. Test. Mol. Biomarkers* **14,** 533–7 (2010).

44. Sasarman, F., Antonicka, H. & Shoubridge, E. a. The A3243G tRNALeu(UUR) MELAS mutation causes amino acid misincorporation and a combined respiratory chain assembly defect partially suppressed by overexpression of EFTu and EFG2. *Hum. Mol. Genet.* **17,** 3697–707 (2008).

45. Conforto, A. B. *et al.* Screening for MELAS mutations in young patients with stroke of undetermined origin. *Arq. Neuropsiquiatr.* **65,** 371–6 (2007).

46. Goodfellow, J. A. *et al.* Mitochondrial myopathy, encephalopathy, lactic acidosis and stroke-like episodes: an important cause of stroke in young people. *Postgrad. Med. J.* **88,** 326–34 (2012).

3.7 Tables

Table 3.1.Characteristics of study subjects.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | SVIS | ICH | AllCase | AllControl | All StrokeP-value | SVIS P-value | ICH P-value |
| N | 85 | 100 | 185 | 185 | - | - | - |
| Men, n (%) | 49(57.6%) | 53(53%) | 102(55.1%) | 102(55.1%) | 1 | 1 | 1 |
| Age(SD) | 50.1(9.79) | 45.6(12.55) | 47.6(11.55) | 74.1(8.36) | <2.2x10-16 | <2.2x10-16 | <2.2x10-16 |
| Hypertension(%) | 46(54.12%) | 49(49%) | 95(51.35%) | 106(57.30%) | 0.30 | 1 | 0.12 |
| Diabetes(%) | 23(27.06%) | 12(12%) | 35(18.92%) | 27(14.59%) | 0.33 | 0.20 | 1 |
| Current Smoker(%) | 31(36.47%) | 21(21%) | 52(28.11%) | 17(9.19%) | 3.94x10-6 | 4.08 x10-5 | 0.03 |
| BMI(SD) | 26.9(4.88) | 27.7(5.51) | 27.3(5.23) | 26.1(4.83) | 0.02 | 0.14 | 0.08 |
| APOB/APOA1(SD) | 0.90(0.29) | 0.87(0.28) | 0.88(0.28) | 0.80(0.31) | 0.01 | 0.05 | 0.12 |
| Parental History(%) | 23(27.1%) | 7(7%) | 30(16.2%) | 25(13.5%) | 0.56 | 0.47 | 1 |

Table 3.2.Gene-based association results for all stroke.(Top 5 most significant genes and biological candidates (bolded) within the top 50 genes).

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Gene | ICH Count | SVISCount | All Stroke Case Count | All StrokeControl Count | Effect | P-value | Rank |
| *CDK5RAP2* | 21 | 10 | 31 | 9 | RISK | 1.2x10-4 | 1 |
| *EEF1D* | 0 | 1 | 1 | 15 | PROTECTIVE | 1.7x10-4 | 2 |
| *GTSF1L* | 1 | 1 | 2 | 9 | PROTECTIVE | 2.2x10-4 | 3 |
| *BCAN* | 0 | 1 | 1 | 14 | PROTECTIVE | 3.6x10-4 | 4 |
| *POLG2* | 8 | 2 | 10 | 0 | RISK | 5.4x10-4 | 5 |
| *DOK2* | **0** | **4** | **4** | **16** | **PROTECTIVE** | **2.9x10-3** | **22** |
| *THBS4* | **1** | **1** | **2** | **12** | **PROTECTIVE** | **3.2x10-3** | **31** |
| *VEGFC* | **0** | **0** | **0** | **5** | **PROTECTIVE** | **6.6x10-3** | **38** |
| *PEAR1* | **0** | **3** | **3** | **14** | **PROTECTIVE** | **4.8x10-3** | **47** |

