Introduction

This document describes the cleaning procedure for genotype array data. The data was cleaned using the GWASTools[1] framework of objects. 789 samples were genotyped using the Infinium Psychip 1.1 which contains 603,132 probes. 11 of the samples were duplicate HapMap samples. Samples were loaded in two batches. Genotypes were called separately for each batch using Illumina's Gencall[2] algorithm.

Poor Quality Probes

Probes that are not included in the newest version of the Illumina PsychArray manifest (InfiniumPsychArray-24v1-1_A1) were removed (14,504 probes).

Probe and Sample Call Rate

Samples and probes were filtered out due to low call rate. Call rate for probes and samples were calculated using the "missingGenotypeBySnpSex" and "missingGenotypeByScanChrom" functions, accordingly, in the GWASTools package[1]. Probes with less than 97% call rate were removed (26,141 probes), see Table 1 for probes removed at each step. All samples had high call rate so none were removed.

Batch Effects

Two sources of batch effects that were examined: plate effects (11 plates) and batch effects (2 batches). Call rates were consistent across all plates and batches. A linear model between the number of samples per plate and the mean sample call rate of that plate was calculated and was not significant (p-value = 0.15).

Allele frequencies were compared across batches using the function "batchFisherTest" in GWASTools[1] (the fisher test was used instead of the chi squared test because the minor allele frequency for many of the probes was small). The p-values from the fisher exact tests were corrected for multiple testing using the Benjamini and Hochberg method[3]. Probes whose corrected p-value was significant (<0.05) were removed from further consideration (46 probes).

Chromosomal Call Rate

Sample call rates by each chromosome were calculated using the "missingGenotypeByScanChrom" function from GWASTools[1]. All autosomal probes had similar call rates. Probes on sex chromosomes had lower call rate than autosomal probes, however no probes were removed for this reason.

Sex Inference

Mean intensity and heterozygosity for probes on the X and Y chromosomes was calculated for each sample and compared to the self-reported sex to ensure sample quality and consistency, see Figure 1.

Relatedness Analysis

Samples were examined for interrelatedness by principal component analysis. First, the remaining probes were LD pruned using the snpgdsLDpruning function from the SNPRelate[4] package with the parameters: autosome.only=TRUE, maf=0.05, method="corr", slide.max.bp=1e7, and ld.threshold=sqrt(0.1). The resulting pruned probe set was used in the KING-robust[5] method (implemented in SNPRelate[4] with the "snpgdsIBDKING" function) to obtain an initial estimate of a kinship matrix between all samples. This kinship matrix was used to initialize an iterative process of refinement with the PC-AiR[5] and PC-Relate[6] methods from the GENESIS[7] package in order to obtain convergent matrices of genetic principal components and sample relatedness. The predicted relatedness was compared to the annotated relatedness, and any samples which differed were flagged for possible removal in further statistical analyses (Figure 2). This method was tested for accuracy by combining the GWAS samples with 1,207 HapMap genotypes, repeating the method, and plotting the resulting principal components (Figure 3) to see if the GWAS samples clustered appropriately with the HapMap samples.

Clinical Variable Trends

A two-factor ANOVA test was performed for each principal component calculated in the relatedness analysis in order to find if the mean value differed between self-reported ethnicity and ADHD status. This procedure also informs on how many genetic principal components should be included in order to control for genetic variability during statistical analyses. Additionally, a t-test is performed to test for differences in sample call rate between control ADHD positive patients. Probes with greater than 2% difference in mean call rate between control and ADHD positive patients were removed from further consideration (1,490 probes).

