GlaxoSmithKline

Alzheimer's Disease Study

Quality Control and Analysis Procedures

Greta Linse Peterson, G. Bryce Christensen, and Christophe G. Lambert

Golden Helix, Inc.
PO Box 10633
Bozeman, MT 59719
http://www.goldenhelix.com

February 18, 2010

Contents

1 Introduction

Golden Helix, Inc (GHI) performed the analyses described in this document in collaboration with data provider GlaxoSmithKline (GSK). All analyses were performed using Golden Helix SNP & Variation Suite (SVS) unless otherwise indicated. The final results are contained in the SVS project directory titled **SVS_GSK_Alzheimers_Revised_Feb2010.** All spreadsheets and plots referred to below are contained within this project and can be referenced using the given node name and ID number. A *free* SVS project viewer is available from GHI at http://www.goldenhelix.com/SNP_Variation/svstrial.html.

During the course of analysis and quality control, several samples were found to have discrepancies in genotype and/or phenotype data. These discrepancies included mismatches of NSP and STY arrays, discrepancies in reported gender, sample contamination and unexpected duplicates of genotype data for which the associated phenotypes ruled out the possibility of MZ twins. The source of these inconsistencies could not be directly identified from the data, and it was impossible therefore to determine the informed consent status of the human subjects represented by the associated phenotype and genotype data. As a result, all phenotype and genotype data relating to those subjects has been withheld from this submission. Any researchers wishing to exactly replicate the analysis described here should be aware that this document contains references to the methods used to identify the discrepant subjects, but no data for those subjects is included in the public data distribution. Graphical illustrations of quality assurance measurements contained in this document include representations of the discrepant subjects, but corresponding figures in the accompanying SVS project files do not include those subjects.

2 Data Preparation and initial quality assurance

GHI received raw CEL files from GSK for Alzheimer's disease cases and controls genotyped with Affymetrix 500K SNP genotyping system, based on the 250K NSP & STY arrays. Samples were collected via venous blood collection into a blood tube with K_2EDTA anticoagulant. DNA was extracted by a modified salting-out precipitation method and resuspended in TE (10mM Tris, 1mM EDTA). 3290 CEL files were provided to GHI, consisting of data from 1646 NSP arrays, and 1644 STY arrays. Subject-sample matching data indicated 1628 matched NSP-STY pairs. One subject (SUBJID 1075) had 2 instances of STY CEL files. GSK2239_STY_E8_RETRY_1.CEL was chosen because it came from the same vendor as the NSP CEL file and had a higher call rate than EA05062_0027-08E_STY250_2239-61.CEL, which was dropped from analysis. The remaining STY and NSP arrays were unpaired and excluded from analysis.

The samples were processed by three different genotyping vendors. In some instances the matching NSP and STY files were not processed by the same vendor. Two markers (rs429358 and rs7412) not found in the Affymetrix 500k array were genotyped separately and added to the analysis. These markers are of special interest due to their position in the APOE gene, for which earlier studies indicated association with Alzheimer's disease.

The primary phenotype data provided by GSK included variables for subject ID, gender, Alzheimer's case/control status, investigator site, sex, age, age at onset, and ethnic origin. Additional phenotypic and QA data as provided

by GSK or determined by GHI are listed in the accompanying data dictionary files. All samples were reportedly white/Caucasian, collected at eight sites in Canada, including Quebec. After removing samples that failed quality control as described below, there were 1577 subjects, consisting of 778 controls and 799 cases. The ages for all of the samples ranged from 42.6 to 100 years old. The mean age for controls was 73.4 (min=47.6, max=93.9), and the mean age for cases was 78.0 (min=42.6, max=100). There were 615 males and 962 females; the reported gender for these samples matched the imputed gender.

The quality control steps to exclude samples and SNPs are outlined below in the order in which they were performed.

2.1 Calculate genotypes using CRLMM

We ran the CRLMM[2] calling algorithm from the Bioconductor 2.4 oligo[1] (version 1.8.0) package, which has been demonstrated[2] to generate superior calls to the BRLMM[3] and Birdseed[4] algorithms and to be less sensitive to batch effects. We ran CRLMM on the NSP and STY files separately and calculated call rates on autosomes. All 1646 NSP CEL files were processed in one undivided batch, and all 1644 STY CEL files in a second undivided batch so as to not introduce computational batch effects with the CRLMM algorithm. The CRLMM genotype calls for the NSP markers are in the spreadsheet **crlmm-calls NSP – Sheet 1** which corresponds to node ID 600, the calls for the STY markers are in the spreadsheet **crlmm-calls STY – Sheet 1**, node ID 603.

CRLMM generates genotype calls as well as per-call confidences. We replaced all genotype calls whose CRLMM-reported confidence was less than 0.95 with missing values.

