A Genome-wide Association Study of Schizophrenia in the South African Xhosa and Generalizability of Polygenic Risk Score across African populations

Lerato Charlotte Majara, M.Sc. (Med)

This dissertation is submitted in fulfilment of the requirements for the degree of Doctor of Philosophy in the Division of Human Genetics, Department of Pathology in the Faculty of Health Sciences at the University Of Cape Town

Supervisor: Prof Raj Ramesar

Co-supervisors: Prof Dan Stein & Prof Emile Chimusa

March 2021

Plagiarism Declaration

I, LERATO CHARLOTTE MAJARA, hereby declare that the work on which this thesis is based is my original work (except where acknowledgements indicate otherwise) and that neither the whole work nor any part of it has been, is being, or is to be submitted for another degree in this or any other university. I authorise the University to reproduce for the purpose of research either the whole or any portion of the contents in any manner whatsoever.

The work done in chapter 4 of this thesis has resulted in a publication that is currently in preprint (https://www.biorxiv.org/content/10.1101/2021.01.12.426453v1)

Signature: Signed by candidate

Date: 05 March 2021

Acknowledgements

SAX Study participants: Without you, this study would not be possible.

Medical Research Council South Africa: Bongani Mayosi National Health Scholars Programme for funding.

Supervisors: Professors Ramesar, Stein and Chimusa; thank you for affording me the opportunity to do this PhD, and lending your expertise.

My mom and sister: Thank you for being my biggest cheerleaders.

To the friends that completed their PhDs (Drs Lorna Gcanga and Stacey Moses), I am so inspired by your strength and tenacity. Things were not always easy, but you soldiered on. You are true exemplars of work-life balance.

To the friends who are still on the journey (Anathi Nkayi and Amanda Gcanga), I am grateful for the moral support you've provided. Being able to share PhD woes with you kept me sane. Wishing all the best.

Clement Tawanda Nhunzvi: We started this PhD journey around the same time. I am grateful for the encouragement and support you provided throughout. The daydreams about post-PhD life were important in helping to see this through.

Alicia Martin: I feel incredibly lucky to have worked with you for the past two years, and look forward to continue working with you going into the future. You've become a big part of my academic journey. I am so grateful for all that I have learned from you: from the hard to the soft skills. Your patience is unrivalled - I don't know how you do it. With you, I have felt safe to make mistakes and learn from them. I look up to you in so many ways. I'd be privileged if I became half the mentor that you are.

The GINGER program: Things happen just as they are meant to. I have learned so much through this programme. I'm eternally grateful to Lori Chibnik and Bizu Gelaye for regularly checking in, and for the support. A big thank you to Kristi Post, Wairimu Mwaura and Courtney White for all your efforts. You make/made running a programme like GINGER look easy.

Table of Contents

List of figures

List of tables

Abbreviations

°	degrees Celsius
µg	microgram
µl	microlitre
1KGP3	1000 Genomes Project Phase 3
AFR	African
AGVP	African Genome Variation Project
AMR	Admixed American
ASD	autism spectrum disorder
BBJ	Biobank Japan
BDNF	brain-derived neurotrophic factor
bp	base pair
CI	confidence interval
CNS	central nervous system
COMT	catechol-o-methyltransferase
CPGR	Centre for Proteomic and Genomics Research
CSA	Central South Asia
CT	childhood trauma
CTQ	Childhood Trauma Questionnaire
DALY	disability-adjusted life years
DCHS	Drakenstein Child Health Study
DNA	deoxyribonucleic acid
DQC	dish quality control
DSM	diagnostic and statistical manual of mental disorders
EAS	East Asian
EDTA	Ethylenediaminetetraacetic acid
EUR	European
FDR	false discovery rate
FRS	first rank symptoms
FUMA	Functional Mapping and Annotation of Genome-Wide Association Studies
g	gravity
GINGER	Global Initiative for Neuropsychiatric Genetics Education and Research
GPC	General Population Cohort
GRM	genotypic relatedness matrix
GWAS	genome-wide association study
GWS	genome-wide significance
HGDP	Human Genome Diversity Project
HWE	Hardy-Weinberg Equilibrium
IBD	identity by descent
ICD	International Classification of Diseases
ID	intellectual disability
ISC	International Schizophrenia Consortium

K	thousand
kb	kilobase
kya	thousand years ago
LD	linkage disequilibrium
LoF	loss of function
M	Molar
MAD	median absolute deviation
MAF	minor allele frequency
Mb	megabases
MC	multi-channel
MEGA	Multi-Ethnic Global Array
Met	methionine
MHC	Major histocompatibility complex
MID	Middle Eastern
min	minutes
ml	millilitre
NaCl	sodium chloride
NeuroGAP	Neuropsychiatric Genetics in African Populations
NMDAR	n-methyl-d-aspartate receptor
OR	odds ratio
PAGE	Population Architecture using Genomics and Epidemiology
PC	principal component
PCA	principal component analyses
PCR	polymerase chain reaction
PGC	Psychiatric Genomics Consortium
PRS	polygenic risk score(s)
Q-Q	quantile-quantile
QC	quality control
QOL	quality of life
RA	relative accuracy
REML	restricted maximum likelihood
rpm	revolutions per minute
SAX	South African Xhosa
SCID	Structural and Clinical Interview for Diagnostic and Statistical Methods
SDS	sodium dodecyl sulphate
SE	South East
SNP	single nucleotide polymorphism
SW	South West
TAD	topologically associated domains
TE	tris EDTA
UBACC	University of California, San Diego Brief Assessment of Capacity to Consent
UCT	University of Cape Town
UKBB	United Kingdom Biobank

USA	United States of America
Val	valine
VCF	variant call format
WES	whole exome sequencing
WGS	whole genome sequencing
WHO	World Health Organization

Abstract

African populations are vastly underrepresented in genetic studies despite having the most genetic variation globally and facing wide-ranging environmental exposures. Most of these studies have been conducted in populations of European (EUR) ancestry using GWAS arrays that represent the genetic variation in these populations. Thus, the prediction accuracy of polygenic risk scores (PRS) derived from EUR ancestry populations is less accurate in populations of non-European ancestry, and least accurate in African (AFR) ancestry populations. The extent to which PRS prediction accuracy varies within AFR ancestry populations has not, however, been previously investigated.

This study had two aims: the first was to investigate the contribution of common variants to the risk of schizophrenia in the South African Xhosa (SAX) population through genome-wide association study (GWAS) analysis, and to determine if PRS derived from EUR and East Asian (EAS) ancestry populations from the Psychiatric Genomics Consortium (PGC) Schizophrenia Working Group were generalizable to SAX. The second aim was to assess the generalizability of PRS for non-psychiatric phenotypes that were derived from EUR ancestry individuals from the UK Biobank (UKB, n = ~350,000) in the Uganda General Population Cohort (GPC, n = 4,778) and the South African Drakenstein Child Health Study (DHCS, n = 638).

To address the first aim, a GWAS was conducted in 2,086 Xhosa individuals from South Africa with and without schizophrenia (n_{cases} = 1,038; $n_{controls}$ = 1,048) using a custom-designed Affymetrix GWAS array designed to capture variation in the Xhosa population. The schizophrenia GWAS in SAX yielded one SNP (rs35172303 ; *P* = 4.74e-08, OR = 0.6004, 95%CI:[0.499,0.721]) in *ZFP3* that met genome-wide significance. The association of variants in *ZFP3* from the schizophrenia GWAS is consistent with those from an earlier exome-sequence study in SAX undertaken by colleagues, but this gene has not previously been associated with schizophrenia in large-scale schizophrenia GWAS of predominantly EUR ancestry.

After characterizing the genetic architecture of schizophrenia in SAX, it was found that the heritability was enriched across functional categories involved in the regulation of gene expression. Then, the accuracy of PRS derived from PGC Schizophrenia Working Group from both EUR and EAS ancestries in predicting schizophrenia in SAX was quantified. There was low PRS prediction accuracy using PGC-derived summary statistics in SAX (PGC-EUR: max R^2 = 0.0057, *P* = 0.008; PGC-EAS: max R^2 = 0.0059, *P* = 0.007). These findings are consistent with previous findings that showed that PRS predication accuracy is low when discovery and target cohorts come from different ancestral backgrounds.

For the second aim, PRS prediction accuracy was quantified in simulations using data from the African Genome Variation project (AGVP) to represent continental AFR diversity. Samples were categorised by geographical region into West, East and South Africa cohorts. Each cohort was divided into a discovery and target datasets. The West and East African discovery data was used to predict the simulated phenotype in the three target cohorts. Using UKB EUR ancestry individuals, PRS prediction accuracy was assessed for 34 anthropometric and blood panel traits in the Uganda GPC, and then meta-analysed UKB with PAGE (Population Architecture using Genomics and Epidemiology, comprising about 50,000 Latino/Hispanic and African-American individuals) and BBJ (Biobank Japan, n = ~162,000) to assess how the inclusion of diverse sample impacts PRS prediction accuracy.

Simulations were limited by sample size but showed that PRS prediction accuracy was highest when the discovery and target cohorts were matched by African region, and for phenotypes with the sparsest genetic architecture. Using empirical data from UKB and the Uganda GPC, a low prediction accuracy was observed across all 34 quantitative traits in GPC when using GWAS data from UKB. There was differential prediction accuracy across AFR ancestry groups within UKB, i.e. the prediction accuracy was highest for the Ethiopian and admixed populations, and lowest for southern African populations. When comparing PRS prediction accuracy of East African individuals from the UKB to that of individuals from GPC, the prediction accuracy was lowest in the Ugandan GPC population, indicating that the difference in environments between the two groups may be contributing to the difference in PRS accuracy. Moreover, the cross-ancestry meta-analyses showed that the inclusion of diverse samples in large scale studies improves PRS prediction accuracy, most especially for phenotypes with population-enriched variants.

It was demonstrated for the first time in this thesis that EUR ancestry-derived PRS prediction accuracy varied within continental AFR ancestry groups, and tracks with population history and the evolution of humans. The higher prediction accuracy observed in Ethiopians can be explained by their genetic proximity to Europeans as a result of the back to Africa migration, whereas the southern African populations (including SAX) are more proximal to the ancestral populations that never left the continent. It is therefore imperative to not only include more African samples in future large-scale studies, but to have samples that adequately represent the genetic and environmental diversity on the African continent.

Preface

The majority of genetics studies are conducted in populations of European ancestry. In this thesis, the candidate investigates the genetics of schizophrenia in the indigenous South African Xhosa cohort, and explores how well polygenic risk scores computed from European and East Asian ancestry populations predict schizophrenia, psychological and anthropometric traits in and across various African populations.

The thesis comprises five chapters. Chapter 1 is a review of the genetic basis for schizophrenia and the findings from genetic studies. The chapter is concluded by highlighting the need to include African ancestry participants in genetics studies, and the anticipated benefits that may result.

Chapter 2 represents an investigation of the genetics of schizophrenia in the SAX cohort using genome-wide association study (GWAS) analyses. In chapter 3, the distribution of heritability of schizophrenia in SAX is explored by partitioning the genome into its constituent chromosomes and 24 functional categories, as well as minor allele frequency bins. Further, the prediction accuracy of polygenic risk scores derived from the European and East Asian populations from the Psychiatric Genomics Consortium is explored.

Given that no large genetics datasets exist for psychiatric disorders in African populations, the candidate uses publicly available datasets for non-psychiatric traits including 34 anthropometric blood panel traits to assess the transferability of European ancestry-derived polygenic risk scores in and across diverse African populations. This work is covered in chapter 4.

Chapter 5 is a general discussion summarizing findings from the thesis, noting the limitations of the thesis and providing considerations for future research.

Chapter 1: Introduction

1.1 Abstract

This chapter is a review of the epidemiological background of schizophrenia. The societal and economic impact are briefly discussed. Next, the review focusses on the risk factors for schizophrenia from the perspective of both genetic (according to family-based studies) and environmental factors (according to epidemiological studies), including the proposed mechanisms by which disease is caused. Moreover, a historic perspective of the diagnostic criteria that are currently in use is provided. This is proceeded by a review of genetic studies that have been conducted to date, most of which have been conducted in populations of European ancestry. Specific gaps in the research are illuminated; and lastly, a case for the need to include diverse populations in large-scale genomics research is made.

1.2 Background

Mental disorders, such as schizophrenia and major depression, contribute significantly to the rise in the global burden of disease. According to the global burden of disease report, mental disorders account for 6% of the global disease burden, and are responsible for over 50% of the number of disability adjusted life-years (DALYs) — which are a measure of reduction in life expectancy, where one DALY is the equivalent of one healthy life year lost to disease (Vos et al., 2020). Schizophrenia is listed among the most important mental disorders contributing to DALYs (Global Health Estimates, 2016). In South Africa, the burden of disease due to mental disorders is similarly high (Whiteford et al., 2015), with a representative survey indicating that mental disorders have a lifetime prevalence of 30.3% (Herman et al., 2009). To reduce this impact, it imperative that focus is placed on mental health research to better understand the biology of these conditions in order to develop effective therapies and better management strategies.

1.3 Epidemiology and burden of schizophrenia

The worldwide lifetime prevalence of schizophrenia appears to be about 0.5%, although most relevant data are from high-income countries (Perälä et al., 2007). Although the prevalence is relatively low, schizophrenia accounts for great personal, societal, and economic burden. Indeed, schizophrenia is among the top 25 leading causes of disability around the world (Vos et al., 2020).

Schizophrenia affects people between the ages of 15 and 49 years, although the disorder may occur in early childhood (Driver et al., 2013). Affected individuals may present with impaired cognition; positive symptoms such as hallucinations, delusions, and disorganized speech; or negative symptoms such as amotivation, apathy and diminished speech (Andreasen, 1982; Andreasen et al., 1995; Kay et al., 1987). The presentation of symptoms differs between affected individuals.

Schizophrenia is associated with high rates of mortality. The mortality rate of people with schizophrenia is two- to three-fold higher than the general population, due in part to high suicide rates among this cohort (Hayes et al., 2017; Saha et al., 2007). Under-diagnosis or undertreatment also leads to the high rates of mortality observed in people with schizophrenia (Kugathasan et al., 2019). Additionally, there is high comorbidity with chronic disorders such as cardiometabolic disease, stroke, type II diabetes and cancers which may occur as a result of poor lifestyle choices in schizophrenia patients, such as excessive drinking of alcohol, poor diet, and inadequate physical activity, as well as side effects from medications (Jayatilleke et al., 2017).

Individuals with schizophrenia have a poor quality of life (QOL) (Hayes et al., 2015), with factors such as stigma contributing significantly to poor QOL (Angermeyer et al., 2015). Stigmatization is mainly characterized by stereotypes of dangerousness that are rooted in the lack of awareness or misinformation about mental disorders in the public. Patients with schizophrenia tend to feel socially isolated, generally have low self-esteem and may experience feelings of depression (Crisp et al., 2000; Pellet et al., 2019; Zäske et al., 2019). Self-stigmatization, which relates to the patients' reaction to stigma (Gerlinger et al., 2013), leads to patients having a negative attitude towards their medication, this in turn may negatively affect their clinical outcome (Corrigan et al., 2014; Maeng et al., 2016). Unfortunately, stigma is not diminishing despite interventions such as anti-stigma and awareness campaigns (Gronholm et al., 2017; Schomerus et al., 2012).

The management and treatment of schizophrenia places huge economic burden on health systems, especially in resource-limited countries (Awad & Voruganti, 2016). This burden can be divided into direct and indirect costs. The associated direct costs include clinic visits, hospitalization, procedures and medications, as well as non-medical costs such as the transportation required to get to healthcare facilities, among others. The indirect costs include loss of income due to patient losing employment due to reduced or non- productivity, a caregiver's unpaid time taking care of the schizophrenia patient, costs incurred by disruptive behaviour such as damage to property, as well as medical costs for any physical injury. The leading indirect cost of schizophrenia in many countries around the world is loss of income due to unemployment (Cloutier et al., 2016; Marcellusi et al., 2018; Oloniniyi et al., 2019; Teoh et al., 2017; Xu et al., 2019).

The tremendous burden of disease associated with schizophrenia indicates that concerted effort is required to provide better healthcare for patients to alleviate the pressure on families and society and provide better quality of life for patients. There is a critical need to coordinate physical and mental health services to close the mortality gap between schizophrenia patients and the general population.

1.4 Risk factors for schizophrenia

Schizophrenia has a heritability rate of up to 80%, suggesting that genetic factors are a strong contributor to the development of the disorder (Sullivan et al., 2003). The remaining 20% of heritability may be attributed to many factors including those from the environment (Brown, 2011; van Os et al., 2010). Below, individual genetic and environmental factors that contribute to the risk of schizophrenia are reviewed.

1.4.1 Genetic factors

Family studies show that an offspring born to parents with schizophrenia has an elevated risk of developing schizophrenia (Ellersgaard et al., 2018; Rasic et al., 2014) as well as other psychiatric disorders (Mortensen et al., 2010; Rasic et al., 2014; Sanchez-Gistau et al., 2015). This indicates that the genetic factors contribute substantially to the risk of developing schizophrenia. The risk of developing schizophrenia as a function of being related to an individual affected with schizophrenia is depicted in **Figure 1.1**.

The Danish adoption register has served as a valuable resource for assessing the influence of genetic and/or environmental factors on schizophrenia (Petersen & Sørensen, 2011). In adoption studies – where there is little genetic similarity between the adopted child and the adoptive family, but a shared and similar environment – a higher prevalence of schizophrenia was observed if the child came from a biological family with schizophrenia (Peterson & Sørensen, 2011; Heston, 1966; Higgins et al., 1997; 2011; Tienari, 1991; Tienari et al., 1994; Wender et al., 1974). Conversely, the prevalence rate was lower in adopted children if the adoptive family had schizophrenia, but the biological family of the child did not.

Twin studies investigating concordance rate, that is the likelihood of one twin developing the disorder if the first twin examined had it, show that the concordance rate in monozygotic twins is between 40 - 60% (Cannon et al., 1998; Cardno et al., 1999; Franzek & Beckmann, 1998; Mortensen & Kyvik, 1996; Sullivan et al., 2003; Tsujita et al., 1992). In dizygotic twins, the concordance rate is between 0 - 28%. Together, these studies have demonstrated the strong genetic liability to schizophrenia, and therefore studying the underlying genetic mechanism of the disorder may help better understand the disease aetiology.

1.4.2 Environmental factors

There is a gap between the heritability of schizophrenia and the total variance in disease liability which may be attributed to environmental factors. Growing evidence that environmental factors play a crucial role is based on cohort studies and the improvements in methodologies to systematically collect data and test environmental factors.

The neurodevelopmental hypothesis of schizophrenia indicates that the abnormal development of the nervous system occurs during the early stages of development. This has been supported by findings from studies that show impairment of neurocognition and neuromotor function, physical anomalies of the craniofacial area, and morphological brain abnormalities in MRI studies at the onset of the disorder. In addition, the evidence suggests that pre- and perinatal factors contribute to the pathophysiology. Several such factors, including those that occur during childhood and the adolescence stages are outlined in **Table 1.1** below.

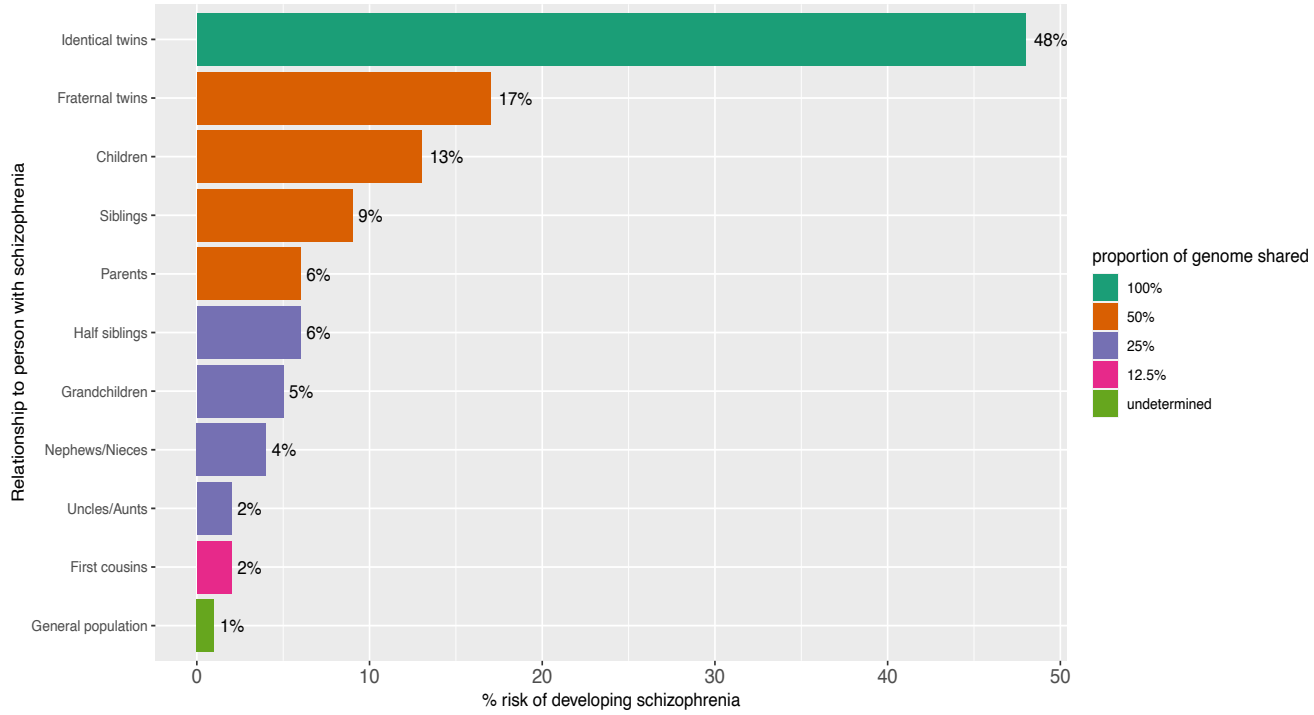


Figure 1.1 - A figure showing the percentage likelihood of an individual developing schizophrenia by the degree of relatedness to a person with schizophrenia.
The bars are coloured by the genes shared with the schizophrenic proband. The figure was adopted from (Gottesman, 1991)

Table 1.1 - Environmental risk factors that are associated with schizophrenia during different stages of life
The 'proposed mechanism' column for each of the risk factors provides some examples of how exposure to the risk factors may lead to schizophrenia, and is not an exhaustive list.