Table 3.3Gene-based association results for ICH.(Top 5 most significant genes and biological candidates (bolded) within the top 50 genes).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Gene | ICH Count | Control Count | Effect | P-value | Rank |
| *CDK5RAP2* | 21 | 4 | RISK | 1.4x10-4 | 1 |
| *GGT5* | 1 | 7 | PROTECTIVE | 4.4x10-4 | 2 |
| *MYOM1* | 6 | 22 | PROTECTIVE | 5.6x10-4 | 3 |
| *CPAMD8* | 8 | 24 | PROTECTIVE | 1.0x10-3 | 4 |
| *BCAN* | 0 | 9 | PROTECTIVE | 1.1x10-3 | 5 |
| *DOK2* | **0** | **9** | **PROTECTIVE** | **1.1x10-3** | **5** |
| *PEAR1* | **0** | **8** | **PROTECTIVE** | **1.1x10-3** | **7** |
| *AARS2* | **17** | **4** | **RISK** | **1.4x10-3** | **8** |
| *PODN* | **8** | **0** | **RISK** | **2.0x10-3** | **9** |
| *NOTCH3* | **22** | **8** | **RISK** | **2.8x10-3** | **17** |
| *EPN1* | **1** | **8** | **PROTECTIVE** | **4.7x10-3** | **34** |
| *ALDH1L1* | **1** | **10** | **PROTECTIVE** | **4.9x10-3** | **36** |
| *MTOR* | **6** | **0** | **RISK** | **6.4x10-3** | **47** |

Table 3.4.Gene-based association results for SVIS.(Top 5 most significant genes and biological candidates (bolded) within the top 50 genes).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Gene | SVIS Count | Control Count | Effect | P-value | Rank |
| *KCNH6* | 9 | 0 | RISK | 1.0x10-3 | 1 |
| *TRANK1* | 9 | 0 | RISK | 1.0x10-3 | 1 |
| *DENND2C* | 8 | 0 | RISK | 1.8x10-3 | 3 |
| *FOCAD* | 9 | 0 | PROTECTIVE | 1.9x10-3 | 4 |
| *STXBP2* | 8 | 0 | PROTECTIVE | 1.9x10-3 | 5 |
| *EIF2AK3* | **0** | **8** | **PROTECTIVE** | **6.6x10-3** | **38** |
| *NBEAL2* | **12** | **3** | **RISK** | **7.5x10-3** | **39** |
| *NINJ2* | **8** | **0** | **RISK** | **8.3x10-3** | **42** |
| *MPL* | **1** | **8** | **PROTECTIVE** | **8.3x10-3** | **42** |
| *PPARGC1B* | **1** | **8** | **PROTECTIVE** | **8.3x10-3** | **42** |

Table 3.5.Pathway-based association results.Biologically relevant pathways are bolded.

|  |  |  |
| --- | --- | --- |
| All Stroke | ICH | SVIS |
| Pathway | **P-value** | **Pathway** | **P-value** | **Pathway** | **P-value** |
| Renin-angiotensin system | **8.1x10-4** | Galactose metabolism | 3.5x10-3 | Glycine, serine and threonine metabolism | 4.1x10-3 |
| Galactose Metabolism | 1.3x10-3 | Synthesis and degradation of ketone bodies | 4.6x10-3 | **Renin-angiotensin system** | **1.2x10-2** |
| Synthesis and degradation of ketone bodies | 2.6x10-3 | Nucleotide excision repair | 5.3x10-3 | Retinol Metabolism | 1.5x10-2 |
| Glycine, serine, and threonine metabolism | 9.8x10-3 | Non-small cell lung cancer | 8.1x10-3 | Dilated cardiomyopathy | 1.7x10-2 |
| Retinol Metabolism | 1.1x10-2 | Pantothenate and CoA biosynthesis | 8.4x10-3 | Phototransduction | 1.8x10-2 |
| SNARE interactions in vesicular transport | 1.3x10-2 | **Type II diabetes mellitus** | **1.1x10-2** | Mismatch Repair | 2.5x10-2 |
| Type II diabetes mellitus | **1.9x10-2** | **Renin-angiotensin system** | **1.3x10-2** | Huntington’s disease | 2.8x10-2 |
| Nucleotide Excision repair | 2.4x10-2 | Taurine and hypotaurine metabolism | 1.4x10-2 | Basal Cell Carcinoma | 3.1x10-2 |
| Vitamin B6 metabolism | 2.6x10-2 | Progesterone-mediated oocyte maturation | 1.5x10-2 | Porphyrin and chlorophyll metabolism | 3.6x10-2 |
| Cytosolic DNA-sensing pathway | 3.6x10-2 | Natural killer cell mediated cytotoxicity | 1.6x10-2 | Pentose phosphate pathway | 3.8x10-2 |

