Chapter 4

Imputation of SNPs to identify susceptibility loci for migraine

4.1 Introduction

To identify common susceptibility variants for migraine, we carried out a GWAS, described in the previous chapter. This study provided evidence of association for a SNP on chromosome 8q22.1 (rs1835740). Expression quantitative trait (eQTL) analysis revealed this SNP to be a key regulator of astrocyte elevated gene 1 *AEG–1* in lymphoblastoid cell lines [243]. A subsequently published population-based GWAS has identified other three risk loci for migraine (chromosome 1p36.23, chromosome 2q37.1, chromosome 12q13.3) [244].

The hundreds of thousands of SNPs directly assayed represent only a fraction of the millions of SNPs contained in the human genome. Genotype imputation is useful to join together datasets genotyped on different platforms and to evaluate association with a phenotype at variants that are not directly genotyped. The

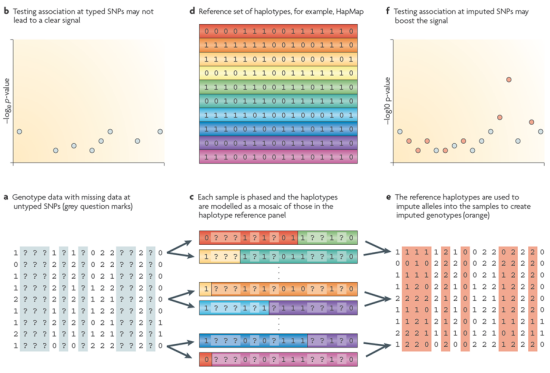

Figure 4.1: **Genotype imputation.** [245]

term imputation means predict genotypes of SNPs, which have not been directly assayed, in a sample using a reference panel of haplotypes including a much larger number of SNPs (Figure 4.1).

Genotype imputation tools involve phasing the typed SNPs in each individual of the study. Then these haplotypes are compared to the haplotypes of the reference panel and missing genotypes are predicted after matching haplotypes with the reference ones. A probability distribution over the possible genotypes is produced for each one of the imputed genotypes [245]. It has been shown that imputation error rate decreases as the minor allele frequency and the size of the reference panel increase [211, 245].

In order to identify novel risk loci for migraine, I have imputed untyped SNPs in migraine cases and population-matched controls from Finland, Germany and the Netherlands, using the 566 haplotypes of 1000 Genomes project (December 2010 release) as reference. The results obtained from the imputed data were replicated in independent migraine case and population-matched controls from Finland, the Netherlands and Spain.

4.2 Results

4.2.1 Initial imputation run

In an initial imputation run, 3279 European individuals affected by migraine with aura (MA) only or by migraine with and without aura (MA/MO) (1124 Finnish, 1276 Germans, and 879 Dutch) and 12369 population-matched controls (Helsinki Birth Cohort study, Health2000 study, KORA study, HNR study, PopGen study, Illumina iControlDB and Rotterdam study I) were included (see Methods).

Study samples had been screened for SNP call rate, presence of population outliers, duplicates and relatedness (see Methods). Overall 2948 cases and 10747 controls passed the quality control filters and remained in the study.

After excluding SNPs which did not pass the quality control filters (see Methods), around 7000000 untyped SNPs were imputed separately in cases and controls of each cohort using the software IMPUTE2 and 1000 Genomes plus HapMap III data as reference [211].

Genotyped and imputed SNPs were tested for association with migraine using a score test, as implemented in SNPTEST v2 [213]. The results of the association tests across the three cohorts (Finnish, German and Dutch) were combined using a fixed effect meta-analysis, as implemented in GWAMA version 2.0.4 [214]. This led to the identification of 62 loci that surpassed the threshold for genome-wide significance ($P = 5 \times 10^{-8}$) (Figure 4.2) [49]. However, quantile-quantile plot of the distribution of the test statistic suggested an overall inflation of P-values (λ= 1.38)(Figure 4.3).

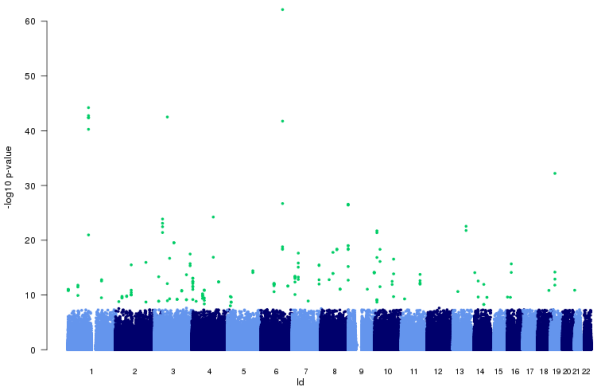

Figure 4.2: **Genome-wide P-values for the initial imputation run.** P-values are log transformed (-log10) (y axis) and plotted against chromosomes (x axis). The signal in green are the ones above the threshold for genome-wide significance.

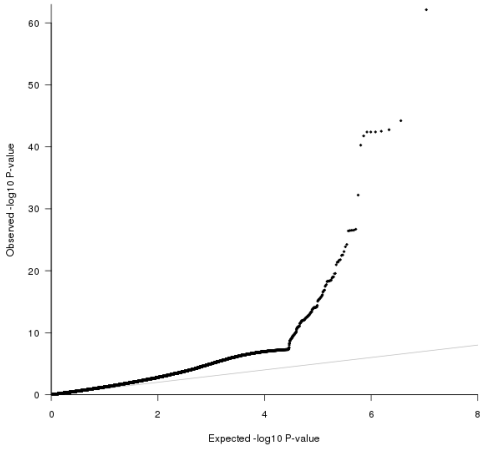

Figure 4.3: **Quantile-quantile plots of the initial imputation run** Plots of the fixed effect meta-analysis results of the initial imputation run.

4.2.2 Discovery stage

Since the number of genome-wide significant loci seemed excessively high, I thought that bias could have been introduced by imputing, in each population, cases and control separately. Therefore, it was decided to repeat the imputation in merged sets of cases and controls for each population. In the meantime a new release of 566 European haplotypes was released by the 1000 Genomes project and, hence, it was decided to use this new set as reference for the new imputation run, since the higher number of reference haplotypes would have improved the imputation accuracy. Moreover, two other migraine data sets became available, including 2490 migraine without aura cases (MO) (1208 German and 1282 Dutch) and 4580 population-matched controls. Therefore, since the two main types of migraine (MA and MO) seem to share a common genetic component, we decided to include them in the discovery stage of our study, to increase the power of detection of migraine risk loci [69].

The discovery stage included 5403 European individuals affected by migraine, of which 2748 were part of our previous GWAS. Diagnoses were made by headache specialists using a combination of questionnaires and individual interviews according to the ICHD-II guidelines [58]. Population-matched controls (15327) were drawn from previously genotyped population-based cohorts previously genotyped (see Methods). Study samples had been screened for SNP call rate, presence of population outliers, duplicates and relatedness (see Methods).

After excluding SNPs which did not pass the quality control filters (see Methods), around 11000000 untyped SNPs were imputed in each cohort using the software IMPUTE2 and 566 European haplotypes from the 1000 Genomes project

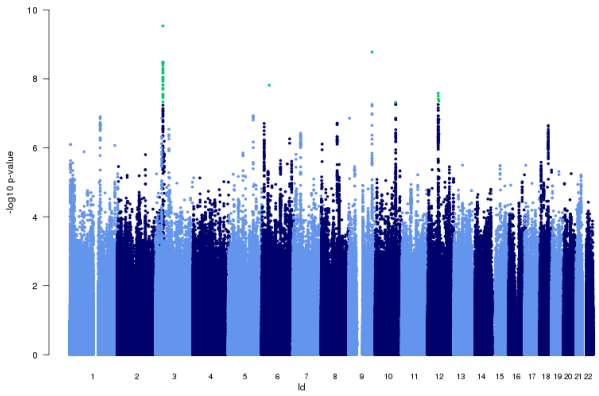

Figure 4.4: **Genome-wide P-values for the discovery phase.** P-values are log transformed (-log10) (y axis) and plotted against chromosomes (x axis). The signal in green are the ones above the threshold for genome-wide significance.

(December 2010 release) as reference [211].

Genotyped and imputed SNPs were tested for association with migraine using a score test, as implemented in SNPTEST v2, to take into account the uncertainty of the imputed genotypes [213]. The results of the association tests across the three cohorts (Finnish, German and Dutch) were combined using a fixed effect meta-analysis, as implemented in GWAMA version 2.0.4 [214].

Six loci that surpassed the threshold for genome-wide significance ($P = 5 \times 10^{-8}$) were identified (Figure 4.4 and Table 4.1) [49]. The genome-wide significant SNPs had the same direction of allelic effect in all the study cohorts. Two were previously identified loci (chromosome 2q37.1 and chromosome 12q13.3) and four were newly identified loci. Quantile-quantile plot of the distribution of the test statistic suggested a modest overall inflation of P-values (λ= 1.09) (Figure 4.5).