2.2 Sample Quality Assurance

As the genotype data was from the dual-array Affymetrix 500k genotyping platform, with sometimes inconsistent matching of the NSP and STY arrays, we analyzed sample quality with the two arrays individually as well as combined.

2.2.1 Assess call rates per individual, exclude low call rate samples

There were 11 samples with a combined call rate for both the NSP or STY arrays less than 0.94 (616, 825, 949, 950, 1108, 2411, 3462, 5162, 5259, 6120, and 6126) that were excluded from the analysis. Figure 1 shows a histogram of the combined NSP & STY call rate for the samples. None of the samples excluded based on call rate were excluded for any other reason. See the spreadsheet, **Phenotype + Sample QC Measures – Sheet 1**, node ID 686. The phenotype spreadsheet has two added columns that report the NSP and STY call rates per sample. These columns are labeled **NSP_AUTOSOME_CALL_RATE** and **STY_AUTOSOME_CALL_RATE**, respectively.

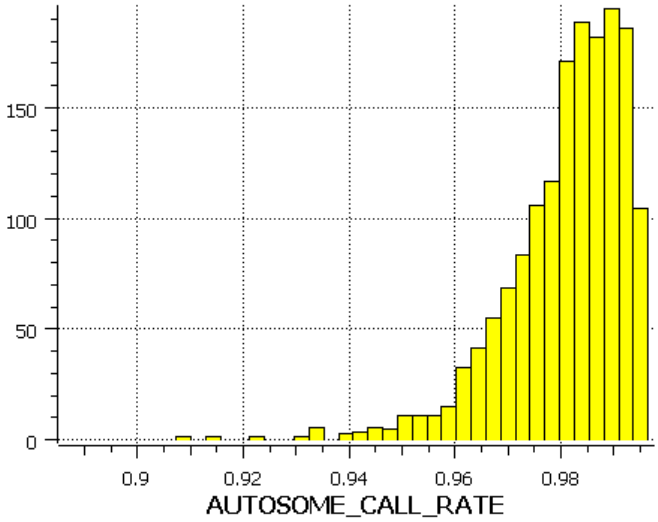

Figure 1: Sample call rate histogram

2.2.2 Verify NSP and STY arrays were matched correctly

In order to ensure the correct NSP and STY arrays were matched together, representing the same sample, a test of the correlation between the genotypes for each sample over both arrays was conducted. The nature of linkage disequilibrium is such that for two neighboring SNPs, the common alleles will usually be inherited together as part of a haplotype block. Similarly, rare alleles are also generally inherited together as part of a haplotype. This phenomenon can be used to test the matching between the NSP and STY genotyping arrays used in this study. Over a large number of closely spaced SNP pairs, with one SNP from each array, we should observe correlation between the genotypes for appropriately matched samples. For each SNP on the NSP array we identified the nearest STY SNP within a maximum distance of 2500 base pairs. We then calculated the correlation of common and rare alleles between all of the resulting SNP pairs for each subject in the data. A histogram of the resulting correlation values revealed two distinct clusters. Five samples had very low correlations and are apparent STY-NSP mismatches. The five NSP-STY mismatched samples are excluded from the analysis. The sample ids, NSP CEL file name, associated STY CEL file name, and the correlation between the genotypes for the paired CEL files are in the spreadsheet **NSP STY correlation – Sheet 1**, node ID 617. The histogram of the correlation values for the original data is shown in **Figure 2**.

