JCZZ8_MSc Project

by Ava Halvai

Submission date: 14-Sep-2020 04:02AM (UTC-0700)
Submission ID: 133285927
File name: 100811_Ava_Halvai_JCZZ8_MSc_Project_2011071_1192091744.pdf (1.29M)
Word count: 20714
Character count: 100701

GWAS and heritability analysis in intellectual disability with comorbid psychiatric disorder

Genetics of Human Disease MSc

Professor Andrew McQuillin

UCL Molecular Psychiatry

Candidate Number: JCZZ8

Abstract

Whilst genetic liability is shared amongst psychiatric conditions, the specific causative variants for different disorders remain largely unknown. Genetic variation associated with common psychiatric disorders has been investigated. These associations are with a range of genetic causes such as copy number variants, aneuploidies and single nucleotide polymorphisms (SNPs). Previous GWAS have identified SNPs associated with psychiatric conditions such as autism spectrum disorder (ASD), schizophrenia and major depression disorder (MDD), but less work has been carried out in the context of intellectual disability (ID).

There is a significant increased prevalence of psychiatric disorders amongst adults with ID compared to the general population, with estimates of an increased risk of up to 10-fold. Polygenic risk score (PRS) analysis was undertaken on a small sample of 242 individuals using genetic data from patients with ID and comorbid psychiatric conditions. Analysis was undertaken for attention deficit hyperactive disorder (ADHD), ASD, anxiety disorders, MDD, schizophrenia and bipolar disorder. PRS analysis was carried out on PRSice2 using GWAS results of individuals with European ancestry, followed by a second set of PRS analysis carried out using a dataset of the samples with ID merged with healthy control samples.

The majority of PRS generated for the ID dataset were not predictive and others showed SNPS to account for a small proportion of the variance of certain conditions. Four PRS results in the schizophrenia analysis (for the merged dataset) survived multiple testing correction, with a highest variance of 4.46% at a p-value threshold of 1. Whilst there is potential for clinical use of PRS for assessing individuals for psychiatric conditions, further studies are required with larger and more diverse samples in order for it to be considered valuable clinically.

1.0 Table of Contents

1.1 Table of figures

1.2 Table of tables

1.3 Abbreviations

ADHD	Attention deficit hyperactive disorder
ASD	Autism spectrum disorders
CACNA1E	Calcium Voltage-Gated Channel Subunit Alpha1 E
CEU	Utah residents with Northern and Western European ancestry from the CEPH collection
CHB	Han Chinese in Beijing, China
CHR	Chromosome
cM	centimorgan
CMA	Chromosomal microarray analysis
CNV	Copy number variant
DSM	Statistical Manual of Mental Disorders
DRD4	Dopamine receptor D4 gene
ERBB4	Erb-B2 Receptor Tyrosine Kinase 4
FDR	False discovery rate
GOSH	Great Ormond Street Hospital
GRCh37	Genome Reference Consortium Human genome build 37
GWAS	Genome wide association study
HWE	Hardy-Weinberg Equilibrium
ICD-10	International Statistical Classification of Diseases and Related Health Problems 10th Revision
ID	Intellectual disability
INFO	Information (relating to the information score)
iPSYCH	Integrative Psychiatric Research
JPT	Japanese in Tokyo, Japan
LD	Linkage disequilibrium
MAF	Minor allelic frequency
MDD	Major depressive disorder

NHS	National Health Service
PAS-ADD	Psychiatric Assessment Schedule for Adults with Developmental Disabilities
PC	Principal component
PCA	Principal component analysis
PGC	Psychiatric Genomics Consortium
PCLO	Piccolo gene
PRS	Polygenic risk score
SNP	Single nucleotide polymorphism
THBS2	Thrombospondin-2 gene
WHO	World Health Organisation
YRI	Yoruba in Ibadan, Nigeria

2.0 Introduction

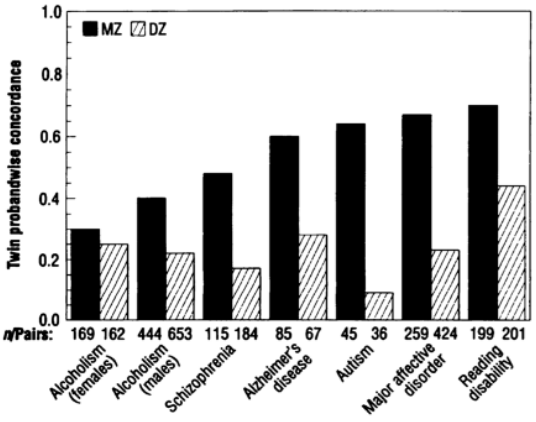
2.1 Intellectual Disability

ID is defined as significant restrictions in intellectual functioning, including, learning, problem solving, reasoning and adaptive behaviour which covers a range of everyday social and practical skills. This disability typically arises before the age of 18. ID is clinically considered a heterogeneous disorder with a range of genetic and environmental risk factors. Genetic causes previously identified include copy number variants, aneuploidies and copy number variants in specific genes. However, it has been estimated that in around 50% of individuals with ID there is no known cause(1). Adults with ID have a much higher prevalence of psychiatric disorders, with estimates of up to 10-fold increased risk in comparison to the general population(2). Investigation of ID and its causes mainly occurs at onset when the patient is under the age of 18, however, there is no formally established system of diagnostic review. In the past few years genetic testing has been increasingly used with ID patients, carried out either by ID psychiatrists or other clinicians such as geneticists and neurologists.

2.2 Psychiatric disorders

Many quantitative genetic tests and studies have been conducted on mental disorders to assess the heritability across these different conditions in comparison to other non-psychiatric traits. Figure 1.1.1 and 1.1.2 summarises the mean concordances of different conditions, determined through many past monozygotic and dizygotic twin studies, both psychiatric (see Figure 1.1.1) and non-psychiatric (see Figure 1.1.2). These findings demonstrate the significant genetic influence on psychiatric disorders and ID in comparison to some other conditions.(3)

Figure 1.1.1: Monozygotic and dizygotic twins average weighted concordances for psychiatric conditions (left). These values were averages taken from multiple past twin studies. Figures taken from R. Plomin (1994)(3).



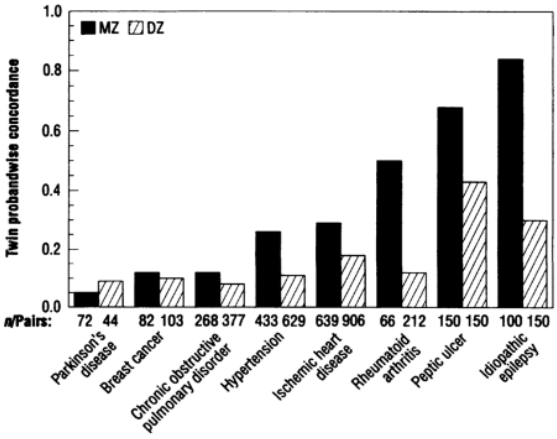


Figure 1.1.2: **Monozygotic and dizygotic twins average weighted concordances for common medical conditions (right).** These values were averages taken from multiple past twin studies and a significant difference can be seen between the concordances of psychiatric and non-psychiatric conditions, emphasising the genetic liability of psychiatric disorders(3).

2.2.1 ADHD

ADHD is a complex psychiatric disorder that develops in childhood. It is described by the World Health Organisation (WHO) in the International Statistical Classification of Diseases and Related Health Problems 10th Revision (ICD-10) as a 'hyperkinetic disorder' with traits such as hyperactivity, impulsivity and a short attention span(4). These characteristics can have a significant impact on the individual's daily life, both at work and at home. The Diagnostic and Statistical Manual of Mental Disorders (DSM) requires at least six specific symptoms of these traits in addition to evidence of the condition negatively affecting daily life in at least two different settings for the classification of ADHD(5).

ADHD has a strong genetic basis, with genetic factors accounting for around 80% of the causes of cases(6) and with over 30 twin studies supporting this estimate(7). Association studies have been carried out in an attempt to identify these genetic risk factors specifically. For example, association between ADHD and the dopamine receptor D4 (*DRD4*) gene were first reported in children and have since been replicated in several studies on adolescents and adults(8)(9). It is a seven-repeat gene found to be overexpressed in those with ADHD and is associated with traits such as seeking behaviour(7). L. Tovo Rodrigues et al. (2013) carried out an association study on *DRD4* rare variants and ADHD, where they identified 7R variants of DRD4 with mutations in the VNTR region which were associated with high hyperactivity and inattention scores(10).

Genome wide association studies (GWAS) for ADHD have had mixed successes, likely due to the genetic overlap between ADHD and other psychiatric conditions such as MDD, ASD, schizophrenia and bipolar disorder(11), as reported previously in family studies by Ghirardi et al. (2018), H. Larsson et al. (2013) and S. Faraone et al. (1997)(12)(13)(14). The first identified genome-wide significant risk loci for ADHD were identified by D. Demontis et al. (2018) through a genome-wide association meta-analysis. The analysis was carried using 20,183 participants diagnosed cases and 35,191 controls, with 304 SNPs on 12 loci on 11 different chromosomes marked for significance for ADHD(15) However, it has been debated whether these early association studies for ADHD were robust enough and reliably replicated.

2.2.2 Autism Spectrum Disorder

ASD is a complex developmental condition characterised by ongoing challenges in speech, communication and social interaction. It is classified in ICD-10 by the following traits: abnormalities in reciprocal social interaction and communication, repetitive behavioural patterns, an over dependency on routine and distress over small changes(4). It is a spectrum disorder and therefore individuals can fall under ASD but express different traits and varying severities. Recent studies of ASD have reported an estimated 50% of the risk of ASD attributed to genetic factors(16), with studies of monozygotic twins reporting a concordance rate of up to 70% as reported in the ASD genetic heritability study carried out by J. Hallmayer (2011)(17).

GWAS have also been carried out to identify common genetic risk variants for ASD. J. Grove et al. (2019) undertook a meta-analysis in GWAS and identified five genome-wide significant risk loci for ASD, in addition to seven additional loci which were shared with other disorders such as MDD and schizophrenia, across 9 chromosomes(18).

Known SNPs associated with ASD were further studied by E. Skafidas et al. (2012) who mapped these SNPs to gene pathways and generated predictive models. This was carried out with the aim to develop a diagnostic test where groups of SNPs, involved in cellular pathways known to ASD, were successfully identified. This process was automated and provided a successful model for a predictor for early ASD diagnosis(19).

2.2.3 Anxiety Disorders

Anxiety is a psychiatric disorder characterised by significant feelings of fear and anxiety. It can have a profound negative impact on the individual's mental and physical health and can disturb both daily life at work and home. In ICD-10 it is classified as 'other anxiety disorders' and encompasses panic disorders, generalised anxiety disorder and other mixed anxiety disorders. It is defined by reoccurring panic attacks, that are not in response to a danger or imminent threat, in addition to physical symptoms such as sweating, numbness, shortness of breath and feeling dizzy.

A meta-analysis by J Hettema et al. (2001) analysed family and twin studies and calculated a genetic heritability of 31.6%. Association studies have also been carried out for anxiety disorders, primarily GWAS and gene candidate studies. GWAS include that carried out by EC. Dunn et al (2016) where the rs78602344 SNP on chromosome 6 in the thrombospondin-2 gene (*THBS2*) was identified as the most significant hit(20) for individuals with Hispanic and Latin American ancestry. However, these results were not replicated in the following meta-analysis the research team carried out. In another meta-analysis, where multiple GWAS on individuals with European ancestry were combined, T. Otowa et al. (2016) identified rs1709393 on the chromosomal band 3q12.3 was associated with lifelong anxiety disorders such as generalised anxiety disorder and panic disorder.(21)

2.2.4 Major Depressive Disorder

MDD is a psychiatric disorder marked by ongoing low moods, lack of energy and interest in usually enjoyable activities and low self-esteem. It is a serious condition which can negatively impact every aspect of life including sleeping, eating and working and it is associated with high levels of morbidity and mortality(22). It is characterised in ICD-10 by a number of symptoms, including a sustained depressed mood for at least two weeks uninfluenced by external influences, fatigue, loss of confidence and/or excessive guilt and suicidal thoughts(4).

It is thought that multiple genetic risk factors, combined with environmental factors, cause MDD and that it has a complex genetic heterogeneity. A heritability of 40-50% has been reported from twin studies(23) and familial studies have demonstrated up to a three-fold risk of developing MDD amongst first-degree relatives(24).

Several GWAS have been conducted for MDD and whilst most have not generated significant genome-wide results, a few significant SNPs have been identified. For example, PF Sullivan (2009), identified 11 SNPs, within a 167-kb region, associated with MDD overlapping the piccolo gene (*PCLO*). PLCO expresses a presynaptic protein which plays a role in monoamine neurotransmitters in the brain(25). P. Muglia et al. (2010) conducted a meta-analysis combining several MDD GWAS and identified the SNP rs4238010, near the cyclin D2 gene. However, similarly they did not achieve genome-wide significance. This is a common occurrence in many of the GWAS and meta-analyses that have been carried out for MDD.

More recently, a GWAS meta-analysis of 135,438 MDD subjects and 344,901 controls on MDD identified 44 risk loci and 153 significant genes, carried out by N. Wray et al. (2018). The PCLO gene was again identified as having a potential role in MDD amongst others such as Calcium Voltage-Gated Channel Subunit Alpha1 E (CACNA1E) associated with calcium signalling. Furthermore, six shared risk loci were found between known risk loci of schizophrenia and several SNPs identified were also significantly associated with schizophrenia and autism(18).

2.2.5 Schizophrenia

Schizophrenia is a severe lifelong psychiatric condition characterised by hallucinations, delusions, thought disorder and social withdrawal. These symptoms heavily impact the individual's quality of life, with up to 50% of diagnosed individuals having reported suicidal ideation at least once throughout their lives(26). Twin studies conducted by R. Hilker et al. (2018) estimated a 79% genetic heritability for schizophrenia, with a third of monozygotic twins both being diagnosed with schizophrenia and only in 7% of the cases of dizygotic twins(27).

There has been some progress with identifying both common and rare significant variants in schizophrenia GWAS and meta-analyses. In the meta-analysis carried out by J. Shi et al. (2010), significant association was observed in a region of linkage disequilibrium on chromosome 6p22.1 which contains a histone gene cluster and several genes involved in immunity. Significant genes were identified too, including Erb-B2 Receptor Tyrosine Kinase 4 (ERBB4), which expresses a receptor of the protein neuregulin-1 which has been linked to schizophrenia in other studies(28)(29). More recently, Ripke et al (2014) carried out an association analysis and identified 128 independent associations spread across 108 loci, with 83 having not been reported before(30).

Rare genetic variants that have been significant for association with schizophrenia include repeated microdeletions and microduplications reported by SE. McCarthy (2009), spanning 600 kb on chromosome 16p11.2. A meta-analysis was also carried out which found that these microduplications and microdeletions were also associated with ASD and the microduplications with bipolar disorder, demonstrating an overlap between these psychiatric disorders(31).

2.2.6 Bipolar Disorder

Bipolar disorder is a psychiatric disorder marked by extremes of mood, consisting of periods of mania and depression and in some cases psychosis. Similarly to other psychiatric disorders, those with bipolar disorder statistically have a lower quality of life and lower productivity, with an increased rate of suicide(32)(33). It is classified in ICD-10 as 'bipolar affective disorder' with the symptoms of hypomania, depressive episodes, psychosis(4). Monozygotic twin studies carried out by A. Bertelson et al. (1977) determined a concordance of 67% for bipolar disorder, up to 20% for dizygotic twins and up to a 20% relative risk in first-degree relatives. These findings suggest that multiple loci are involved in bipolar disorder(34). Linkage analyses have been carried out to identify these, with significant loci for bipolar disorder discovered on chromosome 13q31-33(35), 4p12-13(36) and 18q21-23(37).

More recently, Stahl et al. (2019) conducted a GWAS on bipolar disorder and identified 30 risk loci which were significant genome-wide and of which 20 were novel. These had roles in a range of processes such as neurotransmission and ion channels which suggests how bipolar disorder affects key mechanisms(38).

2.3 Intellectual Disability with Comorbid Psychiatric Disorder

Psychiatric disorders are present at a much higher prevalence in individuals with ID compared to the general population. This is likely due to overlapping genetic risks between ID and various psychiatric disorders. Therefore, studying these disorders in the context of ID could potentially lead to identifying a higher frequency of or novel significant genetic risk factors that have previously not been detectable in a non-ID specific cohort.

Copy number variants (CNVs) have been a known risk factor for disorders such as ASD and schizophrenia. However, these CNVs have been predominantly studied amongst single disorder cohorts. JH. Thygesen et al. (2018) undertook a chromosomal microarray analysis (CMA) in order to establish both the frequency and type of CNVs in adults with both ID and comorbid psychiatric disorders. Pathogenic loci that were known as a disorder risk were found at a much higher rate (13%) than amongst healthy controls (1.2%), supporting the hypothesis that having ID with comorbid psychiatric disorder produces an additive effect. CNVs were

found at 23 of the 63 known neurodevelopmental disorder loci with a total of 58 CNV carriers identified. Interestingly, these carriers expressed a range of severities for both ID and psychiatric disorder, suggesting that genes in these loci play a broader role rather than a risk for specific disorder. A further 25 CNVs were identified in addition to these disorder loci that had been reported as pathogenic by clinical genetic services(1).

CMA carried out by K. Wolfe et al. (2016) also identified an occurrence of pathogenic CVs in the ID sample (13%) at a much higher prevalence than compared to the healthy controls tested for. Five of these CNVs were found at the 16p11.2 locus, which is associated with an increased risk for schizophrenia, MDD and ASD. Furthermore, another five of these CNVS had little or no previously known associations with psychiatric disorders, suggesting the likely identification of rare novel neurodevelopmental CNVs. These consisted of a duplication of 4p16.3, a duplication at Xq24-25, a deletion at 12q21.2-21.31 consisting of 17 gene, a duplication at 13q32.3-13q33.3 consisting of 33 genes and a deletion at 19q13.32 consisting of 56 genes. The findings from both papers demonstrated a potential beneficial clinical use of CMA in CNV screening for adult patients with ID(39).