Stage of life	Risk factor	Proposed mechanism(s)	Citations
Pre- and perinatal	*In-utero* exposure to infections such as Rubella, Herpes Simplex Virus, Influenza and *Toxoplasma gondii*	Infection induces an inflammatory immune response. Over stimulation of the inflammatory response results in an excess of proinflammatory cytokines. These cytokines may: stimulate microglia and astroglia to produce nitric oxide excitatory amino acids, which are toxic to neurons; disturb the maturation of oligodendrocytes causing abnormalities in white matter; lead to brain anomalies thereby increasing the risk of developing neuropsychiatric disorders such as schizophrenia. *Toxoplasma gondii* causes congenital anomalies in the central nervous system	(Brown et al., 2001; Burgdorf et al., 2019; Dickerson et al., 2020; Esshili et al., 2020; Kępińska et al., 2020; Lesh et al., 2018; Wang et al., 2020)

	Maternal stress	Several factors that lead to maternal stress during pregnancy have been reported, such as unwanted pregnancy and grief during pregnancy Stress may lead to low birthweight in the offspring — a known risk factor for schizophrenia. Elevate cortisol is associated with diminished neuromuscular maturation.	(Abel et al., 2010; Dorrington et al., 2014; Ellman et al., 2019; Khashan et al., 2008; Pugliese et al., 2019; Van den Bergh et al., 2017; Weinstein et al., 2018)
	Advanced paternal age	The rate of replication of the spermatogonial stem is accelerated in older males. This results in the rapid accumulation of *de novo* mutations, which may contribute to the risk of schizophrenia. Offspring inherits genome with epigenetic modification through 'imprinting'. These	(Denomme et al., 2020; Fond et al., 2017; Fountoulakis et al., 2018; Gratten et al., 2016; Lan et al., 2020)

		epigenetic modifications, such DNA methylation, silence genes by inhibiting transcription.	
	Nutritional deficiency such as Iron and Folate	Low levels of folate result in elevated levels of homocysteine, an amino acid that has roles that include the antagonism of the N-methyl-d-aspartate receptor (NMDAR). NMDAR antagonists reduce the volume of the subiculum and number of neurons in the hippocampus. Additionally, high levels of homocysteine may impair delivery of oxygen to the foetus — leading to hypoxia. Hypoxia is a known risk factor for schizophrenia Other known risk factors for schizophrenia, that are a result of high levels of homocysteine, include pre-eclampsia, premature birth and low birthweight	(Ayesa-Arriola et al., 2012; Eyles et al., 2018; He et al., 2018; Kim & Moon, 2011; Moustafa et al., 2014; Trześniowska-Drukała et al., 2019)

	Season of birth	Increased incidence of seasonal infectious diseases like Influenza, increases the risk of exposure to these pathogens	(Escott-Price et al., 2018; Karlsson et al., 2019; Kim et al., 2017; Konrath et al., 2016; Wang & Zhang, 2017)
Childhood and Adolescence	Cannabis use	Variation in the catecholamine-O-methyl transferase (*COMT*) gene (e.g. rs4680) results in the substitution of Valine (Val) with Methionine (Met) (Val158Met), thereby increasing the release of prefrontal dopamine into synapses. Individuals with the Val/Val genotype have the highest risk for schizophrenia induced by the use of cannabis compared to those with the Val/Met and Met/Met genotypes	(Arseneault et al., 2002; Bagot et al., 2015; Bosia et al., 2019; Caspi et al., 2005; Estrada et al., 2011; Mustonen et al., 2018; Nieman et al., 2016; Vinkers et al., 2013)

	Socio-economic status	The lack of socioeconomic resources leads to chronic stress, predisposing to psychiatric illnesses. Maternal stress, infections and obstetric complication are known risk factors for schizophrenia and are common among mothers in the lower socioeconomic strata.	(Dohrenwend et al., 1992; Hakulinen et al., 2020; Hatzimanolis et al., 2020; Jensen et al., 2017; Wicks et al., 2005)
	Urbanicity	Crowding in urban areas leads to exposure and spreading of infectious diseases that predispose to schizophrenia Social fragmentation leads to instability and lack of communal support.	(Attademo et al., 2017; Vassos et al., 2012)

	Migration	Discrimination against immigrants leads to social exclusion and chronic stress. Migrants are more likely to occupy lower socioeconomic positions.	(Anderson et al., 2015; Bourque et al., 2011; Cantor-Graae et al., 2003; Kirkbride et al., 2012; Tortelli et al., 2014; Veling et al., 2008)
	Childhood trauma	Trauma in the form of physical, emotional and sexual abuse; and physical and emotional neglect can cause severe stress, which in turn make sufferers more vulnerable to developing schizophrenia.	(Bonoldi et al., 2013; Kilian et al., 2017; Larsson et al., 2013; Li et al., 2017; Shannon et al., 2011; Varese et al., 2012)

1.4.3 The interaction between genetic and environmental risk factors

There are complex interactions between genes and the environment that contribute to the development of schizophrenia, as demonstrated by some of the proposed mechanisms in **Table 1.1**. For example, when considering cannabis use as the environmental exposure, the odds of schizophrenia is markedly increased in Val/Val carriers for the catechol-O-methyltransferase (*COMT*) Val158Met polymorphism, but not in the Met/Met homozygotes (Caspi et al., 2005; Ermis et al., 2015), however these findings have not been replicated (Costas et al., 2011; Gutiérrez et al., 2009; Kantrowitz et al., 2009; van Winkel, 2011; Zammit et al., 2007), partly because candidate gene studies are not the most suitable design for polygenic disorders.

The effects of childhood trauma have been linked to the Val66Met polymorphism in *BDNF*, with the Met allele leading to reduced BDNF (brain-derived neurotrophic factor) expression as well as reduced sub-volumes in the hippocampus (Aas et al., 2014; de Castro-Catala et al., 2016). It has been found that the *BDNF* 66Val allele is associated with higher vulnerability to the effects of childhood trauma in male twins, while the 66Met allele has the same effect in female twins, suggesting a sex-specific effect (de Castro-Catala et al., 2016). Incidentally, both *BDNF* and *COMT* are located on autosomes (chromosome 11 and 22, respectively) and not on the sex chromosomes (X or Y).

Notwithstanding evidence from the gene by environment studies above, it is often difficult to choose which genes and which environmental factors to study (Esposito et al., 2018). The choice of the genes to study is based on findings from either candidate gene studies or genome-wide association studies (GWAS), both methods are reviewed in section 1.6 below. Candidate gene studies are usually conducted in small sample size and thus the results are not replicable across study sites. While GWAS studies are typically conducted in large samples sizes, it is often the case that information on environmental exposure is not acquired during study participant ascertainment. The choice of which environmental exposure to study is limited by lack of standardized measures for environmental factors, and only one end of the environment spectrum is usually investigated, which leads to spurious associations (Duncan & Keller, 2011).

1.5 Diagnostic criteria: an historical perspective

The prevailing classification of schizophrenia is rooted in that of psychiatrists in the mid-nineteenth century. Schizophrenia is diagnosed according to the psychotic disorders module of the Diagnostic and Statistical Manual of Mental Disorders (DSM) and section F2 of the International Classification of Diseases (ICD). There are five editions of the DSM: (DSM-1, DSM-II, DSM-III, DSM-III-R, DSM-IV, DSM-IV-TR and DSM-5) (American Psychiatric American Psychiatric Association, 1952; American Psychiatric Association, 1968; American Psychiatric Association, 1980; American Psychiatric Association, 2000; American Psychiatric Association, 2013), and several versions of the ICD (now available in its 11th edition).

Kraepelinian's *dementia praecox* (1856 - 1926): In the middle of the 19th century, psychiatrists in Europe were grappling with describing the disorders that were largely present in young adults, which typically progressed to cognitive and behavioural decline. In France, the disorder was named "*démence précoce"*, "*adolescent insanity*" in Scotland; and "*hebephrenia"* (the catatonic state) was first described in Germany (Clouston, 1904; Kahlbaum, 1874; Morel, 1860). Kraepelin proposed the integration of these varied symptoms into "dementia praecox" and proposed another category for manic-depressive symptoms (Kraepelin, 1971). His definition emphasized the chronicity of the disorder.

Bleuler's schizophrenias (1857 - 1939): "Schizophrenia" replaced "dementia praecox" after Eugen Bleuler amended Kraepelin's classification by incorporating a clinical state that did not progress to chronic deterioration that Kraepelin considered as the hallmark of the disease (Bleuler, 1958). Bleuler considered schizophrenia to be a group of diseases rather than a single illness, therefore proposing that it be referred to in the plural – the schizophrenias. Further, Bleuler introduced the distinction between obligatory and supplementary symptoms. The obligatory symptoms were designated as the four As: ambivalence, affective incongruity, autistic thoughts and associative disturbance, and were consistently present in affected individuals. Accessory symptoms included delusions and hallucinations, and unlike obligatory symptoms, were nonspecific and variable. Bleuler's emphasis on negative symptoms was incorporated into DSM-I and DSM-II.

Schneider's first rank criteria: Further distinctions of the schizophrenia subcategories were proposed in the ensuing decades. New definitions included schizoaffective disorders, schizophreniform psychoses, process-nonprocess and paranoid-nonparanoid schizophrenia. Schneider claimed that nine categories of psychosis, the "first rank symptoms (FRS)" had a decisive weight in the diagnosis of schizophrenia (Schneider, 1959). These included audible thoughts, voices arguing about the patient, voices commenting on the patient's actions,

thought withdrawal and other interference with thoughts. First rank symptoms were incorporated into DSM-III to DSM-IV and ICD-8 to ICD-10. Nuclear schizophrenia was characterised by the presence of at least three FRS.

The latest diagnostic classification systems incorporate Kraepelin chronicity, Bleuler' negative symptoms, and Schneider's positive symptoms, to varying degrees, as part of their definition. Thus, the DSM-5 defines schizophrenia based on criteria provided in **Table 1.2**.

Table 1.2 - DSM-5 diagnostic criteria

A. Characteristics of symptoms **At least two of the symptoms present for a significant amount of time during a 1-month period (or less if successfully treated). At least one of these must be (1), (2) or (3)**	1) Delusions
	2) Hallucinations
	3) Disorganized speech
	4) Grossly disorganized or catatonic behaviour
	5) Negative symptoms (i.e. diminished emotional expression or avolition)
B. Social or occupational dysfunction **Dysfunctional for a significant amount of time since the beginning of the disturbance. Lower level of functioning in at least two major areas compared to before the onset of symptoms. Or, failure to achieve expected level of functioning in (1), (2) and (3), when the onset is in childhood or adolescence,**	1) Work
	2) Interpersonal relations
	3) Self-care
C. Duration	Persistence of symptoms for at least six months, including those in Category A. This 6-month period must include at least 1 month of symptoms (or less if successfully treated) that meet criterion A (i.e., active-phase symptoms) and may include periods of prodromal or residual symptoms. During these prodromal or residual periods, the

	signs of the disturbance may be manifested by only negative symptoms or two or more symptoms listed in Criterion A present in an attenuated form (e.g., odd beliefs, unusual perceptual experiences).
D. Schizoaffective and mood disorder are ruled out if:	(1) no major depressive or manic episodes have occurred concurrently with the active-phase symptoms, or
	(2) mood episodes have occurred during active-phase symptoms, they have been present for a minority of the total duration of the active and residual periods of the illness.
E. Exclusion of substance or general medical condition	The disturbance is not attributable to the physiological effects of a substance (e.g., a drug of abuse, a medication) or another medical condition
F. Relationship to a pervasive developmental disorder	If there is a history of autism spectrum disorder or a communication disorder of childhood onset, the additional diagnosis of schizophrenia is made only if prominent delusions or hallucinations, in addition to the other required symptoms of schizophrenia, are also present for at least 1 month (or less if successfully treated).

1.6 The genetic aetiology of schizophrenia

Several genetic methods have been employed and are currently in use to identify the genes associated with schizophrenia. In this section the main findings that have emerged from these methodologies are briefly reviewed.

1.6.1 Linkage studies

Linkage studies were designed to identify rare single loci that are highly penetrant, and provide evidence for candidate 'susceptibility' genes involved in schizophrenia. This methodology is best applied to monogenic disorders, and has been useful for mapping the loci for diseases such as Type II diabetes mellitus (T2D), obesity and Alzheimer's disease (Badano & Katsanis, 2002). Previously mapped loci for schizophrenia include those on chromosomes 1p32.3, 2q36.1, 3q28 and 12q23.1(Klei et al., 2005; Paunio et al., 2001; Suarez et al., 2006; Wilcox et al., 2002), however, these findings have not replicated across different study cohorts (Vieland et al., 2014). Reasons for this include lack of statistical power and etiological heterogeneity

between groups. Methods such as genome-wide association studies (GWAS) have proven more useful for polygenic or complex disorders like schizophrenia.

1.6.2 Genome-Wide Association Studies (GWAS)

GWAS is defined as the hypothesis-free measure of the statistical difference in the allele frequency of genetic variants across the genome between affected individuals compared to a set of unaffected matched controls (Hirschhorn & Daly, 2005; Lee et al., 2012). Despite the limitations of this study design (discussed in chapter 2), GWAS studies have indeed led to the discovery of many disease-associated loci and a better understanding of disease aetiology. Additionally, GWAS can be used to find the association of both common and rare variants, although GWAS of common variants are more typical. GWAS of common variants investigates the association of variants with an allele frequency greater than 5% in the population, while GWAS of rare variants tests for variants with a frequency less than 0.1%. GWAS studies of common variants in schizophrenia are reviewed and discussed in chapterChapter 2:

1.6.3 Next generation sequencing

Unlike GWAS of common variants that uses a pre-selected set of known markers that are used as probes on GWAS arrays, high-throughput next generation sequencing methods such as whole exome- and whole genome sequencing (WES and WGS, respectively) allow for identification of rare and potentially novel variants through deep sequencing. The decline in the cost of next generation sequencing technologies, development of analytical tools, as well as increasingly fast computational power have allowed for risk gene discovery towards a better understanding of the genetic aetiology of complex disorders, including schizophrenia.

WES identifies protein-coding variants in about 20,000 genes per individual sequenced (Ng et al., 2009). This allows investigators to identify genetic variants and focus on those which may have functional consequences, e.g. regulation of gene expression, or lead to amino acid substitutions, insertions, deletions and premature termination of the encoded protein, and to categorize genes as part of groups that have specific roles in functional networks related to the phenotype of interest (Avramopoulos, 2010).

WES has been applied to schizophrenia studies to identify *de novo* mutations (i.e. those that are not inherited) by comparing the exome sequence of the probands to that of their parents. WES studies have demonstrated an increased frequency of *de novo* mutations in patients with schizophrenia compared to controls (Fromer et al., 2014; Gulsuner et al., 2013; Purcell et al., 2014; Rees et al., 2020; Rees et al., 2012). The rate of *de novo* variants can be as much as eight times higher in schizophrenia cases than in controls (Xu et al., 2008). These *de novo* mutations are enriched in genes that are intolerable to loss of function (LoF) variants, that is, those variants that introduce premature stop codons, shifts the frame of a protein codon or disrupt messenger RNA splicing (Singh et al., 2016; Singh et al., 2017). Implicated genes emerging from WES studies include *SETDA1A, RBM12, ARC*/*NMDAR* protein complexes and *SLC6A1*, which also overlap with other neurodevelopmental disorders such as autism spectrum disorder (ASD) and intellectual disability (ID) (Fromer et al., 2014; Rees et al., 2020; Singh et al., 2016; Steinberg et al., 2017). The largest exome sequence study to date (24,248 cases and 97,322 controls) has identified 10 genes harbouring ultra-rare and *de novo* variants that are highly expressed in neurons and function in synapses (Singh et al., 2020).

Compared to the variants exposed through WES, WGS identifies about four million variants per individual sequenced (Bentley et al., 2008). WGS studies have proven useful in closing the missing heritability gap between SNP-based and family-based heritability estimates for height and body mass index (Wainschtein et al., 2019). This missing heritability was found in ultra-rare variants with minor allele frequencies (MAF) ranging between 0.0001 – 0.1, that were in regions of low linkage disequilibrium (LD) and outside of coding regions. In light of the fact that GWAS arrays are designed to capture common variants that are in high LD, and WES only capturing known protein-coding variants, WGS provides a comprehensive means of interrogating genomic factors contributing to the phenotype including non-coding regulatory as well as structural variants.

There have only been about ten WGS studies of schizophrenia. The latest and largest study to date (1162 schizophrenia cases and 936 ancestry-matched population controls from Sweden) identified an enrichment of ultra-rare variants in genes that are intolerant to LoF mutations in cases compared to controls (Halvorsen et al., 2020), similar to findings from other WES studies (Genovese et al., 2016; Gulsuner et al., 2020). There was also an increased burden of non-coding structural variants in cases compared to controls in topologically associated domains (TADs) in adult and fetal brain functional elements (Halvorsen et al., 2020). TADs are genome partitions; each partition acts as a regulatory unit within which enhancers and promoters can interact to regulate gene expression. Alteration of TADs may lead to the dysregulation of gene expression. These findings were consistent with those from

other studies showing that disruption of TADs are associated with developmental disorders (Lupiáñez et al., 2015). An interesting finding from the Halvorsen et al. (2020) study was that the SNP-based point estimates for heritability were close to the those estimated from family-based studies (0.52 vs 0.6 - 0.65), albeit with a large standard error (Halvorsen et al., 2020). It is likely that WGS in larger samples will narrow the standard error and provide more precise heritability estimates.

1.6.4 Copy Number Variants

Copy number variants (CNVs), which are chromosomal aberrations of sizes 1kb to several mega-bases (Mb) can be identified using GWAS arrays, WES and WGS. Studies have found an enrichment of large, rare CNVs in schizophrenia cases compared to controls (Kirov et al., 2014; Marshall et al., 2017; Rees et al., 2014; Szatkiewicz et al., 2014; Wainschtein et al., 2019). Among the most penetrant CNVs are those in the chromosomal region 22q11.2, which have been shown to confer the greatest risk for schizophrenia.

The chromosomal region 22q11.2 was suggestively linked to schizophrenia through linkage studies more than a decade ago (Karayiorgou et al., 1995). Subsequent large-scale studies have identified copy number variation in this region to be associated with schizophrenia (Marshall et al., 2017). The 22q11.2 CNV occurs as a hemizygous deletion known as 22q11.2 Deletion Syndrome (or 22qq11DS) that spans about 2.5Mb and impacts about 90 genes (Guna et al., 2015). 22q11DS is the strongest known single genetic risk factor, contributing a 25-fold increased risk of developing schizophrenia. About 0.3% of patients with schizophrenia carry the 22q11.2 deletion (Kirov et al., 2014). Interestingly, duplications at 22q11.2 confer protection from schizophrenia (Rees et al., 2014). Both deletions and duplications at the 22q11.2 locus have been associated with other psychiatric disorders such as ASD and ID (Sanders et al., 2011), indicating their pleiotropic effect.

There are several proposed mechanisms by which 22q11DS may contribute to the schizophrenia phenotype, based on the genes encompassed by the deletion. One of the implicated genes is *DGCR8*, which is part of a complex of microRNAs that regulate gene expression by binding to messenger RNA and inhibiting their translation (Rajman & Schratt, 2017). Functional studies indicate that these microRNAs are involved in neuronal and neurodevelopmental pathways (Merico et al., 2014). Another proposed mechanism involves the *PI4KA* gene which encodes kinase PI4KIIIα, whose function is to recruit and regulate plasma membrane proteins, thereby regulating processes such as ion transportation and the

structural actin cytoskeleton (Falkenburger et al., 2010). Depletion of *PI4KA* has been shown to disrupt the maintenance of synapses (Bulat et al., 2014).

The MHC region on chromosome 6 plays a critical role in the immune system, with genes encoding for a number of antigen-presenting molecules. Following up on a GWAS study that found that variants in the MHC region were the most significantly associated with schizophrenia, Sekar et al. (2016) narrowed this association to the *C4* gene (Sekar et al., 2016). They showed that the C4A and C4B versions of the gene, are present in different combinations of copy number, and that the higher the number of copies of *C4*, the greater their expression and subsequently the greater the association with schizophrenia (Sekar et al., 2016). Functionally, C4 is involved in the complement pathway that undergoes development of the synaptic connections between neurons. Synaptic pruning, or elimination of the synapses, has been connected to individuals with schizophrenia, further strengthening the observation that C4 plays a role in increased schizophrenia risk (Comer et al., 2020; Sekar et al., 2016). Pathways such as those involved in neural structure and signalling processes are vital to the discovery of potential drug targets for the treatment of schizophrenia, and future genetic studies are key in furthering our understanding of the biological processes involved in this complex disorder.

1.7 Need for large scale studies in African populations

There is a gross under-representation of African populations in large-scale genomics studies. African samples make up only about 2% of study participants in GWAS studies (Popejoy & Fullerton, 2016; Sirugo et al., 2019). This is despite Africans constituting 17% of the total global population, and being the most genetically diverse population (Abecasis et al., 2012; Auton et al., 2015; Behar et al., 2008). There is a critical need to include African populations in large-scale genetics studies for the reasons expanded on below.

1.7.1 High genetic diversity

Complex migration patterns, i.e. out of Africa, back into Africa and intra-Africa migrations contribute significantly to the genetic diversity observed in modern African populations (Behar et al., 2008). The out of Africa hypothesis posits that modern humans left the continent through two primary routes towards Europe and Asia, respectively, about 100 kya (kya: thousand years

ago) (Campbell & Tishkoff, 2008; Forster & Matsumura, 2005; Quintana-Murci et al., 1999; Reed & Tishkoff, 2006). These migrations, out of the north-eastern region of Africa, in different directions, were by relatively small subsets of the original continentally-distributed parental or ancestral African populations. The relatively small population size of the emigrants, eventually settling in the new habitats (and a subfraction moving on, recursively) meant there was a fractional representation of the genetic diversity on the continent, resulting in a reduced genetic pool for breeding, which ultimately would have led to significant loss of heterozygosity at some regions of the genome. Several generations of breeding within this relatively limited gene pool led to a state of genetic equilibrium also known as a population bottleneck (Cornuet & Luikart, 1996; Li & Durbin, 2011). A population bottleneck refers to the reduction in the growth of a population and consequently, a reduction in variation of the genetic pool.

Although the phenomenon of the original out of Africa migrations has long been known, evidence for the back to Africa migration of more than 20kya, has only relatively recently been published. These migrations occurred about 30 - 20 kya from the middle east into North Africa, resulting in admixture between North Africans and non-Africans (Forster & Romano, 2007; Henn et al., 2012; Sánchez-Quinto et al., 2012).

The intra-African migration wave that shapes the genetic diversity in sub-Saharan African populations is known as the “Bantu expansion”, which refers to the Bantu people of west Africa migrating to east Africa, and then downwards to southern Africa (the “eastern stream”); and directly from west Africa into southern Africa (the “western stream”) about 5 - 2 kya (Phillipson, 2005; Salas et al., 2002). Thus, the current populations in southern Africa represent ancestries derived from east Africa, west Africa and the Khoi-San population who are native to the southern African region, and belong to the most ancient lineages of modern humans (Henn et al., 2011; Soares et al., 2016). The South African Bantu languages incorporate closely related languages such as Zulu, Xhosa, Sotho and Tswana.

1.7.2 GWAS findings cannot be extrapolated between populations

Findings from GWAS studies conducted in EUR populations cannot always be extrapolated to other populations. Because genetics studies in AFR populations have been limited to small sample sizes and have employed low-cost methodologies involving assessing variation in single genes or short genomic regions to investigate the genetic basis of psychiatric disorders, the comparison of findings between study populations is limited to the genes or genomic regions investigated in these earlier studies.