Table 4.1: **Summary results of the discovery stage**

				Finnish MA and MA/MO (1064/3513)[a]			Germans MA and MA/MO (1029/2317)[a]			Dutch MA and MA/MO (820/4917)[a]			German MO (1208/2564)[a]			Dutch MO (1282/2016)[a]			Meta-analysis (5403/15327)[a]	
Chr	Position	SNP	Alleles minor/major	MAF cases/controls	OR (95% CI)	P	MAF cases/controls	OR (95% CI)	P	MAF cases/controls	OR (95% CI)	P	MAF cases/controls	OR (95% CI)	P	MAF cases/controls	OR (95% CI)	P	OR (95% CI)	P
1	3319777	rs4471209	A / G	0.36 / 0.34	1.05 (0.95 – 1.16)	3.34×10^{-1}	0.31 / 0.27	1.22 (1.09 – 1.37)	4.10×10^{-4}	0.31 / 0.28	1.13 (1.01 – 1.27)	2.45×10^{-2}	0.30 / 0.27	1.17 (1.05 – 1.30)	2.66×10^{-3}	0.29 / 0.28	1.07 (0.96 – 1.19)	2.16×10^{-1}	1.13 (1.08 – 1.19)	2.93×10^{-6}
1	4185770	rs6679136	T / C	0.33 / 0.34	0.95 (0.86 – 1.06)	3.62×10^{-1}	0.36 / 0.38	0.94 (0.84 – 1.04)	2.05×10^{-1}	0.34 / 0.39	0.79 (0.71 – 0.88)	1.49×10^{-5}	0.34 / 0.38	0.86 (0.78 – 0.95)	1.61×10^{-3}	0.37 / 0.39	0.91 (0.82 – 1.00)	5.27×10^{-2}	0.88 (0.84 – 0.93)	7.98×10^{-7}
1	15532130	rs10927720	T / C	0.20 / 0.19	1.08 (0.95 – 1.22)	2.23×10^{-1}	0.25 / 0.20	1.27 (1.13 – 1.44)	1.27×10^{-4}	0.23 / 0.21	1.15 (1.01 – 1.30)	3.01×10^{-2}	0.24 / 0.21	1.19 (1.06 – 1.34)	2.76×10^{-3}	0.21 / 0.21	1.04 (0.92 – 1.18)	5.01×10^{-1}	1.15 (1.08 – 1.21)	2.75×10^{-6}
1	74118855	rs61767191	C / T	0.06 / 0.04	1.30 (1.05 – 1.62)	8.61×10^{-3}	0.06 / 0.05	1.29 (1.03 – 1.61)	9.11×10^{-3}	0.05 / 0.04	1.16 (0.90 – 1.49)	1.63×10^{-1}	0.07 / 0.05	1.27 (1.04 – 1.55)	4.68×10^{-3}	0.04 / 0.04	1.17 (0.92 – 1.50)	1.43×10^{-1}	1.36 (1.20 – 1.55)	1.31×10^{-6}
1	115677183	rs2078371	C / T	0.15 / 0.13	1.21 (1.06 – 1.39)	5.78×10^{-3}	0.11 / 0.11	0.98 (0.83 – 1.15)	7.66×10^{-1}	0.12 / 0.11	1.12 (0.95 – 1.32)	1.64×10^{-1}	0.12 / 0.11	1.06 (0.91 – 1.23)	4.33×10^{-1}	0.13 / 0.10	1.36 (1.17 – 1.59)	8.60×10^{-5}	1.15 (1.07 – 1.23)	1.54×10^{-4}
1	156456301	rs3790455	C / T	0.43 / 0.42	1.04 (1.15 – 0.94)	4.32×10^{-1}	0.35 / 0.33	1.10 (1.22 – 0.98)	1.01×10^{-1}	0.36 / 0.33	1.11 (1.24 – 1.00)	5.56×10^{-2}	0.38 / 0.33	1.24 (1.37 – 1.12)	3.63×10^{-5}	0.37 / 0.32	1.22 (1.35 – 1.10)	1.72×10^{-4}	1.14 (1.20 – 1.09)	1.26×10^{-7}
1	231361110	rs11590715	T / C	0.07 / 0.06	1.17 (0.97 – 1.42)	7.21×10^{-2}	0.06 / 0.05	1.30 (1.04 – 1.62)	8.55×10^{-3}	0.07 / 0.06	1.21 (0.98 – 1.49)	4.53×10^{-2}	0.07 / 0.05	1.40 (1.15 – 1.71)	1.24×10^{-4}	0.06 / 0.06	1.12 (0.91 – 1.38)	2.26×10^{-1}	1.32 (1.18 – 1.48)	8.58×10^{-7}
2	145222038	rs13403907	G / A	0.18 / 0.15	1.16 (1.02 – 1.32)	1.65×10^{-2}	0.13 / 0.10	1.28 (1.09 – 1.51)	1.04×10^{-3}	0.13 / 0.11	1.17 (1.00 – 1.37)	3.35×10^{-2}	0.12 / 0.11	1.15 (0.99 – 1.34)	4.43×10^{-2}	0.13 / 0.11	1.11 (0.95 – 1.29)	1.59×10^{-1}	1.20 (1.12 – 1.30)	1.58×10^{-6}
2	234821445	rs11892538	C / G	0.11 / 0.14	0.81 (0.70 – 0.94)	3.30×10^{-3}	0.14 / 0.17	0.81 (0.70 – 0.93)	2.86×10^{-3}	0.15 / 0.17	0.85 (0.73 – 0.98)	2.07×10^{-2}	0.14 / 0.17	0.82 (0.71 – 0.94)	2.43×10^{-3}	0.14 / 0.17	0.80 (0.70 – 0.92)	9.92×10^{-4}	0.81 (0.76 – 0.86)	2.92×10^{-10}
2	241447428	rs4676436	A / C	0.14 / 0.13	1.10 (0.96 – 1.27)	1.71×10^{-1}	0.13 / 0.11	1.24 (1.06 – 1.45)	7.48×10^{-3}	0.12 / 0.11	1.18 (1.00 – 1.38)	4.59×10^{-2}	0.13 / 0.11	1.20 (1.04 – 1.40)	1.42×10^{-2}	0.14 / 0.11	1.27 (1.09 – 1.48)	1.62×10^{-3}	1.20 (1.12 – 1.29)	6.46×10^{-7}
3	30480085	rs6790925	T / C	0.41 / 0.39	1.09 (0.99 – 1.21)	7.48×10^{-2}	0.37 / 0.35	1.11 (0.99 – 1.23)	6.20×10^{-2}	0.40 / 0.38	1.11 (1.00 – 1.24)	5.13×10^{-2}	0.38 / 0.35	1.16 (1.05 – 1.29)	2.88×10^{-3}	0.41 / 0.36	1.20 (1.09 – 1.33)	3.36×10^{-4}	1.14 (1.08 – 1.19)	1.39×10^{-7}
3	67144706	rs4311165	C / G	0.17 / 0.20	0.81 (0.71 – 0.92)	1.35×10^{-3}	0.24 / 0.27	0.87 (0.77 – 0.98)	2.04×10^{-2}	0.25 / 0.26	0.93 (0.82 – 1.05)	2.30×10^{-1}	0.24 / 0.27	0.87 (0.77 – 0.97)	1.01×10^{-2}	0.25 / 0.27	0.87 (0.78 – 0.97)	1.28×10^{-2}	0.87 (0.82 – 0.92)	2.89×10^{-7}
5	75495849	rs6453209	T / A	0.46 / 0.50	0.85 (0.93 – 0.77)	8.84×10^{-4}	0.45 / 0.47	0.91 (1.01 – 0.82)	6.49×10^{-2}	0.47 / 0.48	0.96 (1.06 – 0.86)	3.94×10^{-1}	0.48 / 0.49	0.93 (1.03 – 0.85)	1.30×10^{-1}	0.43 / 0.47	0.85 (0.93 – 0.76)	8.61×10^{-4}	0.89 (0.93 – 0.85)	1.41×10^{-6}
5	127722107	rs77050147	C / G	0.10 / 0.09	1.24 (1.05 – 1.45)	1.11×10^{-2}	0.10 / 0.08	1.27 (1.06 – 1.52)	8.43×10^{-3}	0.12 / 0.10	1.24 (1.05 – 1.46)	9.28×10^{-3}	0.09 / 0.08	1.16 (0.98 – 1.39)	8.78×10^{-2}	0.12 / 0.09	1.25 (1.07 – 1.47)	5.37×10^{-3}	1.25 (1.15 – 1.35)	1.18×10^{-7}