2.4 Polygenic Risk Score Analysis and Future Clinical Use

Psychiatric disorders are highly polygenic with their genetic risk originating from thousands of genetic variants and often overlapping with other genetic risk factors of other psychiatric disorders. Therefore, polygenic risk score (PRS) analysis is being increasingly employed predict and model risks for psychiatric disorders. PRS is a calculation of an individual's genetic risk for a condition, calculated using their own genotype data and GWAS data for the relevant condition, in addition to any covariates which are taken into account.

AG. Jansen et al. (2019) carried out PRS analysis on clinical samples of children and adolescents with ASD, ADHD and schizophrenia. PRS scores were generated for each disorder using GWAS to weight risk alleles by their effect size. The PRS for ASD and schizophrenia was not significantly associated with ASD or schizophrenia status, however, significant associations were observed for the ADHD PRS with ADHD status and combined ADHD/ASD status. These results align with past findings of genetic overlap between the two disorders, however, the association was primarily driven by ADHD status and so suggests that the ADHD PRS is due to ADHD specific genetic risk factors(40).

S. Lee et al. (2013) carried out PRS analysis on schizophrenia, bipolar disorder, MDD, ASD and ADHD. PRS scores were generated using GWAS data where up to 29% of genetic risks of these psychiatric disorders was found to be explained by these SNPs. Significant associations were found between PRS and the disorder pairs schizophrenia/bipolar disorder, schizophrenia/MDD, bipolar disorder/MDD and ADHD/MDD(41).

Several PRS for psychiatric disorders have had significant associations and studies have been successful in predicting the disease status in both smaller research case-control studies and in wider population based cohort studies(42)(43). These results have proven to be useful in research and give insights into the comorbidity of different psychiatric disorder and different mechanisms involved. In the future, PRS may have the potential to be used in assisting diagnosis or prove beneficial in the field of personalised medicine and shift the way psychiatric disorders are diagnosed and treated clinically. However, high precision of these PRS scores must be ensured in much larger and more diverse, non-European cohorts before they will be considered to have a clinical use.

3.0 Methods

3.1 Participant Recruitment and Criteria

Individuals from England (United Kingdom) were recruited across 32 National Health Service (NHS) trusts and one private clinic by clinical psychiatrists specialising in ID between 2012-2015. Recruitment was based on inpatient and community case-loads of participants with an inclusion criteria of a standard IQ threshold of below 70, age of over 18 and a previous diagnosis of one or more psychiatric disorders or substantial challenging behaviours in addition to comorbid idiopathic ID. Idiopathic ID was classified as ID with no conclusive genetic or environmental cause.

Ethical approval was given for this study and each participant was assessed on their ability to give consent. Written consent was provided by patients who were able to via easy read information sheets. Advice was sought regarding participants who lacked the capacity to give written consent.

3.2 Phenotyping

Detailed behavioural and psychiatric phenotyping was carried out using standardised psychiatric assessment schedules for adults with developmental disabilities (PAS –ADD) and challenging behaviours were completed for all participants. The PAS-ADD is an assessment tool for adults with ID and psychiatric disorders which includes 86 psychiatric symptoms across seven diagnostic areas. It generates a series of scores and provides a threshold score on a range of symptoms which are likely to lead to a psychiatric diagnosis.

This phenotypic data from each participant was classified through ICD-10 and was placed into categories**.** The following categories were included in the analysis: autism spectrum disorders (ASD), anxiety disorders, attention deficit hyperactive disorder (ADHD), bipolar disorder, MDD and schizophrenia **(**see Table 1.2.1**).** These were selected based on their higher prevalence within the sample in comparison to other categories and the availability of GWAS studies carried out on these particular conditions. ASD included autism, autistic traits, atypical

autism and Asperger's syndrome. Anxiety disorders included any anxiety disorders that came under F41 in ICD-10 and MDD included depressive episodes (see Appendix 9.1)

Table 1.2.1: Psychiatric disorder categories used in the analysis, their ICD-10 code and prevalence within the ID samples.

Phenotype	ICD-10	Prevalence (%)
ASD	F84	25
Schizophrenia	F20	28
MDD	F32	14
Anxiety Disorders	F41	10
Bipolar Disorder	F31	8.0
ADHD	F90	7.0

These particular disorders were chosen for their highest prevalence within the sample and specific disorders were grouped together based on their ICD-10 code. These phenotypes were then used in the PRSice2 analysis.

3.3 Intellectual Disability Sample Collection, DNA Preparation, Genotyping

Blood or saliva samples were obtained from each participant and DNA was extracted and quantified at the Great Ormond Street Hospital (GOSH) Regional Genetics Lab. Samples were analysed on the NimbleGen CGX-135K array (Roche) or on the Cytoscan 750K array (Affymetrics). The total cohort consisted of 242 individuals.

Genotyping of the ID sample was carried out on the Illumina PsychArray at Life and Brain GmbH (Bonn, Germany) and the healthy control subjects were genotyped at the Broad Institute (Massachusetts, USA), a customised high density array chip composed of 595,427 markers, 271,000 proven tag SNPs originating from the Infinium Core-24 BeadChip, 277,000 markers obtained from the Infinium Exome-24 BeadChip, in addition to 50,000 other markers known to be associated with various psychiatric conditions, including ADHD, ASD, schizophrenia, bipolar disorder, anxiety disorders and MDD, provided by the PGC.

3.4 Merged Case-Control Sample

Due to the small sample size of the ID dataset, the analysis was also carried out using a merged dataset of both the original sample of ID subjects and healthy control subjects. Healthy control DNA samples were collected by UCL as well as collected from blood donors. Genotyping was carried out using the PsychArray at the Broad Institute (Massachusetts, USA). The dataset contained a total of 1518 subjects, consisting of 234 cases and 1284 controls.

The ID data was provided as raw data and so required QC. However, the control subjects data had already been through imputation and post imputation QC and so could be merged with the ID data without any further QC.

3.5 Dataset File Processing

Merging of the two datasets was performed using scripts run on PLINK (see Appendix 9.2) and written in bash. The imputed PsychArray ID data and the PsychArray control data were merged and only SNPs from the imputed ID data were included. A separate script was used (see Appendix 9.3) to identify and list SNPs with different rates of missingness between the case and control samples and consequently any SNPs with a case-control missingness p-value of less than 0.00001 were removed. It was ensured that SNPs were arranged in chromosomal order before EIGENSTRAT was used to carry out PCA and produce 100 PCs.

A script was run in R (see Appendix 9.4) to create a covariate file containing the covariate values for each individual which could be included in the PRSice2 analysis. The covariate file and the PCA were combined and PRSice2 was run, generating PRS values calculated at various p-value thresholds for the merged sample, with associated plots of the results.

3.6 Quality Control Procedures

Quality control (QC) of the dataset was carried out in PLINK, a genetic toolset developed for analyses related to GWAS using genotype and phenotype data, alongside R. This was carried out to address potential systematic problems in the data, for example, DNA quality issues for specific individuals, assay issues for specific SNPs, batch effects and sample mishandling. The quality control protocol, including the relevant thresholds used, were taken from Johan H. Thygesen and Andries T. Marees et al. (2018). All scripts were run on a server linux using Bourne-Again SHell (Bash) shell. Editing of scripts was performed using nano on the linux server or on downloaded versions of the scripts on a Macbook Air with TextMate.

The dataset used in the analysis was exported from the GenomeStudio software (Illumina). The PsychChip raw data was processed into PLINK text format, .ped and .fam files, which contain both genotypic and familial data. Two scripts were used, the first, ‘pre_qc_stephs.sh’ (see

Appendix 9.5 for script), was written in R. The PED and FAM files were converted into PLINK binary format, BIM, BED and FAM files, which contain compressed forms of the data and therefore allowed for a significantly more efficient analysis.

The script also contained commands to clean the data and format it so it would be compatible with future programs and analyses. The IDs of each participant and their parents were checked and updated in addition to the phenotype status which was updated to set all subjects as cases. Minus strand SNPs were flipped to ensure all SNPs were on the positive strand and multi-matched and SNPs were excluded as their location in the genome could not be confirmed. Blacklisted SNPs, which had been provided by the Broad institute (MIT) upon their analysis of the PsychChip data, were excluded. The SNPs on the "Blacklist" were SNPs that were known to map to multiple regions on the genome or that had produced unreliable genotyping results. Centimorgan (cM) positions listed in the dataset file were overwritten with zeros to allow the program EIGENSTRAT to work. New files were then created with the updated names and positions.

SNPs that had a low minor allelic frequency (MAF) were excluded as they are rare and thus reduce the reliability of their associations with phenotypes. This is particularly significant given the small size of the dataset. Errors in genotyping are also more likely to arise with low MAF SNPs. A threshold of 0.01 was used, with SNPs with MAFs below this value excluded. A threshold was also set for SNP and genome-wide missingness, to reduce both individuals whose DNA samples were of poor quality and decrease the level of missingness that could lead to bias in the analysis. Consequently, SNPs undetected in a large majority of subjects and subjects with large genotype missingness were excluded by a threshold of 0.1 and 0.05 respectively (see Appendix 9.6 for script).

In order to further reduce the chance of including contaminated DNA samples, individuals with both particularly high and low heterozygosity rates were removed, with a threshold of a standard deviation of 3 from the mean heterozygosity rate of the sample. Sex discrepancies were filtered out using homozygosity rates in comparison to participant's records. Those without gender information or sex mismatches were removed using the homozygosity rate estimate of ~1 for males and <0.2 for females. Further samples were excluded if they violated the Hardy-Weinberg equilibrium (HWE), as a noticeable difference in expected and observed genotype frequencies suggests the presence of genotyping errors in the sample. SNPs which had a HWE p-value $< 10^{-6}$ were removed from the sample.

Population stratification was taken into account through principal component analysis (PCA). 17% of the sample was comprised of individuals with non-European ancestry and so the potential varied allele frequencies between the subpopulations within the sample were taken into account to prevent the risk of false positives (see Figure 1.1.5). The EIGENSTRAT method was used to carry out PCA which was based on SNPs with a MAF exceeding 0.05 which were not within known linkage disequilibrium blocks in the genome. This was used to remove SNPs that were outliers in this analysis. A PCA file containing principle components (PC) was generated and was employed to carry out association testing and to model any

ancestral differences in the sample and correct for population structure. A pair-wise identical by descent (IBD) check was carried out to identify any duplicate or related samples which could interfere with the analysis. As the sample contained two nuclear families, this was carried out to prevent overestimates of SNP effect sizes. Duplicate samples were excluded whereas related samples were included but taken into account.

3.6.1 Imputation and INFO Score Filtering

Imputation was carried out on the data to replace any missing genotypes for each subject using statistical models. This was particularly important to increase the power of the analysis given the small size of the dataset. This was carried out on the Sanger Imputation Server (Wellcome Sanger Institute). In order for the imputation to work, the PLINK binary files were converted to VCF format and records were ordered by genomic position (see Appendix 9.7). The dataset was also verified to be compatible with the Genome Reference Consortium Human genome build 37 (GRCh37). The information (INFO) score of a SNP is a measure of what percentage of a dataset of the complete sample size, accurately genotyped, the imputed SNP genotypes are equivalent to. It was calculated to determine how well the SNPs had been imputed and to consequently remove those that had a low INFO score. A threshold of 0.9 was used to keep only SNPs of a frequency <1%. The imputed VCF files were converted to PLINK format using the --recode command on bash to filter the SNPs.

The QC script was run again for a second round to ensure that any erroneous SNPs were removed before the main analysis and checked for any SNP missingness and violations of HWE.

3.7 Polygenic Risk Score Analysis

Polygenic risk score (PRS) analysis was undertaken using PRSice2, a PRS software which generates both PRS and visual models of these results(44). PRSice2 run for each condition included in the analysis and PRS values were calculated at a various number of p-value thresholds using the imputed data. The PRS were generated from the risk alleles and weighted by effect sizes estimated from GWAS results on a base phenotype and the imputed genotype dataset on a target phenotype. The results were also presented visually through PRS models, plotted at a range of p-value thresholds.

PRSice2 requires several input files. The genotype data files (BIM, FAM and BED) from the ID sample, which had undergone quality control procedures and INFO score filtering, were used as the target dataset. A linkage disequilbrium (LD) reference was also included to improve estimates of LD during the clumping process, where SNPs that are in LD with each other are removed. The LD structure in the genotype data from each subject was used. LD clumping was applied with an R2 threshold of 0.1, where R2 is the percentage of the variation in the dependent variable that is accounted for by the regression model (see Appendix 9.8 for script).

A base dataset in the form of a GWAS is required, containing association analysis results for SNPs on the base phenotype. The base dataset file requires information on the effective allele, effect size estimates, SNP ID and the p-values for association. A phenotype file relating to the target dataset is also required in order to generate PRS, in addition to a covariate file containing covariates values for each subject in the sample if necessary.

3.7.1 GWAS Samples and Base Phenotypes

GWAS results for each condition included in the analysis were obtained. The results for ASD were taken from the ASD Integrative Psychiatric Research (iPSYCH) and GWAS, along with samples of European ancestry from the PGC, carried out by J. Grove et al. (2019)(18). The results for ADHD were obtained from the PGC ADHD GWAS carried out by D. Demontis et al. (2018)(15). The MDD data was taken from the 23andMe and PGC GWAS led by N. Wray et al. (2018)(42). The anxiety disorders GWAS was taken from the Ben Neale Lab anxiety GWAS of UK Biobank phenotypes (2018)(20). The bipolar disorder data was taken from the bipolar disorder PGC GWAS conducted by E. Stahl et al. (2019)(38) and the schizophrenia results were obtained from the unpublished PGC-SCZ3 GWAS via personal communication from Professor Andrew McQuillin (UCL).

An INFO score filtering with a threshold of 0.9 was applied to the base dataset to remove SNPs with a frequency below 1%.

3.7.2 Ben Neale Dataset

In PRSice2, the formatting of the columns for chromosome information and the positions of the SNPs are required to match those of the target file. These differed for the target file and the anxiety GWAS base file from the Ben Neale Lab GWAS. The Ben Neale results file originally had a format of “chromosome:position” whereas the target BIM files had a first column of just chromosome position alone (see Table 1.2.2.1-1.2.2.4). A script was made in bash to rectify this and edit the both the base and target file, using the -awk command which searches the files for these differences and edit them accordingly (see Appendix 9.9 and 9.10).

Table 1.2.2.1: The first five lines of the original BIM target file.

CHR	SNP	BPP	A1	A2
1	rs3131972	752721	A	G
1	rs200599638	752918	T	G
1	rs3115860	753405	C	A
1	rs2073813	753541	A	G
1	rs3131969	754182	A	G

Table 1.2.2.2: The first five lines of the reformatted BIM target file

CHR	CHR:BP	BPP	A1	A2
1	1:752721	75721	A	G
1	1:752918	752918	T	G
1	1:753405	753405	C	A
1	1:753541	753541	A	G
1	1:754182	754182	A	G

The target file in its original form and after its columns were reformatted to match the base file on SNP position. CHR=chromosome, BPP=base pair position, A1=minor allele and A2=major allele.

Table 1.2.2.3: The first five lines of the original GWAS base file

Variant	**CHR:BP**	**A1**	**A2**
1:15791:C:T	1:15791	C	T
1:69487:G:A	1:69487	G	A
1:69569:T:C	1:69569	T	C
1:139853:C:T	1:139853	C	T
1:692794:CA:C	1:692794	C	C

Table 1.2.2.4: The first five lines of the reformatted GWAS base file

CHR:BP	A1	A2
1:15791	C	T
1:69487	G	A
1:69569	T	C
1:139853	C	T
1:692794	CA	C

The original and reformatted base file, where CHR:BP=chromosome:base position.

3.8 Multiple Testing Correction

The PRSice2 analysis runs multiple tests in one sitting to form multiple regression models at various thresholds. Therefore, the chance of getting a seemingly significant result simply by chance increases and multiple testing correction is needed to readjust these measures and minimise the expected proportion of false positives. Once the PRSice2 results were generated, false discovery rate (FDR) and Bonferroni corrections were applied in R, with scripts written in bash.

FDR (Benjamini and Hochberg, 1995) makes corrections by controlling the estimated proportion of false positives that are present amongst all the significant results. The Bonferroni corrections is a more stringent correction method which controls the probability of having at least one false positive finding amongst all the significant results (see Appendix .5).