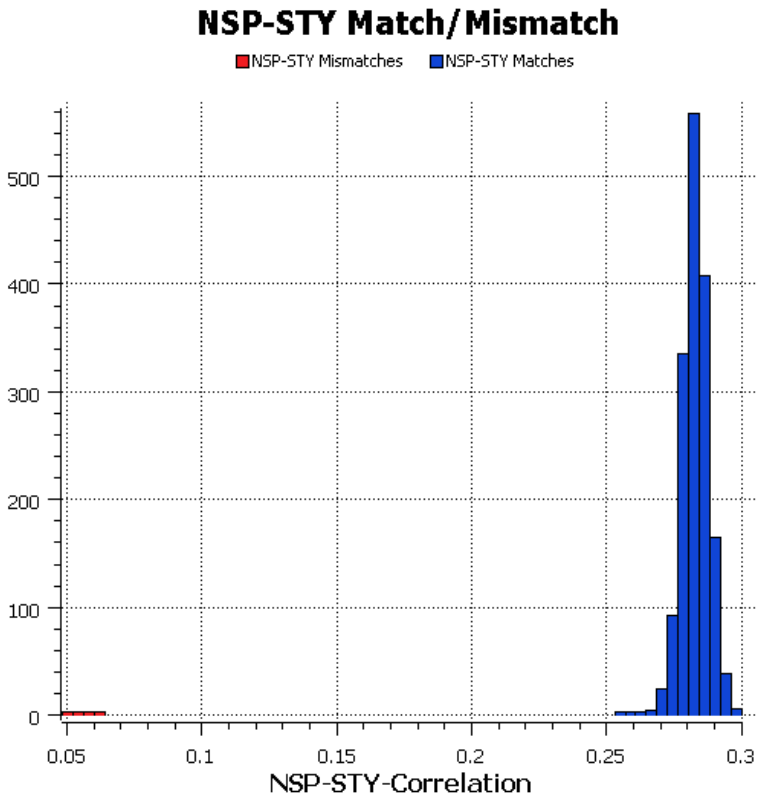

Figure 2: Histogram of NSP-STY-Correlation values

2.2.3 Verify gender using X heterozygosity and X log ratio intensities

Reported gender in the provided phenotype data was verified with two methods, X chromosome heterozygosity and average log ratio of X chromosome intensity.

The first method calculated the heterozygosity rate of the X chromosome, assuming that males should have a heterozygosity rate near zero as they have only one copy of the X chromosome. Similarly, as females have two copies, they should have a higher heterozygosity rate. The male heterozygosity rate departs somewhat from zero due to the inclusion of several hundred pseudoautosomal markers. The heterozygosity was calculated using the NSP and STY markers separately in order to assess data quality from each array. See the spreadsheet **Phenotype + NSP & STY X Heterozygosity – Sheet 1,** node ID 645.

Using this method, 16 samples had mismatches between reported gender and calculated gender based on heterozygosity. For two samples both previously implicated as NSP/STY mismatches, gender was inconsistent for only the NSP markers. The scatter plot of the NSP and STY heterozygosity rates split on reported gender is reproduced in **Figure 3**. Although difficult to see in this figure, there are several reported male samples in green within the blue cluster of imputed females.

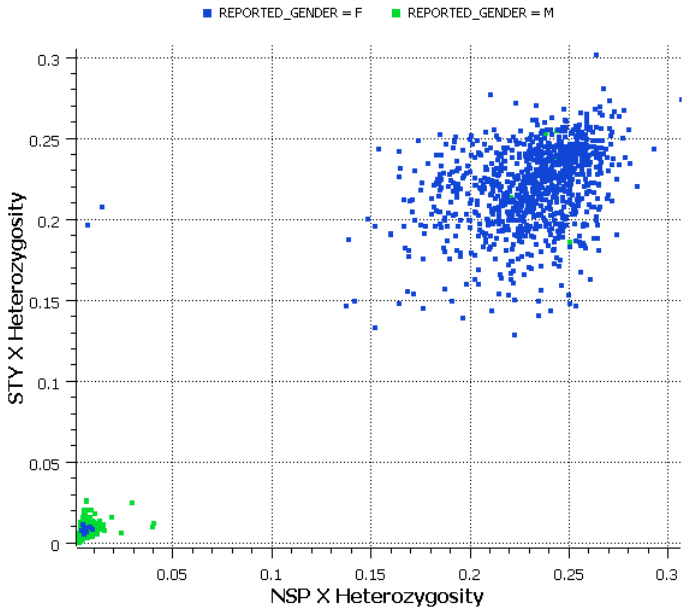

Figure 3: Heterozygosity method for determining gender mismatches

The second method of gender determination required importing the Affymetrix CEL files into SVS, quantile normalizing the data, and calculating the average X chromosome log ratio intensity for each sample. Quantile normalizing centers the intensity data about zero. This places copy number two at zero in autosomes, while copy numbers one and two should be to the left and right of zero, respectively, in the X chromosome as both males and females were used as the reference distribution. The average intensity for females should be higher than for males. The X chromosome average log ratio intensities were calculated using the NSP and STY markers separately. See the spreadsheet **Phenotype + X NSP & STY log ratio averages – Sheet 1**, node 640.

Using this method, mismatches between reported and imputed gender were observed for the same 16 samples identified by the heterozygosity method above. The scatter plot of the NSP and STY average X chromosome log ratio intensities split on reported gender is shown in **Figure 4**. This plot corresponds to node 642, **Plots from Phenotype + X NSP & STY log ratio averages – Sheet 1 against NSP X log ratio average**.

In addition to the samples with gender mismatches found with the Heterozygosity method, there are four additional samples with unusual average intensities for either the NSP or STY markers or both. These appear to be indicative of data quality problems in one of the NSP/STY pairs in three samples and possible XX/XO mosaicism in a fourth. These four samples are excluded from the analysis.