In South Africa, research into the genetics of schizophrenia has been conducted in the Afrikaner population and in some indigenous African (Bantu-speaking) populations and has revealed discordance in association findings between these two ethnic groups. The deletion on chromosome 22q11, that has been implicated in the aetiology of schizophrenia and also known to confer the largest risk to schizophrenia, was reported not to be associated with disease in a cohort of 97 Xhosa individuals (Riley et al., 1996). The frequency of the mutation in the cohort was well below the 2% rate previously reported (Karayiorgou et al., 1995; Koen et al., 2012). Similar to EUR-ancestry findings, this deletion was identified in the South African Afrikaner population (Wiehahn et al., 2004), who are a relatively homogenous population and descendants of the original Dutch, French and German settlers who came to South Africa in the 17th century. These findings suggest an ancestry-dependent association with the 22q11 microdeletion. Larger studies in AFR-ancestry samples are needed to either confirm or reject these initial findings.

Other associations, e.g. those on chromosomes 6 and 22 that have been identified as risk factors in EUR ancestry populations have not been linked to schizophrenia in Bantu-speakers (Riley et al., 1996; Riley et al., 1997), while those on chromosomes 1, 9 and 13 have been replicated in the EUR-derived Afrikaner population (Abecasis et al., 2004; Hovatta et al., 1999; Moises et al., 1995). Further, studies have shown no association of candidate genes *KCNN3* (on chromosome 1q21), *PPP2R2B* (on 5q32) and *SOD2* (on 6q25.3) with schizophrenia in the Xhosa population (Hitzeroth et al., 2007; Laurent et al., 2003) despite associations being found in cohorts of EUR.

WES has been used to identify *de novo* mutations linked to schizophrenia in a small sample of the Afrikaner population, which also identified an enrichment of these mutations in cases compared to controls (Xu et al., 2011; Xu et al., 2012). The largest WES in an AFR cohort comprised about 2,000 Xhosa study participants. Findings from this study were consistent with findings from EUR-ancestry samples in that it identified an enrichment in damaging mutations in genes that are intolerant of LoF variants (Gulsuner et al., 2020). The contrast between this latter study and others, is the association of variants in the *ZFP3* gene with schizophrenia, although the level of significance was below the accepted threshold for exome-wide association testing. Notwithstanding, this study provided evidence that genetic studies in under-represented populations may lead to more variant discovery. Findings from the studies above, highlight that although common genetic risk factors may exist between populations, there may be population-specific variants that contribute to disease risk.

1.7.3 Use of commercially available GWAS arrays is less informative

There are many GWAS arrays available on the market, from companies such as Affymetrix and Illumina, for genetic studies. Many of these arrays were designed by selecting variants that were used as probes from variant analysis from an ascertainment group of people largely of EUR ancestry. Thus, they may not provide comprehensive coverage for variants in AFR genomes.

Most arrays are designed using whole genome reference sequences from few members of a population usually from the HapMap and 1000 Genomes Projects where only West and East African genomes were studied (Abecasis et al., 2010; Gresham et al., 2008). The probes used in these array identify variants which have MAF above 1%. However, these variants may not be present at the same frequency in all AFR populations, and do not account for rare and population-specific variants. The ascertainment of probes from a group genetically divergent from the intended group introduces bias, and thus identifies low genetic variation in the genotyped group.

GWAS arrays that target non-EUR ancestry individuals have been developed including the Multi-Ethnic Genotyping Array (MEGA) from Illumina (Bien et al., 2019) and the PanAFR Array from Affymetrix(http://tools.thermofisher.com/content/sfs/brochures/axiom_panafr_arrayplate_datasheet.pdf). The MEGA array was developed in collaboration with the Population Architecture Using Genome and Epidemiology (PAGE) study. The study sample comprises African Americans, Asian Americans, Native Americans, Native Hawaiians and Latino/Hispanics. The PanAFR array covers the Yoruba from Nigeria, Luhya and Maasai from Kenya and admixed African-Americans with West African ancestry from the 1000 Genomes Project. Both arrays do not efficiently represent the genetic diversity in Africa.

The H3Africa array from the Human Hereditary and Health in Africa (H3Africa) Consortium was developed to fill this gap (Rotimi et al., 2017). The array was designed based on 350 genomes of individuals from all over Africa including those from projects running within the H3Africa consortium. Currently, this array is the most representative of genetic variation on the African continent.

1.7.4 Short LD blocks improve fine-mapping

Linkage disequilibrium (LD) refers to the degree of association between alleles at different loci, and is a function of population size and locus-specific factors such as recombination, natural selection, and mutation. AFR populations have shorter LD blocks compared to non-AFR populations because of the maintenance of an effective population size — the opposite of the population bottleneck that came as result of the out of Africa migration, and more time for recombination to occur. Since GWAS of common variants rely on the association of SNP with a causal SNP in strong LD, shorter LD blocks help to narrow in on the disease-associated locus to identify causal variants, an analytical process called 'fine-mapping' (Reich et al., 2001).

This concept of AFR ancestry genomes being useful for fine-mapping was first proven in Type II diabetes studies (Helgason et al., 2007), and has since been expanded to many other complex traits (Auton et al., 2015; Fernández-Rhodes et al., 2017; Franceschini et al., 2016; Yoneyama et al., 2017; Zubair et al., 2016). For schizophrenia, findings from a study by Bigdeli et al. (2019) showed improved fine-mapping of schizophrenia-associated loci when African American samples were meta-analysed with the larger cohort of EUR ancestry individuals.

1.8 Aims and objectives

This thesis had two aims. The first aim was to investigate the genetic aetiology of schizophrenia in the SAX population using a custom-designed GWAS array. This work done to fulfil this aim are covered in chapters 2 and 3 of the thesis. The specific objectives under this aim are listed below:

1. Investigate the association of common variants with schizophrenia in SAX
2. Determine the distribution of heritability by genomic partitions and functional categories
3. Investigate how well genetic risk scores derived from the largest schizophrenia datasets predict schizophrenia in SAX

Why the Xhosa people

The Xhosa people are the second largest and southernmost indigenous group in South Africa with over eight million individuals (Koen et al., 2012). They belong to the Nguni linguistic group are a culturally and linguistically homogenous. Genetic studies have shown that the Xhosa are a two-way admixture of the Khoisan, a native South African group, as well as the Yoruba from West Africa, and are genetically similar within the group (Chimusa et al., 2015). Due to the polygenicity of schizophrenia, studies in a homogenous group reduces genetic heterogeneity among the study sample and increases the chance of variant discovery. It is also advantageous as more heterogenous groups would require even larger sample sizes.

The prevalence of schizophrenia in this group is yet to be determined, however, hospital records show that over 2,700 patients are admitted to two major psychiatric hospitals annually in the Western Cape Province (of South Africa) with over 10,000 out-patient visits. In the Eastern Cape Province, over 600 patients are admitted annually in major psychiatric hospitals in the larger cities of East London and Port Elizabeth. This study provides an opportunity to explore population-specific variation in an African context.

The second aim was to assess the generalizability of PRS derived from EUR ancestry individuals in AFR ancestry populations, covered in chapter 4. The specific objectives under this aim were to:

1. Investigate the generalizability of PRS derived from EUR ancestry individuals from UKB in the Ugandan General Population Cohort for 34 quantitative traits
2. Investigate the generalizability of PRS derived from EUR ancestry individuals from UKB South African Drakenstein Child Health Study for six traits
3. Investigate how the inclusion of diverse populations in GWAS meta-analysis impacts PRS accuracy.

Chapter 2: Genome-wide association study of Schizophrenia in the South African Xhosa people

2.1 Abstract

GWAS have been used to identify the common genetic variants that are associated with schizophrenia. Over 200 single nucleotide polymorphisms (SNPs) with small effect sizes have been identified in studies conducted mostly in EUR ancestry populations, as contributing to the risk of developing schizophrenia. There is a paucity of data on the genetic and environmental factors that contribute to the risk of schizophrenia in populations of AFR ancestry. In this chapter, GWAS was used to investigate the contribution of common variants to schizophrenia risk in a cohort of 1,038 cases and 1,048 controls from SAX. Further, the impact of childhood trauma and biological sex on the genetic aetiology of schizophrenia outcome were investigated. The initial GWAS yielded one SNP (rs35172303, *P* = 4.74e-08, OR = 0.6004, 95% CI:[0.499,0.721]) in *ZFP3* that met the genome-wide significance threshold. After controlling for childhood trauma, five SNPs in *ZFP3* that were in linkage disequilibrium (r^2 > 0.6) with rs35172303 were significantly associated with schizophrenia. No conclusive results were obtained from the sex-stratified GWAS as the two strata could not be compared because the female strata was significantly underpowered. Results from the childhood trauma analysis suggested that there was an interaction between childhood trauma and genetics that may lead to the development of schizophrenia, highlighting the role of environmental factors in for this disorder. In conclusion, *ZFP3* — although not previously associated with schizophrenia in large-scale GWAS of schizophrenia, may be important for disease risk in AFR populations, and childhood trauma is likely an important contributor to this risk. It is therefore imperative that large scale genomics studies are conducted in populations of AFR ancestry to uncover previously-unidentified variants that may have biological relevance.

2.2 Introduction

As discussed in chapter 1, GWAS are a hypothesis-free approach to investigating variants across the genome to identify those that are significantly overrepresented in cases when compared to controls, or vice versa (Kitsios & Zintzaras, 2009). Such associated variants are useful in determining genetic risk factors for complex disorders. Since the first published GWAS of schizophrenia in 2007, more than 100 studies have been reported to date, implicating at least 3,000 SNPs (https://www.ebi.ac.uk/gwas/). Initial GWAS studies were underpowered to detect true and replicable associations due to their relatively small sample size and the genetic complexity of schizophrenia (Ioannidis, 2005; Kraft et al., 2009)'. Generally, for a complex disorder, an adequately powered GWAS study comprises tens of thousands of individuals (Spencer et al., 2009). Perhaps, because of the established infrastructure and funding for large scale research, the large majority of GWAS studies have thus far been conducted in the United States of America (USA) and in Europe, in populations of EUR ancestry. However, recently, large-scale studies have also been conducted in cohorts of EAS). There are still only a relatively small number of a studies that have been conducted in populations of AFR ancestry.

In this chapter, GWAS studies are reviewed by ancestry and findings from these GWAS are contrasted across ancestries. Then, a GWAS was conducted to investigate the association of common variants with schizophrenia in SAX. First, a standard GWAS was performed, followed by a GWAS controlling for childhood trauma experience, and GWAS stratified by biological sex. The results are presented and discussed in the sections that follow.

2.2.1 GWAS in populations of EUR ancestry

To overcome the sample size limitation, networks of international collaborators have formed consortia to allow for pooling of samples from GWAS studies for meta-analysis. This increases the study sample size increases the potential to identify true variant associations. One of the first consortia to perform case-control comparisons was the International Schizophrenia Consortium (ISC, comprising 3,322 cases and 3,587 controls) in 2009 (International Schizophrenia Consortium, 2009). This study by the ISC was the first to show an association with markers on chromosome 6p in the vicinity of the MHC, implicating this immunological complex as being important in the aetiology of schizophrenia; and also the first to demonstrate the polygenetic nature of schizophrenia. In 2013, the association of the MHC was replicated in an independent GWAS conducted by the PGC which is to date, the largest psychiatric

genetics consortium internationally (Psychiatric Genomics Consortium, 2011; Ripke et al., 2013; Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014).

Fine mapping in the MHC region implicated the *complement 4* (C4) gene, specifically, as a contributor to schizophrenia risk, showing that individuals with the *C4A* version of the gene had a greater chance of developing schizophrenia compared to those with the *C4B* version (Sekar et al., 2016). The landmark 2014 PGC study (comprising 36,989 cases and 113,075 controls) identified 108 independent loci that were associated with schizophrenia (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014) implicating calcium signalling channels, glutamatergic neurotransmission pathways and long- and micro-RNAs, in addition to the MHC. Subsequently, Pardiñas et al. (2018) demonstrated that schizophrenia-associated variants are enriched in mutation-intolerant genes and identified 145 independent loci associated with the disorder. The latest and largest schizophrenia study conducted in 69,369 cases and 236,642 controls identified 270 independent loci associated with schizophrenia. These variants were enriched in genes that have an increased burden of rare deleterious mutations for schizophrenia (Ripke et al., 2020), indicating that both rare and common variants implicate the same genes.

2.2.2 GWAS in populations of EAS ancestry

GWAS studies of schizophrenia in individuals of EAS ancestry, and specifically in the Chinese and Japanese populations, emerged in the last decade. Similar to observations from studies conducted in cohorts of EUR ancestry, variant discovery in the early GWAS studies in EAS populations was limited by sample size. The earliest study, comprising 575 cases and 564 controls from Japan did not find any associated variants meeting the genome-wide significance level of a p-value less than 5e-08, but the study produced a sub-significant signal on chromosome 6p21.31 in the vicinity of *NOTCH4*, thereby implicating this gene (Ikeda et al., 2011).

Numerous studies have since been performed in the Han Chinese population. In a study of 479 cases and 1,599 healthy controls, followed by a replication cohort of 4,027 cases and 5,603 controls, Yue et al. (2011) identified novel schizophrenia risk loci on two chromosomal regions: (i) 11p11.2, implicating *TSPAN18,* and (ii) 6p22.1 implicating *NKAPL*, *ZKSCAN4* and *TSPAN18* (Yue et al., 2011). However, a subsequent study failed to replicate the association of variants in the chromosomal regions reported by Yu et al. (2011) (in the vicinity of either

NKAPL or *TSPAN18*) in 700 cases and 700 controls (Li et al., 2017). The region 6p22.1 spans the MHC coding region, which has been consistently associated with schizophrenia in EUR.

Other genomic regions, such as those on chromosomes 1q24.2 and 8p12 harbouring the genes *BRP44* (on 1q24.2) and *WHSC1L1/NSD3*, *LSM1* (on 8q12) were implicated and validated in a discovery cohort of 3,750 cases and 6,468 controls; and subsequently validated in a replication cohort of 4,383 cases and 4,539 controls (Shi et al., 2011). Another susceptibility locus in Han Chinese was identified on chromosome Xq28 (harbouring *RENBP*, *MECP2*, and *ARHGAP4*); this locus was originally identified in 498 cases and 2,025 controls; and subsequently replicated in 1,027 cases and 1,005 controls (Wong et al., 2014).

More recently, new data has emerged from cohorts of larger sample sizes. Three novel loci on chromosomes 22p16.1 (*VRK2*), 6p22.1 (*GABBR1*) and 10q24.32 (*AS3MT* and *ARL3*) were reported to be associated with schizophrenia in a Han Chinese primary cohort of 4,384 cases and 5,770 controls, and a replication cohort of 4,339 schizophrenia cases and 7,043 controls. In a follow-up study with double the initial sample size (7,699 cases and 18,327 controls, and a replication cohort of 4,384 cases and 18,327 controls and a meta-analysis sample of 12,083 cases and 24,097 controls), Shi et al (2017) identified three loci which had been previously associated with schizophrenia and four that were novel at loci 3p21.31, 6q21, 6q27 and 7q31.1.

The subsequent and largest study of schizophrenia in the EAS population comprised 22,778 cases and 35,362 controls, and identified 21 genome-wide-significant associations at 19 distinct genetic loci (Lam et al., 2019). These findings replicated the four loci reported in an earlier study of ‘Chinese-only’ cohort (Yu et al., 2017). Interestingly, the MHC region was not significantly associated in this study. This was attributed to allele frequency differences in variants within MHC between the EAS and EUR populations (Lam et al., 2019).

2.2.3 EAS-EUR cross-ancestry risk loci

Two strategies have been employed to evaluate whether risk variants identified in EUR studies confer the same risk in EAS samples. The first involves the selection and testing of specific variants at a locus, and the second is through the combination of summary statistics from GWAS studies, followed by association testing. The latter strategy is known as meta-analysis.

In the first approach, loci implicated in EUR-GWAS are specifically tested due to their plausible biological mechanism in the aetiology of schizophrenia. SNPs with the same effect in both populations have been found in genes such as on chromosomal regions 2p16.1 near *VRK2*, 3p21.1 near *ITIH3/4*, 4q26 near *NDST3*, 6p21.32 near *NOTCH4* and 13q31.3 near *MIR17.* However, SNPs on 2q32.1 near *ZNF804A* found to be associated in the EUR study, could not be replicated in EAS individuals (Huang, 2016; Li et al., 2015; Xiao et al., 2017).

Meta-analyses improve statistical power and potential for variant discovery. In a EUR-EAS trans-ancestry sample of 43,175 schizophrenia cases and 65,166 controls, Li et al. (2017), identified 109 loci that were genome-wide significant. Of these, 83 loci had been previously reported and 26 were novel, and 75% of the genome-wide significant variants identified in EUR remained significant in the trans-ancestry meta-analysis (Li et al., 2017). Another meta-analysis of EUR and EAS cohorts, comprising 56,418 cases and 78,818 controls identified 178 loci associated with schizophrenia, 53 of which were novel. Of the 108 loci reported in the landmark 2014 EUR-GWAS, 89 loci remained significant (Lam, Chen, et al., 2019). The latter study also demonstrated that a significant number of gene-sets overlap between EUR and EAS cohorts.

2.2.4 GWAS in populations of AFR ancestry

There is a paucity of data from GWAS studies in populations of AFR ancestry. A GWAS study of schizophrenia in an African-American cohort of 6,152 cases and 3,918 controls by Bigdeli et al. (2019) is the largest such study to date. Although this study did not identify any variants surpassing the genome-wide significance threshold, perhaps due to the inadequate sample size, it reported that EUR-identified variants have similar direction of effect in their AFR dataset. This is consistent with findings from the studies in the EAS individuals. Meta-analysis of an African-American sample with the PGC-EUR cohort yielded 93 loci that reached genome-wide significance, ten of which were unique (i.e. and not among the 108 previously reported) indicating the utility of inclusion of diverse samples in large-scale studies.

Trans-ancestry GWAS studies have been important in identifying shared common disease loci and have demonstrated that the biology of schizophrenia is shared between ancestral groups. Conversely, the studies have also found that there are population-specific risk variants which make it necessary to incorporate population diversity in the design of GWAS studies. An increase in sample size and ethnic diversity of study participants would create the opportunity to explain more of the genetic liability underlying schizophrenia (Shi et al., 2016).

2.2.5 Environmental impact on schizophrenia

While schizophrenia is highly heritable (rate ~80%), the proportion of variance in liability attributable to the genetic variants that have been associated with schizophrenia is only 24% (Ripke et al., 2020). Numerous non-genetic factors, including environmental exposure have been proposed to contribute to risk of schizophrenia (Manolio et al., 2009). The literature on environmental factors, although not as robust as that for genetic studies, has revealed the strong contribution of factors including paternal age, obstetric complications such as preeclampsia, drug use, urbanicity, and childhood trauma among others, and as reviewed in chapter 1 (Gage et al., 2016; McGrath et al., 2014; Stilo & Murray, 2019; van Os et al., 2010; Vaucher et al., 2018), the mechanism by which these environmental exposures predispose to schizophrenia are outlined in **Table 1.1**. Co-workers have shown that childhood trauma contributes to the risk of schizophrenia in the SAX cohort being reported on in this thesis (Mall et al., 2019). The continued improvement in instruments for measuring environmental exposures is leading to a new era of a 'multi-omics analysis' of the contributors to schizophrenia and other psychiatric disorders.

A wide range of data including cannabis use, highest level of education, and HIV status of study participants were collected for the parent study of which the work in this thesis is a part. In this regard, the scope of the present investigation of potential environmental contributors was limited to: (i) biological sex because of the large discrepancy in the numbers of males recruited in the SAX study, compared to the females; and (ii) childhood trauma because childhood trauma has been associated with schizophrenia in this cohort. These two points are further discussed below.

2.2.5.1 Biological sex and schizophrenia

According to previously published reports, schizophrenia affects females and males at a 1:1 ratio. However, data on subject recruitment in the present study reflects a sex-bias in that there is a remarkably larger proportion of recruited males than females [4:1]. Whether this is reflective of an ascertainment bias or a difference in the genetic risk profile between Xhosa men and women is unclear.

Three genetic models, which are not mutually exclusive, have been proposed to explain phenotypic differences between men and women for complex traits. The first or the *Carter effect*, also known as the sex-dependent liability threshold model posits that a sex-specific

genetic liability threshold is required to manifest a disorder (Carter & Evans, 1969). In other words, one of the sexes may require a larger proportion of risk alleles to manifest the disorder. Thus, the trait is likely to have a higher heritability rate in the group with the lower prevalence. However, studies to date show no difference in heritability rates or prevalence between men and women across a hundred complex traits (Ge et al., 2017; Traglia et al., 2017; Vink et al., 2012). For schizophrenia, the earliest sex-specific effects were identified in a study by Shifman et al.(2008) in a sample of Ashkenazi Jews. In this study, the strongest association was on chromosome 7q22, with the variant rs7341475 (in the *RELN* gene) in women, but not in men. This finding has been replicated in subsequent studies (Alfimova et al., 2019; Sozuguzel et al., 2019).

The second model relates to the sex chromosomes (i.e. XX for females and XY for males). These chromosomes have properties and effects that drive phenotypic differences in men and women, such as non-PAR X chromosome (ChrX) genes (one copy in males and two copies in females), parental imprinting of non-par ChrX genes, non-PAR Y genes (genes only expressed in men) and the presence of absence of ChrX inactivation. ChrX inactivation occurs when one copy of the non-par ChrX genes are silenced in females to ensure dosage compensation for the same genes in men. Evasion from ChrX silencing may result in sex-biased gene expression patterns. Furthermore, aneuploidy, the presence of more than two sex chromosomes, may lead to differential gene expression, making it more or less likely to develop disorders (Raznahan et al., 2018). It is important to note that generally, sex chromosomes are removed from GWAS analyses due to the lack of methodologies to handle the aforementioned characteristics of these chromosomes compared to autosomes; inadvertently excluding genetic heritability contributed by the sex chromosomes.

The sex chromosome model implicates gene-by-sex interactions. In this model, sex is considered as both a biological and environmental factor which modulates the endogenous (cellular) environment and response to exogenous factors. In the case of modulation of endogenous factors, sex-specific hormones at different life stages may influence predisposition to diseases. For example, the incidence of asthma is highest in boys, but the incidence in women is double that of men after puberty (Zein & Erzurum, 2015).

2.2.5.2 Childhood trauma and schizophrenia

The neural diathesis-stress model proposes that early-life psychosocial stress acts on a predisposition to schizophrenia and triggers symptoms later in life (Walker & Diforio, 1997). Childhood trauma described in the literature as adversity experienced before the age of 16 years, is characterized as physical, emotional or sexual abuse, and emotional or physical neglect. Physical and emotional neglect, in particular, have previously been associated with schizophrenia (Cutajar et al., 2010; Garcia et al., 2016; Xie et al., 2018).

The relationship or interplay between genetic factors and childhood trauma in contributing to schizophrenia is not straightforward. Evidence exists for two mechanisms: (i) prolonged exposure to childhood trauma predisposes victims to developing schizophrenia later in life, and (ii) childhood trauma interacts with the genetic framework to mediate schizophrenia effects in a dose-dependent manner. A study by Arnsten et al. (2009) encapsulates these two mechanisms. Their study showed that exposure to childhood trauma, on the one hand, disinhibits stress signal pathways leading to impaired responses by neuronal cells and psychotic symptoms; while on the other hand, genetic factors moderate the sensitivity of the cellular types to stimuli (Arnsten, 2009).