Chromosomal Anomaly Detection and Removal

Chromosomal anomalies were detected using circular binary segmentation (CBS)[8], [9] and subsequent filtering as implemented in the GWASTools[1] package. CBS[8], [9] was first applied to B allele frequency (BAF) with the anomSegmentBAF function with the parameters: alpha=0.01, min.width=2, and the results were subsequently filtered with the function anomFilterBAF and parameters: num.mark.thresh=10. The anomFilterBAF function also treats some samples as "low quality", which was set to any sample which had a median standard deviation of B allele frequency greater than 0.05. CBS[8], [9] was also applied to the log R ratio (LRR) statistics and the results were filtered using the GWASTools[1] package with the function anomDetectLOH with default parameters. The SNPs lying inside the set of anomalies found by examining BAF and LRR were set as missing data. In total 509 anomalies were found using the BAF and 809 anomalies were found using LRR, an example can be found on Figure 4.

Duplicate Sample Discordance

Eleven duplicate HapMap samples were analyzed with the clinical samples, with one HapMap sample on each plate. Genotype calls among the HapMap samples were compared, and any probe which had any inconsistency was removed (30 probes) from further consideration.

Hardy-Weinberg Equilibrium

Hardy-Weinberg equilibrium (HWE) p-values were calculated for all probes using only the self-reported White/Middle Eastern samples (the largest self-reported ethnic group), see Figure 5. 17 probes with HWE p-values of less than 1e-06 were removed from further consideration.

Ethnicity Subset Calculation

The samples were subset into White and Middle Eastern ethnicity for certain analyses by calculating the centroid of the self-reported White/Middle Eastern samples of the first two principal components from the relatedness analysis. The RMSE was calculated between the PC coordinates of self-reported White and Middle Eastern samples and the centroid. Any sample whose distance to the centroid was less than the RMSE was annotated as White and Middle Eastern.

Genetic Association Study

A preliminary genetic association study was performed for each probe independently by applying a logistic regression model with ADHD status as the explanatory variable, with sex and the first four principal components from PC-AiR as covariates. A QQ plot of the resulting Wald Test p-values is shown in Figure 6.

Imputation

Imputation of nongenotyped SNPs was performed with IMPUTE2[10] using 1000 genomes (1KG phase 3) reference panel. SHAPEIT[11] was used to phase and preprocess autosomal chromosomes. Missing and mismatched SNPs were removed (115,543). Imputation was performed on 3Mb chunks with 1Mb buffers on both sides. The highest probability was used for each imputed SNP, unless the probability was < 0.8, in which it case it was set to missing. The final imputed data set consisted of 16,284,035 SNPs.

Tables

Reason for SNP removal	Number of SNPs removed	Number of SNPs remaining	Percentage of SNPs remaining
Poor quality according to Illumina	14,504	588,628	97.60%
SNP call rate <97%	26,141	562,487	93.26%
Allele frequencies differed between batches	46	562,441	93.25%
Call rate difference between status >0.02	1,490	560,951	93.01%
Discordance between replicate samples	30	560,921	93.01%
HWE p-value < 1e-06	17	560,904	93.00%

Table 1: Number of probes removed at each cleaning step

Number of expected and predicted related pairs				
	Duplicates	Full Siblings	Half Sibs	Degree 3
Expected Related Pairs	55	108	0	0
Inferred Related Pairs	61	102	3	9

Table 2: Number of expected and inferred related pairs as determined by PC-AiR and PC-Relate