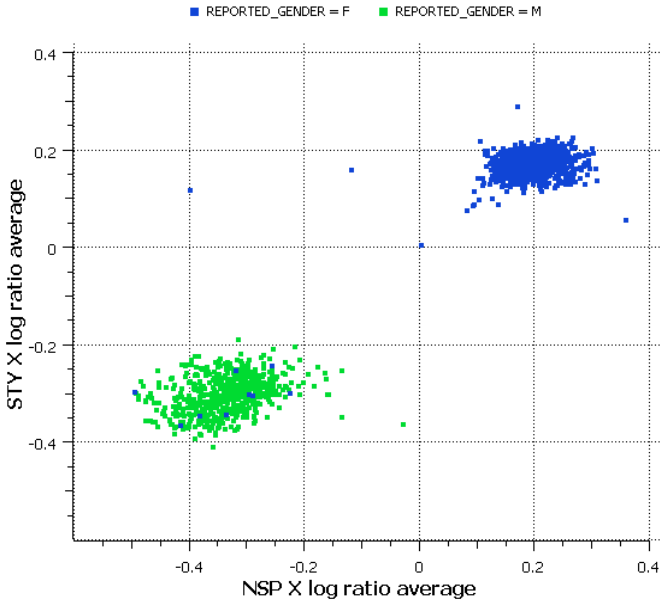

Figure 4: Average X chromosome log ratio method for determining gender mismatches

2.2.4 Identify and exclude duplicate and closely related samples by genotype concordance

NSP and STY autosomal data were separately compared for cryptic relatedness using PLINK[5, 6] version 1.0.6. The PI_HAT statistic calculated by PLINK assesses the degree of genetic identity on a scale from 0 to 1. PI_HAT scores close to 1 correspond with either monozygotic twins or duplicate samples. The spreadsheet **Identity by Descent – Sheet 1**, node 653, contains a row for each sample, with columns that list the most similar samples based on both the NSP and STY arrays, along with the PI_HAT scores and flags indicating probable cryptically related samples. 18 samples with a PI_HAT close to 1 were found in the NSP data and 16 in STY – the difference is due to the NSP/STY mismatches detected earlier. Ideally we would keep the best performing sample of each pair. However, the reported age for each identical pair is different. Because twins cannot be born years apart, the data inconsistency is most likely due to inadvertent repeated measures on the same sample. Because we have two different phenotypes in terms of age (and sometimes case/control status) we removed all 18 samples from the analysis. The maximum PI_HAT values for NSP and STY are plotted in the scatter plot, **Plots from Identity by Descent + Phenotype – Sheet 1 against PI_HAT_NSP**, node 655, which is reproduced in **Figure 5** below. There are a number of samples who appeared to be first and second degree relatives with PI_HAT scores around 0.5 and 0.25 respectively. However, a graph-based analysis revealed these were a few isolated pairs, not involved in any larger family structure and we elected to leave these samples in the analysis.

Two samples appear to be contaminated for their NSP genotypes as they had high Identity by Descent scores with many of the other NSP samples. According to the PI_HAT statistic, one contaminated sample was the closest match for 1289 samples, and the second contaminated sample was the nearest match for 219 other samples. No such trends were observed for the STY genotypes.

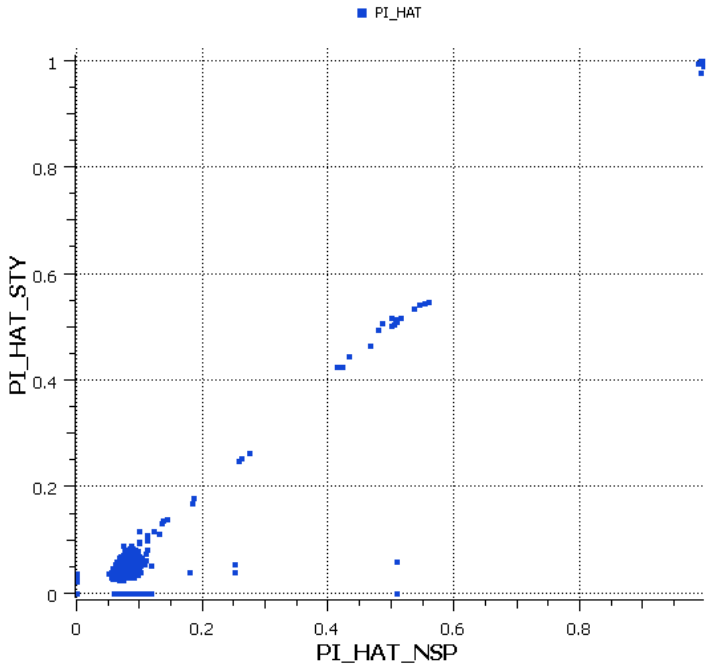

Figure 5: Cryptic Relatedness as revealed by PI_HAT values from NSP and STY. PI_HAT values near 1 indicate a duplicate sample or monozygotic twins.

2.2.5 Identify population structure using PCA/Eigenstrat – exclude samples that depart significantly from target population

Autosomal genotypes from 270 HapMap samples from the Affymetrix 500k combined NSP and STY arrays were used as a reference to verify the reported ethnicity of the samples. The HapMap samples were from four distinct ethnic groups: Caucasian (CEPH), Chinese, Japanese, and Yoruban. Principal components analysis via the Eigenstrat methodology[7] implemented in SVS suggests that sample 6145 has mixed Caucasian and Asian ancestry, but it was not removed from the analysis. All remaining samples from the Alzheimer's study (both cases and controls) were fairly consistent with a white/Caucasian ethnicity, though a few modest outliers exist. See **Figure 6**.

