

Uncovering the role of common and rare variants in migraine

Maria Stella Calafato, MD

Supervisor: Professor Aarno Palotie

Second supervisor: Professor John Todd

Sponsor: Professor David Lomas

Wellcome Trust Sanger Institute

University of Cambridge

Corpus Christi College

This dissertation is submitted for the degree of Doctor Philosophy

December 2011

To my dear parents
Matteo Calafato and Nina Tuccari

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration, except where specifically indicated in the text. This dissertation is not substantially the same as any that they may have submitted for a degree or diploma or other qualification at the University of Cambridge or any other university or similar institution, except for Chapter 3, the content of which has been submitted by Verner Anttila for his PhD at the University of Helsinki. This dissertation does not exceed the word limit (60000 words) set by the Biology Degree Committee.

Abstract

Migraine is a paroxysmal disorder of the nervous system. In order to uncover the genetic architecture underlying migraine, we performed a genome-wide association study (GWAS) of typed variants, a GWAS of imputed variants, and a pilot whole exome sequencing of familial migraine samples. In the GWAS of typed variants, a SNP (single nucleotide polymorphism) on chromosome 8q22.1 reached genome-wide significance in 2748 migraine patients and 10747 population-matched controls. The association was replicated in a further 3202 cases and 40062 population-matched controls. Expression quantitative trait (eQTL) analysis revealed the SNP to be a regulator of astrocyte elevated gene 1 (*AEG-1*). To identify further susceptibility loci for migraine, we carried out a GWAS of imputed SNPs using as reference 1000 Genomes project data (December 2010 release). Testing more than 11000000 SNPs in 5403 migraine patients and 15327 population-matched controls, six loci reached genome-wide significance. In the replication phase, consisting of 3268 cases and 2916 controls, three loci reached the Bonferroni corrected replication threshold. Of these, two loci had been previously identified (*TRPM8* and *LRP1*) and one was a newly identified locus (*C7orf10*). Whole exome sequencing is potentially an effective tool to identify coding variants underlying human diseases. We designed an extended set of baits (GENECODE exome) for capturing the entire human exome. The extended set allowed the coverage of additional 5594 genes and 10.3 Mb compared to the available CCDS-based sets. In order to identify rare variants contributing to migraine, whole-exome sequencing of 88 cases from 44 families with familial hemiplegic migraine (FHM) was performed. On average, we called 22169 variants per exome and we found 31 shared rare functional variants per family. In one family (family 1), we identified a missense variant in *CACNA1A* (rs121908212), which had been previously described as causing FHM. In another family (family 2), we detected a splice-site variant in *EAAT1*. Mutations in this gene had been previously found in a form of episodic ataxia associated with migraine and alternating hemiplegia (EA6). The functional impact of the identified splice-site *EAAT1* variant has still to be verified.

Acknowledgements

I wish to thank Prof. Aarno Palotie for supervising my work and for having given me the possibility to grow scientifically and personally. During the time spent at the Sanger, he was for me like a second father, noticing and supporting me in difficult moments. Moreover, he has been for me a good example of group leader, from whom I learnt how to work in multi-center collaborative projects.

I would like to thank Prof. John Todd, whose support has been essential to progress in my PhD. Prof John Todd has been closely keeping track of progress and he has been very supportive in moments in which things needed to be moved forward.

I need to thank immensely Prof. David Lomas, who gave me the opportunity to do this PhD and who has been always keeping track of my progress.

Dr. Jeffery Barrett, Dr Eli Zeggini, Dr Ines Barroso and Dr Carl Anderson for their suggestions. They have always been available when I needed suggestions and discuss my projects. A particular thank to Eija Hamalainen, whose support was essential. She has not only been great in organizing and performing the Sequenom follow-up for the GWAS projects, but also she has been giving me encouragement throughout this study. She is also the one who I have to thank for having pushed me to cycle, given that since I started cycling in Cambridge my life changed com-

pletely. A big thank to Dr Johannes Kettunen, Dr. Kati Kristiansson and Dr. Kate Morley for their advice; to Verner Anttila for the contribution to the quality checks and data analysis during the initial GWAS.

Most of this project was a multi-center collaboration. I would like to thank Emmanouli Dermizakis' group for performing the expression study, Carol Scott for the exome variants calling and for the interesting discussions, Alison Coffey and Felix Kokocinski for the great work done during the development of the GENECODE exome, Mikko Muona for performing the transcription experiments and all the collaborators for collecting the study samples. Finally, I would like to thank everyone in team 128, team 'awesome', and at the Sanger Institute for all the help they gave me.