6	12908747	rs13197912	T / A	0.30 / 0.29	1.06 (0.95 – 1.17)	3.09×10^{-1}	0.37 / 0.36	1.08 (0.97 – 1.21)	1.44×10^{-1}	0.37 / 0.35	1.09 (0.97 – 1.21)	1.32×10^{-1}	0.40 / 0.35	1.23 (1.12 – 1.36)	3.73×10^{-5}	0.38 / 0.33	1.21 (1.09 – 1.34)	2.42×10^{-4}	1.14 (1.09 – 1.20)	1.98×10^{-7}
6	39177971	rs873690	C / G	0.09 / 0.07	1.26 (1.05 – 1.50)	3.25×10^{-3}	0.09 / 0.08	1.19 (0.99 – 1.44)	1.50×10^{-2}	0.08 / 0.07	1.10 (0.90 – 1.34)	2.18×10^{-1}	0.09 / 0.08	1.20 (1.01 – 1.42)	1.01×10^{-2}	0.09 / 0.07	1.30 (1.09 – 1.56)	1.81×10^{-4}	1.37 (1.23 – 1.52)	1.54×10^{-8}
6	97039665	rs2971609	G / A	0.20 / 0.18	1.15 (1.02 – 1.30)	2.44×10^{-2}	0.26 / 0.24	1.11 (0.98 – 1.25)	8.59×10^{-2}	0.23 / 0.23	1.01 (0.89 – 1.14)	9.24×10^{-1}	0.27 / 0.24	1.19 (1.07 – 1.33)	1.51×10^{-3}	0.27 / 0.23	1.22 (1.08 – 1.36)	7.17×10^{-4}	1.14 (1.08 – 1.21)	2.34×10^{-6}
6	143288832	rs1041655	A / C	0.41 / 0.42	0.93 (0.84 – 1.02)	1.29×10^{-1}	0.37 / 0.41	0.88 (0.79 – 0.98)	1.58×10^{-2}	0.35 / 0.38	0.87 (0.78 – 0.98)	1.79×10^{-2}	0.37 / 0.41	0.86 (0.78 – 0.95)	1.35×10^{-3}	0.36 / 0.39	0.90 (0.81 – 1.00)	3.04×10^{-2}	0.88 (0.84 – 0.93)	5.51×10^{-7}
6	153997563	rs6937294	C / T	0.43 / 0.45	1.08 (1.19 – 0.98)	1.11×10^{-1}	0.40 / 0.43	1.12 (1.24 – 1.01)	3.11×10^{-2}	0.38 / 0.42	1.20 (1.34 – 1.08)	5.28×10^{-4}	0.41 / 0.43	1.07 (1.18 – 0.97)	1.85×10^{-1}	0.37 / 0.40	1.14 (1.26 – 1.03)	9.10×10^{-3}	1.12 (1.18 – 1.07)	2.46×10^{-6}
7	17014115	rs17136988	G / T	0.05 / 0.06	0.84 (0.68 – 1.04)	1.03×10^{-1}	0.08 / 0.10	0.75 (0.62 – 0.91)	1.48×10^{-3}	0.07 / 0.08	0.83 (0.67 – 1.01)	4.84×10^{-2}	0.09 / 0.10	0.87 (0.73 – 1.03)	8.08×10^{-2}	0.07 / 0.09	0.80 (0.66 – 0.96)	8.66×10^{-3}	0.80 (0.73 – 0.88)	1.55×10^{-6}
7	40466200	rs4379368	T / C	0.16 / 0.14	1.17 (1.02 – 1.33)	2.14×10^{-2}	0.12 / 0.11	1.15 (0.97 – 1.35)	9.87×10^{-2}	0.12 / 0.11	1.17 (1.00 – 1.38)	5.24×10^{-2}	0.13 / 0.10	1.29 (1.11 – 1.50)	7.58×10^{-4}	0.13 / 0.11	1.20 (1.03 – 1.40)	1.58×10^{-2}	1.20 (1.12 – 1.29)	3.76×10^{-7}
8	4391037	rs17070498	C / T	0.10 / 0.11	0.90 (0.77 – 1.06)	2.01×10^{-1}	0.12 / 0.15	0.79 (0.67 – 0.92)	1.79×10^{-3}	0.13 / 0.15	0.86 (0.74 – 1.01)	5.99×10^{-2}	0.13 / 0.16	0.84 (0.73 – 0.97)	1.59×10^{-2}	0.12 / 0.14	0.82 (0.71 – 0.95)	6.92×10^{-3}	0.84 (0.79 – 0.90)	7.83×10^{-7}
8	81379656	rs368280	A / C	0.19 / 0.23	0.81 (0.72 – 0.92)	8.82×10^{-4}	0.25 / 0.28	0.85 (0.75 – 0.95)	4.89×10^{-3}	0.27 / 0.27	0.96 (0.86 – 1.09)	5.40×10^{-1}	0.25 / 0.28	0.86 (0.77 – 0.96)	5.80×10^{-3}	0.26 / 0.29	0.87 (0.78 – 0.98)	1.68×10^{-2}	0.87 (0.82 – 0.92)	1.93×10^{-7}
9	3943997	rs7024842	T / C	0.26 / 0.24	1.12 (1.01 – 1.26)	3.47×10^{-2}	0.34 / 0.30	1.18 (1.06 – 1.32)	3.02×10^{-3}	0.36 / 0.32	1.18 (1.06 – 1.32)	2.35×10^{-3}	0.33 / 0.30	1.15 (1.04 – 1.27)	9.03×10^{-3}	0.35 / 0.33	1.08 (0.98 – 1.20)	1.26×10^{-1}	1.15 (1.09 – 1.21)	1.40×10^{-7}
9	119252629	rs6478241	A / G	0.39 / 0.37	1.11 (1.22 – 1.00)	4.96×10^{-2}	0.39 / 0.36	1.12 (1.25 – 1.01)	3.55×10^{-2}	0.40 / 0.37	1.14 (1.26 – 1.02)	1.84×10^{-2}	0.42 / 0.36	1.26 (1.39 – 1.14)	5.45×10^{-6}	0.40 / 0.37	1.16 (1.29 – 1.05)	4.11×10^{-3}	1.16 (1.21 – 1.10)	1.68×10^{-9}
10	105039048	rs1163084	T / C	0.45 / 0.47	0.92 (1.02 – 0.84)	1.02×10^{-1}	0.50 / 0.48	0.92 (1.02 – 0.83)	9.48×10^{-2}	0.50 / 0.47	0.89 (0.99 – 0.80)	2.55×10^{-2}	0.51 / 0.47	0.87 (0.96 – 0.79)	3.99×10^{-3}	0.51 / 0.46	0.81 (0.89 – 0.73)	2.93×10^{-5}	0.88 (0.92 – 0.84)	4.79×10^{-8}
12	53831472	rs12820532	A / G	0.15 / 0.14	1.12 (0.98 – 1.29)	7.90×10^{-2}	0.17 / 0.15	1.14 (0.99 – 1.32)	4.16×10^{-2}	0.19 / 0.17	1.16 (1.01 – 1.32)	2.26×10^{-2}	0.18 / 0.14	1.28 (1.13 – 1.46)	6.42×10^{-5}	0.19 / 0.17	1.18 (1.04 – 1.35)	6.73×10^{-3}	1.21 (1.13 – 1.29)	2.61×10^{-8}
12	57527283	rs11172113	C / T	0.37 / 0.40	0.90 (0.82 – 1.00)	4.43×10^{-2}	0.39 / 0.41	0.92 (0.83 – 1.02)	1.20×10^{-1}	0.38 / 0.42	0.85 (0.76 – 0.94)	2.18×10^{-3}	0.36 / 0.40	0.86 (0.77 – 0.95)	1.90×10^{-3}	0.38 / 0.41	0.86 (0.77 – 0.95)	2.68×10^{-3}	0.88 (0.83 – 0.92)	4.38×10^{-8}
18	6767965	rs4798494	T / C	0.33 / 0.36	1.13 (1.25 – 1.02)	1.95×10^{-2}	0.34 / 0.37	1.14 (1.27 – 1.02)	2.21×10^{-2}	0.32 / 0.34	1.07 (1.20 – 0.96)	2.32×10^{-1}	0.33 / 0.36	1.12 (1.24 – 1.01)	2.82×10^{-2}	0.32 / 0.35	1.16 (1.29 – 1.04)	5.52×10^{-3}	1.12 (1.18 – 1.07)	2.64×10^{-6}
18	43706491	rs28532950	T / C	0.40 / 0.37	1.13 (1.02 – 1.25)	1.43×10^{-2}	0.45 / 0.41	1.15 (1.03 – 1.27)	9.51×10^{-3}	0.46 / 0.43	1.14 (1.02 – 1.26)	1.64×10^{-2}	0.46 / 0.42	1.17 (1.06 – 1.29)	1.20×10^{-3}	0.44 / 0.43	1.06 (0.96 – 1.18)	2.14×10^{-1}	1.13 (1.08 – 1.19)	2.26×10^{-7}

[a](cases/controls)

Position from human NCBI build 37. MAF, minor allele frequency. OR, odds ratio for the minor allele. CI, confidence interval.