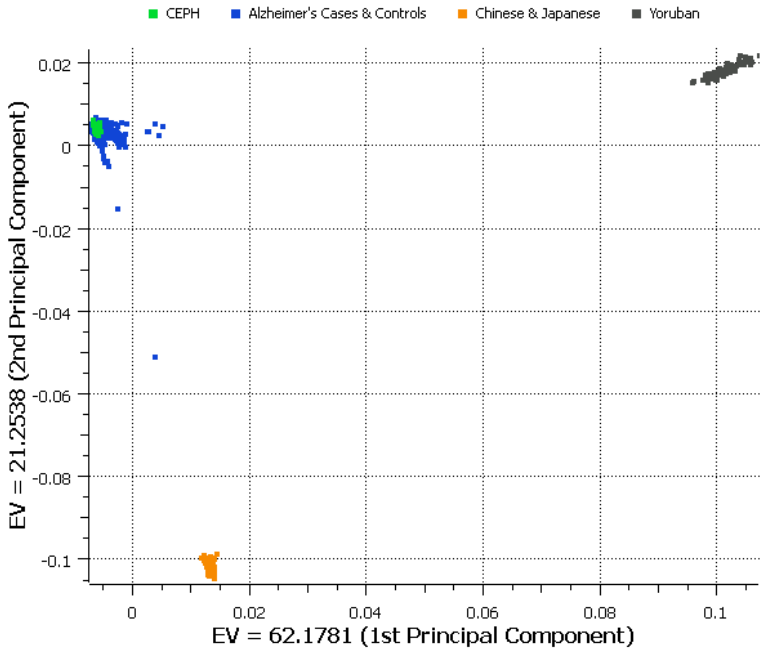

Figure 6: Eigenstrat Population Structure analysis

After all sample quality control procedures, 51 samples of the 1628 samples with an NSP/STY pair were excluded from the analysis. Forty must be dropped due to flagrant problems such as NSP-STY, gender, and phenotype mismatches. Data for these samples has been removed from the public distribution. The other 11 were dropped based on their call rate being below 0.94 – a threshold that could be debated, particularly as the call rate of samples is partly a function of the call confidence threshold (0.95) for importing CRLMM genotypes mentioned earlier.

A Fisher's Exact test of the Alzheimer's case/control status for ***all*** of the markers using only the samples that passed the quality control procedures yielded the log quantile-quantile (QQ) p-value plot in **Figure 7**. This plot (**QQ Plot no SNP quality control**, node 716) shows only two markers with a significant departure from the expected value. Most studies we have observed have significant departures from the line Y=X due to batch effects. The lack of this phenomenon appears to be due to a well-randomized experimental design – batch effects are not systematically correlated with case/control status in this study. Thus a plate-by-plate or batch effect analysis is not warranted.

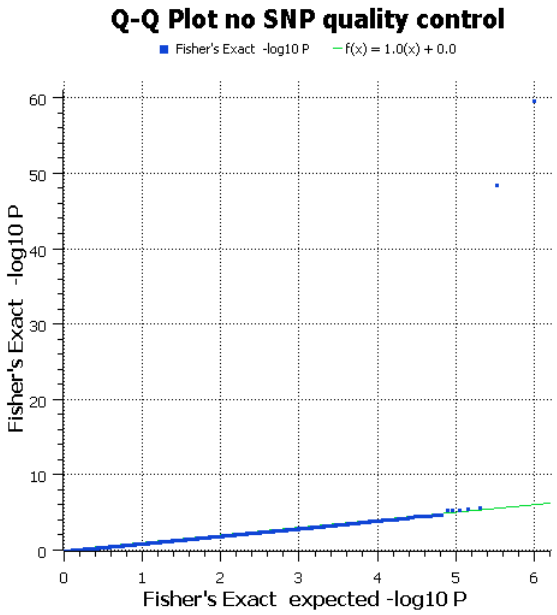

Figure 7: QQ plot of GWAS test results checking for batch effects

2.3 Genotype Quality Assurance

While the above QQ plot indicates that the data follows the expected distribution very closely, it is possible that there are poorly performing SNPs that should be nevertheless be removed from the analysis. The following quality control measures were used to determine which SNPs to exclude from the final analysis.