Contents

1	Introduction	2
1.1	The Human Genome	2
1.1.1	What is a genome?	2
1.1.2	Human genetic variations	2
1.1.3	The Human Genome Project	5
1.1.4	The ENCODE Project	5
1.1.5	The HapMap Project	6
1.1.6	The 1000 Genomes Project	8
1.2	Investigating the role of genetic in complex diseases	9
1.2.1	Recurrence risk ratio and heritability	9
1.2.2	Identification of causative genes	10
1.3	Migraine	18
1.3.1	Clinical features	18
1.3.2	Classification and diagnosis	20
1.3.3	Epidemiology	20
1.3.4	Comorbid disorders	23
1.3.5	Pathogenesis	25

1.3.6	Genetic basis	26
1.4	This thesis	31
2	Material and methods	37
2.1	Genome-wide association study of migraine implicates a common susceptibility variant on 8q22.1	37
2.1.1	Study sample	37
2.1.2	Genotyping	42
2.1.3	Quality control	42
2.1.4	Statistical analysis	43
2.1.5	Imputation	44
2.1.6	eQTL analysis	44
2.1.7	URLs	45
2.2	Imputation of sequence variants to identify susceptibility loci for migraine	46
2.2.1	Study samples	46
2.2.2	Genotyping	46
2.2.3	Quality control	48
2.2.4	Imputation	49
2.2.5	Post imputation quality control	49
2.2.6	Statistical analysis	50
2.2.7	URLs	51
2.3	The GENCODE exome: sequencing the complete human exome . .	51
2.3.1	Bait design	51
2.3.2	Samples	52

2.3.3	Sequence capture and sequencing	53
2.3.4	Sequence alignment and variant calling	54
2.4	Exome sequencing in Familial Hemiplegic migraine	55
2.4.1	Familial Hemiplegic Migraine samples	55
2.4.2	Control exomes	55
2.4.3	Exome library construction	56
2.4.4	Library capture and sequencing	56
2.4.5	Exome data analysis	57
2.4.6	Family 2 variants validation	57
2.4.7	Family 2 complementary DNA (cDNA) analysis	58
2.4.8	Family 3 linkage analysis	58
3	Genome-wide association study of migraine implicates a common susceptibility variant on 8q22.1	59
3.1	Introduction	59
3.2	Results	61
3.2.1	Discovery stage	61
3.3	Conditional and haplotype analysis	65
3.4	Replication stage	68
3.5	eQTL analysis	69
3.6	Discussion	73
4	Imputation of SNPs to identify susceptibility loci for migraine	77
4.1	Introduction	77
4.2	Results	79
4.2.1	Initial imputation run	79

4.2.2	Discovery stage	81
4.3	Replication stage	84
4.3.1	Discussion	86
5	GENCODE exome	91
5.1	Introduction	91
5.2	Results	93
5.2.1	The GENCODE exome features	93
5.2.2	The GENCODE exome performance	97
5.3	Discussion	103
5.4	Notes	106
6	Exome sequencing in Familial Hemiplegic Migraine	107
6.1	Introduction	107
6.2	Results	109
6.2.1	Whole exome capture of 88 FHM cases	109
6.2.2	Potentially pathogenic variants underlying FHM	111
6.2.3	Family 1: known causal <i>CACNA1A</i> mutation	113
6.2.4	Family 2: <i>EAAT1</i> mutation	117
6.2.5	Family 3: Integration of linkage analysis and whole exome sequencing	120
6.3	Discussion	121
7	Concluding remarks	126

List of Figures

1.1	Types of human genetic variants	3
1.2	Organization of the ENCODE project	7
1.3	Comparison of linkage with association analysis for detecting genetic effects	12
1.4	Published GWAS	14
1.5	Published GWAS	15
1.6	Genetic variants frequency and diseases susceptibility	16
1.7	Strategies for identifying rare variants	17
1.8	Migraine attack phases.	19
1.9	Prevalence of migraine in adults of different countries.	23
1.10	Prevalence of different headaches in different age categories.	24
1.11	Anatomical structures involved in migraine pathophysiology.	26
1.12	Pathophysiological mechanisms in the generation of migraine headache.	27
1.13	Roles of proteins encoded by genes involved in FHM at a CNS glutamatergic synapse	30
3.1	Genome-wide P-values for the discovery phase	62
3.2	Quantile-quantile plots of the GWAS discovery phase.	63

3.3	Association signals and recombination rates for the chromosome 8q22.1 locus	65
3.4	Linkage disequilibrium between pair of SNPs at the chromosome 8q22.1 locus	66
3.5	Box-plot of the expression values for AEG1/MTDH based on the rs1835740 genotype	72
4.1	Genotype imputation	78
4.2	Genome-wide P-values for the initial imputation run	80
4.3	Quantile-quantile plots of the initial imputation run.	80
4.4	Genome-wide P-values for the discovery phase	82
4.5	Quantile-quantile plots of the GWAS discovery phase.	84
4.6	Locus-specific association plot: chromosome 2q37.1	87
4.7	Locus-specific association plot: chromosome 12q13.3	88
4.8	Locus-specific association plot: chromosome 7p14.1	90
5.1	Comparison of exon and transcript coverage of the available CCDS- based exome sets and the GENCODE exome set with three current reference gene sets	94
5.2	Cumulative distribution of base coverage for HapMap samples . . .	100
5.3	Cumulative distribution of base coverage for the clinical samples . .	101
5.4	Coverage achieved by the GENECODE based set	105
6.1	Genes with shared rare functional variants in three or more families	113
6.2	Genes with possibly damaging missense, splice site or nonsense vari- ants in two or more families	114