The environmental data collected during the recruitment of the SAX cohort was leveraged to investigate the sex-specific difference in the genetic architecture of schizophrenia based on the first proposed model i.e. the sex-dependent liability threshold. The specific aim of this chapter was to investigate the genetics of schizophrenia in the SAX population and to assess the impact of biological sex and childhood trauma on the genetic risk of schizophrenia in this cohort.

2.3 Methods and Materials

2.3.1 Recruitment

This study was approved by the Human Research Ethics Committee (HREC), the equivalent of an Institutional Review Board of the University of Cape Town (UCT, 149/2018) and forms part of a larger study approved by HREC of UCT (049/2013), the University of Washington in Seattle, USA (29501), Walter Sisulu University (003/2013), Rhodes University in the Eastern Cape Province of South African (2013Q4-7) and Columbia University in New York, USA (049-2013).

Study participants were recruited by psychiatric nurses who were proficient in isiXhosa from psychiatric hospitals and community healthcare centres in the Western and Eastern Cape provinces of South Africa. Potential study participants were identified in an information session conducted by the recruitment nurses. In this session, details of the study aims and expected outcomes were explained, and a flier detailing the study inclusion criteria circulated. Individuals who were interested in taking part in the study underwent an assessment to evaluate their understanding of the material shared in the information session and their capacity to consent using the University of California, San Diego Brief Assessment of Capacity to Consent (UBACC, **Appendix 1**) assessment tool (Jeste et al., 2007).

In order to be included in the study, participants had to be Xhosa-speaking (i.e. first language Xhosa speakers) and between the ages of 21 – 54 years, have a clinical diagnosis of schizophrenia (the cases), have no psychosis and duration of illness longer than two years and pass the UBACC. Participants were excluded if they had an injury requiring hospitalization, a brain injury, self-injury, or any other injury related to the use of alcohol or drugs.

All interviews and questionnaires were translated into isiXhosa according to the World Health Organization (WHO) translation guidelines (Sartorius & Janca, 1996). The schizophrenia cases were screened and diagnosed according to the Structural and Clinical Interview for Diagnostic and Statistical Manual for Axis 1 Disorders (SCID-1) (First et al., 2012). All participants also completed a Childhood Trauma Questionnaire and Discrimination and Stigma Experiences Scale (Bernstein et al., 2003; Brohan et al., 2013).

Consent was three-tiered (**Appendix 2**). The first tier meant that participants were consenting to participate in this study, and that their genetic material should not be shared. In the second tier, participants could consent to participate in this study as well as other studies in the future. In the third tier participants consented to sharing of blood for cell immortalization. After written consent was obtained, whole blood was drawn from each participant and sent to the Genomics Laboratory in the Division of Human Genetics at the University of Cape Town, where samples were catalogued and triaged for: (i) DNA isolation, (ii) DNA storage for long term use and/or (iii) harvesting of cells for immortalization. DNA extraction was conducted in the Human Genetics lab at UCT, whilst the pre-processing and processing of DNA for long term storage and the immortalization of cells were done at Tygerberg Hospital in Cape Town and Rutgers University Cell and Data Repository (RUCDR) in New Jersey, USA, respectively.

2.3.2 DNA isolation

DNA from 5 - 8 ml of whole blood using the 'Salting Out' Method (Miller et al., 1988). Briefly, the red blood cells were lysed by adding red blood cell lysis buffer (**Appendix 3**) up to a total volume of 15ml. The remaining nucleated cells were re-suspended in 600 µl of a white blood cell lysis buffer (**Appendix 3**), 2 µl of 20 µg/ml proteinase K (**Appendix 3**) and 20 µl of 10% sodium dodecyl sulphate (SDS, **Appendix 3**). This mixture was then incubated overnight at 37˚C to allow complete lysis of the white blood cells. After incubation, 400 µl of 6M NaCl (**Appendix 3**) was added to the solution, which was then vortexed and centrifuged for 20 minutes (min). The supernatant was transferred into a sterile 15ml plastic tube containing 2 ml of absolute ethanol, and inverted ten times to precipitate the DNA out of solution. This was followed by centrifugation at 2,500 rpm for 10 min, after which the ethanol was decanted leaving the DNA pellet at the bottom of the tube. The pellet was washed by adding 2 ml of 70% ethanol, followed by centrifugation at 25,000 revolutions per minute (rpm) for 5 min, and discarding of the ethanol. The DNA pellet was allowed to dry by leaving the tube inverted at room temperature for at least an hour, and then resuspended in 300 µl of 1X Tris EDTA (TE) buffer (**Appendix 3**). For further work, an aliquot of the stock DNA was diluted to 20 ng/µl and then stored at -20˚C.

2.3.3 Sample pre-processing and genotyping

DNA sample pre-processing and genotyping was done at the Centre for Proteomic and Genomic Research (CPGR, Cape Town, South Africa) using a custom SAXv2 Affymetrix Axiom™ array (Affymetrix, Santa Clara, California, USA). This array was designed to have probes to capture genetic variants including CNVs that are common in the Xhosa population, as well as those that had previously been associated with schizophrenia. The selection of probes for the array was based on an analysis of 400 Xhosa samples (half with schizophrenia, and the other half without) that were genotyped on the Affymetrix CytoScan™ HD array, variation from exome sequence data from about 1,800 individuals and the African Genome Variation Project and variants from the Infinium PsychArray (Illumina, San Diego, California, USA)

Sample pre-processing was done using the QIAgility System (Qiagen, Hilden, Germany), and genotyping was conducted on the GeneTitan® Multi-Channel instrument. Both pre-processing and genotyping were run in the manner described below and outlined in **Figure 2.1.**

Upon receipt of the DNA samples at a concentration of at least 20ng/µl, the DNA concentration was normalized to 10 ng/µl using the QiAgility System (Qiagen), and then transferred to 2.2 ml deep-well plates in which the polymerase chain reaction (PCR) assays were performed. To perform the assays, samples were denatured and neutralized in the QiAgility Systems Liquid Handler (Qiagen). Amplification master mix (Qiagen) was added to the samples and incubated at 37°C for 23 hours. The amplified product was then fragmented at 37°C for 30 min after addition of the fragmentation mix (Qiagen). The amplification reaction was terminated by adding a 'fragmentation stop solution' (Qiagen). Fragmented DNA was precipitated by addition of precipitation mix (Qiagen) and isopropanol, and incubation at -20°C for at least 16 hours. The fragmented DNA pellet was recovered after the precipitation solution was centrifuged at 3,200x *g* at 4°C for 40 min. The DNA was then resuspended in 100 µl of a propriety elution buffer.

The eluted DNA was quantified using the Nanodrop 1000 spectrophotometer (Nanodrop Technologies Inc., Wilmington, DE, USA). The quality of the DNA was assessed on a 4% e-gel. DNA samples that exceeded 1,000 µg and with fragments between 25 bp and 125 bp were transferred onto PCR plates and mixed with hybridization buffer (Thermofisher, Waltam, Massachusetts, USA) which facilitates the binding of the DNA molecules to the plates. The samples were then denatured at 48°C and transferred to a hybridization tray (Thermofisher), which was then loaded onto the GeneTitan multi-channel (MC) instrument (Thermofisher) for DNA to hybridize for 23.5 to 24 hours. Subsequently, ligation, staining, stabilization and scan reagents (Thermofisher) were added to the plates, before plates were scanned on the GeneTitan MC instrument.

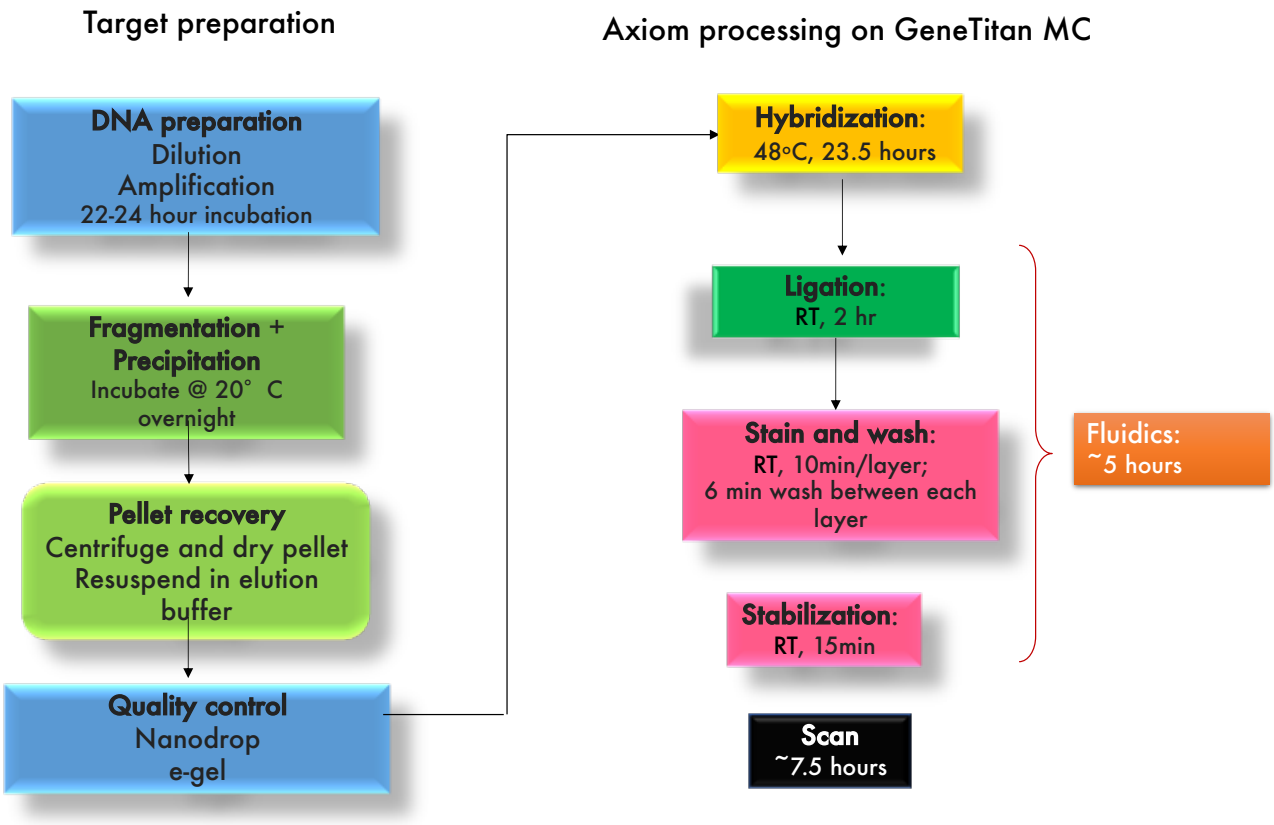


Figure 2.1 - Illustration of the workflow followed to pre-process and genotype DNA samples at the CPGR.
Sample pre-processing was done using the QiAgility System. The process followed is shown on the left panel with the first and last steps in the process shown in blue. The green blocks indicate the intermediary processes. The genotyping process, which is shown on the right panel was conducted on the GeneTitan multi-channel (mc). The process was initiated with hybridization of DNA to plates, and ends with scanning of plates to detect genotypes.

2.3.4 Genotype calling and Quality Control

Genotype data was received from CPGR in the form of .cel (intensity) files. Genotyping quality control (QC) was assessed using the 'Best Practices Genotyping analysis workflow' designed by Affymetrix on the Axiom Analysis Suite Software version 5.0.1 (Affymetrix) implementing apt-genotype-axiom (http://www.affymetrix.com/support/developer/powertools/changelog/apt-genotype-axiom.html). In the workflow poor-quality samples were identified using two metrices: single-sample metric called 'Dish QC' (DQC) and 'Sample QC' call rate.

DQC is a measure of how well signals can be resolved – this is known as contrast and calculated as the difference between the AT and GC signals over total signal. It is based on probe intensities at genomic sites that do not differ between individuals (i.e. those that are not polymorphic). Ideally, probes that are expected to ligate A or T nucleotides should produce signal in the AT channels, and background signal in the GC channel. Conversely, G or C nucleotides should produce signal in GC channels while producing background signal in the AT channel. A high-quality sample will have a high signal in the expected channels and low signal in the background channel. A contrast of 0 indicates poor resolution between signals, whereas a contrast value of 1 indicates perfect resolution. Samples with DQC value of less than 0.82 were excluded.

For 'Sample QC', samples were genotyped using 200,000 autosomal probes, a subset of the comprehensive probe set. Samples with a QC call rate less than 0.97 were excluded. A plate (96-well format) was approved for genotype-calling when the plate QC was above 98.5%, where plate QC is calculated as the number of samples that passed DQC and QC call rate/ total samples on the plate, all multiplied by 100 (equation below)

$$Plate\ QC = \frac{number\ of\ samples\ on\ a\ plate\ that\ passed\ DQC\ and\ QC\ call\ rate}{total\ number\ of\ samples}\ X\ 100$$

The quality of called genotypes was assessed by generating cluster plots. For each plot, SNPs were expected to cluster as AA – major homozygous, AB – heterozygous or BB—minor homozygous, as illustrated in **Figure 2.2**

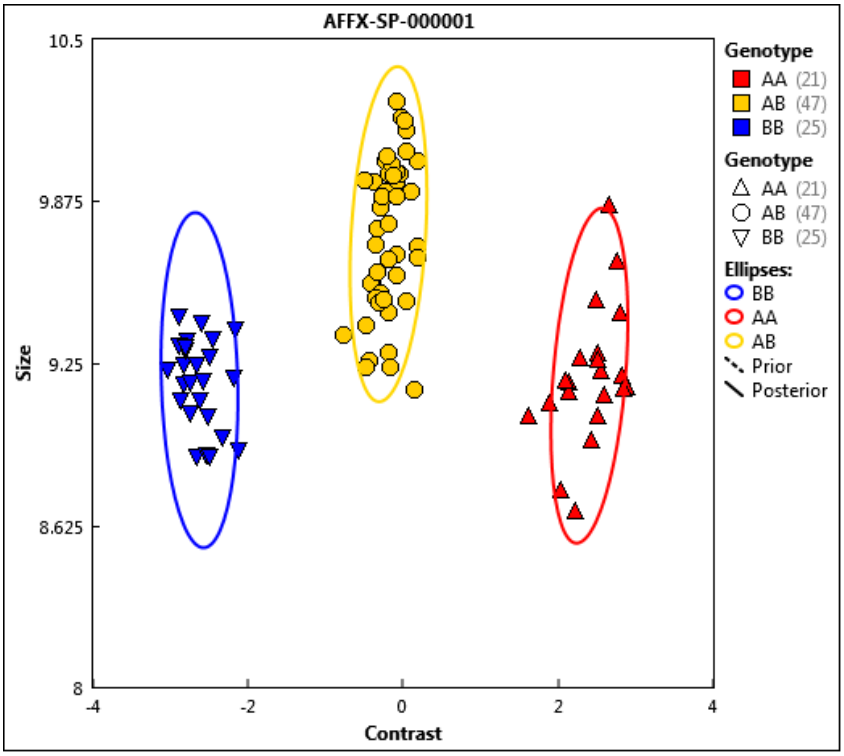


Figure 2.2 - An illustration of SNP genotype clusters.
The blue and red triangle represent the minor homozygous (BB) and major homozygous (AA) alleles, while the yellow circles represent the heterozygous alleles (AB). The ellipses around the shapes represent the prior and posterior ellipses give an indication of where the genotypes are expected compared to where they cluster.

2.3.5 GWAS Quality Control

GWAS studies are susceptible to spurious associations that may occur because of the quality of the input genetic data. Data quality can be impacted by several factors including study design, including ascertainment of cases and controls in the case of binary traits, handling of samples, batch effects and the quality of the genotypes generated. It is therefore important that a quality check is conducted prior to association testing to minimise the potential of false-negative and false-positive associations in GWAS and the likelihood of carrying these into downstream analyses.

Several GWAS QC standards exist in the literature. For this study, GWAS QC was conducted following the recommended best practices by McRae et al. (2017). All quality control steps were carried out by implementing a QC pipeline compiled by Ellingson and Fardo (2016), which runs the steps outlined below in PLINK v1.9b (Chang et al., 2015).

Sample quality control: A high genotype missingness rate indicates low quality or low concentration DNA sample Samples were removed if they had a genotype missingness rate of more than 5%. Such samples tend to have higher genotyping error. Secondly, samples were removed if they had a heterozygosity rate more or less than two standard deviations from the mean. Samples were also removed if the reported sex was different from the genetically determined sex as this may indicate sample swapping or mislabelling. Lastly, samples were also removed if they had an Identity by Descent (IBD) score greater than 1.875 which represents halfway between first- and second-degree relatedness. Relatedness can confound association testing as these tests assume that observations are independent of each other. Population outliers were removed if their principal components were more or less than two standard deviations of the mean.

SNP quality control: SNP quality control was initiated by removing SNPs that had a genotype missingness rate of more than 1%. Hardy-Weinberg equilibrium (HWE) indicates random mating in the population; SNPs that deviate from this equilibrium indicated non-random mating, selection and genotyping error (Ryckman & Williams, 2008). SNPs that deviated from HWE with $P < 10e\text{-}06$ were removed. Additionally, SNPs with a minor allele frequency (MAF) of less than 1% were also removed, since GWAS are statistically underpowered to detect effects from rare variants. GWAS are specifically designed to test for the association of SNPs that are relatively common. SNPs that had a differential missingness rate between cases and controls were also removed. This is important because GWAS tests for allelic frequency difference between cases and controls. Differential missingness of SNPs in cases and controls could lead to spurious association.

2.3.6 Phasing and imputation

Imputation of ungenotyped variants was done on the Wellcome Trust Sanger Institute Imputation Server (https://imputation.sanger.ac.uk/) (McCarthy et al., 2016), including genotype phasing with EAGLE2 (Loh et al., 2016), and imputation by the Positional Burrows Wheel Transform (Rubinacci et al., 2020) tool. Data was prepared according to the instructions on the webpage. Briefly, PLINK-format files were converted to a variant calling format (VCF) using bcftools (Li et al., 2009) and sorted by genomic position. The imputation was done using the African Genome Resources reference panel which contains all the African and non-African populations from 1000 Genomes Project Phase 3 (1KGP3) in addition to samples from

Uganda, Ethiopia, Egypt, and South Africa. This African resource has previously shown superior imputation performance compared to the prevalent 1KGP3 based reference panels for South African samples (Schurz et al., 2019). Post imputation quality filtering was based on population allele frequency cut-off of 1%. This threshold was validated by the distribution of info score values as suggested by Coleman et al. (2015).

2.3.7 Principal component and admixture analysis

Principal component analysis (PCA) was performed using PLINK v1.9b (Chang et al., 2015). Initially, the first 20 principal components (PCs) were computed for the SAX case-control dataset based on 800,000 genotyped SNPs. The imputed data were then merged with 1KGP3 dataset, which consists of 2,504 samples from 26 different populations across the globe, and another dataset containing 220 Khoi-San samples from a study by Schlebusch et al. (2012). Quality control filters were applied to each dataset before merging as follows: filtering out SNPs with MAF < 1% (--maf 0.01, --geno 0.01) and pairwise linkage disequilibrium (r^2) above 0.1 (--indep-pairwise 50 10 0.1). Dataset merging was done based on 210,702 autosomal SNPs that were common in all three datasets using PLINK v1.9b (Chang et al., 2015). The ADMIXTURE (Alexander & Lange, 2011) algorithm was implemented to determine the extent of admixture in the SAX samples. Admixture plots were constructed in pong (Behr et al., 2016). The population codes and descriptions for both PCA and admixture analyses can be found in **Appendix 4.**

2.3.8 Power calculation

The sample size calculation for the current case-control study was based on the requirement to replicate the published schizophrenia significant SNPs from the PGC in EUR samples (PGC-EUR) (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014). The power calculation was carried out using the model implemented by Purcell et al (2003). For each PGC associated SNP, it was determined whether the published odds ratio lies within the 95% CI estimated in the current SAX GWAS dataset. Power analysis was conducted on the basis of a lifetime risk of 1%, the published effect size and MAF of each SNP and the size of the SAX dataset. It was assumed that the level of significance under a one-sided test would be 0.05. The error rate for cases was set at 5% for cases, based on general evidence that misdiagnosis rates are often greater than 5%.

2.3.9 Genome-wide association testing

Several genome-wide association tests were conducted. First, the association of only the QC'd genotyped SNPs were tested. Then the association of the imputed data that passed quality control was tested. For these two GWASs, a logistic regression that included the first 20 PCs as covariates was performed. The PCs were included to control for population stratification. Another GWAS was conducted to test the impact of childhood trauma on the association test result. For this, scores from the CTQ questionnaire as well as the first 20 PCs were included as covariates in the logistic regression model. Although the first 10 PCs have been shown to effectively correct the effects of population stratification, 20 PCs were used in these analyses to correct for population stratification that may result due to cryptic relatedness. This was done to avoid inflating the test statistic. Finally, a sex-stratified GWAS was conducted, which included both PCs and CTQ scores as covariates in the model.

2.3.10 Annotation of GWAS SNPs

For all the generated GWAS summary statistics, FUMA (Functional Mapping and Annotation of Genome-Wide Association Studies) (Watanabe et al., 2017) version 1.3.6 was used to annotate GWAS SNPs. FUMA is a 'one-stop' web-based application (https://fuma.ctglab.nl) that integrates data from across 18 existing biological data repositories (**Appendix 5: Supplementary Table 2.1**) to aid in the prioritization of likely causal variants and genes from GWAS summary statistics.

The SNP2GENE function was used to annotate SNPs according to their biological function and genes to which they map. The African population genomes from 1KGP3 were chosen as the LD reference panel. SNPs that had $P < 1e\text{-}5$ and independent of each other at linkage disequilibrium $r^2 < 0.6$ were defined as independently associated SNPs. For each of these SNPs, correlated SNPs (within the summary statistics files and from the reference panel) with $r^2 > 0.6$ were included for annotation. Independent lead SNPs were defined as independently associated SNPs at $r^2 < 0.1$. A genomic risk locus was defined as an LD block of independently associated SNPs within a 250kb window of each other. Gene mapping and functional consequence of variants on genes is done within FUMA using ANNOVAR (Wang et al., 2010), deleteriousness scores (CADD) (Kircher et al., 2014), potential regulatory function scores (RegulomeDB) (Boyle et al., 2012) and 15-core chromatin state (Ernst & Kellis, 2012). Further, mapped genes were annotated with two scores that indicate the tolerance of those genes to

mutations: the probability of being a loss-of-function intolerance score (pLi) (Lek et al., 2016) and the non-coding residual variation intolerance scores (ncRVIS) (Petrovski et al., 2015).

The GENE2FUNC function was used to functionally annotate the mapped genes, the output from the SNP2GNE function. Gene-based analyses were conducted using MAGMA (de Leeuw et al., 2015). First, a gene analysis is done by computing a gene-based p-value for mapped genes. Then the gene-set p-value was computed using p-values from 4,728 curated gene sets and GO terms obtained from MsigDB v5.2. Bonferroni correction and False Discovery Rate (FDR) were used to correct for multiple testing for the gene and gene-set analyses, respectively (Weisstein, 2004)

2.3.11 Genetic correlation estimation

The genetic correlation (r_g) between schizophrenia and childhood trauma in SAX was estimated from GWAS QC'ed genotypes in GCTA using the bivariate analysis model (Lee et al., 2012). First, total childhood trauma was calculated quantitatively as the sum of the individual trauma domains. Then, GCTA was run using the *--reml-bivar 1 2* flag to determine the genetic correlation between phenotype 1 - childhood trauma and phenotype 2 - schizophrenia. The first 10 PCs were used as covariates.