Figures

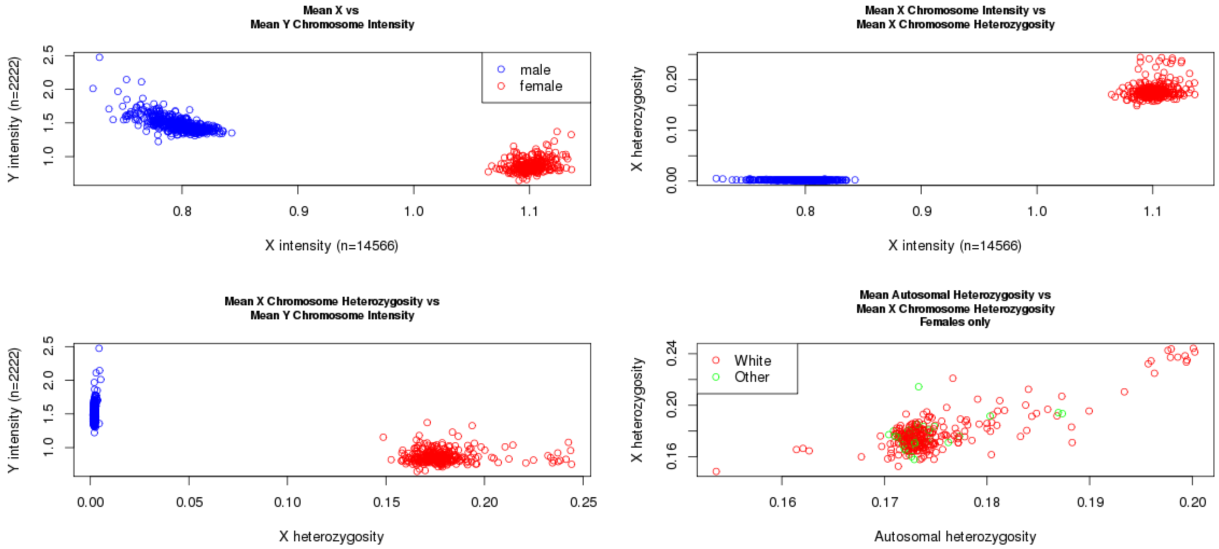

Figure 1: Sex inference using X and Y intensity and heterozygosity

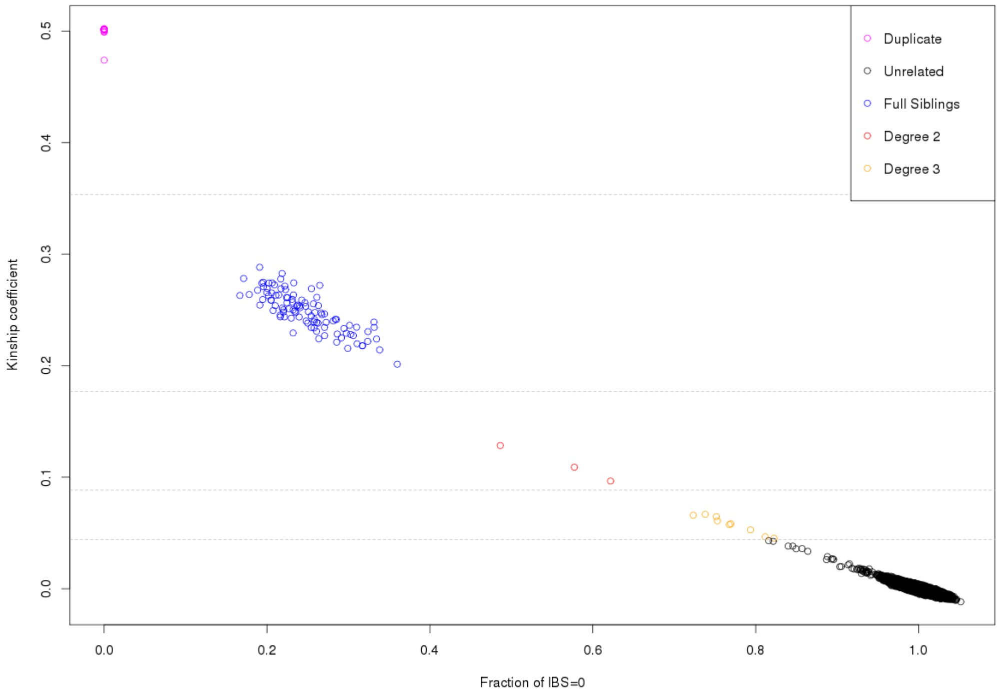

Figure 2: Predicted relatedness among for all possible sample pairings

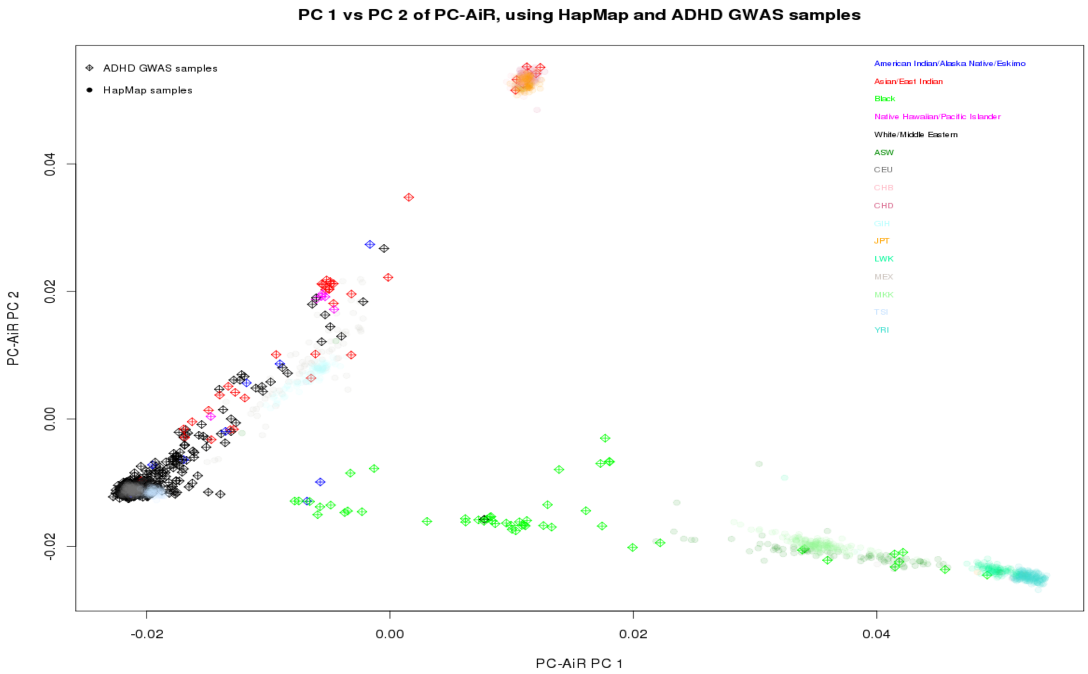

Figure 3: Principal components of combined HapMap and GWAS subjects

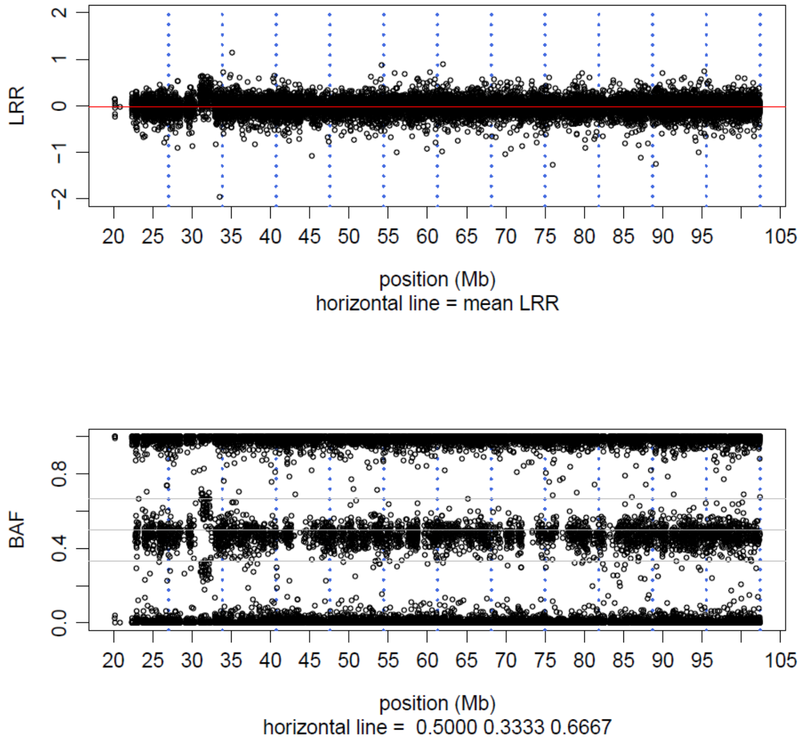

Figure 4: Example of a duplication anomaly

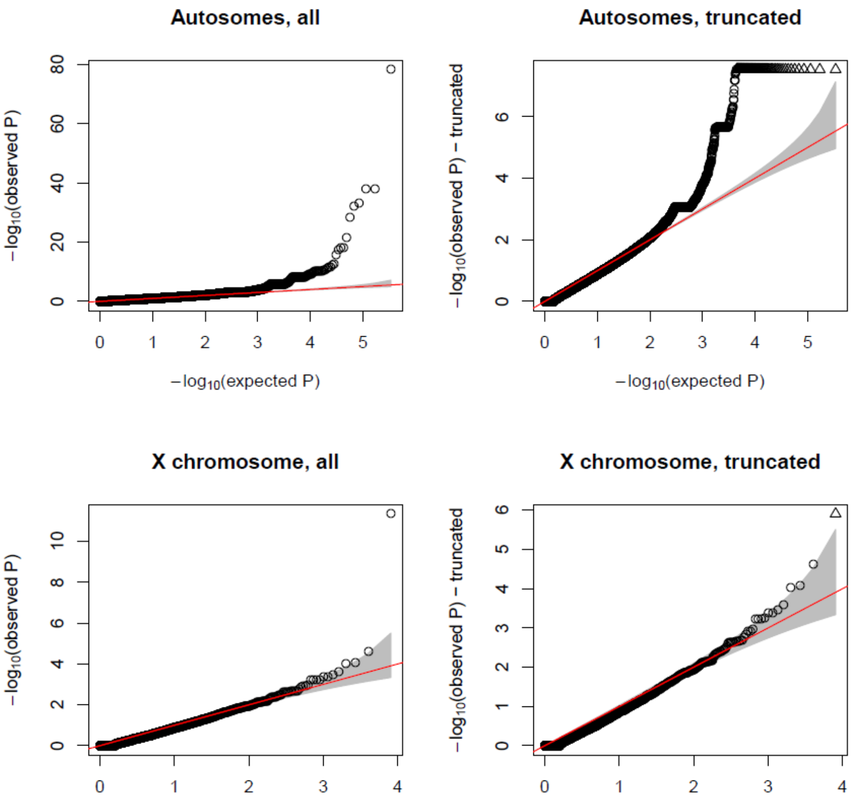

Figure 5: QQ plot of HWE p-values separately for autosomes and X chromosome

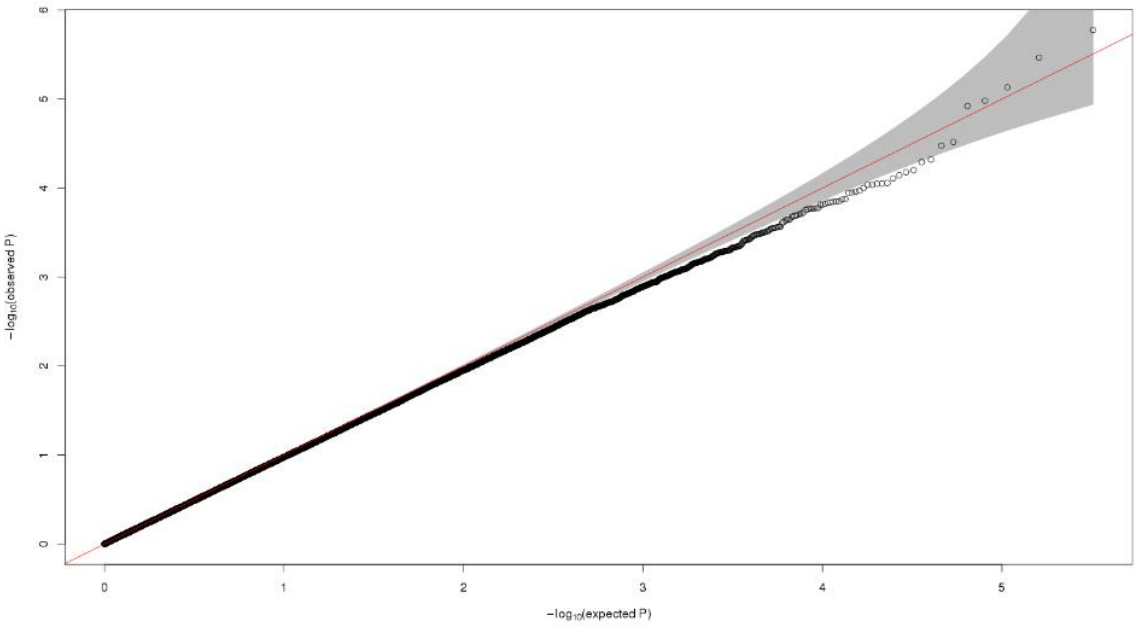

Figure 6: QQplot of Wald test p-values, with sex and first 4 genomic PC's as covariates

References