MA, migraine with aura. MA/MO, migraine with and without aura. MO, migraine without aura.

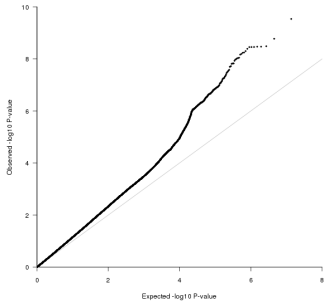

Figure 4.5: **Quantile-quantile plots of the GWAS discovery phase.** Plots of the fixed effect meta-analysis results in the MA discovery phase.

4.3 Replication stage

In the replication stage, SNPs from the top twenty nine loci were genotyped in 3268 migraine case and 2916 control European samples (Finland, The Netherlands and Spain) Of these six had at least one SNP that surpassed the threshold for genome-wide significance ($P = 5 \times 10^{-8}$) and 23 had at least one SNP with a P value lower than 5×10^{-6}.

Among the seventeen SNPs successfully genotyped, three reached the Bonferroni corrected replication threshold ($P \leq 2.94 \times 10^{-3}$): rs11892538 (OR=0.77 , 95% CI = 0.69 – 0.84, $P = 2.74 \times 10^{-5}$) rs4379368 (OR= 1.21, 95% CI = 1.08 – 1.35, $P = 8.68 \times 10^{-4}$) and rs11172113 (OR=0.86 , 95% CI = 0.79 – 0.92, $P = 3.66 \times 10^{-5}$). The effect estimate for rs11892538, rs4379368 and rs11172113 were concordant in direction among all replication cohorts with the discovery cohorts (Table 4.2). All the three SNPs reached genome-wide significance ($P \leq 5 \times 10^{-8}$) in a meta-analysis combining all cohorts (discovery and replication cohorts) (Table 4.3).

Table 4.2: **Summary results of the replication stage**

				Finnish (875/1025)[a]			Dutch (1043/910)[a]			Spanish (1350/981)[a]			Meta-analysis (3268/2916)[a]	
Chr	Position	SNP	Alleles minor/major	MAF cases/controls	OR(95% CI)	P	MAF cases/controls	OR(95% CI)	P	MAF cases/controls	OR(95% CI)	P	OR(95% CI)	P
1	3319777	rs4471209	A / G	0.32 / 0.35	0.88	6.92×10^{-2}	0.29 / 0.28	1.03	0.64	0.30 / 0.29	1.03	0.62	0.98	0.63
1	15532130	rs10927720	T / C	0.20 / 0.20	0.99	0.94	0.23 / 0.22	1.03	0.68	0.22 / 0.23	0.95	0.46	0.99	0.79
2	145222038	rs13403907	G / A	0.17 / 0.16	1.03	0.76	0.13 / 0.13	0.98	0.82	0.13 / 0.12	1.10	0.30	1.04	0.50
2	234821445	rs11892538	C / G	0.13 / 0.16	0.77	5.85×10^{-3}	0.17 / 0.21	0.80	5.20×10^{-3}	0.16 / 0.21	0.74	6.35×10^{-5}	0.77	2.74×10^{-5}
2	241447428	rs4676436	A / C	0.14 / 0.14	1.00	0.99	0.12 / 0.12	0.99	0.94	0.10 / 0.10	1.10	0.35	1.03	0.62
3	67144706	rs4311165	C / G	0.19 / 0.20	0.91	0.24	0.26 / 0.27	0.95	0.48	0.28 / 0.30	0.92	0.18	0.93	0.06
5	127722107	rs77050147	C / G	0.09 / 0.09	1.04	0.72	0.10 / 0.11	0.98	0.86	0.11 / 0.09	1.26	2.48×10^{-2}	1.09	0.14
6	12908747	rs13197912	T / A	0.30 / 0.29	1.02	0.80	0.37 / 0.38	0.97	0.60	0.40 / 0.37	1.13	4.63×10^{-2}	1.04	0.28
6	39177971	rs873690	C / G	0.08 / 0.07	1.03	0.79	0.06 / 0.05	1.05	0.72	0.05 / 0.05	0.95	0.70	1.01	0.89
6	143288832	rs1041655	A / C	0.42 / 0.43	0.97	0.60	0.38 / 0.37	1.07	0.32	0.39 / 0.38	1.05	0.43	1.03	0.46
7	40466200	rs4379368	T / C	0.17 / 0.13	1.31	2.75×10^{-3}	0.13 / 0.11	1.20	0.07	0.09 / 0.09	1.09	0.39	1.21	8.68×10^{-4}
8	4391037	rs17070498	C / T	0.10 / 0.11	0.91	0.38	0.15 / 0.13	1.19	0.06	0.16 / 0.16	1.00	0.96	1.03	0.55
8	81379656	rs368280	A / C	0.20 / 0.20	0.97	0.69	0.28 / 0.25	1.14	0.07	0.35 / 0.36	0.93	0.26	1.00	0.94
9	119252629	rs6478241	A / G	0.41 / 0.36	1.20	6.36×10^{-3}	0.39 / 0.38	1.03	0.69	0.45 / 0.44	1.05	0.38	1.09	2.32×10^{-2}
10	105039048	rs1163084	T / C	0.49 / 0.48	1.02	0.79	0.50 / 0.47	1.12	0.08	0.50 / 0.50	0.99	0.83	0.97	0.33
12	57527283	rs11172113	C / T	0.38 / 0.39	0.97	0.63	0.37 / 0.43	0.77	8×10^{-5}	0.32 / 0.36	0.85	7.61×10^{-3}	0.86	3.66×10^{-5}
18	43706491	rs28532950	T / C	0.38 / 0.38	1.01	0.87	0.45 / 0.42	1.16	2.20×10^{-2}	0.43 / 0.43	0.97	0.61	1.04	0.28

[a](cases/controls)

Position from human NCBI build 37. MAF, minor allele frequency. OR, odds ratio for the minor allele. CI, confidence interval.

MA, migraine with aura. MA/MO, migraine with aura and without aura. MO, migraine without aura.

Table 4.3: **Summary results of the discovery and replication stages**

				Discovery stage (5403/1537)[a]		Replication stage (3268/2916)[a]		Discovery and replication stages (8671/18243)[a]		
Chr	Position	SNP	Alleles minor/major	Meta-analysis OR (95% CI)	P	Meta-analysis OR (95% CI)	P	Meta-analysis OR (95% CI)	P	Gene
2	234821445	rs11892538	C / G	0.81 (0.76 – 0.86)	2.92×10^{-10}	0.77 (0.69 – 0.84)	2.74×10^{-5}	0.80 (0.76 – 0.84)	3.67×10^{-17}	*TRPM8*
7	40466200	rs4379368	T / C	1.20 (1.12 – 1.29)	3.76×10^{-7}	1.21 (1.08 – 1.35)	8.68×10^{-4}	1.20 (1.13 – 1.28)	1.36×10^{-9}	*C7orf10*
12	57527283	rs11172113	C / T	0.88 (0.83 – 0.92)	4.38×10^{-8}	0.86 (0.79 – 0.92)	3.66×10^{-5}	0.87 (0.84 – 0.91)	5.06×10^{-10}	*LRP1*

[a](cases/controls)
Position from human NCBI build 37.
MAF, minor allele frequency.
OR, odds ratio for the minor allele.
CI, confidence interval.

4.3.1 Discussion

In the first GWAS of imputed SNPs for migraine three loci associated with migraine were identified: one on chromosome 2q37.1 (rs11892538), one on chromosome 7p14.1 (rs4379368) and one on chromosome 12q13.3 (rs11172113) (Figure 4.6, 4.7 and 4.8). Two of the three loci (chromosome 2q37.1 and chromosome 12q13.3) had been already identified as associated with migraine in a previous study [244].