All aspects of the methodology, from the isolation of DNA, QC of DNA and genetic data for GWAS, to the analyses were done by the candidate, unless otherwise stated.

2.4 Results

2.4.1 Study participants

A total of 3,000 participants who self-identified as Xhosa (i.e. identify themselves as belonging to the Xhosa ethnic group) were enrolled in this study. The final cohort that was used for the genetic association testing (n = 2,086) comprised of 1,038 cases and 1,048 controls from psychiatric hospitals and community healthcare centres in the Western and Eastern Cape Provinces (**Figure 2.3**). The mean age of study participants was 36 years. Males made up over 80% of the study sample (**Table 2.1**).

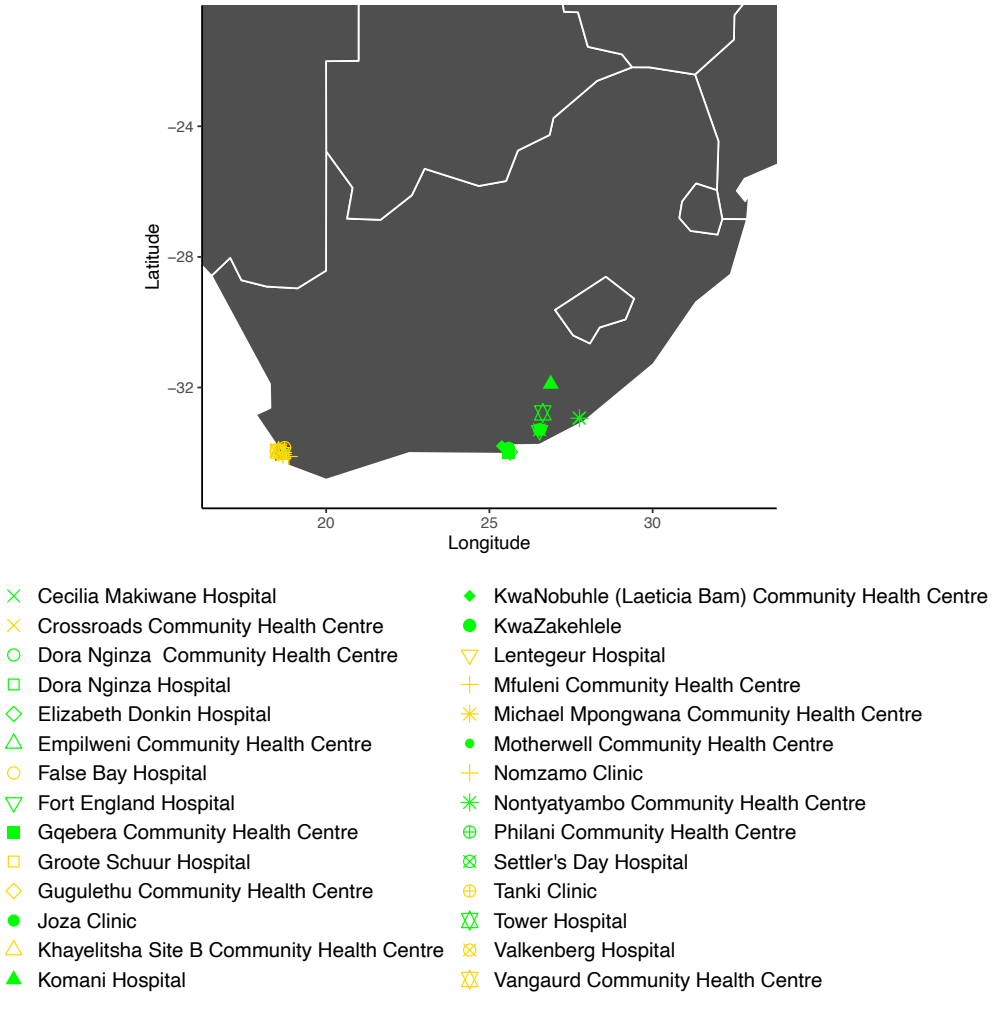


Figure 2.3 - SAX study recruitment sites and study participant demographics.
A) Cases and controls were recruited from the Western (yellow symbols) and Eastern Cape (green) symbols) Provinces of the South Africa. The shapes indicated the different sites from where study participants were recruited. B) Study participant distribution of case-controls status, sex and age.

Table 2.1 - SAX participant demographic information

		Cases	Controls
Total no. of participants		1038	1048
Sex	Male	915 (88)	921 (87)
No.(%)	Female	123 (11)	127 (12)
Age (Years)	Mean	36.17	36.02
	Median	35	35
	SD	9.1	9.0
	Range	21 - 55	21 - 54

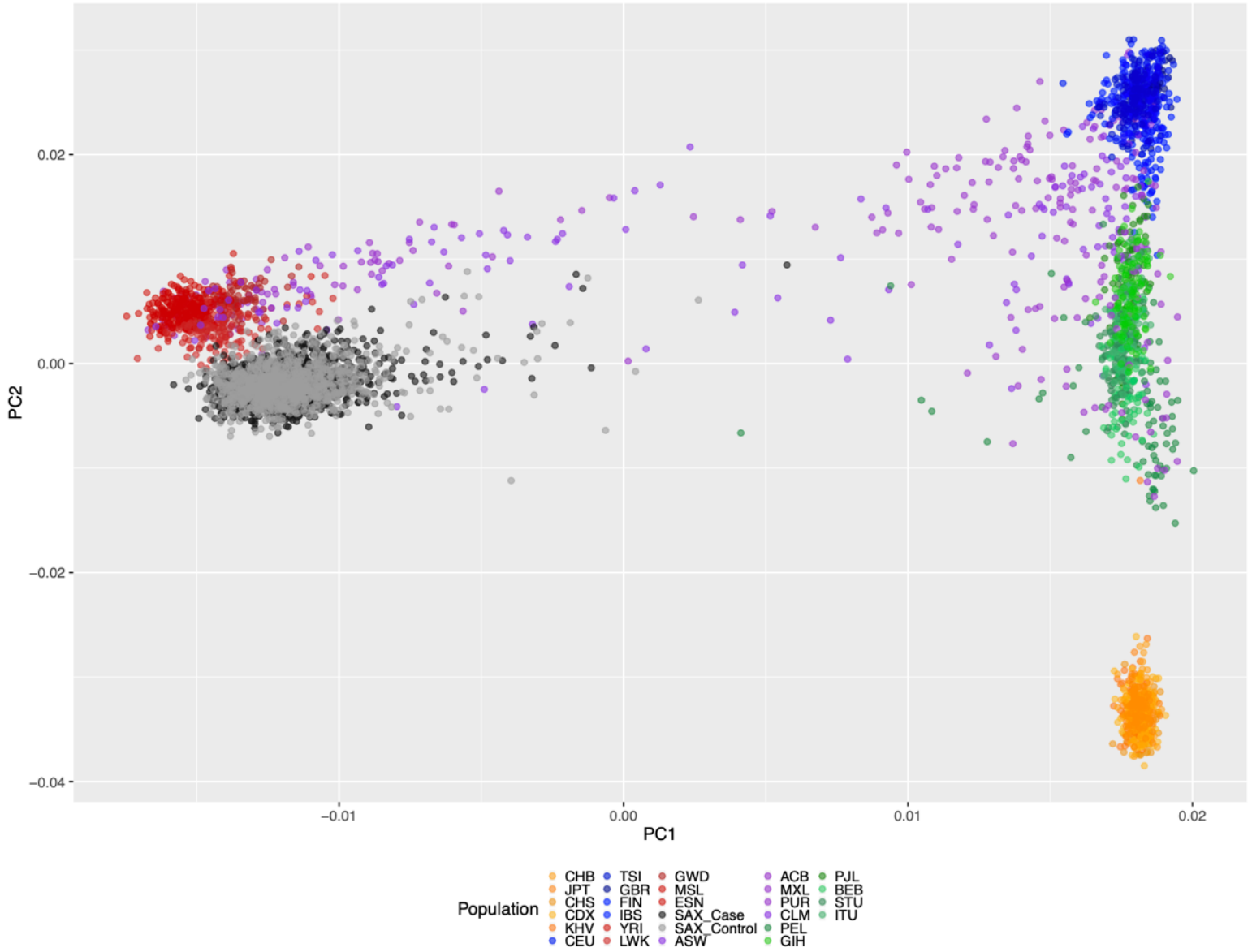
no; number, %; percentage

2.4.2 Population structure

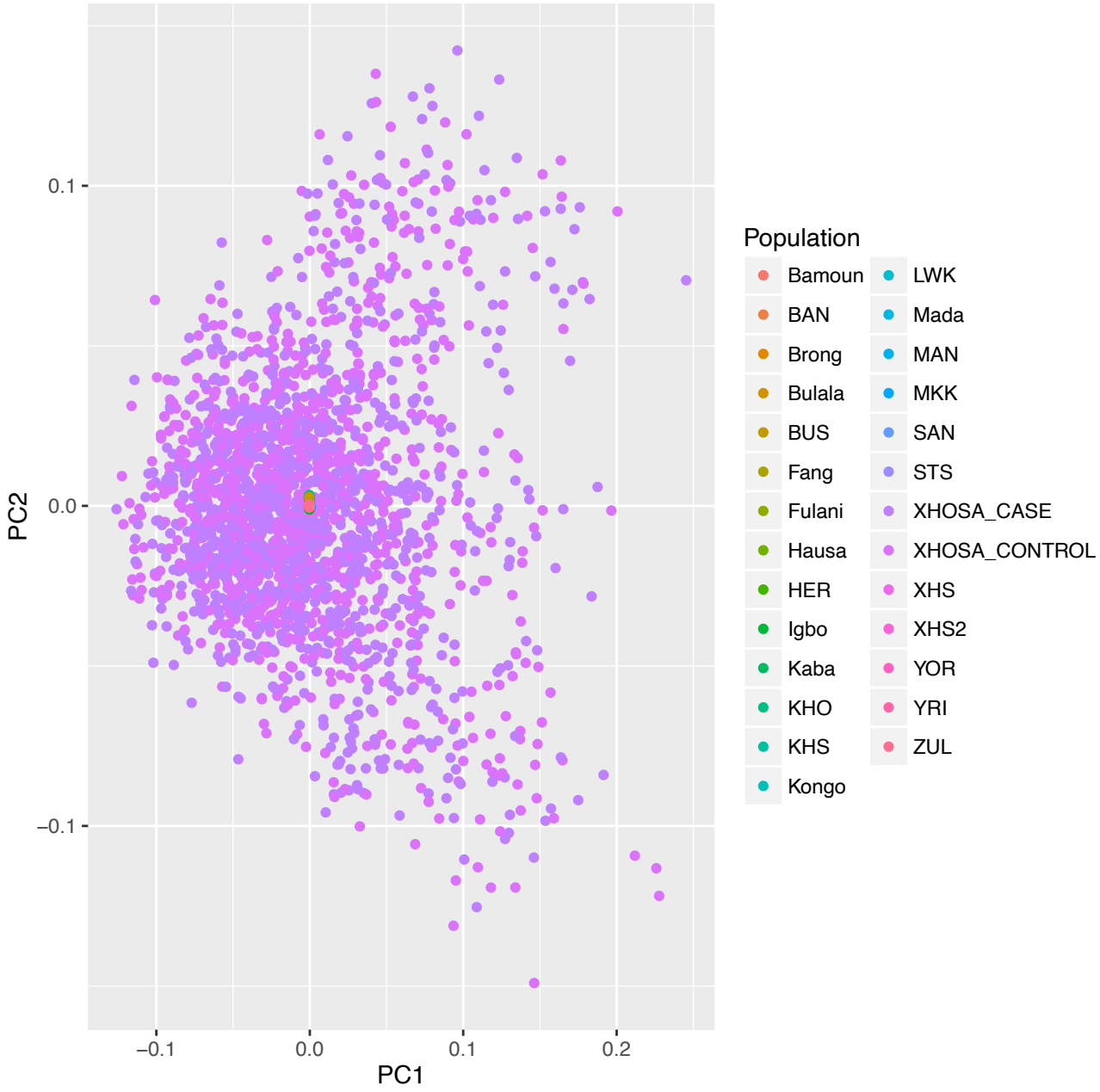
A PCA of the SAX samples, merged with three additional datasets: (i) the southern African Khoi-San samples (Schlebusch et al., 2012), (ii) 1000 Genomes Project Phase 3 (1KGP3) (Auton et al., 2015), and (iii) five South African population including the Xhosa (Chimusa et al., 2015) was plotted. The 1000 Genomes Project data was merged with SAX separately from the other data to evaluate clustering of SAX sample in the context of global diversity. As expected, the principal components (**Figure 2.4A**) showed the clustering of populations by geographical region; that is the EUR samples clustered together, as did the EAS, South Asian, American and AFR samples, respectively. Two close but distinct clusters can be seen among the AFR samples suggesting genetic heterogeneity between SAX and West and East AFR samples. PCA of SW and SE Bantu; and the Khoi-San samples with the Xhosa samples shows no distinction of the Xhosa samples from the AFR samples (**Figure 2.4B**).

To further characterise ancestry, an admixture analysis for up to K=4 (ancestries could not be broken down further than this) was performed. K=4 shows three ancestral lineages. The admixture plot (**Figure 2.4C**) confirms that the Xhosa participants in our study are an admixed population of West African and Khoi-San ancestry.

A


0.02
0.00
−0.02
−0.04
PC2
−0.01
0.00
0.01
0.02
PC1
Population
CHB
JPT
CHS
CDX
KHV
CEU
TSI
GBR
FIN
IBS
YRI
LWK
GWD
MSL
ESN
SAX_Case
SAX_Control
ASW
ACB
MXL
PUR
CLM
PEL
GIH
PJL
BEB
STU
ITU

B


0.1
0.0
−0.1
PC2
−0.1
0.0
0.1
0.2
PC1
Population
Bamoun
BAN
Brong
Bulala
BUS
Fang
Fulani
Hausa
HER
Igbo
Kaba
KHO
KHS
Kongo
LWK
Mada
MAN
MKK
SAN
STS
XHOSA_CASE
XHOSA_CONTROL
XHS
XHS2
YOR
YRI
ZUL

C

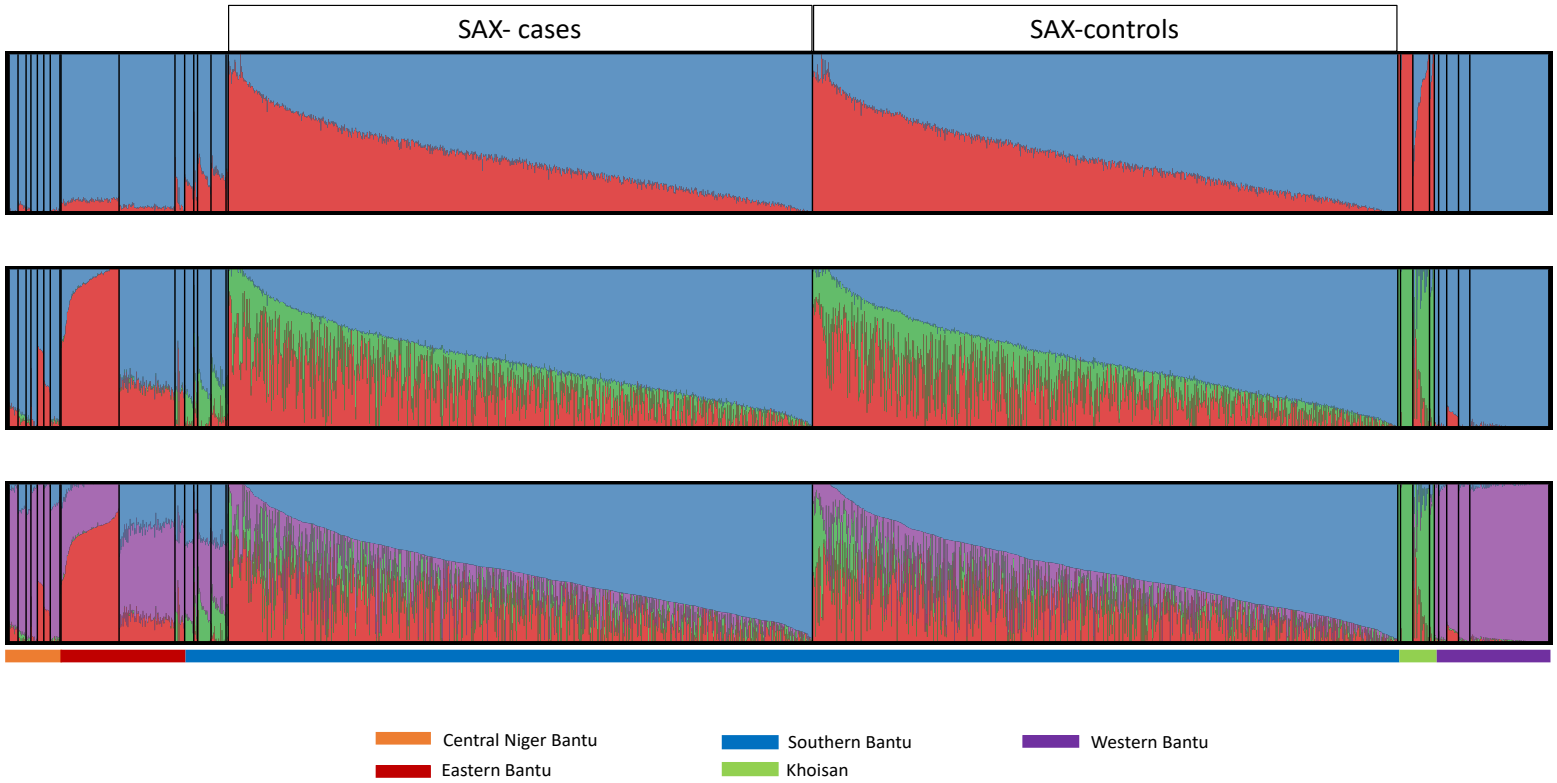


Figure 2.4 - Principal component and admixture analysis of SAX samples against global populations.
Principal component 1 and 2 of SAX samples against (A) 1KGP data and (B) Chimusa et al.(2015) African dataset (C) SAX Admixture mapping with Pagani dataset for K=2, K=3 and K=4.

2.4.3 GWAS of schizophrenia in SAX

After GWAS quality control, 566,146 SNPs and 2,086 individuals were considered for genome-wide association testing. A principal component analysis was done to investigate whether there were genetic differences between cases and controls. **Figure 2.5** shows no genetic difference between cases and controls (no more than 3 standard deviations) confirming no population stratification. The first 20 PCs were used as covariates in the logistic regression model to test for the association of the genotyped SNPs with schizophrenia. Although no SNP achieved the accepted GWAS level of significance, p-value < 5e-8, there were 69 SNPs that were suggestively significant (i.e. p-values < 1e-4). The genomic inflation factor (λ_{gc}) of 1.06, indicated no influence from population stratification on these GWAS findings (**Figure 2.5**).

To improve statistical power to detect association between the genotyped SNPs and schizophrenia, the GWAS quality-verified genetic data was imputed. Following imputation, SNPs with INFO less than 0.8 and MAF of less than 1% were removed from the dataset leaving 15 million SNPs and 2,086 samples remaining for further analysis. Genome-wide association testing was done, including the first 20 principal components. The top five SNPs, where top SNPs refers to the SNPs with the smallest p-value, were all in high LD and in the gene *ZFP3* (on chromosomal region 17p13.2, with rs35172303 significantly associated with schizophrenia (*P* = 4.74e-08, OR = 0.6004, 95%CI:[0.499,0.721], **Table 2.2, Figure 2.6)**.

Functional annotation of SNPs with *P* < 1e-05 indicated that a large proportion of these SNPs were intergenic and mapped to 60 genomic loci, where a genomic locus is defined as SNPs that are in linkage disequilibrium ($r^2 > 0.6$) with the SNP with the smallest p-value and were within a 250kb window of that SNP. The SNP that met the GWS, rs35172303 on chromosome 17p13.2, was in high LD with 14 other SNPs in this region.

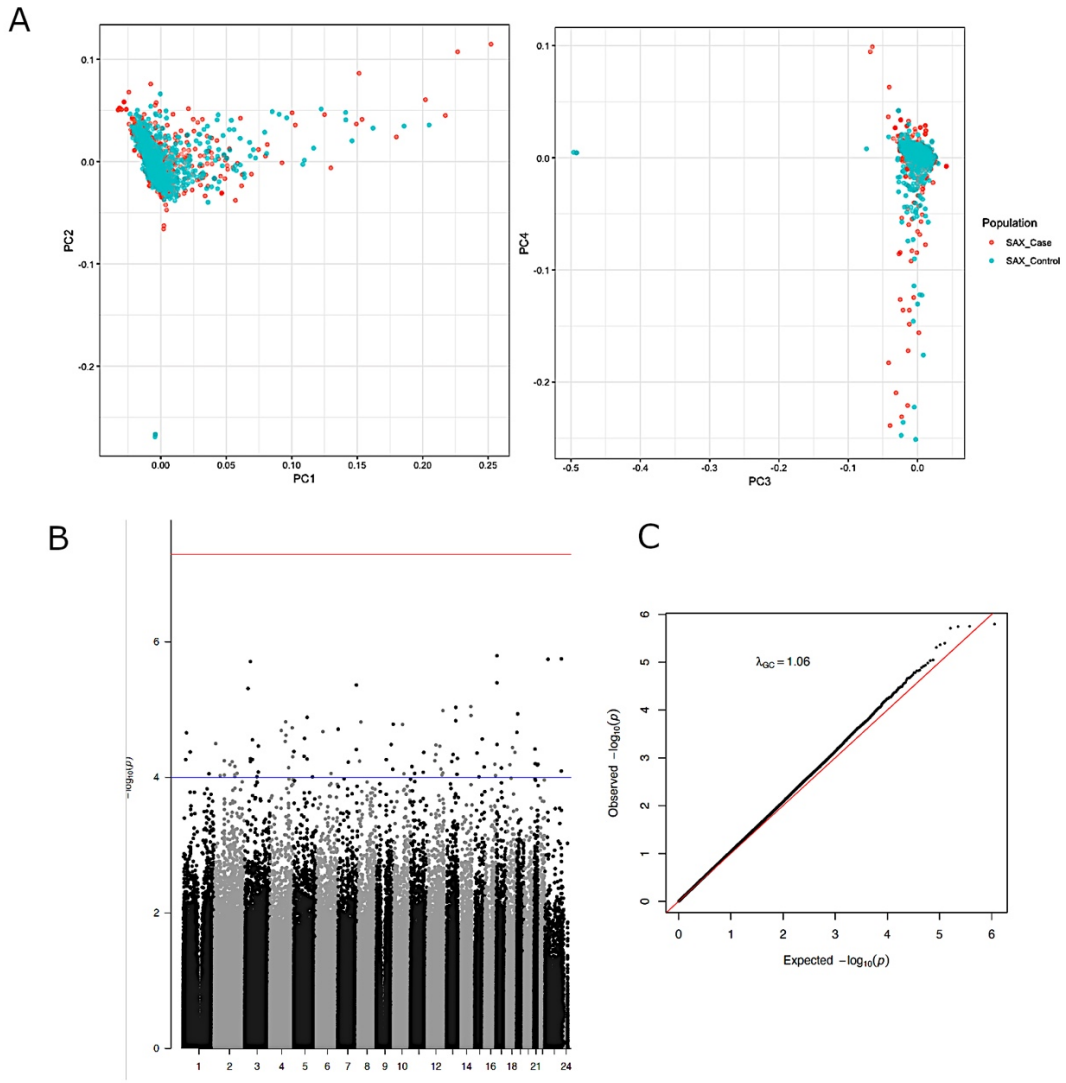


Figure 2.5 - Genome-wide association analysis of unimputed SAX genotype data.
A) Principal components 1 vs 2, and 3 vs 4 for case and control data. Cases are in the pink shade, controls are green. B) Manhattan plot of the GWAS association test. The red line denotes the genome-wide significance thresholds ($-\log_{10}P = 7$). The blue line denotes the level of suggestive significance at $-\log_{10}P = 4$). C) Quantile-Quantile (Q-Q) plot of expected vs observed $-\log_{10}P$ values. The insert is the genomic inflation factor (lambda GC).