3.8 Supplementary

Supplementary Table 3.1.Quality control metrics for alignment and variant calling.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Cases | Controls | All | P-value |
| Total Reads | 75,774,471 | 77,439,871 | 76,607,171 | 0.32 |
| Unpaired Reads | 395,255 | 393,674 | 394,464 | 0.93 |
| Read Pairs | 37,689,608 | 38,523,098 | 38,106,353 | 0.32 |
| % Unmapped | 0.99% | 1.00% | 0.99% | 0.80 |
| % Duplicate Reads | 8.87% | 9.28% | 9.08% | 0.17 |
| Total Unique Reads | 68,712,133 | 70,032,113 | 69,372,123 | 0.36 |
| Unique Reads on Target | 39,608,006 | 40,235,092 | 39,921,549 | 0.47 |
| % Reads on Target | 57.54% | 57.36% | 57.45% | 0.44 |
| Mean Target Coverage | 51 | 52 | 52 | 0.47 |
| % Target > 2x | 94.33% | 94.37% | 94.35% | 0.54 |
| % Target > 20x | 73.32% | 73.96% | 73.64% | 0.48 |
| # Coding SNPs (# Known/# Novel) | 17,599(17,349/249) | 17,670(17,415/255) | 17,635(17,382/252) | 0.65 |
| % Known SNPs | 98.6% | 98.6% | 98.6% | 0.72 |
| Ti/Tv SNPs | 3.4 | 3.4 | 3.4 | 0.81 |
| Het/Hom SNPs | 1.5 | 1.5 | 1.5 | 0.77 |
| # Coding INDELs(# Known / # Novel) | 282(239/43) | 284(241/43) | 283(240/43) | 0.36 |
| % Known INDELs | 85.1% | 85.4% | 85.2% | 0.94 |
| INDEL Ratio | 1.3 | 1.3 | 1.3 | 0.74 |
| Exome Chip Concordance\* | 99.8% | 99.9% | 99.8% | 0.34 |

Supplementary Table 3.2.Variant counts by functional class.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Cases | Controls | All | P-value |
| Synonymous | 9,588 | 9,624 | 9,606 | 0.69 |
| Missense | 7,932 | 7,968 | 7,950 | 0.60 |
| Stopgain | 52 | 53 | 52 | 0.43 |
| Stoploss | 73 | 74 | 73 | 0.43 |
| Splicing | 5 | 5 | 5 | 0.77 |
| Frameshift | 133 | 134 | 134 | 0.53 |
| NonFrameshift | 149 | 150 | 149 | 0.34 |
| Total | **17,880** | **17,954** | **17,917** | **0.64** |

Supplementary Table 3.3. Biologically relevant genes among top 50 genes.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Gene | Associated Stroke Type | Effect | P-value | Related Function |
| *DOK2* | ICH, All Stroke | Protective | 1.1x10-3 | Thrombus formation |
| *PEAR1* | ICH, All Stroke | Protective | 1.1x10-3 | Platelet aggregation |
| *AARS2* | ICH | Risk | 1.4x10-3 | Early Leukodystrophy |
| *PODN* | ICH | Risk | 2.0x10-3 | Atherosclerosis |
| *NOTCH3* | ICH | Risk | 2.8x10-3 | Mendelian stroke |
| *THBS4* | All stroke | Protective | 3.2x10-3 | Atherosclerosis |
| *EPN1* | ICH | Protective | 4.7x10-3 | Angiogenesis |
| *ALDH1L1* | ICH | Protective | 4.9x10-3 | Ischemic Stroke GWAS loci |
| *MTOR* | ICH | Risk | 6.4x10-3 | Stroke protection |
| *EIF2AK3* | SVIS | Protective | 6.6x10-3 | Early-onset diabetes |
| *VEGFC* | ICH | Protective | 6.6x10-3 | Angiogenesis |
| *NBEAL2* | SVIS | Risk | 7.5x10-3 | Mendelian bleeding disorder |
| *NINJ2* | SVIS | Risk | 8.3x10-3 | Ischemic Stroke GWAS loci |
| *MPL* | SVIS | Protective | 8.3x10-3 | Thrombocythemia |
| *PPARGC1B* | SVIS | Protective | 8.3x10-3 | Diabetes |

Supplementary Table 3.4. Gene-based association results using SKAT-O.