On chromosome 2q37.1, the most significantly associated marker rs11892538 maps to an intergenic region less than 5 kb away from the transient receptor potential cation channel 8 gene *TRPM8*. The second closest gene, encoding for Holliday junction recognition protein *HJURP*, maps 58.2 kb away from rs11892538. Given the current knowledge, *TRPM8* could be involved in migraine pathogenesis. *TRPM8* is a cold and menthol modulated ion channel with a role in the detection of cold in the mammals [246]. *TRPM8* is expressed in subpopulations of sensory

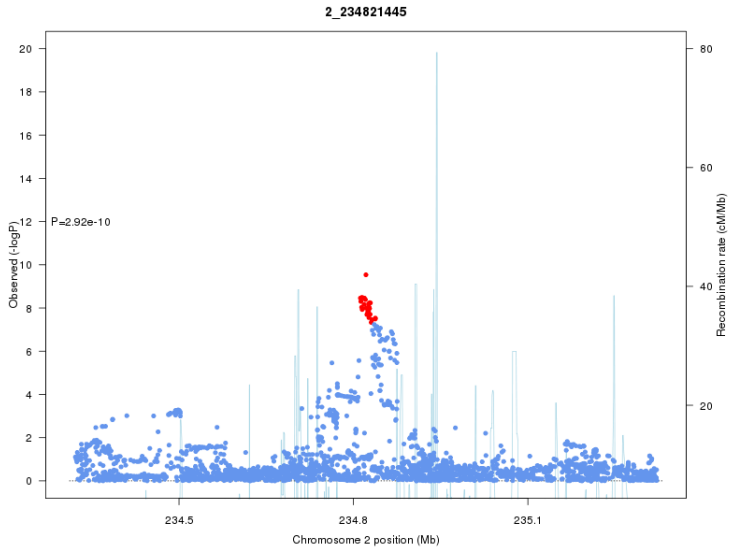

Figure 4.6: **Locus specific association plot: chromosome 2q37.1** The region +/- 500 kb around the most strongly associated SNP is shown. The diamond represents the most strongly associated SNP. P values are shown for the discovery stage. The blue line shows the recombination rate based on HapMap Phase II data. SNP and gene position are based on built 37.

neurons [247]. There is evidence suggesting that *TRPM8* may play a role in inflammatory and neuropathic pain [248]. Given that the migraine headache has some features in common with inflammatory and neuropathic pain, it is possible that *TRPM8* may play a role in its pathogenesis [249]. There is evidence suggesting that the in vivo antagonism of TRPM8 constitutes a possible strategy for treating neuropathic pain [250].

On chromosome 12q13.3, the most significantly associated marker rs11172113 maps to the first intron of low density lipoprotein receptor-related protein gene *LRP1*. *LRP1* is a cell surface receptor member of the low-density lipoprotein (LDL)-receptor family [251,252]. It is expressed in the vasculature, central nervous system, macrophages and adipocytes [253]. *LRP1* seems to play a role in various biological processes including lipoprotein metabolism [253]. Boucher et al. (2003)

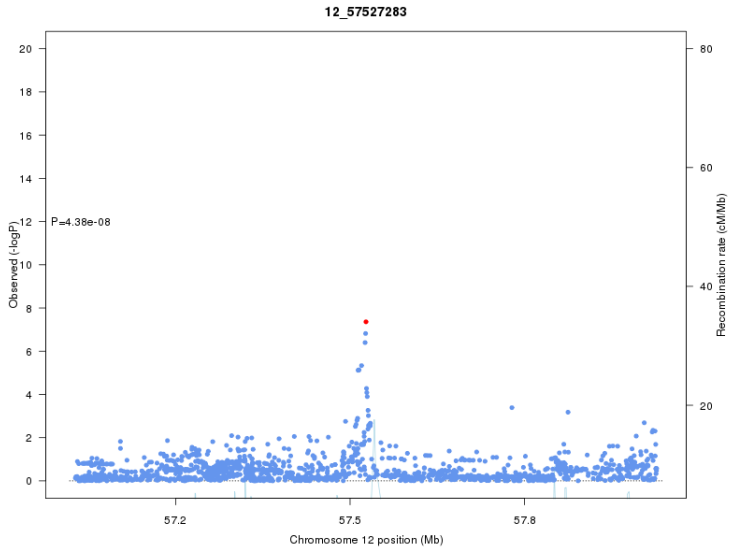

Figure 4.7: **Locus specific association plot: chromosome 12q13.3** The region +/- 500 kb around the most strongly associated SNP is shown. The diamond represents the most strongly associated SNP. P values are shown for the discovery stage. The blue line shows the recombination rate based on HapMap Phase II data. SNP and gene position are based on built 37.

have shown that inactivation of *LRP1* in vascular smooth muscle cells of mice leads to marked susceptibility to cholesterol-induced atherosclerosis [254]. Liu et al. (2010), performing neuronal lrp1 knockout in mice, have shown that the levels of glutamate receptors are reduced in lrp1 knockout neurons and that they are partially rescued by restoring neuronal cholesterol [255]. Glutamate is the main excitatory neurotransmitter in the central nervous system. Data from animal and human studies support a role of glutamate in the pathophysiology of migraine [256]. The second closest gene in the region, signal transducer and activator of transcription 6 (*STAT6*), maps 23.1 kb away from rs11172113. *STAT6* is a member of the STAT family of transcription factors, which plays a role in differentiation and function of T helper 2 (Th2) cells [257].

On chromosome 7p14.1, the most significantly associated marker rs4379368

maps to the dermal papilla derived protein 13 gene *C7orf10*. *C7orf10* is a peroxisomal glutaryl-CoA oxidase [258]. Mutations in this gene have been associated with glutaric aciduria type III, characterized by abnormal amounts of urinary glutaric acid [258]. Bennett et al. (1991) described a lack of peroxisomal glutaryl-CoA oxidase activity in a 1-year-old girl with failure to thrive and hematologic evidence of thalassemia. Sherman et al. (2008) reported three children homozygous for a nonsynonymous variant in C7orf10, who excreted large quantities of glutarate in the urine and remained healthy during a 15 years follow-up period [259]. The second closest gene to rs4379368 , encoding for cell division cycle 2-like 5 *CDC2L5*, maps 331 kb away. This gene encodes for a member of the cyclin-dependent serine/threonine protein kinase family. Members of cyclin-dependent serine/threonine protein kinase family have an important role in cell cycle control. The exact function of the protein encoded by *CDC2L5* has not been defined yet, but it has been suggested that it may have a role in mRNA splicing regulation [260].

In conclusion, in this first GWAS of imputed SNPs for migraine, three loci (one on chromosome 2q37.1, one on chromosome 7p14.1 and one on chromosome 12q13.3), associated with migraine were identified. Two of these three loci (one on chromosome 2q37.1 and chromosome 12q13.3) had already been found associated with migraine in a recent population based study [244]. The most significantly associated marker on chromosome 2q37.1 (rs11892538), maps 5 kb away from the *TRPM8* gene, which encodes for a ion channel with a role in pain pathogenesis [248]. On chromosome 12q13.3, the most significantly associated marker (rs11172113) maps to the first intron of the *LRP1* gene. *LRP1* has been shown to modulate the neural glutamate receptor levels, and therefore, the association of

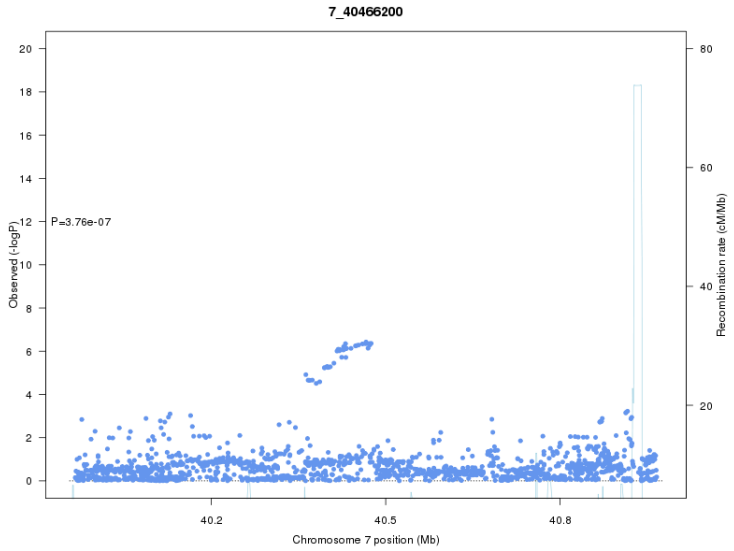

Figure 4.8: **Locus specific association plot: chromosome 7p14.1** The region +/- 500 kb around the most strongly associated SNP is shown. The diamond represents the most strongly associated SNP. P values are shown for the discovery stage. The blue line shows the recombination rate based on HapMap Phase II data. SNP and gene position are based on built 37.

LRP1 with migraine provides further support to the role of glutamate in migraine pathogenesis [255]. The functional role of the new third locus on chromosome 7p14.1 is not currently definable. Future functional studies on the role of genes present in the locus (*C7orf10* and *CDC2L5*) will be provide an understanding of its functional link with migraine. Larger GWAS, currently underway, will allow the identification of further genetic variants underlying the pathophysiology of migraine.

Chapter 5

GENCODE exome

5.1 Introduction

Genome-wide association studies have been successful in identifying common variants associated with complex human diseases and traits [194]. However, in most of the cases the portion of the heritability explained by these variants is modest [51–53]. It has been suggested that rare variants, copy number variations, gene-gene interactions and epigenetic mechanisms may be the source of the 'missing heritability' [54].