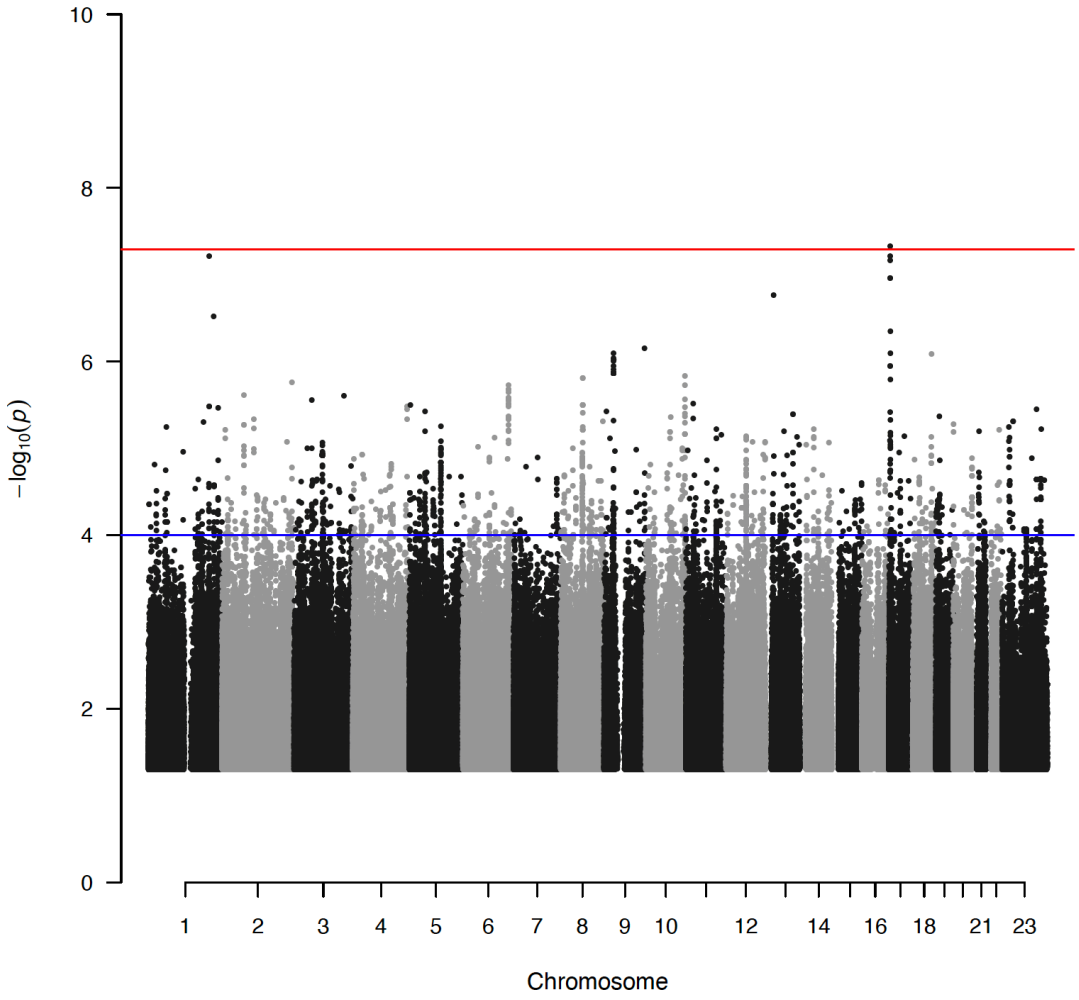


Figure 2.6 - Manhattan plot of GWAS of SAX imputed data
Only SNPs with P < 0.05 are plotted. The red line denotes the genome-wide significance level ($-\log_{10}P = 7$). The blue line denotes the level of suggestive significance at $-\log_{10}P = 4$)

Table 2.2 - Top 20 SNPs for the SAX schizophrenia GWAS

rsID	CHR	POS	Non effect allele	Effect allele	MAF	*P*	OR	95% CI	r^2	IndSigSNP	Nearest Gene	func	genomic locus
rs35172303	17	4987203	C	T	0.09758	4.722e-08	0.6043	[0.4999, 0.721]	1	rs35172303	ZFP3	intronic	17p13.2
rs12600437	17	4997433	G	A	0.09758	5.57e-08	0.604	[0.4987, 0.7212]	1	rs35172303	ZFP3	UTR3	17p13.2
rs79923411	1	205830974	G	A	0.02194	6.234e-08	2.805	[1.946, 4.143]	1	rs79923411	PM20D1	intergenic	1q32.1
rs12941688	17	4985515	G	A	0.09758	6.921e-08	0.6083	[0.5031, 0.7256]	1	rs35172303	ZFP3	intronic	17p13.2
rs79397102	17	4992172	G	T	0.07489	9.79e-08	0.567	[0.4553, 0.6951]	0.740	rs35172303	ZFP3	intronic	17p13.2
17:4992173	17	4992173	A	T	0.07489	9.79e-08	0.567	[0.4553, 0.6951]	0.740	rs35172303	ZFP3	intronic	17p13.2
rs7994406	13	27054702	C	G	0.1271	1.658e-07	1.852	[1.477, 2.355]	1	rs7994406	CDK8	intergenic	13q12.13
rs1335930	1	221094253	G	A	0.4304	2.467e-07	1.395	[1.233, 1.594]	1	rs1335930	HLX	intergenic	1q41
rs112066605	17	5868828	A	G	0.1362	3.866e-07	0.6403	[0.5372, 0.7595]	1	rs112066605	WSCD1	intronic	17p13.2
rs2797838	9	136538052	C	T	0.2337	5.422e-07	1.472	[1.266, 1.714]	1	rs2797838	SARDH	intronic	9q34.2
rs12605747	18	63501290	T	C	0.07791	6.684e-07	1.576	[1.32, 1.896]	1	rs12605747	CDH7	intronic	18q22.1
rs8077075	17	5870139	T	C	0.1369	6.803e-07	0.6463	[0.543, 0.7671]	0.994	rs112066605	WSCD1	intronic	17p13.2
rs17195085	9	32055943	T	C	0.009834	7.093e-07	0.5973	[0.4849, 0.7307]	1	rs17195085	RNA5SP281	intergenic	9p21.1
rs35010287	9	32113781	C	T	0.01891	8.218e-07	0.6146	[0.5027, 0.7434]	1	rs35010287	RNA5SP281	intergenic	9p21.1
rs144828873	9	32052642	C	T	0.009834	8.812e-07	0.5972	[0.4818, 0.7305]	1	rs17195085	RNA5SP281	intergenic	9p21.1
rs34272446	9	32521954	T	C	0.2277	9.276e-07	0.6667	[0.5671, 0.784]	1	rs34272446	DDX58	intronic	9p21.1
rs12236182	9	32056297	T	C	0.01664	1.007e-06	0.6039	[0.4912, 0.7378]	0.795	rs35010287	RNA5SP281	intergenic	9p21.1
rs10970713	9	32053690	T	G	0.009834	1.108e-06	0.6001	[0.4887, 0.737]	1	rs17195085	RNA5SP281	intergenic	9p21.1
rs1490316	9	32057628	C	T	0.009834	1.2e-06	0.6034	[0.4921, 0.739]	1	rs17195085	RNA5SP281	intergenic	9p21.1
rs10813709	9	32055833	G	A	0.009834	1.2e-06	0.6034	[0.4921, 0.739]	1	rs17195085	RNA5SP281	intergenic	9p21.1

BP, genomic position in HG38; OR, odds ratio; MAF, minor allele frequency; func, function; r^2, linkage disequilibrium correlation of variant to the independently associated SNP (IndigSNP); IndSigSNP, independently associated SNP based on $r^2 < 0.1$ in relation to any other SNP.

2.4.4 GWAS of schizophrenia and childhood trauma

Childhood trauma has previously been associated with schizophrenia in this cohort (Mall et al., 2019) Study participants completed the childhood trauma questionnaire, with questions based on five domains of childhood trauma: physical, sexual and emotional abuse; and physical and emotional neglect. Childhood trauma experience data was collected on an ordinal scale: a score of 0 representing not having ever experienced the childhood trauma domain, while the maximum score of 25 represented frequently experiencing the childhood trauma domain. In general, emotional neglect was the most frequently reported childhood trauma domain in the SAX cohort, followed by physical neglect (**Figure 2.7)**

The objective behind this aspect of the study was to investigate whether childhood trauma impacts the genetics of schizophrenia. In this regard, a GWAS controlling for the first 20 principal components (as was previously done), as well as the five childhood trauma domains, was carried out. Five SNPs, all on chromosome 17p13.2 and in *ZFP3* achieved GWS (**Figure 2.8,Table 2.3**). These SNPs were the same five SNPs that had the smallest p-value in the prior GWAS. The improvement of the GWAS test statistics after adjusting for childhood trauma implies that there is an interaction between childhood trauma and genetics that results in schizophrenia. Additionally, a positive genetic correlation (r_g = 0.19 , se = 0.06) was observed between schizophrenia and childhood trauma. This suggests a 'causal' path from childhood trauma to schizophrenia, specifically that the more childhood trauma an individual experiences, the more likely they are to develop schizophrenia.

2.4.5 GWAS of schizophrenia by sex

The male-only GWAS yielded five genome-wide significant SNPs (**Table 2.4**) which emerged from the analysis done in the entire dataset (**Table 2.3** above), indicating that the association was driven by the genetic risk burden from males. The characterization of SNPs with *P* < 1e-05 (n = 539 SNPs) showed that the largest proportion of these were intronic (**Figure 2.9B**). These SNPs mapped to 34 genomic loci and 57 genes. The most significantly associated locus on chromosome 17 contained 12 SNPs that were correlated with the lead SNP 17:4992173 (**Figure 2.9A**)

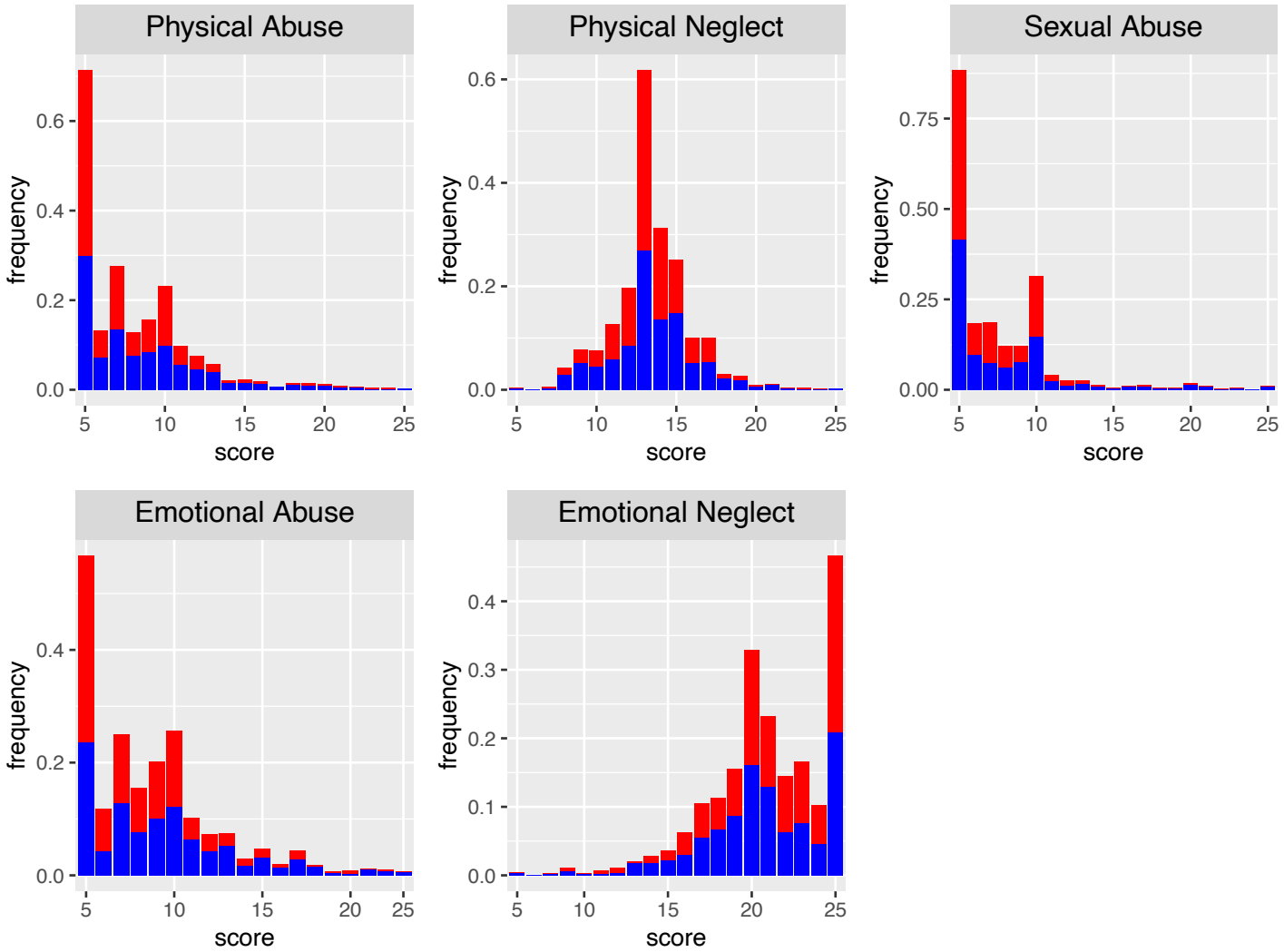


Figure 2.7 - The distribution of childhood trauma in SAX cases and controls
The frequency of childhood trauma experience categorised by the type of childhood trauma. Cases are filled in blue, while controls are filled in red.

Table 2.3 - Top 20 SNPs from SAX GWAS controlling for childhood trauma

rsID	CHR	POS	non effect allele	Effect allele	MAF	*P*	OR	95% CI	r^2	IndSigSNP	Nearest Gene	func	genomic locus
rs79397102	17	4992172	G	T	0.07489	9.086e-10	0.4446	[0.3445, 0.5795]	1	17:4992173	ZFP3	intronic	17p13.2
17:4992173	17	4992173	A	T	0.07489	9.086e-10	0.4446	[0.3445, 0.5795]	1	17:4992173	ZFP3	intronic	17p13.2
rs12600437	17	4997433	G	A	0.09758	4.275e-09	0.5159	[0.4176, 0.6511]	0.740	17:4992173	ZFP3	UTR3	17p13.2
rs35172303	17	4987203	C	T	0.09758	4.966e-09	0.5204	[0.4219, 0.655]	0.740	17:4992173	ZFP3	intronic	17p13.2
rs12941688	17	4985515	G	A	0.09758	6.227e-09	0.5231	[0.4244, 0.6585]	0.740	17:4992173	ZFP3	intronic	17p13.2
rs112464729	17	4979486	G	A	0.06657	1.075e-07	0.5009	[0.3939, 0.6575]	0.877	17:4992173	RP11-46I8.3	ncRNA intronic	17p13.2
rs111965854	17	4991932	A	G	0.05749	1.105e-07	0.4464	[0.3354, 0.6096]	0.749	17:4992173	ZFP3	intronic	17p13.2
rs647586	18	6001128	T	C	0.3094	1.541e-07	0.5625	[0.4531, 0.697]	1	rs647586	L3MBTL4	intronic	18p11.31
rs112625518	17	4992957	C	A	0.05749	2.176e-07	0.4527	[0.3395, 0.6194]	0.749	17:4992173	ZFP3	intronic	17p13.2
rs1749554	1	236440255	C	T	0.2133	3.005e-07	1.535	[1.309, 1.819]	1	rs1749554	ERO1LB	intronic	1q43
rs9577600	13	112709730	A	G	0.4478	6.411e-07	1.66	[1.321, 1.969]	1	rs9577600	SNORD44	intergenic	1q43
rs146742640	19	2819810	C	T	0.03707	6.438e-07	2.738	[1.834, 4.048]	1	rs146742640	ZNF554	upstream	19p13.3
rs12603507	17	4978638	G	T	0.1649	6.758e-07	0.6033	[0.4957, 0.7401]	1	rs12603507	RP11-46I8.3	downstream	17p13.2
rs7994406	13	27054702	C	G	0.1271	6.868e-07	2.008	[1.531, 2.659]	1	rs7994406	CDK8	intergenic	13q12.13
rs9701452	1	229290907	G	T	0.1271	7.342e-07	2.224	[1.619, 3.059]	1	rs9701452	RP5-1065P14.2	intergenic	1q42.13
rs10964321	9	19867614	A	C	0.208	1.129e-06	0.6226	[0.5076, 0.7448]	1	rs10964321	AL158077.1	intergenic	9p22.1
rs7597776	2	155765313	G	A	0.152	1.155e-06	1.817	[1.416, 2.289]	1	rs7597776	CBX3P6	intergenic	2q24.1
rs10776921	9	137872239	C	T	0.1808	1.234e-06	0.5112	[0.3888, 0.669]	1	rs10776921	RP11-447M12.2	intergenic	9q34.3
rs9652265	13	112688048	C	T	0.3321	1.367e-06	1.462	[1.239, 1.687]	1	rs9652265	SNORD44	intergenic	1q43

BP, genomic position in HG38; OR, odds ratio; MAF, minor allele frequency; func, function; r^2, linkage disequilibrium correlation; IndSigSNP, independently associated based on r2 < 0.1 in relation to any other SNP.

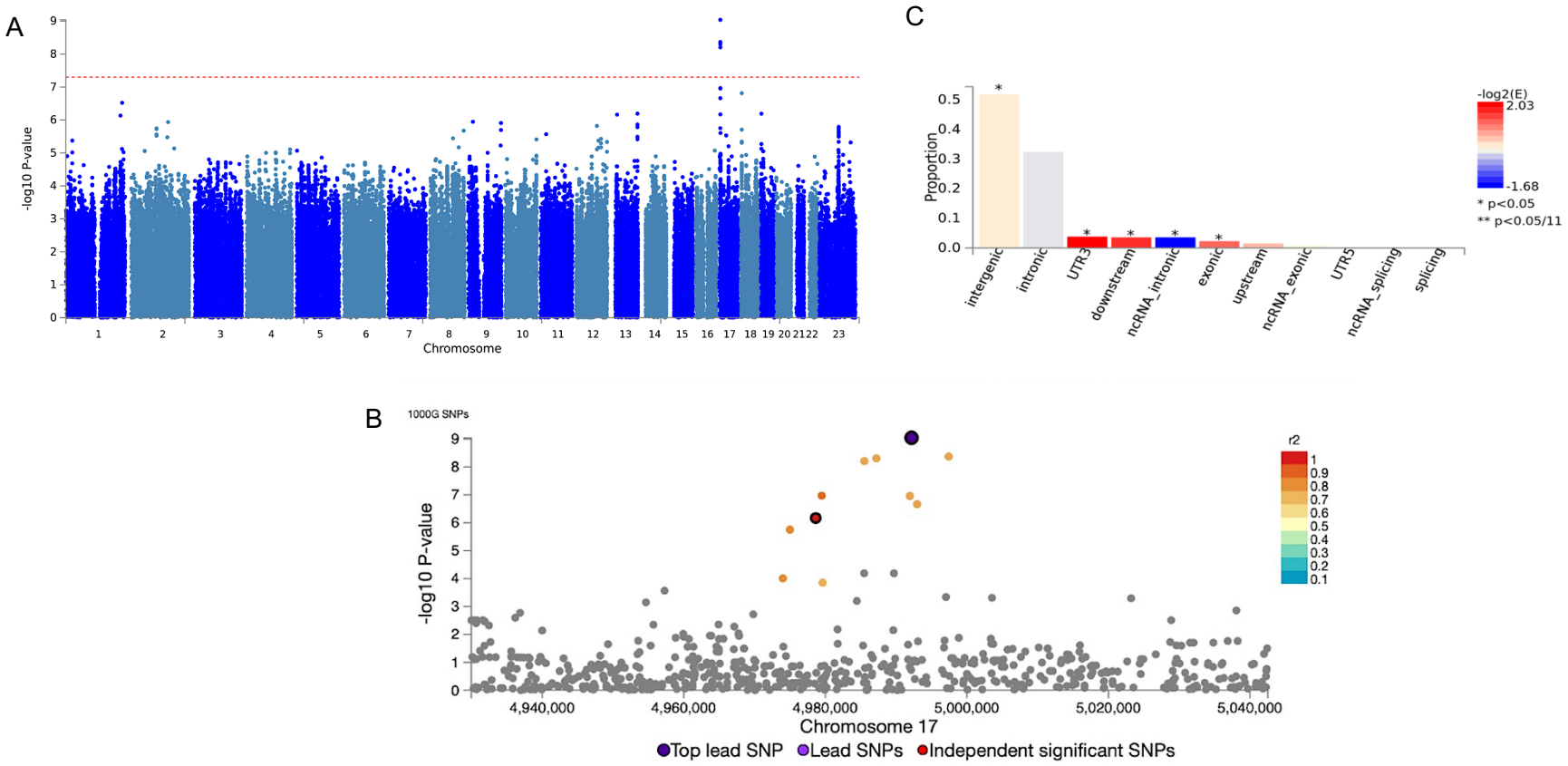


Figure 2.8 - GWAS of schizophrenia controlling for childhood trauma experience.
A) Manhattan plot of GWAS. Broken red line denotes the genome-wide significance thresholds ($-\log_{10}P = 7$). The top associated SNP (rs79397102) is labelled. B) Locus zoom plot of genomic loci on chromosome 17. Plotted variants are coloured according to their correlation (r^2) with the top SNP at the locus, including variants from 1000 Genomes Project. Only SNPs with $r^2 > 0.6$ are coloured. C) The proportion of SNPs with $P < 1e\text{-}05$ for each of the functional categories specified on the y-axis.