4.0 Results

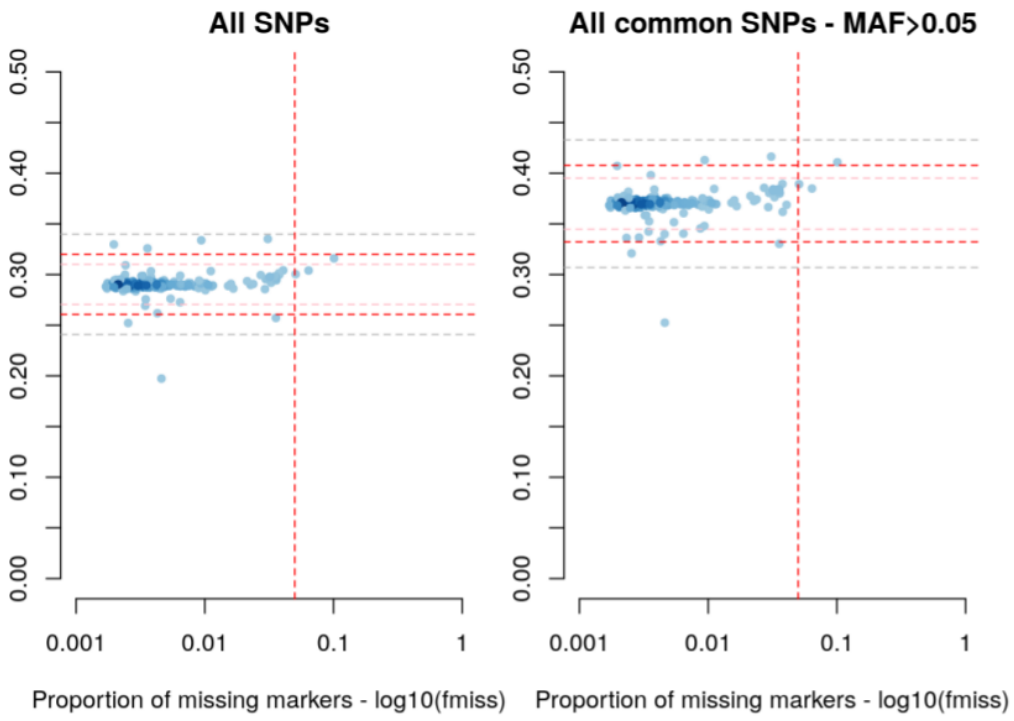
4.1 QC Results

There were originally 242 subjects in the PyschArray ID sample. Heterozygosity was determined using all 515850 SNPs, with a threshold of 3 standard deviations away from the mean measured. This was carried out to as SNPs outside of these thresholds are more likely to generate false positives due to their lower power and a total of six samples were excluded due

to their excessive heterozygosity. Genome-wide missingness was determined using all 241273 SNPs with MAF > 0.05. This was carried out to eliminate potential genotyping errors and three samples were excluded that had more than 5% of the SNPs genotyped with missing information. 32 samples had missing data for more than 1% of the genotyped markers but were included. 4371 SNPs were marked for exclusion as they were detected in the sample at a percentage below the threshold of 0.1 (see Figure 1.1.3).

HWE tests were carried out with a threshold of p-value < 10^{-6} to exclude SNPs that could potentially have been derived from erroneous genotyping or that could be a sign of population stratification. 271 SNPs were marked for exclusion as they failed HWE in all samples. Duplicate samples and tri-allelic markers were identified in the dataset and examined for merge errors, where carriers had non-matching alleles. 6648 SNPs were removed, of which 6628 were duplicates, 20 were triplicates and 20 were triplicates in duplicates.

Figure 1.1.3: Plots of all SNPs and all common SNPs with MAF < 0.05.

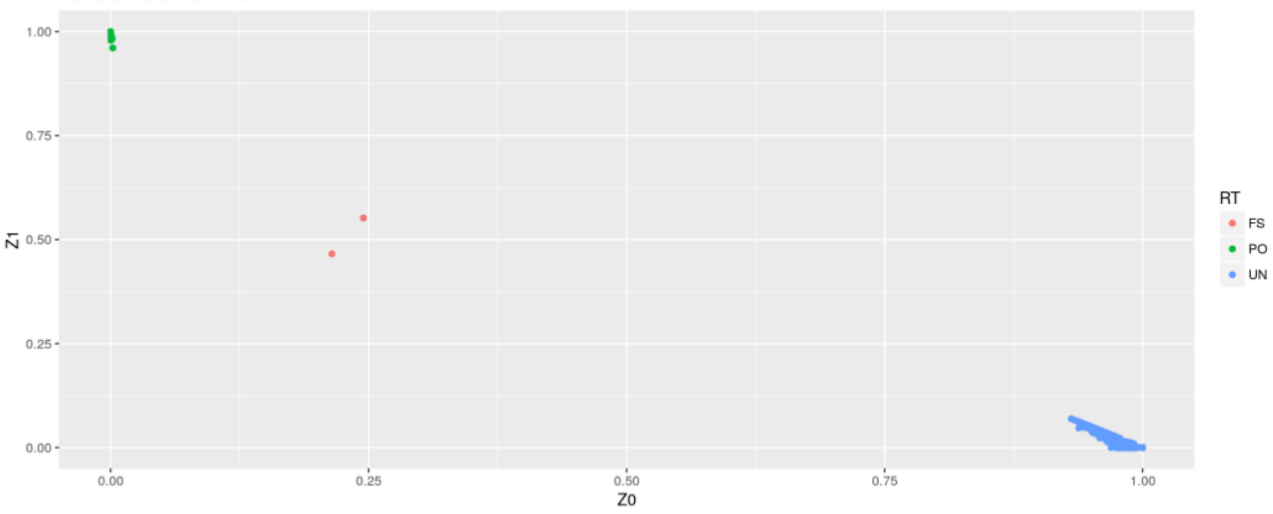


Outlier SNPs in the sample with a high MAF or missingness can be seen beyond the red dotted lines (representing thresholds set), far from the cluster of SNPs on the graph.

Upon checking for sex discrepancies, a single gender issue was identified in the data. Individual LD0248 was recognised as a female with a homozygosity rate of 0.0675. However, the QC recognised a discrepancy and upon looking at the individuals' medical records, they were confirmed to be male. This was due to the individual having a duplication on the X chromosome, which explained the higher X linked heterozygosity than would be expected.

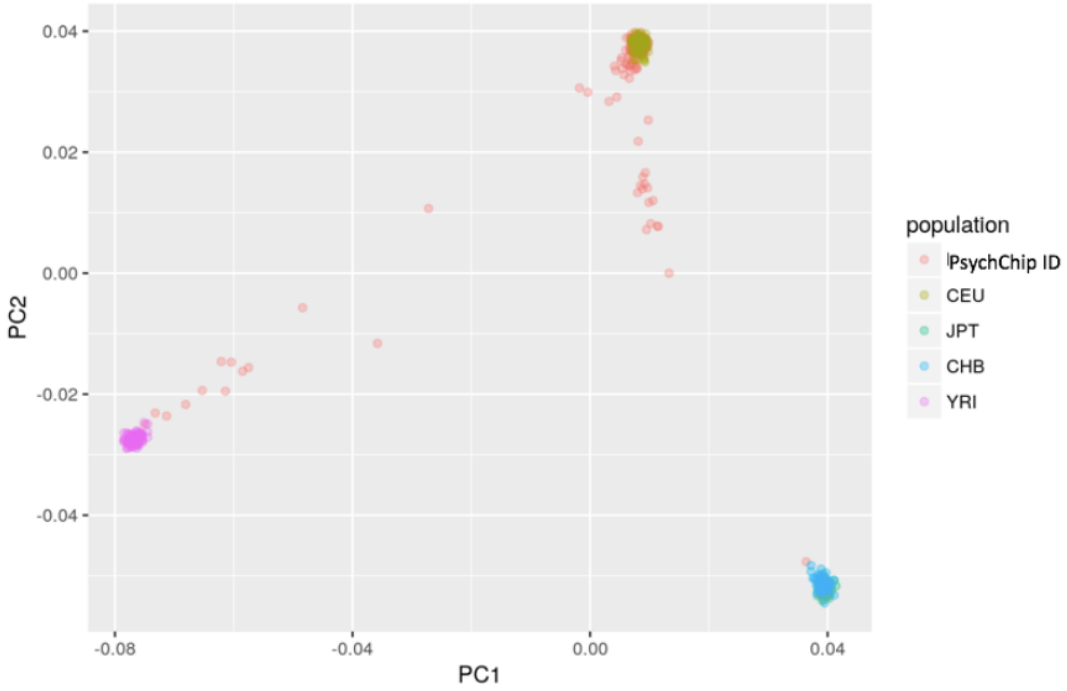
It was known beforehand that the sample contained two nuclear families and the pair-wise IBD check identified 18 samples that showed relatedness consisting of eight parent offspring pairs and two sibling pairs. IBD plots were generated in order to see how these groups were identified, in addition to 1000 random pair-wise plots showing no genetic relatedness for comparison (see Figure 1.1.4). Population stratification was accounted for and samples which were outliers from the PCA analysis were excluded. The top 10 PCs were used for outlier removal and 29 individuals were excluded. Plots were also generated showing the location of the outliers in relation to four HAPMAP populations: Utah residents with Northern and Western European ancestry from the CEPH collection (CEU), Japanese in Tokyo, Japan (JPT), Han Chinese in Beijing, China (CHB) and Yoruba in Ibadan, Nigeria (YRI).

Figure 1.1.4: IBD plots of the ID sample.



Related groups were identified IBD plots were generated. Out of the 18 samples which showed relatedness, eight were parent offspring pairs and two were sibling pairs as shown on the graph, where FS=first siblings, PO=parent offspring and UN=unrelated.

Figure 1.1.5: PCA plot showing clusters of samples based on their similarity.



The plot shows the location of the outlier samples in relation to four HAPMAP populations, where CEU= Utah residents with Northern and Western European ancestry from the CEPH collection (subjects of European ancestry), JPT= Japanese in Tokyo, CHB=Han Chinese in Beijing, YRI= Yoruba in Ibadan, Nigeria (all non-European ancestry) and PsychChip ID is the ID dataset used in the analysis.

A preliminary association analysis was finally run to inspect the allelic cluster plots of the top associated markers. SNPs for which distinct clusters could not be identified and thus had not been clustered to satisfaction were marked for exclusion and consequently 3 SNPs were removed.

4.1.1. INFO Score Filtering

INFO scores were calculated for both the ID cases and controls to determine how well the SNPs had been imputed, set at 0.9. Those with a frequency of less than 1% in the sample were marked for exclusion and 29 samples were removed.

4.2 PRSice Analysis

PRSice2 analysis was carried out on the following conditions: ADHD, ASD, anxiety disorders, MDD, schizophrenia and bipolar disorder, and PRS were generated at various p-value thresholds, in addition to bar plots displaying the model fit of the PRS.

4.2.1 ADHD Analysis

The PRS at each p-value (ranging from 5.00E-08 to 1) did not significantly predict ADHD status in the ID sample, with the observed p-value for each of over 0.05. The highest variance accounted for was only 1.22% by 116 SNPs at a p-value threshold of 0.001 and 3.01% when taking into account the covariates (the PCA). However, the observed p-value for this result was insignificant too at a value of 0.25. The full PRSice2 analysis results for the ID dataset can be seen in Table 1.2.3.1.

The analysis for the merged ID and control dataset also generated no significant PRS (Table 1.2.4.1)

4.2.2 Anxiety Disorders Analysis

The PRS for the ID subjects at each p-value threshold was also insignificant, with the exception of one result at a p-value threshold of 0.01, where 782 SNPs account for 9.36% at an observed p-value of 0.023 (see Table 1.2.3.2)

Similarly, for the merged dataset, the majority of the PRS for the anxiety analysis were insignificant but there was one significant finding for the result of 107 SNPs accounting for 1.23% of the variance at a threshold of 0.001. The observed p-value was 0.031 (Table 1.2.4.2).

4.2.3 ASD Analysis

The PRS for ASD was significant at several p-value thresholds. Four SNPs at a p-value threshold of $1x10^{-6}$ account for 3.80 of the variance, with a p-value of 0.0142199, five SNPs at a threshold of $1x10^{-5}$ account for 2.53% of the variance at a p-value of 0.0445561 and 9997

SNPs at a threshold of 1 account for 2.58% of the variance at an observed p-value of 0.0453 (see Table 1.2.3.3).

None of the merged ID and control data results for the ASD phenotype were significant, with observed p-values all above 0.05. The highest variance accounted for was only 0.66% by 310 SNPs at a p-value threshold of 0.01 and 4.75% upon taking into account the covariates – and this was nonetheless an insignificant result with an observed p-value of 0.070 (see Table 1.2.4.3)

4.2.4 MDD analysis

The PRS at every threshold was insignificant except for at the p-value of 0.05, at which 1307 SNPs account for 1.32% of the variance (p-value=0.02). Taking into account the covariates, they accounted for 4.89% of the variance (see Table 1.2.3.4). None of the PRS for MDD in the merged ID and control sample were significant (see Table 1.2.4.4).

4.2.5 Bipolar Disorder Analysis

The PRS for bipolar disorder for the ID dataset were insignificant at each p-threshold (see Table 1.2.3.5), however, the PRS was significant at several p-value thresholds for the merged dataset, including 457 SNPs account for 2.13% of the variance at a p-value threshold of 0.01 (p-value = 0.034) and 1358 SNPs account for 2.80% of the variance but at a p-value threshold of 0.05 (p-value = 0.539828). The highest variance accounted for was 5.62% by 10,906 SNPs at a p-value threshold of 1 (See Table 1.2.4.5) and 15.53% upon taking into account the PCA. Figure .4A and B demonstrates the PRS for the ID dataset and the merged dataset, respectively.

4.2.6 Schizophrenia Analysis

None of the PRS scores for schizophrenia for the ID dataset were significant (see Table 1.2.3.6). However, the PRS for schizophrenia for the merged dataset had several significant results at various p-value thresholds. 359 SNPs account for 2.20% of the variance at a p-value threshold of 0.001 (p-value = 0.00350837), 868 SNPs account for 3.38% of the variance at a p-value threshold of 0.01 (p-value = 0.539828), 797 SNPs account for 3.05% of the variance at a threshold of 0.005 (p-value = $6x10^{-4}$) and 5629 SNPs account for 4.46% of the variance at a threshold of 1. The highest variance accounted for was 4.46% by 7403 SNPs (see Table 1.2.4.6) at a p-value threshold of 1 and 14.80% upon taking into account the PCA. Figure .5 C and D shows the comparison of PRS for the two datasets.

Table 1.2.3.1: PRS analysis results for ADHD on the ID dataset

Threshold	R2	P	Coefficient	Standard Error	Number of SNPs	FDR	BF
$5.00x10^{-8}$	0.0018	0.65	0.10	0.23	4	0.92	1.00
$1.00x10^{-6}$	0.0015	0.68	0.097	0.24	5	0.92	1.00
$1.00x10^{-5}$	0.0016	0.66	0.097	0.23	9	0.92	1.00
0.00010	$7.02x10^{-6}$	0.98	0.0062	0.22	22	1.00	1.00
0.0010	0.012	0.25	-0.26	0.22	116	0.75	1.00
0.010	$1.83x10^{-5}$	0.96	-0.0099	0.22	515	1.00	1.00
0.050	0.0033	0.54	-0.13	0.21	1584	0.87	1.00
0.50	0.0042	0.49	-0.15	0.21	7198	0.87	1.00
1.00	0.0077	0.35	-0.20	0.21	9843	0.75	1.00

Table 1.2.3.2: PRS analysis results for anxiety on the ID dataset

Threshold	R2	P	Coefficient	Standard Error	Number of SNPs	FDR	BF
$5.00x10^{-8}$	0.00	1.00	0.00	$5.43x10^{-315}$	12	1.00	1.00
$1.00x10^{-6}$	0.00	1.00	0.00	$1.34x10^{-314}$	16	1.00	1.00
$1.00x10^{-5}$	0.0064	0.34	-0.16	0.17	23	0.75	1.00
0.00010	0.063	0.61	-47.45	93.00	46	0.91	1.00
0.0010	0.069	0.30	-28.38	27.31	146	0.75	1.00
0.010	0.094	0.023	-33.54	0.19	782	0.50	1.00
0.050	0.073	0.20	-13.77	10.68	2716	0.75	1.00
0.50	0.065	0.51	-4.54	6.94	20037	0.87	1.00
1.00	0.073	0.23	-7.56	6.30	36296	0.75	1.00

Table 1.2.3.3: PRS analysis results for ASD on the ID dataset

Threshold	R2	P	Coefficient	Standard Error	Number of SNPs	FDR	BF
$1.00x10^{-6}$	0.038	0.014	0.36	0.15	4	0.50	0.065
$1.00x10^{-5}$	0.025	0.045	0.30	0.15	5	0.50	1.00
0.00010	0.0023	0.54	0.088	0.14	9	0.87	1.00
0.0010	0.017	0.094	0.27	0.15	81	0.56	1.00
0.010	0.0086	0.24	0.18	0.16	460	0.75	1.00
0.050	0.011	0.19	0.20	0.15	1524	0.75	1.00
0.50	0.020	0.071	0.27	0.15	7206	0.53	1.00
1.00	0.026	0.043	0.31	0.15	9997	0.50	1.00

Table 1.2.3.4: PRS analysis results for MDD on the ID dataset

Threshold	R2	P	Coefficient	Standard Error	Number of SNPs	FDR	BF
$1.00x10^{-8}$	0.00055	0.76	0.045	0.15	2	1.00	1.00
$1.00x10^{-5}$	0.00078	0.72	-0.053	0.15	3	0.98	1.00
0.00010	$3.64x10^{-5}$	0.94	-0.012	0.15	26	0.87	1.00
0.0010	0.00036	0.81	-0.036	0.15	109	0.75	1.00
0.010	0.014	0.13	-0.28	0.18	536	1.00	1.00
0.050	$3.11x10^{-6}$	0.98	-0.0065	0.29	1748	0.89	1.00
0.50	0.011	0.20	0.64	0.50	8516	0.83	1.00

Table 1.2.3.5: PRS analysis results for schizophrenia disorder on the ID dataset

Threshold	R2	P	Coefficient	Standard Error	Number of SNPs	FDR	BF
$1.00x10^{-7}$	$1.33x10^{-3}$	0.70	0.087	0.22	1	0.97	1.00
$1.00x10^{-6}$	$1.83x10^{-5}$	0.96	-0.010	0.22	6	0.75	1.00
$1.00x10^{-5}$	0.0036	0.52	0.14	0.23	14	0.75	1.00
0.00010	0.015	0.19	0.30	0.23	47	0.92	1.00
0.0010	0.0087	0.32	0.24	0.24	156	1.00	1.00
0.010	0.029	0.072	0.54	0.30	634	0.87	1.00
0.050	0.00070	0.78	0.10	0.36	1765	0.759	1.00
0.50	$3.68x10^{-6}$	0.98	0.0070	0.34	8922	0.75	1.00
1.00	0.038	0.10	-1.43	0.87	13303	0.54	1.00