2.3.1 Exclude SNPs with low call rates, low MAF, and large departures from HWE in controls

Fisher's exact test was used to test Hardy-Weinberg Equilibrium (HWE) for autosomal SNPS in the control samples. SNPs with a p-value less than 5e-07 were excluded. Markers with a minor allele frequency (MAF) of less than 0.05 were excluded if they also had a call rate of less than 0.99. All markers with a call rate of less than 0.95 were excluded. Finally, all markers with unknown location based on the Affymetrix na28 marker map were dropped from the analysis. Out of the original 500,568 markers, 74,539 markers were dropped. The marker statistics used for this procedure are contained in the **Marker Statistics** spreadsheet (node 723). All retained markers are in the spreadsheet **Markers passing QC** (node 744).

2.3.2 Augment data with 3 additional SNPs

One 500k marker, SNP_A-2236481, was retained in the analysis despite not having a valid Affymetrix map position. The marker passed all other quality assurance measures and further investigation revealed that it is located near APOC1, a gene previously implicated in Alzheimer's disease. It appears to be unmapped because the oligo probe overlaps two consecutive SNPs, one base pair apart at positions chr19:50114785-50114786. Two markers not found in the Affymetrix 500k array (RS429358 and RS7412) were genotyped individually and added to the analysis. These markers are of special interest due to their position in the APOE gene, which is associated with Alzheimer's disease.

3 Genome-wide association testing

We performed a Fisher's Exact test for association using a genotypic model on all markers that passed quality control. The Q-Q plot for these associations versus their expected values is in **Figure 8**, which is substantially the same as **Figure 7** which excludes no SNPs.

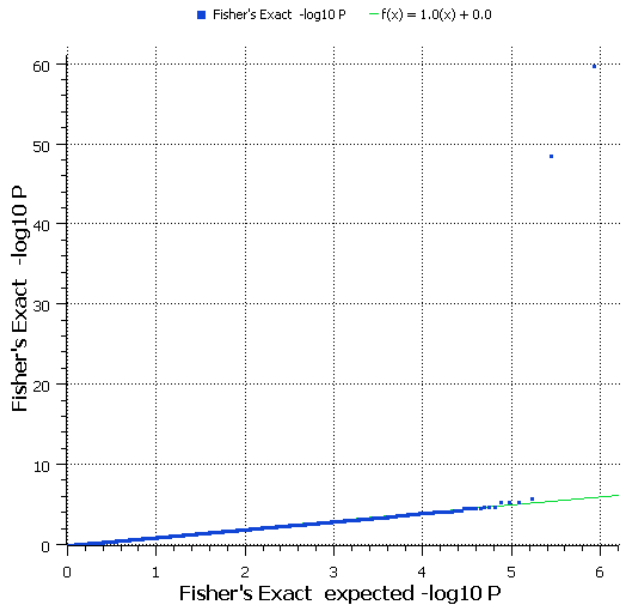

Figure 8: Log quantile-quantile (QQ) plot after SNP quality control

The most significant findings are RS429358 in the APOE gene and SNP_A-2236481 near to the APOC1 gene in Chromosome 19. A Manhattan plot of the $-\log_{10}$(p-value) from the Fisher's Exact Test is shown in **Figure 9**, and with a zoomed y-axis in **Figure 10**. Complete results of the association tests can be seen in the **Association Tests (Genotypic Tests)** spreadsheet, node 759.

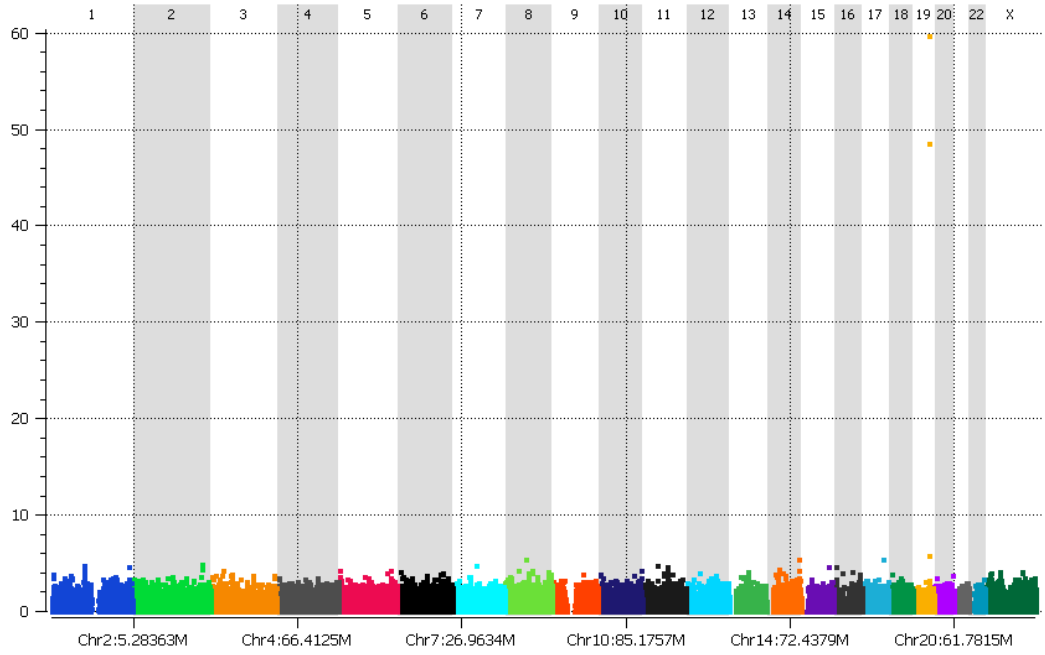

Figure 9: Manhattan Plot of the association test results after augmenting for APOE & APOC1 SNPs