As expected, no GWS significant SNPs were obtained from the female-only GWAS (**Table 2.5**). The SNP with the smallest p-value was rs8051198 (*P* = 2.847e-06; OR = 7.695) on chromosome 16 (**Figure 2.10A**), which surpassed the suggestive association threshold. There were 202 SNPS that met the *P* < 1e-05 threshold selected for the annotation. A large proportion of these were intergenic (**Figure 2.10B**) and mapped to seven genomic loci (**Figure 2.10C**).

2.5 Discussion

This GWAS of schizophrenia in the Xhosa population represents the first application of this method to dissect the genetic underpinnings of a psychiatric phenotype in an African cohort of this size. The genetic variants associated with schizophrenia, as well as the impact of childhood trauma and biological sex on the aetiology of schizophrenia in the Xhosa population, were investigated.

Population structure in SAX

Only individuals who self-identified as Xhosa were recruited for this study. The inclusion criteria was limited to a single ethnic group in South Africa, despite there being so many, to limit genetic heterogeneity between study participants. This was because the sample size was small; even larger samples are required when working with heterogenous populations. Homogeneity in SAX was demonstrated through population structure analyses, i.e. PCA and admixture.

ZFP3 is associated with schizophrenia in SAX

Given that the sample size was considerably smaller than that of other studies of complex disorders, including schizophrenia, it was unexpected that the GWAS conducted here would yield SNPs associated at GWS level. However, intronic variants at chromosomal position 17p13.2, notably in the gene *ZFP3,* were consistently associated with schizophrenia. These finding were consistent with those reported by Gulsuner et al. (2020) for the same cohort using WES data. The report showed that the SNP rs12600437, which is in high LD (r^2 = 0.74) with the associated variant in this study, rs35172303, was associated with schizophrenia, albeit below the exome-wide level of significance.

Table 2.4 - Top 20 SNPs from the male GWAS

rsID	CHR	BP	Non Effect allele	Effect allele	Freq_A	Freq_U	*P*	OR	95% CI	se	r^2	IndSigSNP	Nearest Gene	function	genomic locus
rs79397102	17	4992172	G	T	0.08507	0.1436	2.469e-09	0.4308	[0.3266, 0.5681]	0.1412	1	17:4992173	ZFP3	intronic	17p13.2
17:4992173	17	4992173	A	T			2.469e-09	0.4308	[0.3266, 0.5681]	0.1412	1	17:4992173	ZFP3	intronic	17p13.2
rs35172303	17	4987203	C	T	0.1187	0.182	2.923e-08	0.5127	[0.4049, 0.6492]	0.1205	0.740495	17:4992173	ZFP3	intronic	17p13.2
rs12941688	17	4985515	G	A	0.1193	0.1823	3.809e-08	0.5163	[0.4079, 0.6534]	0.1202	0.740495	17:4992173	ZFP3	intronic	17p13.2
rs12600437	17	4997433	G	A	0.1177	0.1806	4.219e-08	0.5146	[0.4058, 0.6526]	0.1212	0.740495	17:4992173	ZFP3	UTR3	17p13.2
rs111965854	17	4991932	A	G	0.06278	0.1046	1.435e-07	0.4213	[0.3053, 0.5814]	0.1643	0.749337	17:4992173	ZFP3	intronic	17p13.2
rs1749554	1	236440255	C	T	0.3279	0.2557	1.839e-07	1.594	[1.338, 1.899]	0.08941	1	rs1749554	ERO1LB	intronic	1q43
rs112464729	17	4979486	G	A	0.0881	0.1362	2.456e-07	0.4843	[0.3677, 0.6378]	0.1405	0.877624	17:4992173	RP11-46I8.3	ncRNA intronic	17p13.2
rs112625518	17	4992957	C	A	0.06222	0.1037	2.607e-07	0.4267	[0.3086, 0.5901]	0.1654	0.749337	17:4992173	ZFP3	intronic	17p13.2
rs146742640	19	2819810	C	T	0.06301	0.03517	6.682e-07	2.856	[1.888, 4.32]	0.2111	1	rs146742640	ZNF554	upstream	19p13.3
17:39477002	17	39477002	A	C			6.802e-07	1.495	[1.275, 1.752]	0.08092	1	17:39477002	TBC1D3P7	intergenic	17q21.2
rs647586	18	6001128	T	C	0.1858	0.2478	7.822e-07	0.5616	[0.4467, 0.7061]	0.1168	1	rs647586	L3MBTL4	intronic	18p11.31
rs4790738	17	4974992	G	T	0.1703	0.2345	8.096e-07	0.5964	[0.4857, 0.7324]	0.1048	1	rs4790738	RP11-46I8.3	intergenic	17p13.2

rs761314	23	102793333	C	T	0.237	0.1613	8.352e-07	2.071	[1.55, 2.766]	0.1477	1	rs761314	RAB40A	intergenic	Xq22.2
rs12603507	17	4978638	G	T	0.1548	0.2116	9.684e-07	0.5862	[0.4734, 0.7259]	0.1091	0.811711	rs4790738	RP11-46I8.3	downstream	17p13.2
17:39507419	17	39507419	C	T			9.855e-07	0.6492	[0.5461, 0.7718]	0.08825	1	17:39507419	KRT33A	upstream	17q21.2
rs9701452	1	229290907	G	T	0.07424	0.04723	1.467e-06	2.286	[1.633, 3.2]	0.1717	1	rs9701452	RP5-1065P14.2	intergenic	7q21.3
rs114070035	19	2817555	C	T	0.09008	0.05029	1.636e-06	2.241	[1.42, 2.304]	0.1684	1	rs114070035	ZNF554	intergenic	19p13.3
17:39507970	17	39507970	G	T			1.733e-06	0.6615	[1.42, 2.304]	0.08643	0.825001	17:39507419	KRT33A	upstream	17q21.2
17:39504593	17	39504593	C	G			1.779e-06	0.6618	[0.5587, 0.784]	0.0864	0.696126	17:39507419	KRT33A	intronic	17q21.2

BP, genomic position in HG38; OR, odds ratio; MAF, minor allele frequency; Freq_A, effect allele frequency in affected individuals (cases); Freq_U, effect allele frequency in unaffected individuals (controls) func, function; r^2, linkage disequilibrium correlation; IndSigSNP, independently associated based on $r^2 < 0.1$ in relation to any other SNP.

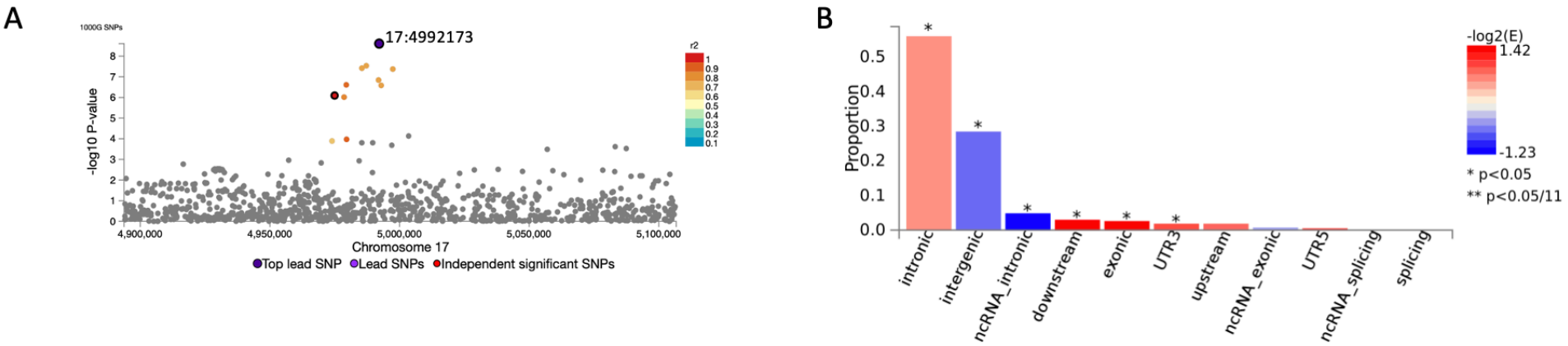


Figure 2.9 - Annotation of SNPs from male GWAS.
A) Locus zoom of genomic loci on chromosome 17. B) The consequence of GWAS SNPs with P < 1-e5 on gene function. C) Distribution of GWAS SNPs with P < 1e-5 across genomic loci in terms of size, number (#SNPs), number of genes (#mapped genes) and number of genes located in loci

Table 2.5 - Top 20 SNPs from female GWAS

rsID	CHR	POS	Non Effect allele	Effect allele	Freq_A	Freq_U	*P*	OR	95% CI	se	r^2	IndSigSNP	Nearest Gene	function	genomic locus
rs8051198	16	6541644	T	G	0.3136	0.1652	2.847e-06	7.695	[3.275, 18.08]	0.4359	1	rs8051198	RP11-420N3.2:RBFOX1	ncRNA intronic	16p13.3
rs61531769	21	33291090	T	C	0.3475	0.2059	3.295e-06	5.72	[2.743, 11.93]	0.3749	1	rs61531769	HUNK	intronic	21q22.11
rs28382800	19	10501017	A	G	0.1626	0.2677	7.232e-06	0.1227	[0.0490, 0.3067]	0.4677	1	rs28382800	CDC37	downstream	19p13.2
rs9378874	6	4420162	A	G	0.3583	0.4843	7.383e-06	0.2192	[0.1129, 0.4256]	0.3386	1	rs9378874	RNA5SP202	intergenic	6p25.1
rs6922499	6	4421361	T	A	0.3607	0.4843	7.383e-06	0.2192	[0.1129, 0.4256]	0.3386	1	rs9378874	RNA5SP202	intergenic	6p25.1
rs10023721	4	178649639	C	T	0.3319	0.452	8.048e-06	0.1988	[0.0978, 0.4041]	0.3619	1	rs10023721	LINC01098	upstream	4q34.3
rs6837148	4	178649809	A	G	0.3319	0.452	8.048e-06	0.1988	[0.0978, 0.4041]	0.3619	0.939905	rs10023721	LINC01098	upstream:downstream	4q34.3
rs6837168	4	178649835	C	G	0.3319	0.452	8.048e-06	0.1988	[0.0978, 0.4041]	0.3619	1	rs10023721	LINC01098	upstream:downstream	4q34.3
rs6837028	4	178649926	T	C	0.3319	0.4516	8.048e-06	0.1988	[0.0978, 0.4041]	0.3619	1	rs10023721	LINC01098	UTR5	4q34.3
rs6837951	4	178650206	A	G	0.3319	0.4516	8.048e-06	0.1988	[0.0978, 0.4041]	0.3619	1	rs10023721	LINC01098	intronic	4q34.3
rs6815385	4	178650457	C	T	0.3319	0.4516	8.048e-06	0.1988	[0.0978, 0.4041]	0.3619	1	rs10023721	LINC01098	intronic	4q34.3
rs10015021	4	178650548	T	C	0.3319	0.4516	8.048e-06	0.1988	[0.0978, 0.4041]	0.3619	1	rs10023721	LINC01098	intronic	4q34.3
rs34618406	11	116163151	G	T	0.4425	0.2899	8.61e-06	5.027	[2.468, 10.24]	0.3629	1	rs34618406	snoU13	intergenic	11q12.1

rs10491125	17	38895500	T	A	0.07965	0.190	9.594e-06	0.0536	[0.0146, 0.1958]	0.6611	1	rs10491125	KRT25	intergenic	17q21.2
rs12646325	4	178650931	A	G	0.3263	0.4458	9.997e-06	0.2038	[0.1006, 0.4128]	0.3601	1	rs10023721	LINC01098:LINC01099	ncRNA intronic	4q34.3
rs7664256	4	178650948	C	T	0.3263	0.4458	9.997e-06	0.2038	[0.1006, 0.4128]	0.3601	1	rs10023721	LINC01098:LINC01099	ncRNA intronic	4q34.3
rs2168039	4	178649581	G	C	0.3391	0.4556	1.07e-05	0.1982	[3.159, 19.84]	0.3676	1	rs10023721	LINC01098	upstream	4q34.3
rs9504074	6	4411482	G	A	0.395	0.5244	1.413e-05	0.2047	[0.0964, 0.4074]	0.3653	0.961674	rs9378874	RNA5SP202	intergenic	6p25.1
rs6597055	6	4405454	G	A	0.3833	0.5119	1.645e-05	0.2169	[0.0620, 0.3472]	0.3547	0.961674	rs9378874	RNA5SP202	intergenic	6p25.2
rs9392596	6	4411009	A	G	0.3833	0.5119	1.645e-05	0.2169	[0.1082, 0.4347]	0.3547	0.955778	rs9378874	RNA5SP202	intergenic	6p25.3

BP, genomic position in HG38; OR, odds ratio; MAF, minor allele frequency; Freq_A, effect allele frequency in affected individuals (cases); Freq_U, effect allele frequency in unaffected individuals (controls) func, function; r^2, linkage disequilibrium correlation; IndSigSNP, independently associated based on r2 < 0.1 in relation to any other SNP

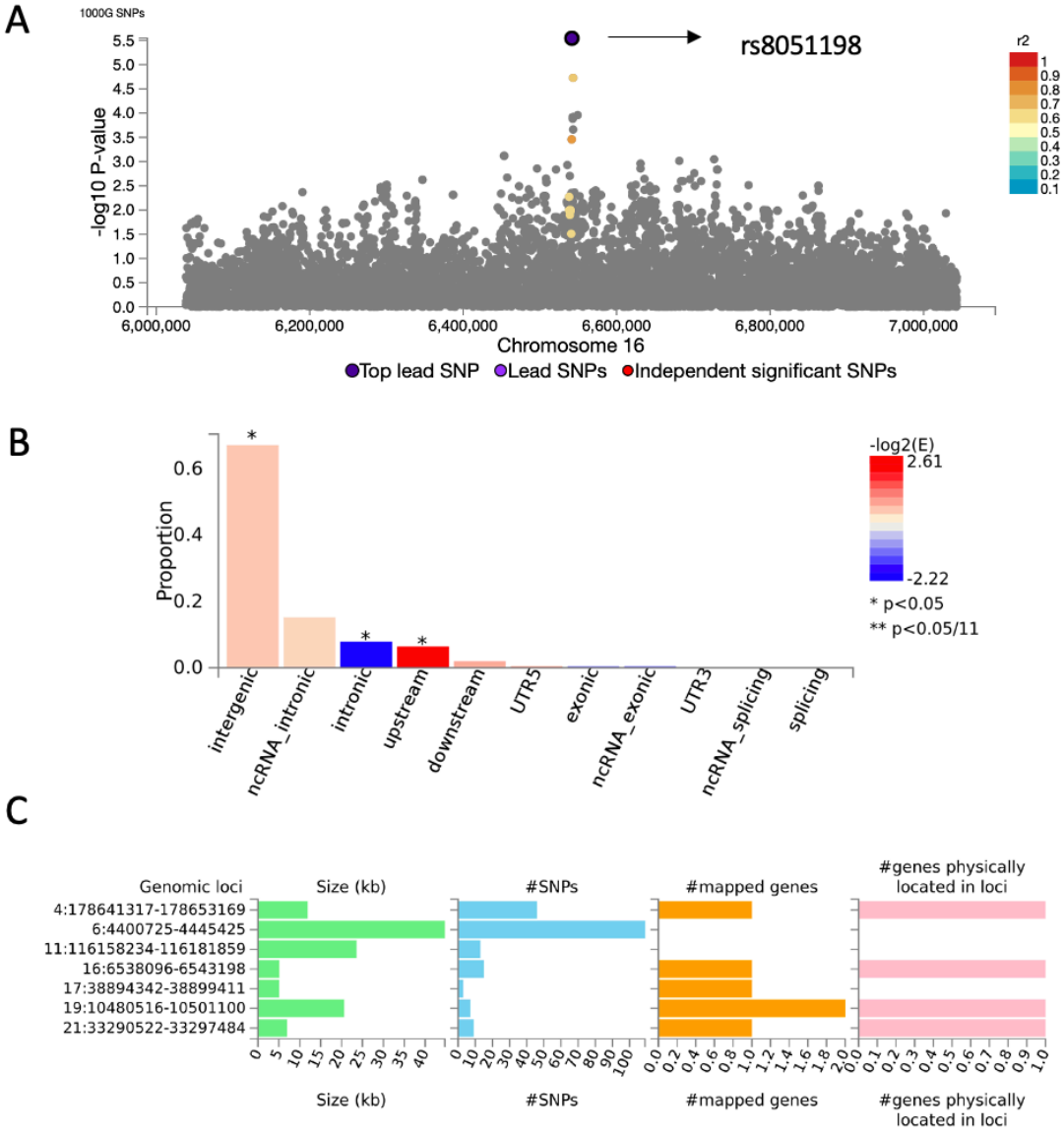


Figure 2.10 - Annotation of SNPs from SAX female GWAS
A) Locus zoom of genomic locus on chromosome. The top leading SNP is labelled.
B) Consequence of SNP with P < 1e-5 on the function of genes. C) Distribution of SNPs with *P* < 1e-5 across genomic loci.

The discovery of genome-wide significant loci in this cohort of modest size is reflective of how the genetic diversity in AFR genomes, and therefore the shorter LD blocks may be leveraged to discover previously unidentified disease-associated loci. GWAS catalogue shows that AFR ancestry populations contribute about 7% of GWAS associations across all traits, despite only making up 2.4% of the individuals included in the GWAS studies (Morales et al., 2018).

Although *ZFP3* has not been previously associated with schizophrenia, even in the published large-scale GWAS studies on schizophrenia, the largest EUR meta-analysis identified variants in *ZFP3* that were associated with Alzheimer's disease (Jansen et al., 2019). Given that there is a genetic correlation between Alzheimer's and schizophrenia, and both disorders affect some of the same regions of the brain (Creese et al., 2019), these data suggest that *ZFP3*, as a candidate gene, may be more prominent in the aetiology of schizophrenia in South African populations than has been the case for its EUR- or EAS counterparts. Larger sample sizes and replication cohorts of AFR ancestry samples are required for further investigation.

Childhood trauma and schizophrenia

It is uncommon for GWAS studies of schizophrenia to include childhood trauma as a covariate, as has been done in this chapter, perhaps because data on childhood trauma is not routinely collected during study participant recruitment. The SAX recruitment protocol was designed to allow for the collection of as much phenotypic data as possible, especially phenotypes that are known to be associated with schizophrenia. These phenotypes included childhood trauma experience.

Childhood trauma has been associated with schizophrenia in the SAX cohort and others (Mall et al., 2019). After controlling for childhood trauma in the GWAS model, the original association signal on chromosome 17p13.2 (*ZFP3*) was enhanced, suggesting that an interaction exists between childhood trauma and genetics that contributes to the xrisk of schizophrenia. To further disentangle this interaction, the schizophrenia affected group would have to be divided into two groups; one with childhood trauma and the other without similar to a study by Coleman et al. (2020)

The positive genetic correlation between schizophrenia and childhood trauma (rg = 0.19, se. = 0.066) was similar to that reported for schizophrenia ASD (r_g = 0.16, se = 0.16) (Lee et al., 2013). Individuals with childhood trauma are more likely to develop schizophrenia. However, it should be noted that non-genetic factors were not accounted for in the computation for the correlation between schizophrenia and childhood trauma in this study.

Previous studies have shown that the severity of schizophrenia symptoms is mediated by variants in candidate genes that interact with childhood trauma including those in brain derived neurotrophic factor (*BDNF*), FK506-binding protein 5 (*FK506*) and catechol-O-methyltransferase (*COMT*) (Aas et al., 2013; Green et al., 2014; Green et al., 2015). BDNF is important for neurodevelopment in the hippocampus. Stress, which may result from experiencing childhood trauma, reduces BDNF levels. Lower levels of BDNF have been observed in patients with schizophrenia compared to those without (Thompson Ray et al., 2011). Further, epigenetic studies show that childhood trauma is associated with differential DNA methylation of various promoter regions (Labonté et al., 2012). These changes in DNA methylation may be the mechanism by which trauma leads to the neurobiological abnormalities associated with schizophrenia.

Inconclusive evidence for the genetic effects of sex in SAX

Schizophrenia is known to be more common and have an earlier onset in males than in females. The widely accepted ratio of affected males to females is 1.4-to-1, however these reports are from epidemiological studies in high income countries. There is reason to believe that there is a bigger male-to-female disparity in African population based on the SAX study, and the ongoing Neuropsychiatric Genetics in African populations (NeuroGap, https://www.broadinstitute.org/stanley-center-psychiatric-research/stanley-global/neuropsychiatric-genetics-african-populations-neurogap), as well as undocumented observations in psychiatric wards in Ethiopia.

The case sample in the present study constituted 88% male participants (915 males versus 123 females). Ascertainment bias was ruled out based on the fact that the cohort was recruited from institutions where the male-predominance already existed, which suggested that sociological (e.g. specific roles males and females play in their society that may be tied to cultural norms), or pathological factors (e.g. presentation and severity of disease) may influence institutionalization.

It has been shown that females are more responsive to antipsychotics and thus more likely to be hospitalized for shorter periods of time than males (Grossman et al., 2008). This may explain why fewer females were encountered in the institutions where recruitment took place for the SAX study. Additionally, the fact that females present with less severe symptoms than males may explain explain why fewer females present to hospitals in the first place, as they are able to maintain an active role in society until much later in life.

As expected, the exploratory analysis undertaken in this chapter to differentiate the genetic effects of schizophrenia in the SAX cohort did not yield any conclusive results due to the limited sample size, especially in the female strata. However, large-scale epidemiological and genetics studies are warranted to detangle the disproportionate representation of schizophrenia in African males. It may be worth expanding inclusion criteria to be able to recruit more women, especially menopausal females who have are more prone to schizophrenia than those who are not. It may also be beneficial to not only recruit cases from psychiatric hospitals but from the general community as females are less likely to be hospitalized.

Trans-ancestry genetic effects

Although the present study cohort is small in comparison to the largest published GWAS, it is adequately powered for the current set of investigations. From findings presented here, it provides important clues about the potential genetic risk differences between ancestral populations. In general, findings from this chapter provide distinct insights:

First, the majority of the GWAS SNPs that were annotated in this study fell in intergenic regions of the genome, whereas the majority of the associated SNPs in other studies were intronic. Intergenic SNPs are typically assigned pathogenicity based on the function of the most proximal gene, whereas the function of intronic variants are the regulation of gene expression.

Second, the top associated variants in the SAX study were not significantly associated in either the PGC-EUR nor PGC-EAS studies. This is likely due the difference in the frequency of the effect allele. For example, the frequency of the effect allele (T) of rs35172303 is 16% in EUR- and 48% in EAS populations, with p-values of 0.5558 and 0.5446 in PGC-EUR and PGC-EAS, respectively, compared to 10% AFR populations. Conversely, the highest associated PGC-EUR variant rs115329265 in the MHC (P = 3.48e-31) has a frequency of 15% in EUR, 44% in AFR and 2% in EAS populations. It is also likely that there are population-specific patterns of linkage disequilibrium in the MHC region. A larger sample sizes may help delineate the genetic contribution of variants in the MHC to the risk of schizophrenia in individuals of AFR ancestry. The largest published GWAS study in individuals of AFR ancestry showed that PGC-EUR associated variants had the same direction of effect in that cohort (Bigdeli et al., 2019), implying that there are shared genetic risk factors between populations.