Table 1.2.3.6: PRS analysis results for bipolar on the ID dataset

Threshold	R2	P-value	Coefficient	Standard Error	Number of SNPs	FDR	BF
$1.00x10^{-8}$	0.0046	0.18	0.20	0.15	32	0.97	1.00
$1.00x10^{-7}$	0.0030	0.28	0.16	0.15	37	1.00	1.00
$1.00x10^{-6}$	0.0031	0.27	0.16	0.14	55	0.56	1.00
$1.00x10^{-5}$	0.0068	0.10	0.24	0.15	89	0.74	1.00
0.00010	0.0084	0.071	0.27	0.15	162	0.84	1.00
0.0010	0.022	0.0035	0.43	0.15	359	0.72	1.00
0.010	0.034	0.00034	0.54	0.15	868	0.46	1.00
0.050	0.031	0.00064	0.51	0.15	1797	0.97	1.00
0.50	0.044	$4.41x10^{-5}$	0.64	0.16	5629	0.47	1.00
1.00	0.045	$4.27x10^{-5}$	0.64	0.16	7403	0.50	1.00

Table 1.2.4.1: PRS analysis results for ADHD on the merged ID and control dataset

Threshold	R2	P-value	Coefficient	Standard Error	Number of SNPs	FDR	BF
$1.00x10^{-8}$	0.00095	0.65	0.097	0.21	3	0.83	1.0
$1.00x10^{-5}$	0.0026	0.45	0.16	0.22	4	0.64	1.0
0.00010	0.0043	0.33	0.20	0.21	15	0.54	1.0
0.0010	0.00019	0.84	-0.044	0.21	74	0.93	1.0
0.010	0.00070	0.69	-0.083	0.21	367	0.86	1.0
0.050	$4.37x10^{-5}$	0.92	-0.021	0.21	1162	1.00	1.0
0.50	$1.85x10^{-4}$	0.84	-0.044	0.21	5792	0.93	1.0
1.00	$6.14x10^{-4}$	0.71	-0.079	0.21	8325	0.86	1.0

Table 1.2.4.2: PRS analysis results for anxiety disorders on the merged ID and control dataset

Threshold	R2	P-value	Coefficient	Standard Error	Number of SNPs	FDR	BF
$5.00x10^{-8}$	-	1	0	$1.77x10^{-314}$	11	1.0	1.0
$1.00x10^{-6}$	-	1	0	$5.56x10^{-315}$	13	1.0	1.0
$1.00x10^{-5}$	-	1	0	$1.35x10^{-314}$	17	1.0	1.0
0.00010	0.0075	0.078	-0.38	0.22	31	0.27	1.0
0.0010	0.012	0.031	-0.56	0.26	107	0.16	1.0
0.010	0.011	0.053	-0.37	0.19	562	0.23	1.0
0.050	0.084	0.094	-0.32	-.19	2077	0.28	1.0
0.50	0.0092	0.18	-0.24	0.18	16146	0.38	1.0
1.00	0.0053	0.18	-0.96	0.72	29247	0.38	1.0

Table 1.2.4.3: PRS analysis results for ASD on the merged ID and control dataset

Threshold	R2	P-value	Coefficient	Standard Error	Number of SNPs	FDR	BF
1.00x10-6	0.0037	0.17	0.16	0.12	2	0.38	1.0
1.00x10-5	0.0010	0.48	0.085	0.12	4	0.66	1.0
0.00010	0.00059	0.59	0.064	0.12	7	0.77	1.0
0.0010	0.0043	0.15	0.18	0.12	53	0.37	1.0
0.010	0.0066	0.070	0.22	0.12	310	0.26	1.0
0.050	0.0015	0.39	0.10	0.12	1144	0.58	1.0
0.50	0.0030	0.22	0.14	0.12	5862	0.42	1.0
1.00	0.0040	0.16	0.17	0.12	8468	0.38	1.0

Table 1.2.4.4: PRS analysis results for MDD on the merged ID and control dataset

Threshold	R2	P-value	Coefficient	Standard Error	Number of SNPs	FDR	BF
$1.00x10^{-5}$	0.0016	0.38	0.11	0.12	2	0.58	1.0
0.0001	0.00083	0.53	0.074	0.12	18	0.71	1.0
0.001	0.0014	0.42	0.099	0.12	77	0.62	1.0
0.01	0.00017	0.77	0.034	0.12	363	0.92	1.0
0.05	0.013	0.020	0.37	0.16	1307	0.11	1.0
0.5	0.012	0.094	0.54	0.32	6905	0.28	1.0
1	0.011	0.099	0.52	0.32	10125	0.28	1.0

Table 1.2.4.5: PRS analysis results for bipolar disorder on the merged ID and control dataset

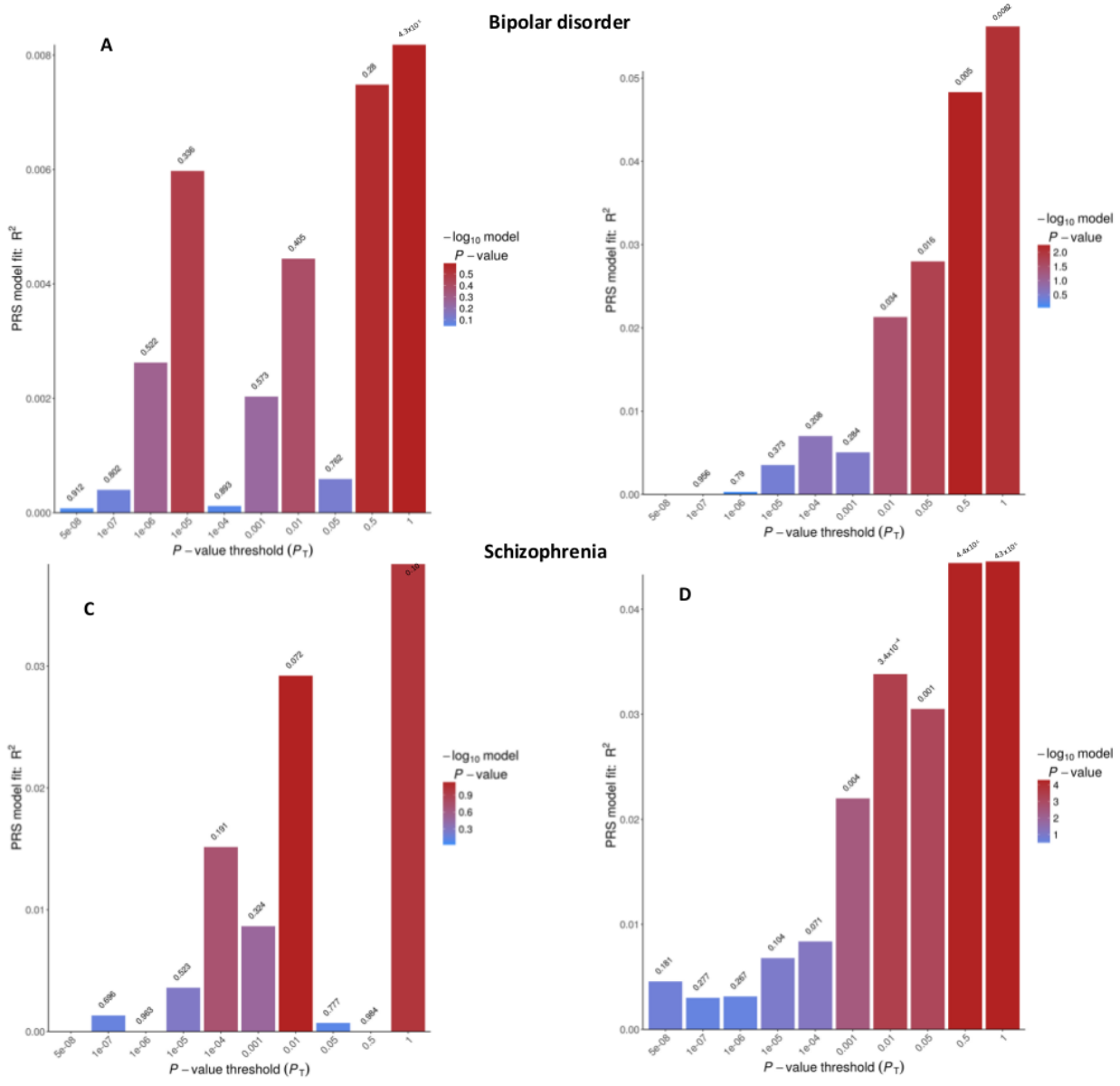
Threshold	R2	P-value	Coefficient	Standard Error	Number of SNPs	FDR	BF
$1.00x10^{-7}$	$1.36x10^{-5}$	0.96	0.011	0.21	1	1	1.00
$1.00x10^{-6}$	0.00031	0.79	-0.056	0.21	6	0.92	1.00
$1.00x10^{-5}$	0.0035	0.37	0.20	0.22	13	0.58	1.00
0.00010	0.0070	0.21	0.28	0.22	37	0.41	1.00
0.0010	0.0051	0.28	0.23	0.21	113	0.48	1.00
0.010	0.021	0.34	0.40	0.19	457	0.16	1.00
0.050	0.028	0.016	0.51	0.21	1358	0.10	0.83
0.50	0.048	0.0051	0.71	0.25	7077	0.043	0.26
1.00	0.056	0.0082	1.30	0.49	10906	0.060	0.42

Table 1.2.4.6: PRS analysis results for schizophrenia on the merged ID and control dataset

Threshold	R2	P-value	Coefficient	Standard Error	Number of SNPs	FDR	BF
$1.00x10^{-8}$	0.0046	0.18	0.20	0.15	32	0.38	1.0
$1.00x10^{-7}$	0.0030	0.28	0.16	0.15	37	0.48	1.0
$1.00x10^{-6}$	0.0030	0.27	0.16	0.15	55	0.48	1.0
$1.00x10^{-5}$	0.0068	0.10	0.24	0.15	89	0.27	1.0
0.00010	0.0083	0.071	0.27	0.15	162	0.26	1.00
0.0010	0.022	0.0035	0.43	0.15	359	0.036	0.18
0.010	0.033	0.00034	0.54	0.15	868	0.0057	0.017
0.050	0.034	$6.4x10^{-4}$	0.51	0.15	1797	0.0081	0.033
0.50	0.031	$4.4x10^{-5}$	0.64	0.15	5629	0.0011	0.0022
1.00	0.045	$4.3x10^{-5}$	0.64	0.15	7403	0.0011	0.0021

Table 1.2.3.1-1.2.4.6: PRSice2 results for each phenotype. PRSice2 calculated the number of SNPs that account for a value of variance (R2) at a certain p-value threshold (Threshold), in addition to the observed p-value, coefficient, standard error and false discovery rate and Bonferroni (BF) correction values, on both datasets.

Figure 1.2.4A-D: Barplots of PRS model fit at various p-value thresholds for schizophrenia and bipolar disorder.



Barplots of PRS model fit for (A) bipolar disorder for the ID dataset, (B) bipolar disorder for the merged ID and control dataset, (C) schizophrenia for the ID dataset (D) and schizophrenia for the merged ID and control dataset.

4.3 Multiple Testing Correction

After carrying out FDR on the ID dataset results, none of the p-values were significant and after carrying out Bonferroni corrections all the p-values were changed to 1 except for one value (which was still >0.05 and were therefore no longer considered to be significant) (Table .6).

For the merged ID and controls dataset results, four results were significant once FDR corrections were carried out. For the bipolar disorder analysis, the significant results were the 7077 SNPs accounting for 4.83% of the variance at a p-value threshold of 0.05. For the schizophrenia analysis, the 868 SNPs account for 3.34% of the variance at a 0.01 threshold, the 1797 SNPs accounting for 3.05% of the variance at a 0.05 threshold, the 5629 SNPs accounting for 4.44% of the variance at a p-value threshold of 0.5 and the 7403 SNPs accounting for 4.46% of the variance at a threshold of 1, which remained significant after FDR was carried out. These four results were the only to remain significant after the more stringent Bonferroni corrections were carried out.

5.0 Discussion

The aim of the project was to determine whether PRS for psychiatric conditions such as ADHD, ASD, anxiety disorders, MDD, schizophrenia and bipolar disorder is predictive of these conditions in people with ID, using a cohort of patients with intellectual disability and comorbid mental illness, with which this method could have potential use in the prediction of risk of psychiatric disorders in patients and diagnosis in clinical settings.

A high proportion of the results from the PRS analysis were either not significant or later deemed insignificant upon not surviving multiple testing correction. No significant results were obtained for ADHD as the PRS was insignificant at each p-value threshold, both for the ID dataset and the merged ID and healthy controls dataset. Significant PRS results found in the MDD and ASD analysis explained only a small variance of 2% or less and did not pass either of the multiple testing corrections.

An initial prominent result was produced in the PRS analysis for anxiety, where 782 SNPs accounted for 9.36% of the variance of the ID dataset at a p-value threshold of 0.001 (p-value = 0.023), suggesting that this PRS is somewhat a predicctor of anxiety. However, this failed multiple testing correction too (FDR = 0.50, Bonferroni = 1) and was not observed at all the other p-value thresholds.

Other notable results were the 10,906 SNPs that accounted for 5.62% of the variance of the merged ID and controls dataset for bipolar disorder, but yet again this is PRS was not considered a good predictor of bipolar disorder as the result did not pass either of the multiple testing corrections.

The only results that remained significant throughout the analysis were those for the schizophrenia analysis using the merged dataset. 359 SNPs account for 2.20% of the variance at a 0.01 threshold, the 1797 SNPs accounting for 3.05% of the variance at a 0.05 threshold and 5629 SNPs accounting for 4.44% of the variance at a p-value threshold of 0.5. This suggests that this PRS is somewhat of a predictor for schizophrenia, however, these findings demonstrate that identifying causative SNPs for these psychiatric disorders is in the early stages of analysis and that the use of PRS for psychiatric conditions cannot be justified based on these results.

The PRS analysis using the merged dataset consisting of genetic data from both ID cases and controls produced more significant results overall, particularly for the schizophrenia PRS. The lack of controls in the first set of PRS analysis (with just the ID samples), where the other ID subjects acted as controls, appeared to negatively affect the analysis. This was particularly evident from the FDR and Bonferroni corrections, with which every corrected p-value showed the PRS results to be insignificant. This could likely be the result of the ID of the subjects was interfering in the background and making the other conditions harder to pick up on.

However, a limitation was presented in the merged dataset as despite the ID and healthy control samples' data originating from the same array chip, the PsychArray, the healthy control samples were genotyped on a different occasion which bears the risk of interfering with the reliability of the analysis.

Furthermore, Bonferroni corrections were arguably too stringent for this particular study as many of the PRS analyses were not independent statistical tests, for example, any analysis for a particular condition at each p-value threshold. FDR alone would likely be a more suitable multiple testing correction method to employ for this study.

Another clear limitation is the small sample size. The sample consists of only 242 individuals and this was reduced to 233 after QC had been undertaken. A larger sample would have given more reliable results with greater precision and power, amongst allowing for other additional tests to take into account and include samples that were otherwise removed by QC. With a larger sample size, linear mixed models could be used to conduct analyses on the related individuals in the sample and to take into account this relatedness rather than remove all related individuals.

The PCA posed another limitation for the PRS analysis. As only GWAS containing individuals with European ancestry was used in the PRS analysis, individuals within the ID sample with non-European ancestry were marked for exclusion from the dataset. The original sample contained individuals with Nigerian, Japanese and Han Chinese ancestry, however, they were removed from the sample due to the fact that LD patterns differ across populations and would

have disrupted the PRS results. It would have been beneficial to include individuals of non-European ancestry in order to have a more representative picture of different genetic variation in different sub-populations, particularly as a higher level of genetic variation is often found in non-European populations (as well as to include individuals of non-European ancestry in genetic studies for ethical reasons). Including these individuals would have also benefitted the study by reducing the size of the dataset by less. Future analysis could incorporate methods to take into account individuals of different ancestry such as the sub-Population Comparison (PopCorn) method, a program for estimating the correlation of causal variant effect sizes across populations in GWAS. It allows for the identification of sub-populations within a sample whilst simultaneously performing sub-populations mapping across the sample, as carried out by B. Brown et al. (2016) in their study on transethnic genetic correlation estimates.

Future studies should incorporate other measures of psychiatric conditions other than diagnosis-based analysis, such as symptoms based analysis using the Mini PAS-ADD which measures symptoms and thresholds rather than categorical diagnoses. Different covariates should also be explored including covariates of sex and age.

5.1 Conclusion

This study aimed to calculate PRS for several psychiatric conditions in individuals with ID with both an ID dataset and a merged dataset of both ID cases and healthy controls. Whilst the PRS generated showed that several groups of SNPs accounted for a small percentage of the variance of these conditions, the majority of these results were insignificant and even more were after they had been subjected to multiple testing corrections. The PRS for the schizophrenia analysis on the merged dataset appeared to be a good predictor for schizophrenia, with the highest variance of 4.44% being accounted for by 5629 SNPs at a p-value threshold of 0.5, and surviving both FDR and Bonferroni corrections. However, the other PRS did not prove to be good predictors of any of the other five psychiatric disorders analysed in ID subjects. Calculating PRS have proved before to be a valuable tool in predicting the risk of psychiatric disorders in patients and has potential even in clinical settings, with PRS for other conditions having been considered reliable enough for use in the clinic. However, this particular study, in its early stages, requires much further testing and validation to be reliable enough to be considered for use in genetic screenings and risk assessments.