6.3	Pedigree of family 1	115
6.4	<i>CACNA1A</i> missense mutations causing hemiplegic migraine (HM) .	116
6.5	Pedigree of family 2	117
6.6	The splicing code	119
6.7	Pedigree of family 3	120

List of Tables

1.1	Estimates of heritability	10
1.2	Migraine classification	21
1.3	Migraine diagnostic criteria	22
1.4	FHM diagnostic criteria	28
1.5	Ion transportation genes and familial hemiplegic migraine	34
1.6	Migraine linkage studies	35
1.7	Migraine candidate-gene association studies	36
2.1	Study samples and genotyping platforms	41
2.2	Quality control	43
2.3	Study samples and genotyping platforms	47
2.4	Quality control: Illumina arrays	48
2.5	Quality control: Sequenom genotyping	49
3.1	Summary results of the GWAS discovery phase for rs1835740	64
3.2	Conditional analysis	67
3.3	Haplotype analysis	68
3.4	Summary results of the GWAS replication phase for rs1835740	70
3.5	SNPs correlated with AEG-1expression levels	71

3.6	Correlation between rs1835740 and gene expression levels	72
3.7	Summary results of the GWAS discovery phase	76
4.1	Summary results of the discovery stage	83
4.2	Summary results of the replication stage	85
4.3	Summary results of the discovery and replication stages	86
5.1	Comparison of the three different exome capture sets	93
5.2	Exon and transcript coverage of the three different exome capture sets	95
5.3	Comparison of the three different exome capture sets with the GENECODE design target	96
5.4	Assessment of repeat and low-complexity coverage of the three ex- ome sets.	96
5.5	Bait/probe covered CTR (Capture Target Regions) regions assessed using a uniqueness mask	98
5.6	Mapping statistics for clinical and HapMap samples using GEN- CODE and Agilent CCDS exome captures.	99
5.7	Variant calling statistics for clinical and HapMap samples using GENCODE and Agilent CCDS exome captures.	102
6.1	Mapping statistics.	110
6.2	Variant calling statistics	112
6.3	Rare functional variants shared by the two cases of each family . . .	112
6.4	Regions with a LOD score greater than zero.	122
6.5	Rare functional variants shared by the two cases of family 2	123

List of abbreviations

ASP	affected sib pair
ATP1A2	ATPase, Na ⁺ /K ⁺ transporting, alpha 2 polypeptide
CACNA1A	Calcium channel, voltage-dependent, P/Q type, alpha 1A subunit
CDH	Chronic daily headache
ChIP	Chromatin immunoprecipitation
ChIP-Seq	chromatin immunoprecipitation followed by sequencing
CMH	Cochran-Mantel-Haenszel
CNP	Copy number polymorphism
CNS	Central nervous system
CNV	Copy number variant
CSD	Cortical spreading depression
CVD	Cardiovascular disease
DBH	deCODE Migraine Questionnaire

DMQ	Dopamine beta-hydroxylase
DOE	Department of Energy
DRD2	Dopamine receptor D2
EA	Episodic ataxia
EAAT1	Excitatory amino acid transporter 1 gene
ECRs	Expressed cluster regions
ELSI	Ethical, legal and social issue
ENCODE	Encyclopedia of DNA Elements
ESR1	Estrogen receptor 1
FHM	Familial hemiplegic migraine
GRR	Genotype relative risk
GWAS	Genome-wide association study
HBCS	Helsinki birth cohort study
HM	Hemiplegic migraine
ICHD	International Classification of Headache Disorders
IHS	International Headache Society
INSR	Insulin receptor
LCLs	Lymphoblastoid cell lines

MA	Migraine with aura
MA/MO	Migraine with and without aura
MAF	Minor allele frequency
MDS	Multidimensional scaling
MO	Migraine without aura
MTHFR	Methylentetrahydrofolate reductase
NGS	Next-generation sequencing
NHGRI	National Human Genome Research Institute
PFO	Patent foramen ovale
PBMCs	Peripheral blood mononuclear cells
PGR	Progesterone receptor
SCA	spinocerebellar ataxia
SCN1A	Sodium channel, voltage-gated, type I, alpha subunit
SHM	Sporadic hemiplegic migraine
SLC6A3	Solute carrier family 6 member 3
SNP	Single nucleotide polymorphism
SV	Structural variant
TF	Transcription factor

TG	Trigeminal ganglia
TNC	Trigeminal nucleus caudalis
TSC	The SNP Consortium
TTH	Tension-type headache
VDCC	Voltage-dependent calcium channel