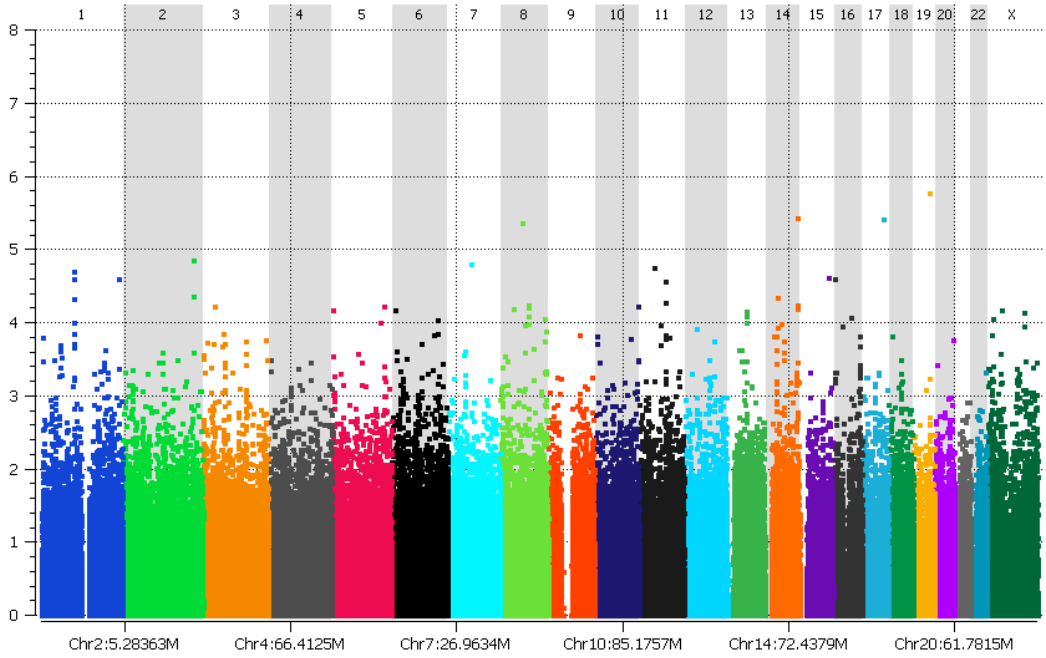

Figure 10: Manhattan Plot scaled to view nominally significant results

Table 1 lists in chromosome and position order the SNPs whose Fisher's Exact P is less than 1e-4. Note that with 410,972 association tests, the alpha level for genome wide significance using Bonferroni correction is 1.2e-7. The two genome-wide significant results are displayed in bold font (RS429358 and SNP_A-2236481). It is interesting to note that in addition to the 19q13.32 APOE region, there are a number of regions that have several localized

nominally significant SNPs in genes that could be of interest for replication studies. These include 1p21.2, 2q35, 8q21.13, 8q24.22, 11q14.1 and 14q32.2.