[1] S. M. Gogarten *et al.*, “GWASTools: an R/Bioconductor package for quality control and analysis of genome-wide association studies,” *Bioinformatics*, vol. 28, no. 24, pp. 3329–3331, Dec. 2012.

[2] B. G. Kermani, “Artificial intelligence and global normalization methods for genotyping,” US7467117B2, 16-Dec-2008.

[3] Y. Benjamini and Y. Hochberg, “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing,” *J. R. Stat. Soc. Ser. B Methodol.*, vol. 57, no. 1, pp. 289–300, 1995.

[4] X. Zheng, D. Levine, J. Shen, S. M. Gogarten, C. Laurie, and B. S. Weir, “A high-performance computing toolset for relatedness and principal component analysis of SNP data,” *Bioinforma. Oxf. Engl.*, vol. 28, no. 24, pp. 3326–3328, Dec. 2012.

[5] A. Manichaikul, J. C. Mychaleckyj, S. S. Rich, K. Daly, M. Sale, and W.-M. Chen, “Robust relationship inference in genome-wide association studies,” *Bioinformatics*, vol. 26, no. 22, pp. 2867–2873, Nov. 2010.

[6] M. P. Conomos, A. P. Reiner, B. S. Weir, and T. A. Thornton, “Model-free Estimation of Recent Genetic Relatedness,” *Am. J. Hum. Genet.*, vol. 98, no. 1, pp. 127–148, Jan. 2016.

[7] M. P. Conomos *et al.*, *GENetic EStimation and Inference in Structured samples (GENESIS): Statistical methods for analyzing genetic data from samples with population structure and/or relatedness: UW-GAC/GENESIS*. GAC, 2018.

[8] A. B. Olshen, E. S. Venkatraman, R. Lucito, and M. Wigler, “Circular binary segmentation for the analysis of array-based DNA copy number data,” *Biostat. Oxf. Engl.*, vol. 5, no. 4, pp. 557–572, Oct. 2004.

[9] E. S. Venkatraman and A. B. Olshen, “A faster circular binary segmentation algorithm for the analysis of array CGH data,” *Bioinforma. Oxf. Engl.*, vol. 23, no. 6, pp. 657–663, Mar. 2007.

[10] B. N. Howie, P. Donnelly, and J. Marchini, “A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies,” *PLOS Genet.*, vol. 5, no. 6, p. e1000529, Jun. 2009.

[11] O. Delaneau, J.-F. Zagury, and J. Marchini, “Improved whole-chromosome phasing for disease and population genetic studies,” *Nat. Methods*, vol. 10, no. 1, pp. 5–6, Jan. 2013.