The development of next generation sequencing (NGS) has allowed the systematic discovery of rare variants in thousands of samples [55]. The cost of whole-genome sequencing has been falling dramatically over the last couple of years, however it is still too expensive to be applied to large scale genomic studies aimed at identifying variants associated with complex diseases. Currently, the combination of next generation sequencing technologies with efficient methods of sequence capture has enabled the widespread targeting of the exome [195–199]. Exome re-

sequencing constitutes an effective tool for discovering coding variants underlying monogenic diseases and for identifying coding variants associated with complex diseases [261–263].

As part of a pilot study aimed at identifying rare variants associated with complex neurological diseases, we sequenced the exomes of five individuals with epilepsy. Analyzing the called variants we realized that genes with a potential impact on the studied phenotype, such as ion channel subunits, were not captured by the used CCDS based capture array.

The two most widely used commercial kits for capturing the exome, available at that time, (NimbleGen Sequence Capture 2.1M Human Exome Array, http://www.nimblegen.com/products/seqcap/ and Agilent SureSelect Human All Exon Kit, http://www.genomics.agilent.com) targeted exons from genes in the consensus coding sequence (CCDS) consortium database, in addition to a selection of miRNAs and non-coding RNAs [264]. Although the CCDS database contains a high-quality set of consistently annotated protein-coding genes, many annotated genes, with solid evidence of transcription, are not part of this set yet. In addition, in the CCDS database only 21% of the genes have alternative spliced variants included.

To address this shortcoming, a more complete set of target regions for the human exome, based on the GENCODE annotation was designed and experimentally tested [200]. The GENCODE is part of the Encode project and responsible for the annotation and experimental validation of gene loci on the human genome.

Table 5.1: **Comparison of the three different exome capture sets**

	[a]NimbleGen CCDS	[b]Agilent CCDS	GENECODE exome
Number of bait regions	197218	316000	406539
Genome coverage (Mb)	34.1	37.6	[c]47.9 [d](35.2)
ECRs covered [e](%)	150529 (72.7)	164225 (79.3)	205031 (99.0)
Transcripts covered [e](%)	66828 (81.0)	71279 (86.4)	81204 (98.4)
Genes covered [e](%)	28203 (76.5)	30030 (81.5)	35989 (97.7)

[a]NimbleGen Sequence Capture 2.1M Human Exome Array
[b]Agilent SureSelect Human All Exon Kit
[c]Total length of bait regions including flanking regions.
[d]Design target length without flanking regions.
[e]Percentage of the GENCODE exome design target

5.2 Results

5.2.1 The GENCODE exome features

A comparison of the coverage of the bait/oligonucleotide positions of available CCDS-based exome sets and our GENCODE exome set with the GENCODE design target showed an increased coverage of our set (Table 5.1).

The GENCODE exome baits covered 99% of the design target, resulting in additional 59600 exons available for capture, which were not present in either one of the CCDS-based sets [265]. The missing 1% were regions for which reliable bait design was not possible.

A comparison of exon and transcript coverage of the available CCDS based exome sets and the GENCODE exome set with three current reference gene sets (CCDS, RefSeq, Gencode), showed that the GENCODE exome set covered a

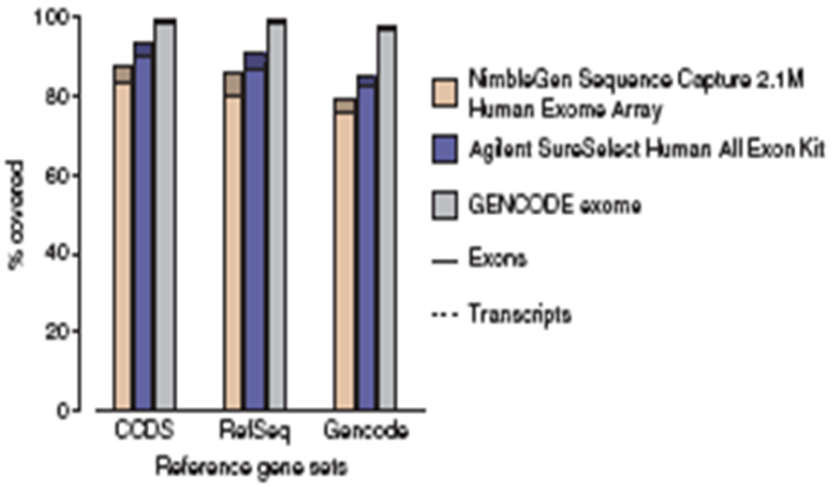

Figure 5.1: **Comparison of exon and transcript coverage of the available CCDS-based exome sets and the GENCODE exome set with three current reference gene sets.** The histograms show the near-complete coverage by the GENCODE exome set of all reference sets.CCDS database March 2010, RefSeq genes March 2010 and GENCODE version 3c.

greater percentage of the reference genes (Table 5.2, figure 5.1). For example, the GENCODE exome set covered an additional 9% of the exons from the CCDS database and 12% of the exons from RefSeq compared to the CCDS-based exome sets.

The content present exclusively in the GENCODE exome set consists of 38933 exome cluster regions, which contain 5594 additional genes of the GENECODE exome design target. Of these additional genes, the 4363 Ensembl-53-based genes include 1881 (43%) genes with an official HGNC identifier, 711 (16%) with an OMIM entry and 1410 (32%) with a Gene Ontology annotation (Table 5.3) [265]. The content of repetitive/low-complexity sequence in the GENCODE exome set is similar to the available CCDS-based exome sets (Table 5.4).

A comparison with a sequence uniqueness mask supports these findings (Table

Table 5.2: **Exon and transcript coverage of the three different exome caputure sets**

Exome set	CCDS exons	CCDS transcripts	RefSeq exons	RefSeq transcripts	GENECODE exons	GENECODE transcripts
NimbleGen Sequence Capture 2.1M Human Exome Array	82.80%	87.19%	79.47%	85.28%	75.29%	79.16%
Agilent SureSelect Human All Exon Kit	90.05%	93.16%	86.39%	90.91%	82.15%	84.67%
GENCODE exome	99.18%	99.63%	98.75%	99.19%	97.31%	96.51%
Additional content of the GENCODE exome	9.12%		12.36%			

Table 5.3: **Comparison of the three different exome caputure sets with the GENECODE design target**

	Genes	Transcripts	Exons
[a]Totals	36853	82522	463778
Nimblegen CCDS [b](%)	28203 (76.5%)	66828 (81.0%)	307866 (66.4%)
Agilent CCDS [b](%)	30030 (81.5%)	71279 (86.4%)	334191 (72.1%)
GENCODE exome [b](%)	35989 (97.7%)	81204 (98.4%)	397856 (85.8%)
Covered by all 3 sets	27828	65943	303483
Covered by Nimblegen CCDS only	10	12	318
Covered by Agilent CCDS only	0	0	0
Covered by Gencode exome only	5594	9052	59600
Covered by Nimblegen and Agilent CCDS only	0	0	0
Covered by Nimblegen CCDS and Gencode exome only	365	873	4065
Covered by Agilent CCDS and Gencode exome only	2202	5336	30708

[a]Referred to the GENECODE design target
[b]Percentage of the GENCODE exome design target

Table 5.4: **Assessment of repeat and low-complexity coverage of the three exome sets.** Repeats and low-complexity regions identified with RepeatMasker (parameters: -nolow -species homo -s), Dust and TRF using Ensembl 53 data.

Exome set	Total base pairs	Base pairs with repeats	Ratios
NimbleGen CCDS	34108810	884080	38.6
Agilent CCDS	37640396	799357	47.1
GENCODE exome	47933967	1303879	36.8

5.5).

The list of 5594 additional genes and regions targeted by the GENCODE exome exclusively, data for the final GENCODE exome and the initial design target is available on our ftp site (http://ftp.sanger.ac.uk/gencode/exome).

The 406539 bait locations are supplied as a Distributed Annotation System data source (das.sanger.ac.uk/das/Exome) and they can be displayed in genome browsers such as Ensembl (version 53; http://tinyurl.com/browse-exome) [265].

5.2.2 The GENCODE exome performance

To evaluate the performance of the GENCODE exome set, DNA from three HapMap samples (NA12878, NA07000 and NA19240) was captured using both the Agilent SureSelect Human All Exon Kit and GENCODE exome baits.

Moreover, to evaluate the performance of the GENCODE exome set using DNA from clinical samples, DNA from seven individuals recruited from a clinical neurological unit was captured using the GENCODE exome baits. Samples were sequenced as described in the methods section. Variants were called using SAMtools v0.1.71 and GATK, the intersection of the resulting calls in the GENECODE target regions (39.3 Mb) with a sequence read depth of $\geq$ 8x was reported and compared to known variants in all the samples.