6.0 Author's Contribution:

Genotyping was carried out at the Broad Institute (Massachusetts, USA), the two QC scripts (see Appendix 9.5 and 9.6) and the PRSice2 script (Appendix 9.7) were written and provided Johan H. Thygesen and Andries T. Marees et al. (2018) but were edited by myself with the assistance of Prof. Andrew McQuillin. The imputation was carried out on the Sanger Imputation Server (Wellcome Sanger Institute). Healthy control DNA samples were collected

by UCL and the schizophrenia results were obtained from the unpublished PGC-SCZ3 GWAS via personal communication from Professor Andrew McQuillin.

7.0 Acknowledgments

I would like to thank my research supervisors, Professor Andrew McQuillin and Dr Nick Bass for their continued guidance and support throughout the project and for ensuring that we were all well and in good spirits throughout the lockdown. Both their academic and pastoral support throughout the research project was greatly appreciated. I would also like to thank my professors, colleagues and friends at UCL for their support throughout the year and for making my last year at UCL a really enjoyable one. I would particularly like to extend this thanks to all the members of the Molecular Psychiatry Laboratory.

Lastly, I give my endless thanks to my mother and sister. I am certain that I could not have achieved half of what I have so far without their unwavering encouragement, guidance and support.

8.0 References

1. Thygesen JH, Wolfe K, McQuillin A, Viñas-Jornet M, Baena N, Brison N, et al. Neurodevelopmental risk copy number variants in adults with intellectual disabilities and comorbid psychiatric disorders. Br J Psychiatry. 2018 May;212(5):287–94.
2. Whitaker S, Read S. The Prevalence of Psychiatric Disorders among People with Intellectual Disabilities: An Analysis of the Literature. J Appl Res Intellect Disabil. 2006 Nov 14;19:330–45.
3. Plomin R, Owen MJ, McGuffin P. The genetic basis of complex human behaviors. Science (80-) [Internet]. 1994 Jun 17;264(5166):1733 LP – 1739. Available from: http://science.sciencemag.org/content/264/5166/1733.abstract
4. World Health Organization Geneva. The ICD-10 Classification of Mental and Behavioural Disorders Diagnostic criteria for research. 10th Revis Int Classif Dis. 1993;10:188.
5. Epstein JN, Loren REA. Changes in the Definition of ADHD in DSM-5: Subtle but Important. Neuropsychiatry (London). 2013 Oct;3(5):455–8.
6. Faraone S V, Biederman J, Mick E, Williamson S, Wilens T, Spencer T, et al. Family study of girls with attention deficit hyperactivity disorder. Am J Psychiatry. 2000 Jul;157(7):1077–83.
7. Franke B, Faraone S V, Asherson P, Buitelaar J, Bau CHD, Ramos-Quiroga JA, et al. The genetics of attention deficit/hyperactivity disorder in adults, a review. Mol Psychiatry [Internet]. 2012;17(10):960–87. Available from: https://doi.org/10.1038/mp.2011.138
8. Cardon LR, Smith SD, Fulker DW, Kimberling WJ, Pennington BF, DeFries JC. Quantitative trait locus for reading disability on chromosome 6. Science. 1994 Oct;266(5183):276–9.

9. Cook EHJ, Stein MA, Krasowski MD, Cox NJ, Olkon DM, Kieffer JE, et al. Association of attention-deficit disorder and the dopamine transporter gene. Am J Hum Genet. 1995 Apr;56(4):993–8.
10. Tovo-Rodrigues L, Rohde LA, Menezes AMB, Polanczyk G V, Kieling C, Genro JP, et al. DRD4 rare variants in Attention-Deficit/Hyperactivity Disorder (ADHD): further evidence from a birth cohort study. PLoS One. 2013;8(12):e85164.
11. Neale BM, Medland SE, Ripke S, Asherson P, Franke B, Lesch K-P, et al. Meta-analysis of genome-wide association studies of attention-deficit/hyperactivity disorder. J Am Acad Child Adolesc Psychiatry. 2010 Sep;49(9):884–97.
12. Ghirardi L, Brikell I, Kuja-Halkola R, Freitag CM, Franke B, Asherson P, et al. The familial co-aggregation of ASD and ADHD: a register-based cohort study. Mol Psychiatry. 2018 Feb;23(2):257–62.
13. Larsson H, Rydén E, Boman M, Långström N, Lichtenstein P, Landén M. Risk of bipolar disorder and schizophrenia in relatives of people with attention-deficit hyperactivity disorder. Br J Psychiatry [Internet]. 2018/01/02. 2013;203(2):103–6. Available from: https://www.cambridge.org/core/article/risk-of-bipolar-disorder-and-schizophrenia-in-relatives-of-people-with-attentiondeficit-hyperactivity-disorder/6D7424E3DAC69F129DA235B0E3BFA4F0
14. FARAONE S V, BIEDERMAN J. Do Attention Deficit Hyperactivity Disorder and Major Depression Share Familial Risk Factors? J Nerv Ment Dis [Internet]. 1997;185(9). Available from: https://journals.lww.com/jonmd/Fulltext/1997/09000/Do_Attention_Deficit_Hyperactivity_Disorder_and.1.aspx
15. Demontis D, Walters RK, Martin J, Mattheisen M, Als TD, Agerbo E, et al. Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity disorder. Nat Genet [Internet]. 2019;51(1):63–75. Available from: https://doi.org/10.1038/s41588-018-0269-7
16. De Rubeis S, Buxbaum JD. Genetics and genomics of autism spectrum disorder: embracing complexity. Hum Mol Genet. 2015 Oct;24(R1):R24-31.
17. Hallmayer J, Cleveland S, Torres A, Phillips J, Cohen B, Torigoe T, et al. Genetic Heritability and Shared Environmental Factors Among Twin Pairs With Autism. Arch Gen Psychiatry [Internet]. 2011 Nov 1;68(11):1095–102. Available from: https://doi.org/10.1001/archgenpsychiatry.2011.76
18. Grove J, Ripke S, Als TD, Mattheisen M, Walters RK, Won H, et al. Identification of common genetic risk variants for autism spectrum disorder. Nat Genet. 2019 Mar;51(3):431–44.
19. Skafidas E, Testa R, Zantomio D, Chana G, Everall IP, Pantelis C. Predicting the diagnosis of autism spectrum disorder using gene pathway analysis. Mol Psychiatry [Internet]. 2014;19(4):504–10. Available from: https://doi.org/10.1038/mp.2012.126
20. Dunn EC, Sofer T, Gallo LC, Gogarten SM, Kerr KF, Chen C-Y, et al. Genome-wide association study of generalized anxiety symptoms in the Hispanic Community Health Study/Study of Latinos. Am J Med Genet Part B, Neuropsychiatr Genet Off Publ Int Soc Psychiatr Genet. 2017 Mar;174(2):132–43.
21. Otowa T, Hek K, Lee M, Byrne EM, Mirza SS, Nivard MG, et al. Meta-analysis of genome-wide association studies of anxiety disorders. Mol Psychiatry. 2016 Oct;21(10):1391–9.
22. Lohoff FW. Overview of the genetics of major depressive disorder. Curr Psychiatry Rep. 2010 Dec;12(6):539–46.
23. Sullivan PF, Neale MC, Kendler KS. Genetic epidemiology of major depression: review

and meta-analysis. Am J Psychiatry. 2000 Oct;157(10):1552–62.
24. Weissman MM, Wickramaratne P, Adams PB, Lish JD, Horwath E, Charney D, et al. The relationship between panic disorder and major depression. A new family study. Arch Gen Psychiatry. 1993 Oct;50(10):767–80.
25. Sullivan PF, de Geus EJC, Willemsen G, James MR, Smit JH, Zandbelt T, et al. Genome-wide association for major depressive disorder: a possible role for the presynaptic protein piccolo. Mol Psychiatry. 2009 Apr;14(4):359–75.
26. Breier A, Schreiber JL, Dyer J, Pickar D. National Institute of Mental Health longitudinal study of chronic schizophrenia. Prognosis and predictors of outcome. Arch Gen Psychiatry. 1991 Mar;48(3):239–46.
27. Hilker R, Helenius D, Fagerlund B, Skytthe A, Christensen K, Werge TM, et al. Heritability of Schizophrenia and Schizophrenia Spectrum Based on the Nationwide Danish Twin Register. Biol Psychiatry. 2018 Mar;83(6):492–8.
28. Shi J, Levinson D, Duan J, Sanders A, Zheng Y, Pe'er I, et al. Common variants on chromosome 6p22.1 are associated with schizophrenia. Nature. 2009 Aug 1;460:753–7.
29. Li B, Woo R-S, Mei L, Malinow R. The neuregulin-1 receptor erbB4 controls glutamatergic synapse maturation and plasticity. Neuron. 2007 May;54(4):583–97.
30. Ripke S, Neale BM, Corvin A, Walters JTR, Farh K-H, Holmans PA, et al. Biological insights from 108 schizophrenia-associated genetic loci. Nature [Internet]. 2014;511(7510):421–7. Available from: https://doi.org/10.1038/nature13595
31. McCarthy SE, Makarov V, Kirov G, Addington AM, McClellan J, Yoon S, et al. Microduplications of 16p11.2 are associated with schizophrenia. Nat Genet. 2009 Nov;41(11):1223–7.
32. Tondo L, Baldessarini RJ. Reduced suicide risk during lithium maintenance treatment. J Clin Psychiatry. 2000;61 Suppl 9:97–104.
33. Namjoshi MA, Buesching DP. A review of the health-related quality of life literature in bipolar disorder. Qual life Res an Int J Qual life Asp Treat care Rehabil. 2001;10(2):105–15.
34. Bertelsen A, Harvald B, Hauge M. A Danish twin study of manic-depressive disorders. Br J Psychiatry. 1977 Apr;130:330–51.
35. Blackwood DH, He L, Morris SW, McLean A, Whitton C, Thomson M, et al. A locus for bipolar affective disorder on chromosome 4p. Nat Genet. 1996 Apr;12(4):427–30.
36. McMahon FJ, Hopkins PJ, Xu J, McInnis MG, Shaw S, Cardon L, et al. Linkage of bipolar affective disorder to chromosome 18 markers in a new pedigree series. Am J Hum Genet. 1997 Dec;61(6):1397–404.
37. Stine OC, Xu J, Koskela R, McMahon FJ, Gschwend M, Friddle C, et al. Evidence for linkage of bipolar disorder to chromosome 18 with a parent-of-origin effect. Am J Hum Genet. 1995 Dec;57(6):1384–94.
38. Stahl EA, Breen G, Forstner AJ, McQuillin A, Ripke S, Trubetskoy V, et al. Genome-wide association study identifies 30 loci associated with bipolar disorder. Nat Genet. 2019 May;51(5):793–803.
39. Wolfe K, Strydom A, Morrogh D, Carter J, Cutajar P, Eyeoyibo M, et al. Chromosomal microarray testing in adults with intellectual disability presenting with comorbid psychiatric disorders. Eur J Hum Genet. 2016 Oct 21;25.
40. Jansen AG, Dieleman GC, Jansen PR, Verhulst FC, Posthuma D, Polderman TJC. Psychiatric Polygenic Risk Scores as Predictor for Attention Deficit/Hyperactivity Disorder and Autism Spectrum Disorder in a Clinical Child and Adolescent Sample. Behav Genet [Internet]. 2020;50(4):203–12. Available from: https://doi.org/10.1007/s10519-019-09965-8

41. Lee SH, Ripke S, Neale BM, Faraone S V, Purcell SM, Perlis RH, et al. Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. Nat Genet [Internet]. 2013;45(9):984–94. Available from: https://doi.org/10.1038/ng.2711
42. Wray NR, Ripke S, Mattheisen M, Trzaskowski M, Byrne EM, Abdellaoui A, et al. Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. Nat Genet. 2018 May;50(5):668–81.
43. Khera A V, Chaffin M, Aragam KG, Haas ME, Roselli C, Choi SH, et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nat Genet. 2018 Sep;50(9):1219–24.
44. Choi SW, Mak TS-H, O'Reilly PF. Tutorial: a guide to performing polygenic risk score analyses. Nat Protoc [Internet]. 2020;15(9):2759–72. Available from: https://doi.org/10.1038/s41596-020-0353-1

9.0 Appendix

9.1 phenotype_file.txt

FID	BID	Autism	Schizophrenia	Depressive	Anxiety	ADHD	Bipolar
LD0002	LD0002	1	1	2	1	1	1
LD0004	LD0004	1	1	1	1	1	2
LD0007	LD0007	2	1	1	1	1	2
LD0008	LD0008	1	2	2	1	1	1
LD0009	LD0009	2	2	2	2	1	1
LD0011	LD0011	2	1	2	2	1	1
LD0012	LD0012	1	1	1	1	1	1
LD0013	LD0013	2	1	2	1	1	1
LD0014	LD0014	1	2	1	2	1	1
LD0015	LD0015	2	2	2	2	1	1
LD0016	LD0016	2	1	2	2	2	1
LD0017	LD0017	1	1	1	2	1	1
LD0018	LD0018	1	1	1	2	1	1
LD0020	LD0020	1	1	1	1	1	1
LD0021	LD0021	1	2	1	1	1	1
LD0022	LD0022	1	2	1	1	1	1

LD0023	LD0023	1	1	1	1	1	1
LD0025	LD0025	1	1	1	1	1	1
LD0026	LD0026	1	1	2	1	1	1
LD0027	LD0027	2	2	1	2	1	1
LD0028	LD0028	2	1	1	1	1	1
LD0029	LD0029	1	1	2	2	1	1
LD0030	LD0030	1	1	1	2	1	1
LD0032	LD0032	1	1	2	1	1	1
LD0035	LD0035	2	1	2	1	1	1
LD0036	LD0036	1	1	2	1	1	1
LD0037	LD0037	1	2	2	1	1	1
LD0038	LD0038	1	2	1	1	1	2
LD0039	LD0039	1	2	1	1	1	1
LD0043	LD0043	1	1	1	1	1	2
LD0044	LD0044	1	2	1	2	1	1
LD0046	LD0046	2	1	1	2	2	1
LD0047	LD0047	2	2	1	2	1	1
LD0048	LD0048	2	1	1	2	1	1
LD0051	LD0051	2	1	2	1	2	1
LD0052	LD0052	1	2	2	1	1	1
LD0053	LD0053	2	2	1	2	1	1
LD0054	LD0054	1	1	1	1	1	1
LD0055	LD0055	2	2	1	1	1	1
LD0056	LD0056	1	2	2	1	1	1
LD0059	LD0059	2	2	1	2	1	1
LD0061	LD0061	1	2	2	2	1	1
LD0062	LD0062	1	1	1	1	1	2

LD0064	LD0064	1	1	2	1	1	1
LD0068	LD0068	1	2	1	2	1	1
LD0069	LD0069	1	1	1	1	1	1
LD0070	LD0070	1	1	2	1	1	1
LD0071	LD0071	2	1	2	1	1	1
LD0073	LD0073	1	1	1	1	1	2
LD0074	LD0074	2	1	2	1	2	1
LD0076	LD0076	1	1	1	1	2	1
LD0077	LD0077	1	1	1	1	1	2
LD0078	LD0078	1	1	2	2	1	1
LD0079	LD0079	2	1	1	1	2	1
LD0080	LD0080	1	1	2	2	1	1
LD0081	LD0081	1	1	1	2	1	1
LD0082	LD0082	2	1	1	1	2	1
LD0083	LD0083	2	1	2	1	1	2
LD0086	LD0086	1	2	1	1	1	1
LD0090	LD0090	1	1	2	2	1	1
LD0091	LD0091	2	1	1	1	1	1
LD0092	LD0092	2	2	2	2	1	1
LD0093	LD0093	2	1	1	1	1	1
LD0094	LD0094	2	1	1	1	1	2
LD0095	LD0095	2	1	1	1	1	1
LD0096	LD0096	1	2	1	1	2	1
LD0098	LD0098	1	2	1	1	1	1
LD0099	LD0099	1	1	2	1	1	1
LD0102	LD0102	2	1	1	2	1	1
LD0103	LD0103	1	1	1	1	1	1

LD0104	LD0104	1	2	1	1	1	1
LD0110	LD0110	2	1	1	1	2	1
LD0111	LD0111	1	1	1	2	1	1
LD0112	LD0112	2	1	1	1	2	1
LD0113	LD0113	1	2	2	1	1	1
LD0114	LD0114	2	1	1	2	1	1
LD0115	LD0115	1	1	1	1	2	1
LD0116	LD0116	1	2	1	1	1	1
LD0117	LD0117	1	1	1	1	2	1
LD0119	LD0119	1	1	1	2	1	1
LD0120	LD0120	1	1	2	1	1	1
LD0121	LD0121	1	1	1	1	1	1
LD0122	LD0122	1	1	1	1	2	1
LD0123	LD0123	1	1	2	1	1	1
LD0124	LD0124	1	1	2	1	1	1
LD0125	LD0125	1	1	2	1	1	1
LD0126	LD0126	1	1	1	1	1	2
LD0127	LD0127	1	2	1	1	1	1
LD0128	LD0128	1	2	1	1	1	1
LD0129	LD0129	1	1	1	1	1	2
LD0130	LD0130	1	1	1	1	1	1
LD0131	LD0131	1	1	2	2	1	1
LD0132	LD0132	2	1	1	1	1	1
LD0133	LD0133	2	1	2	2	2	1
LD0134	LD0134	2	1	1	1	1	1
LD0135	LD0135	2	1	2	2	1	1
LD0136	LD0136	2	1	1	1	1	1