|  |  |  |
| --- | --- | --- |
| All Stroke | ICH | SVIS |
| Gene | **P-value** | **Gene** | **P-value** | **Gene** | **P-value** |
| *GTSF1L* | 2.5x10-4 | *MYOM1* | 1.6x10-4 | *KCNH6* | 5.2x10-4 |
| *EEF1D* | 4.0x10-4 | *CDK5RAP2* | 3.1x10-4 | *TRANK1* | 1.4x10-3 |
| *CDK5RAP2* | 4.5x10-4 | *GGT5* | 4.7x10-4 | *STXBP2* | 2.1x10-3 |
| *CHD3* | 5.4x10-4 | *KRT7* | 4.7x10-4 | *IQCH* | 2.1x10-3 |
| *BICC1* | 6.4x10-4 | *DOK2* | 1.6x10-3 | *MRPL36* | 3.0x10-3 |
| *BCAN* | 7.9x10-4 | *PEAR1* | 1.7x10-3 | *NINJ2* | 3.4x10-3 |
| *POLG2* | 1.1x10-3 | *PODN* | 1.7x10-3 | *CNTRL* | 3.6x10-3 |
| *DIP2C* | 1.1x10-3 | *BCAN* | 1.8x10-3 | *PLA2G4F* | 4.5x10-3 |
| *FANCC* | 1.4x10-3 | *BICC1* | 2.4x10-3 | *ANK3* | 4.6x10-3 |
| *VWDE* | 2.2x10-3 | *POLG2* | 2.4x10-3 | *TULP4* | 4.8x10-3 |

Supplementary Table 3.5. Pathway-based association results for all stroke. Biologically relevant pathways are bolded.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Pathway | ICH Count | SVIS Count | All Case Count | Control Count | Effect | P-value |
| Renin-angiotensin system | **28** | **23** | **51** | **81** | **PRO** | **8.1x10-4** |
| Galactose metabolism | 63 | 41 | 104 | 75 | RISK | 1.3x10-3 |
| Synthesis and degradation of ketone bodies | 3 | 2 | 5 | 18 | PRO | 2.6x10-3 |
| Glycine, serine, and threonine metabolism | 49 | 44 | 93 | 64 | RISK | 9.8x10-3 |
| Retinol Metabolism | 47 | 44 | 91 | 71 | RISK | 1.1x10-2 |
| SNARE interactions in vesicular transport | 17 | 15 | 32 | 40 | PRO | 1.3x10-2 |
| Type II diabetes mellitus | **63** | **45** | **108** | **88** | **RISK** | **1.9x10-2** |
| Nucleotide Excision repair | 64 | 38 | 102 | 83 | RISK | 2.4x10-2 |
| Vitamin B6 metabolism | 6 | 1 | 7 | 16 | PRO | 2.6x10-2 |
| Cytosolic DNA-sensing pathway | 53 | 32 | 85 | 66 | RISK | 3.6x10-2 |

Supplementary Table 3.6. Pathway-based association results for ICH. Biologically relevant pathways are bolded.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Pathway | ICH Count | Control Count | Effect | P-value |
| Galactose metabolism | 63 | 44 | RISK | 3.5x10-3 |
| Synthesis and degradation of ketone bodies | 3 | 13 | PROTECTIVE | 4.6x10-3 |
| Nucleotide excision repair | 64 | 46 | RISK | 5.3x10-3 |
| Non-small cell lung cancer | 59 | 42 | RISK | 8.1x10-3 |
| Pantothenate and CoA biosynthesis | 28 | 13 | RISK | 8.4x10-3 |
| Type II diabetes mellitus | **63** | **47** | **RISK** | **1.1x10-2** |
| Renin-angiotensin system | **28** | **44** | **PROTECTIVE** | **1.3x10-2** |
| Taurine and hypotaurine metabolism | 12 | 24 | PROTECTIVE | 1.4x10-2 |
| Progesterone-mediated oocyte maturation | 70 | 55 | RISK | 1.5x10-2 |
| Natural killer cell cytotoxicity | 82 | 69 | RISK | 1.6x10-2 |