On average 97% of reads could be successfully mapped back to the genome, 67% of the mapped reads were uniquely mapped and 82% 0f the unique mapped reads derived from the capture target regions (Table 5.6).

The average coverage of the HapMap samples was 73-fold from 9.2 Gb of se-

Table 5.5: **Bait/probe covered CTR (Capture Target Regions) regions assessed using a uniqueness mask** The used uniqueness mask was developed by Heng Li for the 1000 Genomes project (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/pilot_data/technical/reference/README_hg36_uniqueness_mask)

Exome set	CTR count	Total target size (bp)	[a]Type 0 bases	[a]Type 1 bases	[a]Type 2 bases	[a]Type 3 bases
Nimblegen CCDS	165637	37640396	0 [b](0%)	1570956 [b](4.17%)	838459 [b](2.23%)	35230981 [b](93.60%)
Agilent CCDS	176159	34108807	17688 [b](0.05%)	1236875 [b](3.63%)	745619 [b](2.19%)	32108625 [b](94.14%)
GENCODE exome	206275	47933967	0 [b](0%)	2811712 [b](5.87%)	1177904 [b](2.46%)	43944351 [b](91.68%)

[a]Sequenceability/uniqueness measures used in the method:
- Type 0 : all 35mers covering this site cannot be mapped back due to "N"s in the reference
- Type 1 : otherwise (>35×0.5 reads are exact repeats)
- Type 2 : if not 3, $\geq 35 \times 0.5$ reads 1-away unique
- Type 3 : $\geq 35 \times 0.5$ reads 2-away unique

[b]Percentage of total target bases

Table 5.6: **Mapping statistics for clinical and HapMap samples using GENCODE and Agilent CCDS exome captures.**

Sample name	Sanger 1	Sanger 2	Sanger 3	Sanger 4	Sanger 5	Sanger 6	Sanger 7	NA12878	NA070000	NA19240	NA12878	NA070000	NA19240
Bait set	GENECODE										Agilent CCDS		
Lane sequenced	3	3	3	3	3	3	3	4	3	3	3	4	4
Reads	105712968	115540748	107420522	101912732	129221672	205597102	203472750	249957386	216080714	175087650	131337304	174698654	206110412
Reads mapped	101784111 [a](96.28%)	112290675 [a](97.19%)	104444274 [a](97.23%)	99877068 [a](98.00%)	126463612 [a](97.87%)	197922085 [a](96.27%)	196009151 [a](96.33%)	237675276 [a](95.09%)	212171108 [a](98.19%)	168827968 [a](96.42%)	123007651 [a](93.66%)	169580062 [a](97.07%)	198938967 [a](96.52%)
Unique reads mapped	77800441 [b](76.44%)	81310815 [b](72.41%)	74873012 [b](71.69%)	74979882 [b](75.07%)	97666019 [b](77.23%)	145520170 [b](73.52%)	143615416 [b](73.27%)	134894523 [b](56.76%)	122795772 [b](57.88%)	77220371 [b](45.74%)	91946948 [b](74.75%)	97248119 [b](57.35%)	117233076 [b](58.93%)
Unique reads mapped to CTR +/- 250bp	59625505 [c](76.64%)	69956576 [c](80.04%)	65714990 [c](87.77%)	63735274 [c](85.00%)	68908856 [c](70.56%)	86495628 [c](59.44%)	86910473 [c](60.52%)	124450772 [c](92.26%)	113776953 [c](92.66%)	71794822 [c](92.97%)	83772118 [c](91.11%)	89399731 [c](91.93%)	99555617 [c](84.92%)
Unique reads mapped to GENCOODE ECRs	40507262 [c](52.07%)	47280119 [c](58.15%)	46480228 [c](62.08%)	43032073 [c](57.39%)	45129213 [c](46.21%)	66486708 [c](67.96%)	66344444 [c](46.20%)	84566996 [c](62.69%)	81625999 [c](66.47%)	51683594 [c](66.93%)	56713564 [c](61.68%)	66088857 [c](67.96%)	69982605 [c](59.70%)
[d]Mean depth in the GENCODE ECRs	46.38	54.08	53.45	49.83	51.02	76.96	76.61	95.37	92.37	58.81	65.08	75.92	80.14
[d]Median depth in the GENCODE ECRs	34	39	35	34	36	55	55	63	62	36	43	47	57
Bases covered $\geq 8x$ in the GENECODE design target	32914133 [e](83.66%)	33102756 [e](84.14%)	32201108 [e](81.85%)	32536951 [e](82.70%)	32220983 [e](81.90%)	35330571 [e](89.80%)	34963040 [e](88.87%)	32415365 [e](82.39%)	32910800 [e](83.65%)	31502723 [e](80.07%)	28721876 [e](73.01%)	28903556 [e](73.47%)	29429202 [e](74.80%)

[a]Percentage of reads
[b]Percentage of reads mapped
[c]Percentage of unique reads mapped
[d]Calculated after duplicate read removal
[e]Percentage of GENECODE exome design target bases
CTR, Capture Target Regions; ECRs, Exome Cluster Regions

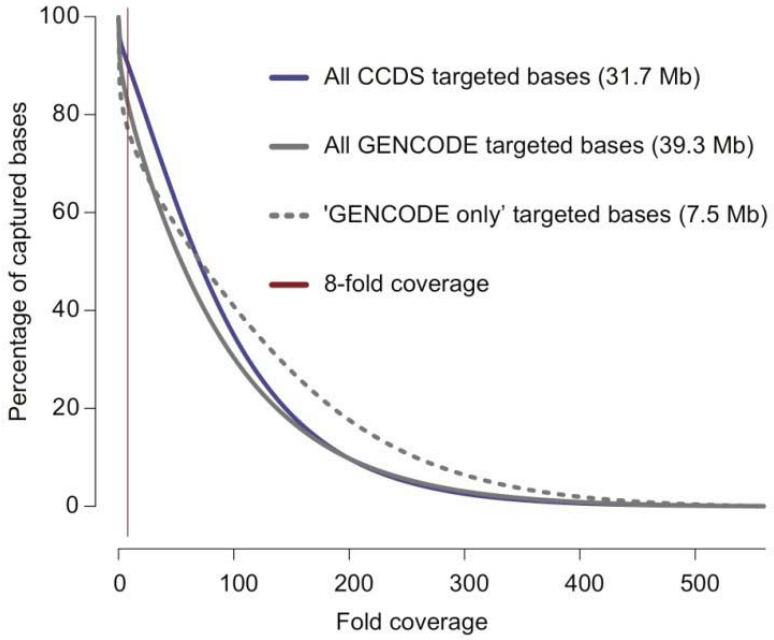

Figure 5.2: **Cumulative distribution of base coverage for HapMap samples.** The blue line represents the cumulative distribution of base coverage for the CCDS based captures. The continuous grey line represents the cumulative distribution of base coverage for the GENCODE exome based captures, and the dashed grey line represents the cumulative distribution of base coverage for the GENCODE exome based captures in the regions covered only by the GENCODE exome baits. The thin red vertical line indicates a coverage of eightfold, which is the coverage commonly required for variant calling.

quence for the CCDS-based captures and 82-fold from 11.5 Gb of sequence for the GENCODE exome captures. The average coverage for the clinical samples, captured only with the GENCODE exome baits, was 58-fold from 7.5 Gb of sequence. The coverage has been calculated only using reads with a mapping quality of 10.

For the HapMap samples, on average, 96% of targeted bases were covered at least once and 90% were covered eightfold or more for the CCDS-based captures. Similar figures were obtained for the the GENCODE exome based captures with 92% of targeted bases covered at least once and 83% covered eightfold or more (Figure 5.2).

For the clinical samples, on average 95% of targeted bases were covered at least once and 88% covered eightfold or more (Figure 5.3).

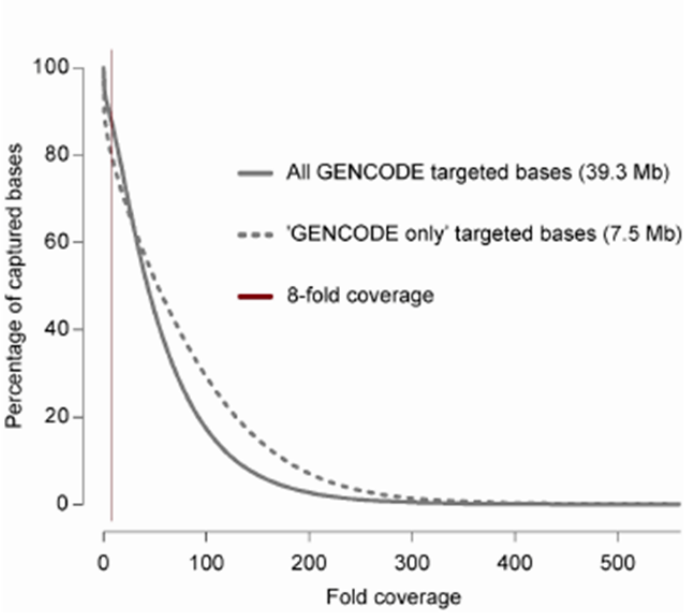

Figure 5.3: **Cumulative distribution of base coverage for the clinical samples.** The continuous grey line represents the cumulative distribution of base coverage in all the targeted regions, and the dashed grey line represents the cumulative distribution of base coverage in the regions covered only by the GENCODE exome baits. The clinical samples were captured only using the GENECODE based Agilent SureSelect Human All Exon Kit. The thin red vertical line indicates a coverage of eightfold, which is the coverage commonly required for variant calling.