LD0137	LD0137	2	1	1	1	1	1
LD0138	LD0138	1	1	1	1	1	1
LD0139	LD0139	2	1	1	1	1	1
LD0140	LD0140	1	2	1	1	1	1
LD0141	LD0141	1	1	1	1	1	1
LD0142	LD0142	1	1	1	1	1	1
LD0144	LD0144	2	1	1	1	1	1
LD0146	LD0146	2	1	1	1	1	1
LD0148	LD0148	2	1	1	2	1	1
LD0149	LD0149	1	2	1	2	1	1
LD0151	LD0151	2	2	2	1	1	1
LD0152	LD0152	1	1	2	1	1	1
LD0153	LD0153	1	1	2	1	1	1
LD0154	LD0154	2	1	2	1	1	1
LD0155	LD0155	2	1	2	1	1	1
LD0156	LD0156	1	2	1	1	1	1
LD0157	LD0157	1	1	2	1	1	1
LD0158	LD0158	1	1	1	1	1	1
LD0161	LD0161	2	1	1	1	1	1
LD0162	LD0162	2	1	1	1	1	1
LD0164	LD0164	1	1	2	2	1	1
LD0165	LD0165	1	1	1	1	1	1
LD0166	LD0166	2	1	1	1	1	1
LD0167	LD0167	2	1	1	1	2	1
LD0168	LD0168	1	1	1	1	1	1
LD0171	LD0171	2	1	1	1	1	1
LD0172	LD0172	2	1	1	1	1	1

LD0174	LD0174	2	1	1	1	1	1
LD0179	LD0179	1	1	2	1	1	1
LD0180	LD0180	1	1	1	1	1	2
LD0181	LD0181	2	1	1	1	1	1
LD0182	LD0182	2	1	2	1	2	1
LD0184	LD0184	1	1	2	1	1	1
LD0189	LD0189	1	1	1	2	1	1
LD0190	LD0190	1	1	1	2	1	1
LD0191	LD0191	1	1	1	2	1	1
LD0192	LD0192	2	1	1	1	2	1
LD0200	LD0200	2	1	1	1	1	1
LD0201	LD0201	1	1	1	1	1	2
LD0204	LD0204	1	1	2	2	2	1
LD0205	LD0205	1	1	2	1	1	1
LD0209	LD0209	2	2	1	1	1	1
LD0210	LD0210	1	2	1	1	1	1
LD0211	LD0211	1	1	2	2	1	1
LD0212	LD0212	1	1	2	2	1	1
LD0213	LD0213	2	1	1	1	1	1
LD0214	LD0214	2	1	1	1	1	1
LD0215	LD0215	1	1	2	1	1	1
LD0216	LD0216	1	2	1	1	1	1
LD0217	LD0217	2	1	1	1	2	2
LD0221	LD0221	2	1	1	1	1	1
LD0222	LD0222	2	1	1	1	1	1
LD0223	LD0223	1	1	2	1	1	1
LD0224	LD0224	1	2	2	1	1	1

LD0225	LD0225	1	1	1	1	1	1
LD0231	LD0231	1	1	2	2	1	1
LD0232	LD0232	1	1	2	1	1	1
LD0233	LD0233	2	1	2	2	1	1
LD0234	LD0234	1	1	2	1	1	1
LD0235	LD0235	2	1	2	1	1	1
LD0236	LD0236	1	2	1	1	1	1
LD0240	LD0240	1	1	2	1	1	1
LD0244	LD0244	1	1	1	1	1	1
LD0245	LD0245	1	1	2	2	1	1
LD0246	LD0246	1	2	1	1	1	1
LD0247	LD0247	1	1	1	1	1	2
LD0248	LD0248	1	1	2	1	1	1
LD0249	LD0249	1	1	2	1	1	1
LD0251	LD0251	2	1	1	1	1	2
LD0255	LD0255	1	2	1	1	1	1
LD0266	LD0266	1	1	2	1	1	1
LD0267	LD0267	2	1	1	1	1	1
LD0268	LD0268	1	2	2	1	1	1
LD0269	LD0269	1	2	1	1	1	1
LD0271	LD0271	1	1	1	1	1	2
LD0272	LD0272	1	1	1	1	1	1
LD0274	LD0274	1	1	1	1	1	1
LD0275	LD0275	2	1	1	1	1	1
LD0276	LD0276	1	2	1	1	1	2
LD0277	LD0277	1	1	2	1	1	1
LD0278	LD0278	2	1	1	1	2	1

LD0279	LD0279	1	2	2	1	1	1
LD0280	LD0280	2	1	1	1	1	1
LD0283	LD0283	2	1	1	1	1	2
LD0284	LD0284	1	1	1	1	1	2
LD0285	LD0285	1	2	1	1	1	1
LD0286	LD0286	1	1	1	1	1	2
LD0289	LD0289	2	1	1	1	1	1
LD0290	LD0290	2	1	1	1	1	1
LD0296	LD0296	1	2	1	1	1	1
LD0297	LD0297	1	1	2	2	1	1
LD0298	LD0298	1	1	1	1	1	2
LD0299	LD0299	1	2	2	1	1	2
LD0300	LD0300	2	1	2	1	1	1
LD0301	LD0301	1	1	1	1	1	1
LD0302	LD0302	1	1	2	1	1	1
LD0303	LD0303	1	1	2	2	1	1
LD0304	LD0304	1	1	2	1	1	1
LD0305	LD0305	2	2	1	1	1	1
LD0311	LD0311	1	2	1	1	1	1
LD0312	LD0312	1	1	1	1	2	1
LD0313	LD0313	1	1	1	1	1	1
LD0314	LD0314	1	1	1	1	1	1
LD0315	LD0315	1	1	1	1	1	1
LD0316	LD0316	2	1	1	1	2	1
LD0317	LD0317	1	1	1	1	1	2
LD0321	LD0321	2	1	1	1	2	1
LD0336	LD0336	1	1	1	1	1	2

LD0337	LD0337	1	1	2	1	1	1
LD0338	LD0338	1	2	1	1	1	1
LD0339	LD0339	2	1	1	1	1	1
LD0340	LD0340	1	1	1	2	1	1
LD0351	LD0351	1	1	1	1	1	1
LD0352	LD0352	1	1	1	1	1	1
LD0354	LD0354	2	1	1	2	1	1
LD0355	LD0355	1	1	1	1	1	1
LD0356	LD0356	1	2	1	1	1	1
LD0357	LD0357	2	1	2	1	1	1
LD0358	LD0358	1	2	1	1	1	1
LD0359	LD0359	1	2	1	1	1	1
LD0360	LD0360	1	1	1	1	1	1
LD0361	LD0361	2	1	1	1	1	1
LD0362	LD0362	1	1	2	1	1	1
LD0363	LD0363	2	1	1	2	1	1
LD0364	LD0364	1	1	1	1	2	1
LD0369	LD0369	1	1	2	1	1	1
LD0372	LD0372	1	1	1	1	1	1
LD0400	LD0400	1	2	1	1	1	1

9.2 merge_ID_ctrl.txt

Generate a merged data file between imputed PsychChip ID data and PsychChip control data.

Only SNPS from the imputed PsychChip ID data are used and then any SNP with a case control missingness p value of less than 0.00001 is removed

plink2 --bfile /mnt/10tbstore/projects/ava/psychchip_ID_imputed/psychchip_ID_qc_info0.8 \

--bmerge /mnt/10tbstore/projects/alc_symptom_score_gwas/ALC/pchipCtrl_pchipALC_info01/pchipCtrl_pchipALC_info01.qc.hrc.a1.pca.ok.3nd.01 \

--out psychchip_ID_ctrl_merge

plink2 --bfile /mnt/10tbstore/projects/ava/psychchip_ID_imputed/psychchip_ID_qc_info0.8 --exclude psychchip_ID_ctrl_merge.missnp --make-bed --out source1_tmp

plink2 --bfile /mnt/10tbstore/projects/alc_symptom_score_gwas/ALC/pchipCtrl_pchipALC_info01/pchipCtrl_pchipALC_info01.qc.hrc.a1.pca.ok.3nd.01 --exclude psychchip_ID_ctrl_merge.missnp --extract source1_tmp.bim --filter-controls --make-bed --out source2_tmp

plink2 --bfile source1_tmp --bmerge source2ctrl_tmp --geno --make-bed --out psychchip_ID_ctrl_merge

base=psychchip_ID_ctrl_merge

input=input

qcstep=qc_steps

plink2 --bfile $base --missing --out 01.$base

plink2 --bfile $base --het --out 01.$base

plink2 --bfile $base --maf 0.05 --missing --out 01.$base.common

plink2 --bfile $base --maf 0.05 --het --out 01.$base.common

plink2 --bfile $base --allow-no-sex --test-missing --out 06.$base

then use script cc_missing_qc.R to identify SNPs with case control missing differences.

plink2 --bfile psychchip_ID_ctrl_merge --exclude cc_remove_list.txt --make-bed --out psychchip_ID_ctrl_merge.qc

plink2 --allow-no-sex --bfile $base.qc --maf 0.06 --indep 50 5 2 --out 06.$base

plink2 --allow-no-sex --bfile $base.qc --extract 06.$base.prune.in --make-bed --out 07.$base

plink2 --allow-no-sex --bfile 07.$base --exclude range $input/ld_regions_hg19.txt --extract $input/

hapmap3r2_CEU.CHB.JPT.YRI.founders.no-at-cg-snps.lifted.bim --make-bed --out 08.$base

```
cd qc_steps
convertf -p <(printf "genotypename: 08.$base.bed
snpname: 08.$base.bim
indivname: 08.$base.fam
outputformat: EIGENSTRAT
#line below is required if SNPs are not in chromosomal order. Else change to YES
pordercheck: NO
genotypeoutname: 08.$base.eigenstratgeno
snpoutname: 08.$base.snp
indivoutname: 08.$base.ind")
## Run of Eigenstrat SmartPCA
# Produces 100 PCs
smartpca.perl \
-i 08.$base.eigenstratgeno \
-a 08.$base.snp \
-b 08.$base.ind \
-o 08.$base.pca \
-p 08.$base.plot \
-e 08.$base.eval \
-l 08.$base.smartpca.log \
-m 0 \
-t 100 \
-k 100 \
-s 6
mkdir results
mv psychchip_ID_ctrl_merge.qc* results/
```

##now use R script write.covars.R

9.3 SNP_missingness_script.txt

```
## Missingness Case Ctrl
    miss <- read.table("06.psychchip_ID_ctrl_merge.missing", header = T)
    miss <- as.data.frame(miss)
    miss <- miss[order(miss[,"P"],decreasing = T),]
    miss <- miss[,c("SNP","F_MISS_A","F_MISS_U","P")]
    miss$qc[miss$P< 0.000001] <- "FAIL_QC"
    miss$qc[miss$P> 0.000001001] <- "PASS"
cc_remove_list <- miss[which(miss$qc=="FAIL_QC"), ]
write.table(cc_remove_list, "cc_remove_list.txt", row.names = F, col.names = F, quote = F)
```

9.4 covariate_script.txt

```
covar <- read.table("qc_steps/08.psychchip_ID_ctrl_merge.pca.evec", header = F)
covar <- covar[,2:5]
colnames(covar) <- c("IID","PCA1","PCA2","PCA3")
covar$FID <- (covar$IID)
write.table(covar[,c("FID","IID","PCA1","PCA2","PCA3")],
        "results/final.pca.txt", row.names = F, col.names = F, quote = F)
```

####### Fix for the weird scenario where pasticipant FID and IIDs are colon separated in column 1.

```
# sed '1d' qc_steps/08.psychchip_ID_ctrl_merge.pca.evec > tmpfile
# awk -F":" '$1=$1' OFS="\t" tmpfile > tmpfile1
#
```

```
# #covar <- read.table("qc_steps/08.psychchip_ID_ctrl_merge.pca.evec", text = gsub(':', ' '), header = T)

#

# covar <- read.table("tmpfile1", text = gsub(':', ' '), header = F)

#

#

#  #covar <- read.table("qc_steps/08.psychchip_ID_ctrl_merge.pca.evec", sep = ":", header = F)

#  covar <- covar[1:5]

#  colnames(covar) <- c("FID", "IID","PCA1","PCA2","PCA3")

#  #covar$FID <- (covar$IID)

#  write.table(covar[,c("FID","IID","PCA1","PCA2","PCA3")],

#              "results/final.pca.txt", row.names = F, col.names = T, quote = F)
```

9.5 pre_qc_steps.sh

```
#!/bin/bash

plink2 --file /mnt/store/snp_genotypes/psychchip_ID/original_files/2016-123-ILL_IND_N=242/2016-123-ILL_IND_Psych1-1_N=242/PLINK_070317_0920/2016-123-ILL_IND_Psych1-1_N=242 --make-bed --out temp1

## Update IDs (use the normal ID IDs)

Rscript make_id_update_list.R

plink2 --bfile temp1 --update-ids update_ids.txt --make-bed --out temp2

## Update parents

plink2 --bfile temp2 --update-parents update_parents.txt --make-bed --out temp3

## Keep only SNPs given to us in second Broad delivery - Exclude blacklisted SNPs (those removed by BROAD between 1 and 2 delivery of the initial psychchip data)
```

```
cut -f2 /home/molpsych/data/snp_genotypes/psychchip/original_supplied/mcqu1_merge_220217.bim > psychip_pass_snps.txt

plink2 --bfile temp3 --extract psychip_pass_snps.txt --make-bed --out temp4

## Flip minus strand snps

plink2 --bfile temp4 --flip /home/molpsych/data/strand_data/PsychChip_v1-1_15073391_A1-b37.strand.snps_on_minus_strand.txt --make-bed --out temp5

## Exclude multimatched SNPs

plink2 --bfile temp5 --exclude /home/molpsych/data/strand_data/PsychChip_v1-1_15073391_A1-b37.multiple --make-bed --out temp6

cut -d" " -f3 /home/molpsych/data/strand_data/PsychChip_v1-1_15073391_A1-b37.miss > /home/molpsych/data/strand_data/PsychChip_v1-1_15073391_A1-b37.miss.sorted

plink2 --bfile temp6 --exclude /home/molpsych/data/strand_data/PsychChip_v1-1_15073391_A1-b37.miss.sorted --make-bed --out temp7

## Create update files for plink to update SNPname, postions, alleles etc.

Rscript pre_qc_snps_updater.R pchip_id.bim

### Update SNPs

plink2 --bfile temp7 --update-map position_update.txt --make-bed --out temp8

plink2 --bfile temp8 --update-alleles alleles_update.txt --make-bed --out pchip_id

## Remove CM positions (as this makes eignestrat throw errors)

awk 'OFS="\t" {print $1, $2, 0, $4, $5, $6}' pchip_id.bim > temp_pchip_id.bim.new

mv temp_pchip_id.bim.new pchip_id.bim

## Update phenotype status to set all as cases

awk 'OFS="\t" {print $1, $2, $3, $4, $5, 2}' pchip_id.fam > temp_pchip_id.fam.new

mv temp_pchip_id.fam.new pchip_id.fam

# cp pchip_alc.bim pchip_alc.bim.preUpdate

# ## cleanup

# rm temp*
```