Supplementary Table 3.7.Pathway-based association results for SVIS. Biologically relevant pathways are bolded.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Pathway | SVIS Count | Control Count | Effect | P-value |
| Glycine, serine and threonine metabolism | 44 | 27 | RISK | 4.1x10-3 |
| Renin-angiotensin system | **23** | **37** | **PROTECTIVE** | **1.2x10-2** |
| Retinol Metabolism | 44 | 32 | RISK | 1.5x10-2 |
| Dilated cardiomyopathy | 68 | 77 | PROTECTIVE | 1.7x10-2 |
| Phototransduction | 23 | 36 | PROTECTIVE | 1.8x10-2 |
| Mismatch Repair | 34 | 23 | RISK | 2.5x10-2 |
| Huntington’s disease | 71 | 79 | PROTECTIVE | 2.8x10-2 |
| Basal Cell Carcinoma | 43 | 55 | PROTECTIVE | 3.1x10-2 |
| Porphyrin and chlorophyll metabolism | 22 | 33 | PROTECTIVE | 3.6x10-2 |
| Pentose phosphate pathway | 35 | 24 | RISK | 3.8x10-2 |

Supplementary Table 3.8.RAS pathway mutation carrier counts.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Gene | ICH Count | SVIS Count | All Case Count | Control Count |
| *ACE* | 11 | 8 | 19 | 30 |
| *ACE2* | 0 | 0 | 0 | 1 |
| *AGT* | 2 | 3 | 5 | 6 |
| *AGTR2* | 0 | 1 | 1 | 1 |
| *ANPEP* | 7 | 0 | 7 | 7 |
| *CMA1* | 3 | 4 | 7 | 15 |
| *CPA3* | 4 | 2 | 6 | 12 |
| *CTSA* | 0 | 1 | 1 | 2 |
| *ENPEP* | 0 | 0 | 0 | 0 |
| *LNPEP* | 3 | 2 | 5 | 7 |
| *MAS1* | 2 | 0 | 2 | 2 |
| *MME* | 1 | 1 | 2 | 8 |
| *NLN* | 0 | 2 | 2 | 4 |
| *REN* | 0 | 0 | 0 | 2 |
| *THOP1* | 1 | 2 | 3 | 6 |

Supplementary Table 3.9.T2DM pathway mutation carrier counts.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Gene | ICH Count | SVIS Count | All Case Count | Control Count |
| *ABCC8* | 2 | 2 | 4 | 5 |
| *ADIPOQ* | 3 | 0 | 3 | 2 |
| *ANGPTL7* | 4 | 2 | 6 | 1 |
| *CACNA1A* | 4 | 5 | 9 | 5 |
| *CACNA1C* | 5 | 4 | 9 | 10 |
| *CACNA1D* | 3 | 0 | 3 | 4 |
| *CACNA1G* | 9 | 6 | 15 | 14 |
| *HK1* | 0 | 1 | 1 | 2 |
| *HK2* | 5 | 3 | 8 | 4 |
| *HK3* | 23 | 11 | 34 | 29 |
| *INSR* | 4 | 2 | 6 | 1 |
| *IRS1* | 3 | 4 | 7 | 6 |
| *IRS2* | 0 | 1 | 1 | 1 |
| *IRS4* | 3 | 5 | 8 | 7 |
| *KCNJ11* | 6 | 1 | 7 | 8 |
| *MAFA* | 1 | 0 | 1 | 0 |
| *MAPK3* | 2 | 0 | 2 | 2 |
| *MAPK9* | 1 | 1 | 2 | 1 |
| *MTOR* | 2 | 1 | 3 | 1 |
| *PIK3CB* | 1 | 1 | 2 | 2 |
| *PIK3CD* | 4 | 4 | 8 | 4 |
| *PIK3CG* | 7 | 1 | 8 | 1 |
| *PIK3R1* | 1 | 0 | 1 | 0 |
| *PIK3R2* | 0 | 0 | 0 | 2 |
| *PIK3R5* | 4 | 1 | 5 | 5 |
| *PKLR* | 4 | 1 | 5 | 5 |
| *PKM* | 4 | 4 | 8 | 7 |
| *PRKCD* | 2 | 0 | 2 | 0 |
| *PRKCE* | 2 | 3 | 5 | 1 |
| *PRKCZ* | 2 | 3 | 5 | 0 |
| *SLC2A2* | 0 | 2 | 2 | 1 |
| *SLC2A4* | 1 | 0 | 1 | 2 |
| *SOCS4* | 0 | 1 | 1 | 1 |