Marker	Chromosome	Position	Cytoband	RSID	(Nearby) Gene	Fisher's Exact P	Call Rate	MAF
SNP_A-4287403	1	100384159	p21.2	rs12733952	CCDC76	4.78E-05	0.99	0.15
SNP_A-1943051	1	100410296	p21.2	rs11166407	LRRC39	2.01E-05	1.00	0.15
SNP_A-4213932	1	100462710	p21.2	rs4143055	DBT	2.54E-05	1.00	0.15
SNP_A-1935530	1	235959268	q43	rs10925500	RYR2	2.58E-05	1.00	0.25
SNP_A-4272479	2	215269797	q35	rs12615863	(BARD1)	4.46E-05	0.99	0.24
SNP_A-4287818	2	215270178	q35	rs1914516	(BARD1)	1.42E-05	0.99	0.24
SNP_A-4209409	3	35970870	p22.3	rs3996	(ARPP-21)	6.12E-05	1.00	0.37
SNP_A-4271508	5	2501146	p15.33	rs370672	(IRX2)	6.66E-05	1.00	0.30
SNP_A-1984300	5	159837307	q33.3	rs883517	DQ658414	5.94E-05	1.00	0.14
SNP_A-1955167	6	11297073	p24.1	rs11751998	NEDD9	6.75E-05	1.00	0.18
SNP_A-1987847	6	138811968	q23.3	rs4895529	NHSL1	9.24E-05	1.00	0.04
SNP_A-1938267	7	70357158	q11.22	rs11772787	WBSCR17	1.56E-05	1.00	0.26
SNP_A-2306756	8	41799010	p11.21	rs4466386	ANK1	6.52E-05	0.99	0.22
SNP_A-2212974	8	65706679	q12.3	rs10808738	CYP7B1	4.28E-06	1.00	0.42
SNP_A-2243269	8	84758151	q21.13	rs6473464	(RALYL)	8.19E-05	0.99	0.47
SNP_A-2290589	8	84793776	q21.13	rs4524788	(RALYL)	5.67E-05	1.00	0.42
SNP_A-4228988	8	84867791	q21.13	rs7006609	(RALYL)	6.24E-05	1.00	0.41
SNP_A-1832762	8	134608378	q24.22	rs2978012	ST3GAL1	8.73E-05	1.00	0.38
SNP_A-2303440	8	134609061	q24.22	rs2978015	ST3GAL1	8.81E-05	0.98	0.38
SNP_A-2311589	10	130941167	q26.3	rs541392	(MGMT)	6.05E-05	0.98	0.24
SNP_A-1868369	11	46218598	p11.2	rs11038830	(CREB3L1)	1.81E-05	1.00	0.12
SNP_A-2203443	11	78768875	q14.1	rs489257	ODZ4	5.34E-05	1.00	0.36
SNP_A-2072092	11	78789470	q14.1	rs472186	ODZ4	2.71E-05	1.00	0.36
SNP_A-1829193	13	57974731	q21.2	rs9538078	?	7.05E-05	1.00	0.22
SNP_A-1845715	13	59171299	q21.2	rs7336489	DIAPH3	8.33E-05	1.00	0.17
SNP_A-2298429	14	38380773	q21.1	rs17108400	?	4.63E-05	1.00	0.12
SNP_A-1811851	14	99345451	q32.2	rs4905898	EML1	5.84E-05	0.97	0.45
SNP_A-1903242	14	99346460	q32.2	rs10141863	EML1	3.75E-06	0.97	0.45
SNP_A-1944770	14	99348608	q32.2	rs12891247	EML1	6.53E-05	0.97	0.44
SNP_A-1903094	15	86967765	q26.1	rs2028389	AEN	2.41E-05	1.00	0.13
SNP_A-2228926	16	6523126	p13.2	rs17822719	A2BP1	2.54E-05	1.00	0.38
SNP_A-4218124	16	55881571	q13	rs8044834	PLLP	8.68E-05	1.00	0.41
SNP_A-2214249	17	63073548	q24.2	rs2537828	PITPNC1	3.82E-06	1.00	0.20
RS429358	**19**	**50103781**	**q13.32**	**rs429358**	**APOE**	**1.80E-60**	0.99	0.26
RS7412	19	50103919	q13.32	rs7412	APOE	1.70E-06	1.00	0.06
SNP_A-2236481	**19**	**50114785**	**q13.32**	**?**	**(APOC1)**	**2.72E-49**	1.00	0.29
SNP_A-2047950	X	19096599	p22.13	rs5955711	(GPR64)	9.05E-05	1.00	0.12
SNP_A-2295670	X	43699027	p11.3	rs2238973	NDP	6.88E-05	1.00	0.44
SNP_A-2110751	X	111994780	q23	rs648170	(AMOT)	7.30E-05	1.00	0.22

Table 1: List of most significant SNPs

4 References

[1] Benilton Carvalho, Rafael A. Irizarry, Ben Bolstad with contributions from Vince Carey, Robert Gentleman and Wolfgang Huber. oligo: Oligonucleotide Arrays. R package version 2.8.1

[2] Shin Lin, Benilton Carvalho, David J Cutler, Dan E Arking, Aravinda Chakravarti and Rafael A Irizarry (2008) "Validation and extension of an empirical Bayes method for SNP calling on Affymetrix microarrays", Genome Biol. 2008; 9(4): R63.

[3] BRLMM: an Improved Genotype Calling Method for the GeneChip Human Mapping 500K Array Set. http://www.affymetrix.com/support/technical/whitepapers/brlmm_whitepaper.pdf

[4] Birdseed. http://www.affymetrix.com/products/software/specific/birdseed_algorithm.affx

[5] Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81: 559–575.

[6] Package: PLINK version 1.06 (24-Apr-2009) Author: Shaun Purcell. URL: http://pngu.mgh.harvard.edu/purcell/plink

[7] Price, Alkes L., Patterson, Nick J. Plenge, Robert M. Weinblatt, Michael E. Shadick, Nancy A. Reich, David. (2006). 'Principal Components Analysis Corrects for Stratification in Genome-Wide Association Studies'. Nature Genetics 38, 904-909.