9.6 qc_steps.sh

```
# #!/bin/bash

#### updated to hg38 ######

### Script to run run QC on SNP-assay data generate final QC-files and report outcome in Rmarkdown report

## Check argumets

if [ "$#" -ne 4 ]; then

    clear

    echo -e "\n Script to run run QC on SNP-assay data; generate final QC-files and report details in Rmarkdown report"

    echo "-------------------------------------------------------------------------"

    echo -e "Please provide 2 arguments: \n 1) Full fill path to the plink files to run QC on \n 2) Base filename of the file-set \n 3) Project title (no spaces please) \n 4) Chiptype (also no spaces please) \n Example: ./gwas_qc_report /home/molpsych/data/snp_genotypes/psychchip mcqu1_pchip-merge Schizophrenia_bipolar_ctrl psychchip\n"

    exit 1

fi

## Pipeline for GWAS QC and reporting

## Settings

hwe_threshold=0.000001     # HWE threshold for autosomes

hwe_threshold_chrX=0.0001 # HEW threshold for X chromosomes (only women)

miss_threshold=0.000001    # Case control difference in missingness p-value threshold

het_sd=3                   # Standard deviation threshold for heterozygosity

het_upper_only=T                  # Only apply the upper heterozygosity threshold (to much heterozygosity) - or both upper and lower

final_gender_error_exclude=true # In final gender check exclude miss matching genders - true or flase

miss_max=0.05              # Genome wide misssingness threshold
```

```
snp_miss_max=0.1          # Per SNP missingness threshold

### Folders and variables

project_title=$3

chiptype=$2

filepath=$1

base=$2

qcsteps=qc_steps

plots=plots

results=results

input=input

gender_up=$input/gender_update.txt

pheno_up=$input/phenotype_update.txt

idb_error=$input/ibd_exclude.txt

### Setup

mkdir -p $qcsteps

mkdir -p $plots

mkdir -p $results

mkdir -p $input

echo -e "\n############################ START OF GWAS QC REPORT ############################\n"

echo "(1) Sort chromosomal order and split pseudoautosomal region on X in to separate chr XY if needed"

echo "--------------------------------"

## Fixing Error: .bim file has a split chromosome & heterozygotic markers on male X

xypresent=`cut -f1 $filepath/$base.bim | grep 25 | wc -l`

if [ $xypresent -eq 0 ]

then
```

```
    plink2 --allow-no-sex --bfile $filepath/$base --split-x hg38 no-fail --make-bed --out $qcsteps/01.$base ## IF XY does not exsist run this else run

else

    plink2 --allow-no-sex --bfile $filepath/$base --make-bed --out $qcsteps/01.$base ## IF XY does not exsist run this else run

fi

echo -e "\nIndividual and SNP QC"

##################################################

echo -e "\nAllelic frequency"

echo "--------------------------------"

plink2 --bfile $qcsteps/01.$base --freq --out $qcsteps/01.$base

### Remove and exclude

#### Individuals missing to much data or have to much heterozyogisity

echo "Heterozygosity and missingnsess"

echo "--------------------------------"

plink2 --bfile $qcsteps/01.$base --missing --out $qcsteps/01.$base

plink2 --bfile $qcsteps/01.$base --het --out $qcsteps/01.$base

plink2 --bfile $qcsteps/01.$base --maf 0.05 --missing --out $qcsteps/01.$base.common

plink2 --bfile $qcsteps/01.$base --maf 0.05 --het --out $qcsteps/01.$base.common

## find heterozygous outliers

## Command args 1: qc_folder, 2: base_name, 3: sd defined as outliers, 4: remove only upper outliers? if F lower outliers are also removed!

Rscript scripts/make_het_remove_list.R $qcsteps $base $het_sd $het_upper_only

plink2 --bfile $qcsteps/01.$base --allow-no-sex --mind $miss_max --remove $qcsteps/01.heterozygous_outliers.txt --make-bed --out $qcsteps/01a.$base

#### SNPs missing

plink2 --bfile $qcsteps/01a.$base --missing --out $qcsteps/01a.$base
```

```
plink2 --bfile $qcsteps/01a.$base --allow-no-sex --geno $snp_miss_max --make-bed --out $qcsteps/01b.$base

#### SNPs out of HWE - with more strigient threshold for women on X (see http://onlinelibrary.wiley.com/doi/10.1002/gepi.21782/full)

plink2 --bfile $qcsteps/01b.$base --hardy --out $qcsteps/01b.$base

awk -v var1=${hwe_threshold} -v var2=${hwe_threshold_chrX} '(($1<23 && $3=="ALL" && $9<var1)||($1>=23 && $3=="ALL" && $9<var2))' $qcsteps/01b.$base.hwe > $qcsteps/01b.$base.hwe.exclude

plink2 --bfile $qcsteps/01b.$base --allow-no-sex --exclude $qcsteps/01b.$base.hwe.exclude --make-bed --out $qcsteps/01c.$base

#### If zcalls have been performed overwrite current SNPs with zcall SNPs for which MAF < 0.1 ##########

if [ -e ../zcall/zcalls.bed ]

then

   ## List SNPs with MAF < 0.1

   plink2 --bfile $qcsteps/01c.$base --freq --out $qcsteps/01c.$base

   awk '$5<0.01 {print $2}' $qcsteps/01c.$base.frq > $qcsteps/01c.$base.SNPs_maf0.01

   ## Extract all of these from the zcalls

   plink2 --bfile ../zcall/zcalls --extract $qcsteps/01c.$base.SNPs_maf0.01 --make-bed --out $qcsteps/01c.zcalls_maf0.01

   ## Remove zcalls with failing HWE and geno

   plink2 --bfile $qcsteps/01c.zcalls_maf0.01 --hardy --out $qcsteps/01c.zcalls_maf0.01

   awk -v var1=${hwe_threshold} '($9<var1)' $qcsteps/01c.zcalls_maf0.01.hwe > $qcsteps/01c.zcalls_maf0.01.hwe.exclude

   plink2 --bfile $qcsteps/01c.zcalls_maf0.01 --exclude $qcsteps/01c.zcalls_maf0.01.hwe.exclude --keep $qcsteps/01c.$base.fam --geno $snp_miss_max --make-bed --out $qcsteps/01c.zcalls_maf0.01.qc

   ## Remove SNPs from original set which have good zcalls and merge the zcalls in!

   cut -f2 $qcsteps/01c.zcalls_maf0.01.qc.bim > $qcsteps/01c.zcalls_maf0.01.qc.snps

   plink2 --bfile $qcsteps/01c.$base --exclude $qcsteps/01c.zcalls_maf0.01.qc.snps --make-bed --out $qcsteps/01c.$base.no_rare
```

```
    plink2 --bfile $qcsteps/01c.$base.no_rare --bmerge $qcsteps/01c.zcalls_maf0.01.qc --merge-mode 2 --out $qcsteps/01d.$base

    ## Save a list of SNPs called with zcall to result folder

    cp $qcsteps/01c.zcalls_maf0.01.qc.snps $results/zcalled_snps.txt

else

    plink2 --bfile $qcsteps/01c.$base --make-bed --out $qcsteps/01d.$base

fi

#### Identify duplicate and triplicate SNPs

################################################################################################################################

plink2 --bfile $qcsteps/01d.$base --allow-no-sex --list-duplicate-vars --out $qcsteps/01d.$base

## List various duplicates

awk '{print $4}' $qcsteps/01d.$base.dupvar > $qcsteps/01d.$base.dupvar.originals.txt

awk '{print $5}' $qcsteps/01d.$base.dupvar > $qcsteps/01d.$base.dupvar.duplicates.txt

awk '($6!=""){print $6}' $qcsteps/01d.$base.dupvar > $qcsteps/01d.$base.dupvar.triplicates.txt

awk '($6!=""){print $5}' $qcsteps/01d.$base.dupvar > $qcsteps/01d.$base.dupvar.dupe_names_for_triplicates.txt

awk '($6!=""){print $4,$5,$6}' $qcsteps/01d.$base.dupvar > $qcsteps/01d.$base.dupvar.all_trip_ids.txt

awk '($6==""){print $4,$5,$6}' $qcsteps/01d.$base.dupvar > $qcsteps/01d.$base.dupvar.all_dupe_ids.txt

## Generate a list of all IDs for markers which are non-unique

cat $qcsteps/01d.$base.dupvar.originals.txt $qcsteps/01d.$base.dupvar.duplicates.txt $qcsteps/01d.$base.dupvar.triplicates.txt > $qcsteps/01d.$base.dupvar.non_unique.txt

## Exclude markers which are genotyped multiple times under different names from dataset

plink2 --bfile $qcsteps/01d.$base --exclude $qcsteps/01d.$base.dupvar.non_unique.txt --make-bed --out $qcsteps/01d.$base.dupvar.unique_only

## Generate separate binary filesets for originals, duplicates and triplicates (and subset of duplicates that are triplicates)
```

plink2 --bfile $qcsteps/01d.$base --extract $qcsteps/01d.$base.dupvar.originals.txt --make-bed --out $qcsteps/01d.$base.dupvar.non_unique_originals

plink2 --bfile $qcsteps/01d.$base --extract $qcsteps/01d.$base.dupvar.duplicates.txt --make-bed --out $qcsteps/01d.$base.dupvar.non_unique_duplicates

plink2 --bfile $qcsteps/01d.$base --extract $qcsteps/01d.$base.dupvar.triplicates.txt --make-bed --out $qcsteps/01d.$base.dupvar.non_unique_triplicates

plink2 --bfile $qcsteps/01d.$base --extract $qcsteps/01d.$base.dupvar.dupe_names_for_triplicates.txt --make-bed --out $qcsteps/01d.$base.dupvar.triplicates_in_dupes

Generate conflicting genotype reports by attempting to merge binary fileset pairs (1v2, 1v3, 2v3 and 3s_in_2v1)

plink2 --bfile $qcsteps/01d.$base.dupvar.non_unique_triplicates --bmerge $qcsteps/01d.$base.dupvar.non_unique_duplicates --merge-equal-pos --merge-mode 7 --out $qcsteps/01d.$base.dupvar.trips_v_dupes

plink2 --bfile $qcsteps/01d.$base.dupvar.non_unique_triplicates --bmerge $qcsteps/01d.$base.dupvar.non_unique_originals --merge-equal-pos --merge-mode 7 --out $qcsteps/01d.$base.dupvar.trips_v_orig

plink2 --bfile $qcsteps/01d.$base.dupvar.triplicates_in_dupes --bmerge $qcsteps/01d.$base.dupvar.non_unique_originals --merge-equal-pos --merge-mode 7 --out $qcsteps/01d.$base.dupvar.trips_in_dupes_v_orig

plink2 --bfile $qcsteps/01d.$base.dupvar.non_unique_originals --bmerge $qcsteps/01d.$base.dupvar.non_unique_duplicates --merge-equal-pos --merge-mode 7 --out $qcsteps/01d.$base.dupvar.dupes_v_orig

Generate counts of conflicting genotypes per marker

tail -n+2 $qcsteps/01d.$base.dupvar.trips_v_orig.diff | awk '{print $1}'| sort | uniq -c | sort -b -k2,2 > $qcsteps/01d.$base.dupvar.trips_v_orig_conflicts.txt

tail -n+2 $qcsteps/01d.$base.dupvar.trips_v_dupes.diff | awk '{print $1}' | sort | uniq -c | sort -b -k2,2 > $qcsteps/01d.$base.dupvar.trips_v_dupes_conflicts.txt

tail -n+2 $qcsteps/01d.$base.dupvar.trips_in_dupes_v_orig.diff | awk '{print $1}'| sort | uniq -c | sort -b -k2,2 > $qcsteps/01d.$base.dupvar.trips_dupes_v_orig_conflicts.txt

tail -n+2 $qcsteps/01d.$base.dupvar.dupes_v_orig.diff | awk '{print $1}' | sort | uniq -c | sort -g -k1,1 > $qcsteps/01d.$base.dupvar.dupes_v_orig_conflicts.txt