**Chapter 4: Conclusion**

4.1 Summary of Main Findings

The candidate analysis (Chapter 2) confirmed the hypothesis that rare disruptive mutations among Mendelian stroke genes increase risk of stroke (ORStroke=1.62; 95% CI, 1.02-2.62; P=0.04) (ORICH=2.82; 95% CI, 1.46-5.59; P=0.001). However, when only considering previously reported disease-causing mutations, there was no significant difference in their frequency between cases (3.2%) and controls (2.7%) (P=1). Although 10 known genes were evaluated, our findings only support a role for *NOTCH3*, the CADASIL-causing gene. *NOTCH3* mutations were associated with increased risk of ICH (OR=3.23; 95% CI: 1.29 to 8.87; P=0.009). CADASIL mutations were enriched in cases (3.2%) as compared to controls (0%) (P=0.03). The six CADASIL mutation carriers all presented with stroke but secondary features of migraine and depression were rare. A new role for other *NOTCH3* mutations was also uncovered. Non-CADASIL *NOTCH3* mutations represented a common and significant risk factor for ICH, being present in 22% of ICH cases and 8% of controls (OR=2.86; 95% CI, 1.13 to 7.93, P=0.02). Many of these mutations would not be detected by standard diagnostic CADASIL testing. Therefore, our findings recommend *NOTCH3* screening of the entire gene in young stroke patients and their families.

To elucidate the role of rare variants beyond known stroke genes, we systematically assessed all protein-coding genes for novel associations (Chapter 3). While no genes or pathways were significant at the genome-wide level, *NOTCH3* was among the top ICH genes out of more than 10,000 tested (P=2.8x10-3). Many of the top associations were also biologically relevant. For instance, *NBEAL2* and *AARS2* were associated with elevated risk of SVIS and ICH, respectively, and are intriguing candidates to cause new Mendelian stroke disorders. Pathway analysis revealed nominal associations for the renin-angiotensin system (RAS) and type II diabetes mellitus (T2DM). Rare disruptive mutations in RAS were protective for all stroke (P=8.1x10-4), ICH (P=1.3x10-2), and SVIS (P=1.2x10-2). Conversely, rare disruptive mutations in T2DM conferred risk to all stroke (P=1.9x10-2) and ICH (P=1.1x10-2) but not SVIS (P=0.27). While these findings are hypothesis-generating, they require further validation in additional studies.

4.2 Research Implications

Our findings corroborate the CDRV hypothesis for early stroke. *NOTCH3* mutations were present in 22% of ICH cases and 8% of controls, whereas protective RAS mutations were present in 27.6% of all stroke cases and 43.8% of all controls. Together, these observations support both a monogenic and polygenic burden of rare mutations in early stroke. While a monogenic burden for rare variants was already supported by the existence of Mendelian strokes, there was no evidence for a polygenic model. A polygenic model for common variants, however, has been demonstrated as both blood pressure and oxidative phosphorylation genes are associated with increased risk of deep ICH1,2. Therefore, stroke risk appears to be the result of rare variants in single genes and rare variants dispersed across multiple genes sharing in the same functional pathway.

The CDRV and CDCV hypotheses are not mutually exclusive but rather complementary. Studies have demonstrated that both rare and common mutations within the same genes contribute to stroke risk. For instance, *NINJ2* was one of the first stroke genes identified through GWAS3 and recently, the Cohorts for Heart and Aging in Genomic Epidemiology (CHARGE) consortium replicated this association using deep sequencing, conveying that both common and rare *NINJ2* mutations impact risk of ischemic stroke4. In our study, rare *NINJ2* mutations were also nominally associated with risk of ischemic stroke (P=8.3x10-3).