The results demonstrate that the GENCODE based captures performed equally to the CCDS based captures. Moreover, considering the regions covered only by the GENECODE baits similar figures were obtained, proving that these regions perform equally to the CCDS regions (Figure 5.2).

Only unique reads mapped to the target were used for variants calling. For the HapMap samples, an average of 22271 variants, of which 2.6% were novel, were called in GENCODE based captures compared with an average of 18554 variants, of which 1.7% were novel, called in the CCDS-based captures (Table 5.7). Variants were defined as novel if they were not present either in dbSNP18 (version 130) or 1000 Genomes project (1000 Genomes Project Consortium, http://www.1000genomes.org, released on 26 March 2010).

Table 5.7: **Variant calling statistics for clinical and HapMap samples using GENCODE and Agilent CCDS exome captures.**

Sample name	Sanger 1	Sanger 2	Sanger 3	Sanger 4	Sanger 5	Sanger 6	Sanger 7	NA12878	NA070000	NA19240	NA12878	NA070000	NA19240
Bait set	GENECODE										Agilent CCDS		
Variants	21170	21529	21052	21445	21124	23612	23276	20780	21513	24520	16732	17014	21915
% dbSNP (version 130)	93.7	93.5	93.7	93.8	92.1	93.5	93.5	96.5	94.1	94.8	98.0	95.2	95.5
% dbSNP and/or 1000 Genomes (26/03/10 pilot 1)	96.3	95.1	96.3	96.1	94.7	96.1	96.1	97.7	97.2	97.3	99.0	97.9	98.1
Heterozigous	12604	13241	12938	13153	13297	14476	14321	12675	13588	16121	10094	10583	14414
Ti/Tv	3.029	2.996	3.036	3.025	2.930	3.021	3.120	3.069	3.112	3.138	3.235	3.258	3.322
% [a] Concordant	99.78	99.79	99.89	99.72	99.72	99.83	$	99.31	99.61	99.16	99.34	99.66	99.17
Synonymous	9196	9249	9072	9191	8948	10220	10111	8480	9207	10568	7979	8133	10528
Non synonymous	8608	8804	8692	8828	8696	9634	9385	8758	8703	9958	6863	6976	8918
Stop gained	86	85	80	89	128	87	95	80	83	99	44	40	51
Variants in the GENCODE only [b] ECRs	5179	5212	5117	5162	5162	5414	5424	5017	5319	5887			

[a] Called variants in the exome captures were compared with Illumina 660K chip genotypes for clinical samples and with the HapMap3 genotypes for HapMap samples.

[b] Excluding flanking regions

Ti, transitions; Tv, transversions. $, missing data. ECRs, Exome Cluster Regions.

An average of 21866 variants, of which 4.2% were novel, were called in the clinical samples. The HapMap samples and the clinical samples had been previously genotyped on arrays, therefore the variants called in the GENECODE based captures were compared to the genotyped SNPs. The concordance rate was found to be 99.7%.

Most of the variants, for which genotypes were discordant between array genotyping and sequencing, were discrepant only in one sample, suggesting that the number of systematic either genotyping or sequencing errors was low.

The 22002 variants found on average in the GENCODE exome captures included, 9006 non-synonymous variants, 9424 synonymous variants and 91 stop-gained variants. Meaning that 268 synonymous variants, 256 non-synonymous variants and 2.6 stop-gained variants were found per megabase of the targeted genomic sequence (35.2 Mb) , corresponding to a total of 626.6 variants per megabase. In the CCDS captured samples among the 18554 coding SNPs found on average, there were 7585 non-synonymous variants, 8880 synonymous variants and 45 stop-gained variants, corresponding to a total of 512 variants per megabase.

5.3 Discussion

The GENCODE gene set used as reference for the design of our ECRs (Exome Cluster Regions) provides a more complete set of targets, as it is the result of the merging of the thorough manual Havana and the genome-wide automatic Ensembl annotation. Both Havana and Ensembl are part of the CCDS consortium, and all

of the CCDS annotated genes have been incorporated into the new target set.

The new set set of target regions for exome capture includes genes potentially relevant for the discovery of disease-associated variants. Among the genes captured by our new expanded set, there are members of well-characterised gene families, which have been associated with important medical conditions. For example, 43 genes, captured only by our set, encode for ion channel subunits. Mutations in ion channel genes have previously been found to cause a range of channelopathies, including arrhythmias and inherited paroxysmal neurological disorders [266]. Seventy genes, encoding for proteins with kinase activity are exclusively present in the GENECODE based set. Members of the protein kinase family have been found commonly mutated in cancer and are considered to be candidate targets for the development of new anticancer therapies, therefore it is important that they are covered in cancer sequencing studies. [267].

The coverage over two genes targeted only by the GENCODE based set, *ABCB11* and *XPC*, (Figures 5.4 a and b) demonstrates that we have been able to design baits for genes not represented in the previous CCDS based sets and that these genes are efficiently captured and uniformly covered. Each exon is covered on average more than eightfold, which is most commonly required depth for variant calling. *ABCB11* and *XPC* are medically relevant genes, since they have already been associated with diseases, and provide examples of candidate disease genes that are missing from the existing CCDS based exome capture sets.

The use of the GENCODE based baits has already allowed the identification of a pathogenic mutation in a gene causing an autosomal recessive Dwarfism syndrome, which would not have been discovered using a standard CCDS-based sets [268].

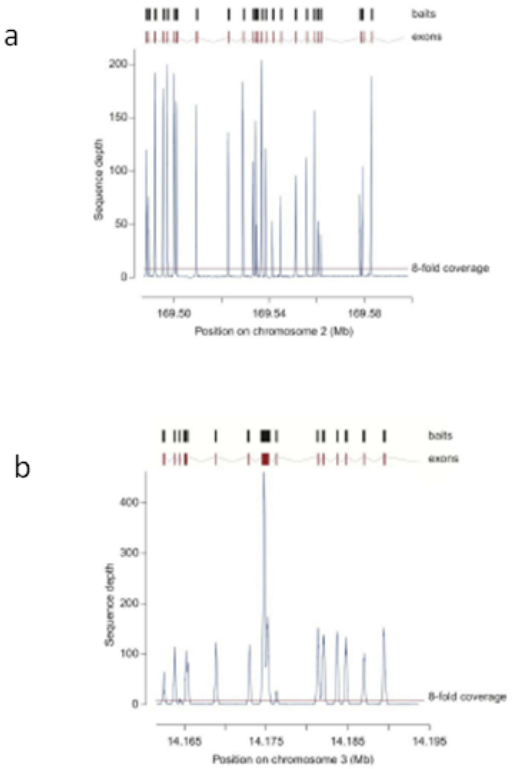

Figure 5.4: **Coverage achieved by the GENECODE based set.** Detailed view of the average depth in the seven clinical samples across two genes that are unique to the GENCODE based set: (a) *ABCB11* and (b) *XPC*. In the upper part of each panel, the positions of the GENCODE exome baits are represented by dark boxes above the gene structure (adapted from the Ensembl genome browser) in red. The increased sequencing depth of the eighth exon of *XPC* is due to high coverage of this larger exon by eight different baits, whereas the other smaller exons are covered by one or two baits. The horizontal thin red line indicates a coverage of eightfold.

The advent of the GENCODE exome represents a substantial improvement to the currently available designs for exome capture, allowing the capture of a more complete target. We estimate that we were able to call variants in 84% of the total GENCODE target regions. The fraction of callable exome regions is likely to increase further with improving sequencing technology and sequencing depth. The GENCODE exome design is currently used by the International Cancer Genome Consortium (ICGC; http://www.icgc.org) for their exome sequencing projects, aimed at obtaining a comprehensive description of different tumor types and by the UK10K project (http://www.uk10k.org).

5.4 Notes

Sequencing data have been deposited at the European GenomePhenome Archive (http://www.ebi.ac.uk/ega/) under accession number EGAS00001000016 and the European Nucleotide Archive (http://www.ebi.ac.uk/ena) under accession ERP000523.