Find duplicated variants(SNPs) that have create conflicting genotypes. Arguments (7): dupvar.all_trip_ids.txt + The four genotype conflict counts generated above + qc_steps + base

```
Rscript scripts/find_conflicting_dupvars.R $qcsteps/01d.$base.dupvar.all_trip_ids.txt \
    $qcsteps/01d.$base.dupvar.trips_v_orig_conflicts.txt \
    $qcsteps/01d.$base.dupvar.trips_v_dupes_conflicts.txt \
    $qcsteps/01d.$base.dupvar.trips_dupes_v_orig_conflicts.txt \
    $qcsteps/01d.$base.dupvar.dupes_v_orig_conflicts.txt \
    $qcsteps \
    $base

#Remove markers failing duplicate analysis from filesets

plink2 --bfile $qcsteps/01d.$base.dupvar.non_unique_originals --exclude $qcsteps/01d.$base.dupvar.conflicting_markers.txt --make-bed --out $qcsteps/01d.$base.dupvar.non_unique_originals_no_conflicts

plink2 --bfile $qcsteps/01d.$base.dupvar.non_unique_duplicates --exclude $qcsteps/01d.$base.dupvar.conflicting_markers.txt --make-bed --out $qcsteps/01d.$base.dupvar.non_unique_duplicates_no_conflicts

plink2 --bfile $qcsteps/01d.$base.dupvar.non_unique_triplicates --exclude $qcsteps/01d.$base.dupvar.conflicting_markers.txt --make-bed --out $qcsteps/01d.$base.dupvar.non_unique_triplicates_no_conflicts

#Merge duplicate markers back in to original fileset using default merge-mode

plink2 --bfile $qcsteps/01d.$base.dupvar.non_unique_originals_no_conflicts --allow-no-sex --bmerge $qcsteps/01d.$base.dupvar.non_unique_duplicates_no_conflicts --merge-equal-pos --make-bed --out $qcsteps/01d.$base.dupvar.non_unique_originals_duplicates_no_conflicts

plink2 --bfile $qcsteps/01d.$base.dupvar.non_unique_originals_duplicates_no_conflicts --allow-no-sex --bmerge $qcsteps/01d.$base.dupvar.non_unique_triplicates_no_conflicts --merge-equal-pos --make-bed --out $qcsteps/01d.$base.dupvar.non_unique_no_conflicts

#Merge non-conflicting markers back into the original dataset

plink2 --bfile $qcsteps/01d.$base.dupvar.unique_only --allow-no-sex --bmerge $qcsteps/01d.$base.dupvar.non_unique_no_conflicts --make-bed --out $qcsteps/02.$base

echo -e "\n(3)(4) Update sex and sex check"

echo "--------------------------------"

################################################

# Examine inital gender issues
```

```
plink2 --bfile $qcsteps/02.$base --check-sex 0.3 0.8 --out $qcsteps/02.$base
sexissues=`grep "PROBLEM" qc_steps/02.$base.sexcheck | wc -l`
if (( $sexissues > 0 ))
then
   if [ -e $gender_up ]
   then
        echo -e "\nGender update data given in input - applying these"
        echo "-------------------------------------------"
        ## Prep update files from input
        ### Arguments are: 1) plink.fam, 2) input file with updates (w header), 3) qcsteps folder
        Rscript scripts/update_gender1.R $qcsteps/02.$base.fam $gender_up $qcsteps
        ## Run update
        plink2 --bfile $qcsteps/02.$base --update-sex $qcsteps/02.update_gender.txt  --make-
bed --out $qcsteps/03.$base
        plink2 --bfile $qcsteps/03.$base --check-sex 0.3 0.8 --out $qcsteps/03.$base
   else
        plink2 --bfile $qcsteps/02.$base --make-bed --out $qcsteps/03.$base
        cp $qcsteps/02.$base.sexcheck $qcsteps/03.$base.sexcheck
   fi
   echo -e "\nUpdate unsupplied genders and exclude any remaining sex mismatches"
   echo "-------------------------------------------"
   # Prep update files from sexcheck
   # Arguments are: 1) plink.sexcheck, 2) qcsteps folder
   Rscript scripts/update_gender2.R $qcsteps/03.$base.sexcheck $qcsteps
   ## Run update
   if [ $final_gender_error_exclude = true ] ; then
```

```
        plink2 --bfile $qcsteps/03.$base --update-sex $qcsteps/03.update_gender.txt --remove $qcsteps/03.gender_exclude.txt --make-bed --out $qcsteps/04.$base
    else
        plink2 --bfile $qcsteps/03.$base --update-sex $qcsteps/03.update_gender.txt --make-bed --out $qcsteps/04.$base
    fi
    plink2 --bfile $qcsteps/04.$base --check-sex 0.3 0.8 --out $qcsteps/04.$base
else
    echo -e "\nNo gender issues found!"
    echo "----------------------------------------"
    plink2 --bfile $qcsteps/02.$base --make-bed --out $qcsteps/04.$base
fi
echo -e "\n (2) Heterozygous haploid genotypes"
echo "--------------------------------"
plink2 --bfile $qcsteps/04.$base --set-hh-missing --make-bed --out $qcsteps/05.$base
echo -e "\n (5) IBD"
echo "--------------------------------"
##################################################
## Calculate Indentical by decent matrix using the complete set of samples and common SNPs from gencal
plink2 --allow-no-sex --bfile $qcsteps/05.$base --exclude range $input/high-LD-regions-hg38.txt --maf 0.05 --genome --out $qcsteps/05.$base
echo "finding outliers"
sed -r 's/^\s,//g' $qcsteps/05.$base.genome | sed -r 's/\s+/\t/g' | awk ' ($7< 0.77) ' > $qcsteps/05.$base.genome.outliers
sed -r 's/^\s,//g' $qcsteps/05.$base.genome | sed -r 's/\s+/\t/g' | awk ' ($7> 0.77) ' | head -n1000 > $qcsteps/05.$base.genome.1000normal
echo "deleting genome file"
rm $qcsteps/05.$base.genome
```

```
## List geneaology problems

### Arguments are: 1) outlier file, 2) plink.fam, 3) qcsteps folder

Rscript scripts/list_geneaology_problems.R $qcsteps/05.$base.genome.outliers $qcsteps/05.$base.fam $qcsteps

## Check the three 05.ibd_* files for problem and asses if any action should be done

### If any problems exsist create the file input/ibd_exclude.txt with two columns "IID" "Exclusion reason"

if [ -e $idb_error ]

then

   ## Arguments are: 1) input samples to be removed due to IBD, 2) fam file, 3) output plink exclusion list

   Rscript scripts/fix_geneaology_problems.R $idb_error $qcsteps/05.$base.fam $qcsteps/05.to_exclude.txt

   ## Exclude samples based on relatives and duplicates identified

   plink2 --allow-no-sex --bfile $qcsteps/05.$base --remove $qcsteps/05.to_exclude.txt --make-bed --out $qcsteps/06.$base

   ## Calculate and plot IBD again to check changes

   plink2 --allow-no-sex --bfile $qcsteps/06.$base --exclude range $input/high-LD-regions-hg38.txt --maf 0.05 --genome --out $qcsteps/06.$base

   echo -e "\nFinding outliers"

   sed -r 's/^\s,//g' $qcsteps/06.$base.genome | sed -r 's/\s+/\t/g' | awk ' ($7< 0.77) ' > $qcsteps/06.$base.genome.outliers

   sed -r 's/^\s,//g' $qcsteps/06.$base.genome | sed -r 's/\s+/\t/g' | awk ' ($7> 0.77) ' | head -n1000 > $qcsteps/06.$base.genome.1000normal

   echo "deleting .genome file - less is more"

   rm $qcsteps/06.$base.genome

else

   echo -e "\nNo IBD outlier removed"

   echo "-----------------------------------------"

   plink2 --allow-no-sex --bfile $qcsteps/05.$base --make-bed --out $qcsteps/06.$base
```

```
fi

echo "Update phenotyes"

echo "--------------------------------"

### Arguments are: 1) Family file to update, 2) file to update from (3 col: FID,IID,Newpheno)

if [ -e $pheno_up ]

then

   echo -e "\n Updateting phenotyes"

   Rscript scripts/update_phenotypes.R $qcsteps/06.$base.fam $pheno_up

fi

echo -e "\nCases - Control missigness"

echo "--------------------------------"

### Test for different genotype call rates between

plink2 --bfile $qcsteps/06.$base --allow-no-sex --test-missing --out $qcsteps/06.$base

echo -e "\n(6)(7)(8) Check Ancestory"

echo "--------------------------------"

### Generate symbolic links for hapmap input data

ln -s /home/molpsych/data/hapmap/hapmap3r2_CEU.CHB.JPT.YRI.founders.no-at-cg-snps.lifted.hg38.bed input/

ln -s /home/molpsych/data/hapmap/hapmap3r2_CEU.CHB.JPT.YRI.founders.no-at-cg-snps.lifted.hg38.bim input/

ln -s /home/molpsych/data/hapmap/hapmap3r2_CEU.CHB.JPT.YRI.founders.no-at-cg-snps.lifted.hg38.fam input/

ln -s /home/molpsych/data/hapmap/hapmap_sample_info.txt input/

ln -s /home/molpsych/data/hapmap/ld_regions_hg38.txt input/

#################################################################

## (6) Prune data set

plink2 --allow-no-sex --bfile $qcsteps/06.$base --maf 0.06 --indep 50 5 2 --out $qcsteps/06.$base
```

```
plink2 --allow-no-sex --bfile $qcsteps/06.$base --extract $qcsteps/06.$base.prune.in --make-bed --out $qcsteps/07.$base

## (7) Exclude long-range LD regions and use only SNPs that overlap with hapmap

## This reduced data set has plenty of information to perform PCA and will not be biased by LD or rare SNPs

plink2 --allow-no-sex --bfile $qcsteps/07.$base --exclude range $input/ld_regions_hg38.txt --extract $input/hapmap3r2_CEU.CHB.JPT.YRI.founders.no-at-cg-snps.lifted.hg38.bim --make-bed --out $qcsteps/08.$base

#### First pass of Eigenstrat no removeal of outliers

## Convert to Eigenstrat

convertf -p <(printf "genotypename: $qcsteps/08.$base.bed

snpname: $qcsteps/08.$base.bim

indivname: $qcsteps/08.$base.fam

outputformat: EIGENSTRAT

#line below is required if SNPs are not in chromosomal order. Else change to YES

pordercheck: NO

genotypeoutname: $qcsteps/08.$base.eigenstratgeno

snpoutname: $qcsteps/08.$base.snp

indivoutname: $qcsteps/08.$base.ind")

## Run of Eigenstrat SmartPCA

# Produces 100 PCs

smartpca.perl \

-i $qcsteps/08.$base.eigenstratgeno \

-a $qcsteps/08.$base.snp \

-b $qcsteps/08.$base.ind \

-o $qcsteps/08.$base.pca \

-p $qcsteps/08.$base.plot \

-e $qcsteps/08.$base.eval \
```

```
-l $qcsteps/08.$base.smartpca.log \

-m 0 \

-t 100 \

-k 100 \

-s 6

## Check for assocation between PCA and case/ctrl status

###  Minor edit to allow import into R

sed -i -e 's/^[ \t]*//' -e 's/:/ /g' $qcsteps/08.$base.pca.evec

Rscript scripts/analyse_pca.evac.R $qcsteps/08.$base.pca.evec

### Second pass of Eigenstrat - Removal of outliers

## NB: -t is set to be equalt to the number of significant PCs found in $qcsteps/08.$base.pca.evec.assoc

nsignif=`cat $qcsteps/08.$base.pca.evec.assoc.nsignif`

echo -e "\nEigenstrat will remove outliers based on $nsignif principal components!"

echo "----------------------------------------------------------"

smartpca.perl \

-i $qcsteps/08.$base.eigenstratgeno \

-a $qcsteps/08.$base.snp \

-b $qcsteps/08.$base.ind \

-o $qcsteps/08.$base.pca_outlier.pca \

-p $qcsteps/08.$base.pca_outlier.plot \

-e $qcsteps/08.$base.pca_outlier.eval \

-l $qcsteps/08.$base.pca_outlier.smartpca.log \

-m 5 \

-t $nsignif \

-k 100 \

-s 6
```

```
## Check for assocation between PCA and case/ctrl status

sed -i -e 's/^[ \t]*//' -e 's/:/ /g' $qcsteps/08.$base.pca_outlier.pca.evec

Rscript scripts/analyse_pca.evac.R $qcsteps/08.$base.pca_outlier.pca.evec

## Create list of samples to remove.

awk '/REMOVED/ {print $3}' $qcsteps/08.$base.pca_outlier.smartpca.log | sed 's/:/ /g' > qc_steps/08.pca_outliers.txt

#### (8) Remove samples and rerun SmartPCA

### Remove samples

plink2 --allow-no-sex --bfile $qcsteps/08.$base --remove $qcsteps/08.pca_outliers.txt --make-bed --out $qcsteps/09.$base

#### Third pass of Eigenstrat - Calculate PCA for covariates

convertf -p <(printf "genotypename: $qcsteps/09.$base.bed

snpname: $qcsteps/09.$base.bim

indivname: $qcsteps/09.$base.fam

outputformat: EIGENSTRAT

#line below is required if SNPs are not in chromosomal order. Else change to YES

pordercheck: NO

genotypeoutname: $qcsteps/09.$base.eigenstratgeno

snpoutname: $qcsteps/09.$base.snp

indivoutname: $qcsteps/09.$base.ind")

smartpca.perl \

-i $qcsteps/09.$base.eigenstratgeno \

-a $qcsteps/09.$base.snp \

-b $qcsteps/09.$base.ind \

-o $qcsteps/09.$base.covariates.pca \

-p $qcsteps/09.$base.covariates.plot \

-e $qcsteps/09.$base.covariates.eval \
```

```
-l $qcsteps/09.$base.covariates.smartpca.log \

-m 0 \

-t 0 \

-k 100 \

-s 6 \

## Check for assocation between PCA and case/ctrl status

sed -i -e 's/^[ \t]*//' -e 's/:/ /g'  $qcsteps/09.$base.covariates.pca.evec

Rscript scripts/analyse_pca.evac.R $qcsteps/09.$base.covariates.pca.evec

## Eigenstrat With HAPMAP - For plot of foreign samples

#########################################################################
######

## (9) Merge the pruned, no long ld, maf > 0.05, overlap with hapmap SNP data set with the
hapmap data set

## !OBS FAILS as multiallelic or FLIPED SNPs are present!! - but produces a list of merge-
missnp that we use for exclusion in next step.

plink2 --allow-no-sex --bfile $qcsteps/08.$base --bmerge
$input/hapmap3r2_CEU.CHB.JPT.YRI.founders.no-at-cg-snps.lifted.hg38.bed \

    $input/hapmap3r2_CEU.CHB.JPT.YRI.founders.no-at-cg-snps.lifted.hg38.bim \

    $input/hapmap3r2_CEU.CHB.JPT.YRI.founders.no-at-cg-snps.lifted.hg38.fam \

    --extract $qcsteps/08.$base.bim --make-bed --out $qcsteps/10.$base


## (9) Remove SNPs block the merging becouse 3 or more alleles are found - i simply exclude
these here (is minor issue) but this should propberbly be checked!!

plink2 --allow-no-sex --bfile $qcsteps/08.$base --exclude $qcsteps/10.$base-merge.missnp --
make-bed --out $qcsteps/10.$base

## (10) Merge our data with the hapmap data set second time - Sucesses

plink2 --allow-no-sex --bfile $qcsteps/10.$base --bmerge
$input/hapmap3r2_CEU.CHB.JPT.YRI.founders.no-at-cg-snps.lifted.hg38.bed \

    $input/hapmap3r2_CEU.CHB.JPT.YRI.founders.no-at-cg-snps.lifted.hg38.bim \
```

```
    $input/hapmap3r2_CEU.CHB.JPT.YRI.founders.no-at-cg-snps.lifted.hg38.fam \
    --extract $qcsteps/10.$base.bim --make-bed --out $qcsteps/11.$base
### Calculate PCA with Eigensoft
## convert to eigensoft
convertf -p <(printf "genotypename: $qcsteps/11.$base.bed
snpname: $qcsteps/11.$base.bim
indivname: $qcsteps/11.$base.fam
outputformat: EIGENSTRAT
#line below is required if SNPs are not in chromosomal order. Else change to YES
pordercheck: NO
genotypeoutname: $qcsteps/11.$base.eigenstratgeno
snpoutname: $qcsteps/11.$base.snp
indivoutname: $qcsteps/11.$base.ind")
smartpca.perl \
-i $qcsteps/11.$base.eigenstratgeno \
-a $qcsteps/11.$base.snp \
-b $qcsteps/11.$base.ind \
-o $qcsteps/11.$base.hapmap.pca \
-p $qcsteps/11.$base.hapmap.plot \
-e $qcsteps/11.$base.hapmap.eval \
-l $qcsteps/11.$base.hapmap.smartpca.log \
-m 0 \
-t 3 \
-k 100 \
-s 6 \
# Convert format to fit R
```

```
sed -i -e 's/^[ \t]*//' -e 's/:/ /g'  $qcsteps/11.$base.hapmap.pca.evec

### (12) Identify SNPs failing cluster inspections - with R script

## Arguments are: 1) base name of project, 2) qc_steps folder

# Rscript scripts/snps_failing_cluster_inspection.R $base $qcsteps

echo "Run QC results generation"

echo "--------------------------------"

### Run reporter script

## Arguments are: 1) qc_steps folder, 2) results folder, 3) base 4) hwe p-value threshold 5) hwe p-value threshold chrX 6) case-ctrl miss p-value threshold 7) heterozyogisity sd threshold

# 8) shuld hetero sd threshold only be used on upper 9) missing_ness max 10) snp_missing_max 11) exclude samples with final gender errors

Rscript scripts/reporter.R $qcsteps $results $base $hwe_threshold $hwe_threshold_chrX $miss_threshold $het_sd $het_upper_only $miss_max $snp_miss_max $final_gender_error_exclude

echo "Exclude based on QC results and generate final QC dataset "

echo "--------------------------------"

plink2 --allow-no-sex --bfile $qcsteps/06.$base --exclude $results/snp_failing_qc.txt --remove $results/id_failing_qc.txt --make-bed --out $results/$base.qc

echo "Create markdown report"

echo "--------------------------------"

echo -e "base <- '$2' \nprojecttitle <- '$project_title'\nchiptype <- '$chiptype'\ngender_error_exclude <- '$final_gender_error_exclude'" >> input/markdown_args.R

Rscript -e "rmarkdown::render('scripts/markdown.Rmd', output_file = '$project_title.qc-report.html', output_dir = 'results')"

echo "END OF SCRIPT"
```

9.7 vcf_to_plink.sh

```
### Script to convert imputed vcf data into stricly biallelic plink format

#### Commands for this are fully stolen from: http://apol1.blogspot.co.uk/2014/11/best-practice-for-converting-vcf-files.html

## Check argumets

if [ "$#" -ne 2 ]; then

   echo -e "\n VCF to Plink - Converts imputed vcf files to plink binary fixing all mulitalleic sites, keeping only info > 0.9"

   echo -e "\n Run after snp_info is generated!"

   echo "------------------------------------------------------------------------"

   echo -e "\n Please provide 2 argument: \n 1) Filepath to imputed data, n\ 2) Filepath to QCed fam file \n\nExample: vcf_to_plink /home/molpsych/snp_genotypes/psychchip/imputed /home/molpsych/snp_genotypes/psychchip/gwas_qc_report/results/pchip_sczbp.qc.fam\n
"

   exit 1

fi

filepath=$1

qcfam=$2

if [ -e $filepath/info/1.snpinfo ]

   then

   ### Exclude all SNPs with INFO < 0.9 and update SNP names where possible and update gender and phenotypes

   awk '{print $2,$2,$5}' $qcfam > $filepath/temp_sex_update.txt

   awk '{print $2,$2,$6}' $qcfam > $filepath/temp_pheno_update.txt

   ## Converte imputed vcf files to plink binary files

   for i in {1..22}

   do

        # Produce second time with vcf to allow vcf-min-gp filtering o f0.9
```

```
        plink2 --vcf $filepath/$i.vcf.gz \
          --keep-allele-order \
          --vcf-idspace-to _ \
          --vcf-min-gp 0.9 \
          --biallelic-only strict \
          --const-fid \
          --allow-extra-chr 0 \
          --split-x b37 no-fail \
          --make-bed \
          --out $filepath/temp_$i
        ## Update RS numbers, 2 arguments: 1) name of bim file 2) chromosome number
        Rscript /home/molpsych/programs/molpsych_toolkit/scripts/update_snps_fromInfo.R $filepath temp_$i.bim $i
        mv $filepath/temp_$i.bim.new $filepath/temp_$i.bim
        ## Update files
        awk '$3>=0.9 {print $filepath}' $filepath/info/$i.snpinfo > $filepath/temp_info_pass$i.txt
        awk '$3>=0.9 {print $filepath,$2}' $filepath/info/$i.snpinfo | awk '$2!="."' > $filepath/temp_info_pass_snpName$i.txt
        plink2 --bfile $filepath/temp_$i --extract $filepath/temp_info_pass$i.txt --make-bed --out $filepath/temp2_$i
        awk '{print $2,$2,$3,$4,$5,$6}' $filepath/temp2_$i.fam > $filepath/temp2_$i.fam.new
        mv $filepath/temp2_$i.fam.new $filepath/temp2_$i.fam
        plink2 --bfile $filepath/temp2_$i --allow-no-sex --pheno $filepath/temp_pheno_update.txt --update-sex $filepath/temp_sex_update.txt --update-name $filepath/temp_info_pass_snpName$i.txt --make-bed --out $filepath/$i
    done
```

```
# rm -f $filepath/temp*

# # ### Only for psychchip - Create subset that only contains Controls

# for i in {1..22}

# do

# plink2 --bfile $filepath/$i --keep /home/molpsych/data/pheno/diagnosis_all_chips_controls.txt --make-bed --out $filepath/$i.ctrl

# done


echo "--------------------------------------------"

echo "Plink binary files have been writen to the input folder"

echo -e "\nEnd of script\n"

else

echo -e "\n$filepath/info/1.snpinfo does not exist"

echo -e "\n\nPlease generate SNP info before running VCF to Plink - info is needed to determine which SNPs are to be included in plink set!\n\n"

fi

	# ####### THIS BLOCK IS NOT IN USE ANY MORE AS IT DOES NOT ALLOW FOR GENOTYPE PROBABILITY FILTERING - NAMING is now done with R script

	## Produce file first with bcftools to get correct file names

	# bcftools norm -Ou -m -any $filepath/$i.vcf.gz |

	# bcftools norm -Ou -f /home/molpsych/data/GRCh37_reference/human_g1k_v37.fasta |

	# bcftools annotate -Ob -x ID \

	#    -I +'%CHROM:%POS:%REF:%ALT' | \

	# plink2 --bcf /dev/stdin \

	#    --keep-allele-order \
```

```
#    --vcf-idspace-to _ \
#    --biallelic-only strict \
#    --const-fid \
#    --allow-extra-chr 0 \
#    --split-x b37 no-fail \
#    --make-bed \
#    --out $filepath/temp_names$i
```

9.8 ID_PRSice_script.txt

```
## PRSice is better with -9/NA (check whether you are using binary or quantatitive phenotype) for missing data

Rscript /home/molpsych/programs/PRSice2/PRSice.R \
    --prsice /home/molpsych/programs/PRSice2/PRSice_linux \
    --bar-levels 5e-08,1e-07,1e-06,1e-05,1e-04,1e-3,0.01,0.05,0.5,1 \
    --base /mnt/10tbstore/projects/alc_symptom_score_gwas/ASPD/PRS/DONE/adhd_eur_jun2017/adhd_eur_jun2017 \
    --clump-kb 250 \
    --clump-p 1.000000 \
    --clump-r2 0.100000 \
    --fastscore \
    --score std \
    --snp SNP \
    --pvalue P \
    --A1 A1 \
    --A2 A2 \
    --bp BP \
```

```
    --chr CHR \
    --stat OR \
    --cov final.pca.txt \
    --cov-col PCA1,PCA2,PCA3 \
    --pheno /mnt/10tbstore/projects/ava/ID_PRSice/ADHD/phenotype_ID_200720.txt \
    --pheno-col ADHD \
    --binary-target T \
    --base-info INFO:0.9 \
    --interval 0.000000050 \
    --lower 5e-08 \
    --model add \
    --out PRSice_ADHD_260720 \
    --seed 3228425514 \
    --target /mnt/10tbstore/projects/ava/psychchip_ID_imputed/psychchip_ID_qc_info0.8 \
        --ignore-fid T
```

9.9 ID_PRSice_script_anxiety.txt

```
## PRSice is better with -9/NA (check whether you are using binary or quantatitive phenotype) for missing data
Rscript /home/molpsych/programs/PRSice2/PRSice.R \
    --prsice /home/molpsych/programs/PRSice2/PRSice_linux \
    --bar-levels 5e-08,1e-07,1e-06,1e-05,1e-04,1e-3,0.01,0.05,0.5,1 \
    --base /mnt/10tbstore/projects/ava/ID_PRSice/anxiety/20002_1287.gwas.imputed_v3.both_sexes.tsv.gz.correct \
    --clump-kb 250 \
    --clump-p 1.000000 \
```

```
--clump-r2 0.100000 \
--fastscore \
--score std \
--snp CHR:BP \
--pvalue pval \
--A1 A1 \
--A2 A2 \
--stat beta \
--beta \
--cov final.pca.txt \
--cov-col PCA1,PCA2,PCA3 \
--pheno /mnt/10tbstore/projects/ava/ID_PRSice/anxiety/phenotype_ID_200720.txt \
--pheno-col Anxiety \
--binary-target T \
--base-info INFO:0.9 \
--interval 0.000000050 \
--lower 5e-08 \
--model add \
--out PRSice_Anxiety_300720 \
--seed 3228425514 \
--target /mnt/10tbstore/projects/ava/ID_PRSice/anxiety//psychchip_ID_qc_info0.8.new \
--extract PRSice_Anxiety_300720.valid \
   --ignore-fid T
```

9.10 Ben_Neale_script.txt

```
### Sort out Ben Neale UKB columns for PRSice
```

```
cp 20002_1287.gwas.imputed_v3.both_sexes.tsv.bgz temp

sed  "s/^variant/CHR:BP:A1:A2/g" temp > temp1

awk -F '[:]' '{ print $0, $1," ",$2," ",$3," ",$4}' temp1 | awk '{print$13":"$14,$15,$16,$1,$2,$3,$4,$5,$6,$7,$8,$9,$10,$11,$12}' > temp2

mv temp2 20002_1287.gwas.imputed_v3.both_sexes.tsv.bgz.correct

## Sort out BIM file

head psychchip_ID_ctrl_merge.qc.bim > temp

awk '{print$1,$1":"$4,$3,$4,$5,$6}' temp > temp1

#cp temp1 psychchip_ID_qc_info0.8.bim
```