Our findings have important implications for future stroke studies investigating the role of rare variants. Firstly, it is imperative to include controls. Previously reported disease-causing mutations were found in a similar proportion of cases (3.2%) and controls (2.7%). If a case-only study design had been adopted, one might conclude that known Mendelian stroke mutations were enriched in young stroke patients because the prevalence of Mendelian strokes in the general population is substantially. However, this conclusion relies on the false premise that previously reported mutations were causal and highly penetrant. Evidently, these allegedly pathogenic mutations appear to be marginally influential or lowly penetrant. Inclusion of controls also enables identification of protective stroke mutations. In our study, many biologically relevant genes reduced risk of stroke, including *DOK2, PEAR1*, *THBS4*, *EPN1*, *ALDH1L1*, *EIF2AK3*, *VEGFC*, *MPL*, and *PPARGC1B*. Other studies also demonstrate strong protective effects: rare *APOC3* variants reduce risk of coronary heart disease by 40%5, and rare *NINJ2* variants reduce risk of incident ischemic stroke by 19%4. Secondly, precise stroke subtyping is necessary to identify genuine associations due to subtype heterogeneity. The *NOTCH3* association was specific to ICH and not SVIS. Additionally, many of the top ICH genes did not overlap with the top SVIS genes in the discovery analysis. Finally, our study validates the use of highly diverse patient populations in genetics research. Inclusion of multiple ethnic groups enhances generalisability of results and improves power to detect associations because mutation frequency varies by ethnicity. Had we limited participants to those of European ancestry, the sample size would have needed to be seven times larger to be as well-powered to detect the higher frequency of CADASIL mutations in cases.

4.3 Future Directions

First and foremost, all findings presented in this thesis require replication in larger studies. The impact of rare variants in different patient populations should also be evaluated in future studies. We investigated the two most common stroke subtypes, SVIS and ICH, which collectively account for approximately 50% of all strokes6. However, the genetic architecture for other stroke subtypes might be vastly different. Cryptogenic strokes may be particularly important to explore as they account for 2% of all strokes but 22% of early-onset strokes7. Furthermore, there may be a genetic basis for cryptogenic stroke as Fabry’s disease is estimated to be responsible for ~0.6-11% of cryptogenic strokes8. Again, these figures are based on case-only studies which are not ideal. As such, conducting a case-control study of Mendelian stroke genes in young cryptogenic stroke would be optimal. Additionally, broader patient populations without stipulations for risk factor profile should be evaluated to improve generalisability of results. Lastly, studies that integrate various types of genetic information (e.g. gene expression and DNA methylation) would aid in defining the precise effects of rare variants.

4.4 Concluding Remarks

Rare genetic mutations have long been recognized as important determinants of disease but mainly in the context of families with severe genetic disorders. Contrary to this view, we demonstrate a broader role for rare mutations in young stroke patients. CADASIL and other *NOTCH3* mutations are important risk factors for stroke in young, unrelated patients. Beyond known genes, novel associations for genes and pathways were found at nominal significance in the first agnostic exploration of rare variants for stroke. Overall, rare mutations are significant risk factors for early stroke.

4.5 References

1. Devan, W. J. *et al.* Heritability Estimates Identify a Substantial Genetic Contribution to Risk and Outcome of Intracerebral Hemorrhage. *Stroke.* (2013). doi:10.1161/STROKEAHA.111.000089

2. Anderson, C. D. *et al.* Common variants within oxidative phosphorylation genes influence risk of ischemic stroke and intracerebral hemorrhage. *Stroke.* **44,** 612–9 (2013).

3. Ikram, M. A. *et al.* Genomewide association studies of stroke. *N. Engl. J. Med.* **360,** 1718–28 (2009).

4. Bis, J. C. *et al.* Associations of NINJ2 Sequence Variants with Incident Ischemic Stroke in the Cohorts for Heart and Aging in Genomic Epidemiology (CHARGE) Consortium. *PLoS One* **9,** e99798 (2014).

5. Heart, N. Loss-of-Function Mutations in APOC3, Triglycerides, and Coronary Disease. *N. Engl. J. Med.* (2014). doi:10.1056/NEJMoa1307095

6. Tsai, C.-F., Thomas, B. & Sudlow, C. L. M. Epidemiology of stroke and its subtypes in Chinese vs white populations: a systematic review. *Neurology* **81,** 264–72 (2013).

7. Wu, C.-H. *et al.* The Binding Affinity and Molecular Basis of the Structure-Binding Relationship between Urinary Tamm-Horsfall Glycoprotein and Tumor Necrosis Factor-α. *Molecules* **17,** 11978–89 (2012).

8. Shi, Q., Chen, J., Pongmoragot, J., Lanthier, S. & Saposnik, G. Prevalence of Fabry Disease in Stroke Patients-A Systematic Review and Meta-analysis. *J. Stroke Cerebrovasc. Dis.* **23,** 985–92 (2014).