Trans-ancestry differences (or between-study heterogeneity in genetic effects) can be attributed to various factors beyond differences in allele effect frequencies, namely (i) sub-phenotypes: there is potential for clinical heterogeneity between studies which means that the phenotype being assessed may not be exactly the same between studies, (ii) geographical background: GWAS study design treats the genome as being static, and does not account for interactions of genes with others or the interaction of genes with the environment.

Limitations

As alluded to throughout this chapter, the most prominent limitation of this study is sample size. The sample size required to conduct a successful GWAS is generally predicted to be tens of thousands of study participants. Although findings from Gulsuner et al (2020) were recapitulated here, those from previous studies in EUR and EAS ancestry cohorts could not be replicated, particularly for variants that have been consistently associated with schizophrenia across different ancestral groups. Further, the sample size limited the number of disease-associated variants that could be discovered, as sample size is directly correlated with the discoverability of risk variants. This has been demonstrated by the PGC studies, where the earliest study in only a few thousand individuals (Psychiatric Genomics Consortium, 2011) led to the identification of three risk loci, compared to the latest study (Ripke et al., 2020) in tens of thousands of individuals that revealed over 248 independently-associated loci.

In the GWAS analyses, the recruitment site was not included as a covariate. This may impact the association results if schizophrenia presented differently between the sites, or may act as a confounder if there were inherent genetic differences between study participants across sites. Because the sample was male-dominated, the question about whether the genetics of schizophrenia differs between males and females could not be thoroughly explored or not a genetic difference exists between males and females. Further, urbanicity and socioeconomic status are among the environmental factors that have been shown to contribute to the risk of schizophrenia as discussed in chapter 1, however their contribution was not investigated in this chapter. It is likely that there are differences in the socioeconomic status between study participants in the Western and Eastern Cape provinces which may impact access to health care and disease outcome.

2.6 Conclusion

The SAX study represents the first GWAS of this size in the southern African indigenous cohort to date. This study identified a genome-wide significant locus on chromosome 17 that has not been found in either EUR and EAS ancestry GWAS studies, suggesting that this locus to be of importance for the aetiology of schizophrenia in the South African population.

Notably, the signal on chromosome 17p13.2, within the *ZFP3* hinted at the biology of this gene and related pathways. *ZFP3* is ubiquitously expressed human tissues; regulates the transcription of RNA polymerase II and may be important in regulating gene expression in the brain. The fact that this locus has not been identified in the large international studies of other EUR and EAS populations suggests that a unique biology underlies the disease in at least a subset of schizophrenia patients in the South African Xhosa population. The identification of childhood trauma as a contributor to schizophrenia in the GWAS indicates that there may be value in investigating environmental components contributing to the aetiology of schizophrenia, at least in the southern African context.

This cohort provides an opportunity for further investigation into population-relevant risk loci, to bridge the knowledge gap that exists in psychiatric genomics research worldwide as a result of the Eurocentric research bias. This will also help minimise the threat of exacerbating the global health disparities in the absence of knowledge from non-AFR populations. Though the present findings are limited by sample size, they suggest that population-specific risk loci for schizophrenia may exist, and highlight the imperative of including more geographically and ethnically diverse samples in psychiatric genomics research, while also trying to ensure that at least threshold numbers of different ethnicities are represented in such global studies.

The next chapter explores the heritability of schizophrenia in SAX and the transferability of polygenic risk scores derived from EUR and EAS populations from the PGC in SAX.

Chapter 3: Heritability and Polygenic Risk Score analyses of Schizophrenia in the South African Xhosa people

3.1 Abstract

GWAS studies in EUR ancestry populations have demonstrated that the heritability of schizophrenia is differentially enriched in functional categories of the genome. These studies have also shown that polygenic risk scores (PRS) derived from EUR ancestry populations do poorly at predicting schizophrenia in AFR ancestry populations. Given that the GWAS conducted in chapter 2 represents the first in the South African Xhosa population, the enrichment of heritability and generalizability of PRS is not known. In this chapter, the heritability of schizophrenia was estimated and further partitioned by chromosome, minor allele frequency (MAF) and 24 functional categories using the Genome-wide Complex Trait Analysis (GCTA) software. To assess the transferability of PRS between ancestral groups, PRS were computed for the SAX cohort using GWAS summary statistics obtained from three ancestral groups: SAX-discovery to assess the within-ancestry transferability of PRS, and EUR- and EAS ancestry GWAS obtained from the PGC to assess trans-ancestral transferability of PRS. The case-control status prediction accuracy of PRS was assessed using Nagelkerke's R^2. The genome-wide SNP heritability of schizophrenia in SAX was 0.288 ± 0.03 (*P* = 2.609e-15). Chromosome 6 explained the most variance (h^2g = 0.109578 ± 0.0292, *P* = 7.0097e-05). Further, heritability was similarly enriched across functional categories involved in the regulation of gene expression, similar to observations from studies in EUR populations. These findings suggested that schizophrenia has a similar biological mechanism across ancestries. The prediction accuracy was low for within- and trans-ancestry PRS, consistent with low PRS prediction accuracy when the target and discovery cohorts are not ancestrally matched. Larger and more diverse samples are required to refine the enrichment of heritability and improve PRS prediction accuracy in populations not currently represented in large scale GWAS studies.

3.2 Introduction

GWAS studies in EUR ancestry populations have demonstrated that the heritability of schizophrenia is differentially enriched in functional categories of the genome. These studies have also shown that PRS derived from EUR ancestry populations do poorly at predicting schizophrenia in AFR ancestry populations. In the previous chapter, the genetic aetiology of schizophrenia was investigated through GWAS for the first time in SAX. The GWAS identified the locus at the genomic region 17p13.2 to be associated with schizophrenia at the genome-wide significance threshold. The sub-genome-wide significant loci are also important as they can be informative of the regions in the genome that harbour disease-relevant loci. Equally, they can be used to quantify the genetic risk of an individual developing schizophrenia. Therefore, this chapter had two aims. The first was to determine in which functional categories the heritability of schizophrenia was enriched. The second was to investigate the prediction accuracy of PRS within the SAX cohort and across ancestries using publicly available GWAS summary statistics.

3.2.1 SNP-based heritability

Heritability (h^2, also known as broad-sense heritability) is described as the proportion of variance explained by genetic effects over the total phenotypic variance. SNP heritability (also known as narrow-sense heritability), denoted as h^2_{SNP} , is the proportion of variance explained by additive genetic effects over the total phenotypic variance. For example, h^2_{SNP} is calculated in GWAS studies as the variance explained by SNPs that have been associated with a phenotype of interest. In some studies, heritability is partitioned ("partitioned heritability") to identify regions on the genome that are enriched with disease-relevant variants in order to gain a better understanding of the genetic architecture of the disorder

3.2.1.1 Partitioned heritability by genomic characteristics

There are numerous ways in which heritability can be partitioned. The most common and perhaps most generic way is to partition by chromosome and MAF. Previously Yang et al. (2011) showed that variance explained by each chromosome is proportional to the length of the chromosome for highly polygenic traits.

3.2.1.2 Partitioned heritability by functional elements

It has long been known that most GWAS associated variants for schizophrenia and other complex disorders occur in non-coding regions, therefore their effects on disease/disorder are mediated through altering gene expression (Nica et al., 2010; Nicolae et al., 2010; Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014). This has led to peaked interest in the field of study of epigenetics, which investigates the impact of non-genetic factors on gene expression. Recently, the term has been used to refer to histone modifications and DNA methylation (Arimondo et al., 2019; Dupont et al., 2009).

The chromatin (**Figure 3.1),** is central to the study of epigenetics, and represents the conserved state in which DNA is packed into a cell. Chromatin has active regulatory elements including promoters, enhancers, silencers and transcription factor binding sites. The nucleosome is the fundamental unit of the chromatin and is composed of four core histones (H2A, H2B, H3 and H4). Several modifications such as acetylation, methylation, and phosphorylation can occur on these histones limiting binding of transcription factors (Klemm et al., 2019; Kouzarides, 2007).

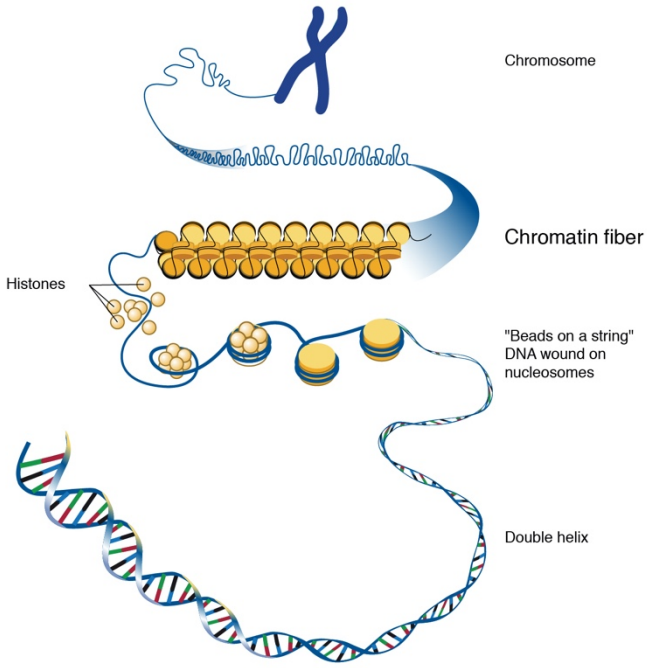


Figure 3.1 - Diagram showing the structure of chromatin
Chromatin is the physical substance making up chromosome, and is made up of DNA and proteins called histones. DNA is tightly wrapped around histones to form the chromatin. Image was obtained from https://www.genome.gov/genetics-glossary/Chromatin

Enhancers are DNA regions to which transcription factors bind, controlling cell-specific gene expression. The human genome has about one million such enhancers (Dunham et al., 2012; Thurman et al., 2012). Super-enhancers are transcription factor complexes comprising key transcription factors and a mediator coactivator. Transcription factors that might form super-enhancers are not known for most cell-types, thus histone modifications H3K27ac, H3K4me1, DNAse hypersensitivity are used as biomarkers of super-enhancers (Hnisz et al., 2013). Disease-specific variation in super-enhancers occurs in cell-types relevant to the disease or disorder.

3.2.2 Polygenic Risk Scores

The computation of PRS allows pooling of multiple loci of small effects to quantify genetic risk. The scores are computed from GWAS summary statistics as the sum of the count of risk alleles weighted by their effect on the phenotype. These scores can then be used to identify 'at risk' populations.

The landmark study conducted by International Schizophrenia Consortium was the first to apply PRS to their dataset of schizophrenia cases and controls of EUR ancestry (International Schizophrenia Consortium, 2009). This study was fundamental in portraying the polygenic nature of schizophrenia, and showing that PRS are specific to groups of disorders. Scores computed in the EUR schizophrenia dataset could accurately predict bipolar disorder status in an independent sample; but had low prediction accuracy for non-psychiatric disorders (International Schizophrenia Consortium, 2009). However, the EUR-derived PRS had poor predictive accuracy for schizophrenia in a sample of African American individuals. This showed that PRS computed from EUR were better able to predict across disorders within ancestry, than across ancestries for the same disorder. Subsequent studies in EUR target populations have shown similar findings of strong within-ancestry PRS prediction accuracy (Derks et al., 2012). Findings have however been inconsistent for non-EUR target populations; while some show strong prediction accuracy across ancestries (Bigdeli et al., 2019; Ikeda et al., 2011; Ikeda et al., 2019), other studies demonstrated poor prediction (Lam et al., 2019; Bigdeli et al., 2017).

Martin et al. (2017) have empirically shown that PRS prediction accuracy diminished with increasing genetic distance between the discovery and target populations. EUR and AFR ancestry populations are the most divergent. The predictive accuracy was 4.5 times less accurate across EUR-AFR cohorts. Since not many studies of schizophrenia have been

conducted in AFR ancestry populations it is to be expected that PRS computed in EUR would perform sub-optimally in African populations.

The aims of the work described in this chapter were two-fold: i) to investigate the distribution of heritability across the chromosomes, allele frequency spectrum as well as 24 functional categories, and ii) assess how well PRS derived from EUR and EAS populations from the PGC can predict schizophrenia in SAX.

3.3 Methods and Materials

3.3.1 Partitioned heritability and functional enrichment of GWAS SNPs

Genome-wide SNP and partitioned heritability estimates were calculated from imputed genotype data using Genome-wide Complex Traits Analysis (GCTA) software (Yang et al., 2011). GCTA was used, instead of the widely used LDSC (Bulik-Sullivan et al., 2015) or LDAK (Speed et al., 2012) because the SAX sample size was smaller than the required size for these software. Heritability estimates for all the partitions were computed for two scenarios: unadjusted analysis — without the inclusion of PCs and exclusion of the MHC region, adjusted analysis — including the first 20 PCs and excluding the MCH region. PCs were adjusted for to account for population structure, to avoid inflation of the heritability estimates

To calculate genome-wide heritability, a kinship matrix was constructed by computing the genotypic relatedness matrix (GRM) from SAX GWAS-QC'd genotype data, which included 2,087 samples and about 15 million SNPs (refer to sections 2.3.5 and 2.3.6). Then, heritability was estimated jointly with GRM using the restricted maximum likelihood (REML) strategy for a case-control design and a disease prevalence of 0.01 (Lee et al., 2011).

The partitioning of the genome by chromosome and MAF, was done in GCTA. The flag *--chr* was used to specify each of the autosomes, and six MAF bins were created, with MAF ranging from 0.01 to 0.5 (**Table 3.1**). GRM was computed for each of the partitions, followed by the computation of the variance explained using REML for a disease prevalence 0.01.

Table 3.1 - The minor allele frequency bins used to partition heritability

MAF bin	MAF range
1	0.01 – <0.05
2	0.05 – < 0.1
3	0.1 – < 0.2
4	0.2 – < 0.3
5	0.3 – < 0.4
6	0.4 – 0.5

The genome was partitioned into genes and exons using the *--cut-genes* flag in the LDAK v5.0 software (Speed et al., 2012) based on the Refseq HG19 annotation. Heritability from genes was estimated using REML in GCTA for a case-control design and a disease prevalence of 0.01 as before. The GRM-models were computed by considering variants within 1,000 bp of a gene, and controlling for the first 20 PCs.

To partition heritability by functional elements, the main 24 functional annotations (**Appendix 6: Supplementary Table 3.1**) were obtained from the LDSC software (Bulik-Sullivan et al., 2015). SNPs were categorized by functional element based on the 1KGP3 African Genome reference panel and annotation files obtained from the LDAK site (http://dougspeed.com/annotations/).

3.3.2 PRS calculation

The strategy used to evaluate the prediction accuracy of PRS within and across ancestries is illustrated in **Figure 3.2**. To assess PRS prediction across within ancestry, the SAX sample was split into two independent groups of equal size (cases = 519 , controls = 524 in each group) by randomly assigning cases and controls to either SAX-discovery or SAX-target groups. This was done using the innate *shuf* function in Linux, and the *--keep-fam* flag in PLINK v2 (Chang et al., 2015). Then, a GWAS was conducted using the SAX-discovery dataset as in section 2.3.9, but without controlling for childhood trauma, to obtain summary statistics. The GWAS summary statistics were then clumped in PLINK v2 (Chang et al., 2015) using the SAX sample as the LD reference panel. In other words, clumping was done using within-sample LD. The clumping of the summary statistics file was done similarly to what was

described by Stefansson et al., (2009) to obtain independently significant SNPs based on $r^2 \leq 0.1$ and a window-size of 500kb. The summary statistics of the independent SNPs within the clumps were used to calculate the PRS scores for the SAX-target group across ten p-value thresholds.

For the trans-ancestry evaluation of PRS predictive accuracy, PGC-EUR and PGC-EAS summary statistics were clumped using 1KGP genetic data from EUR and EAS ancestry populations, respectively (Lam et al., 2019; Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014). Clumping was done in PLINK v2 (Chang et al., 2015) as before: $r^2 < 0.1$ and a window-size of 500kb. The PGC-EAS dataset had 8 million SNPs, and about 360,573 SNPs remained after clumping, and the PGC-EUR dataset had 9.8 million SNPs, and 290,308 remained after clumping. Clumped summary statistics and their corresponding weights were used to calculate PRS in SAX-target cohort across ten different p-value thresholds. For the second part of the trans-ancestry PRS prediction evaluation, a meta-analysis of PGC-EUR, PGC-EAS and SAX-discovery was conducted to investigate whether the inclusion of AFR samples in the meta-analysis improves PRS predictive power. A reference panel was created to reflect the LD structure of the meta-analytic sample. Similar to before, EUR and EAS ancestry samples from 1KGP3 were used to represent PGC-EUR and PGC-EAS LD architecture. In addition to these samples, individuals from SAX were selected to represent the Xhosa individuals in the meta-analysis. The number of individuals included in the reference panel for each cohort was weighted by the sample size of each study, noting that 1KGP3 has a maximum of 500 individuals per population group (**Table 3.2**). The same clumping strategy as before was used to clump SNPs. PRS were computed based on these clumped SNPs.

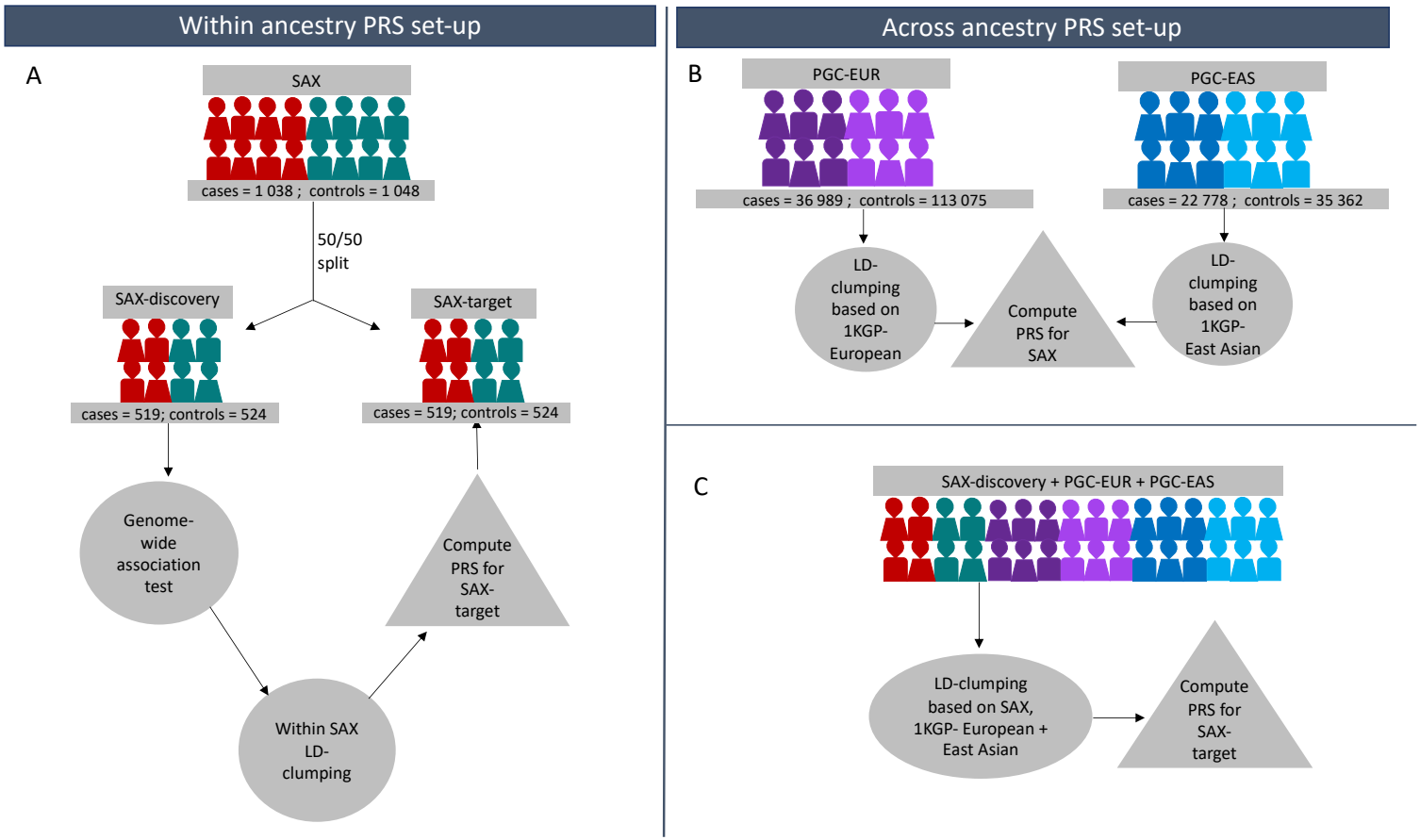


Figure 3.2 - An illustration of the PRS analysis
A) To assess within ancestry PRS, the SAX sample was split to create a SAX-discovery and -target datasets. GWAS was conducted on the SAX-discovery dataset, followed by within-sample clumping and computation of PRS in the SAX-target. B) Summary statistics obtained from PGC-EUR and PGC-EAS were used to calculate scores in the SAX sample. C) All three SAX-discovery, PGC-EUR and PGC-EAS were meta-analysed. Meta-analytic summary statistics were used to compute PRS in the SAX-target

Table 3.2 - Data used to create LD reference panel for the clumping of meta-analytic results

	SAX-discovery	PGC-EUR	PGC-EAS
sample size (cases and controls)	1 043	150 064	58 140
proportion of samples in meta-analysis	0.5%	71.71%	27.28%
n samples in LD panel for clumping	2	358	140

For all PRS predictions, case-control status was predicted using a logistic regression of the scores and the first 20 principal components. The variance explained was assessed using Nagelkerke's R^2 by comparing two models; the first including the first 20 PCs and the computed PRS scores, and the second including the PCs alone. R^2 was computed as the difference between these two models.

3.4 Results

3.4.1 SNP-based heritability

SNP-based heritability was computed in two ways: first without adjusting for population structure and keeping the MHC region; secondly by adjusting for populations structure by including principal components and removing the MHC region. In general, heritability estimates were lower for the adjusted analysis, but the overlap in the 95% CI between the adjusted and unadjusted analyses indicated no difference between the two analyses types,. The unadjusted genome-wide SNP heritability was 0.28 ± 0.03 (*P* = 0), and remained the same after adjustment (0.288 ± 0.03, *P* = 2.609e-15).

Heritability varied across chromosomes (**Figure 3.3**), with the most variance explained by chromosome 6 (h^2g = 0.109578 ± 0.0292, *P* = 7.0097e-05, **Appendix 6: Supplementary Table 3.2**), even after adjustment for population structure. Unexpectedly, the variance explained by chromosome 7 was not different from zero and that explained by chromosome 20 was among the lowest across all. Chromosome 17, which had the significantly associated SNPs from the GWAS, explained 0.055 ± 0.024 (adjusted, *P* = 1.0020e-02, LRT = 5.408) of variance (**Appendix 6: Supplementary Table 3.2**), which was lower than that for chromosomes 6,2,3 and 8

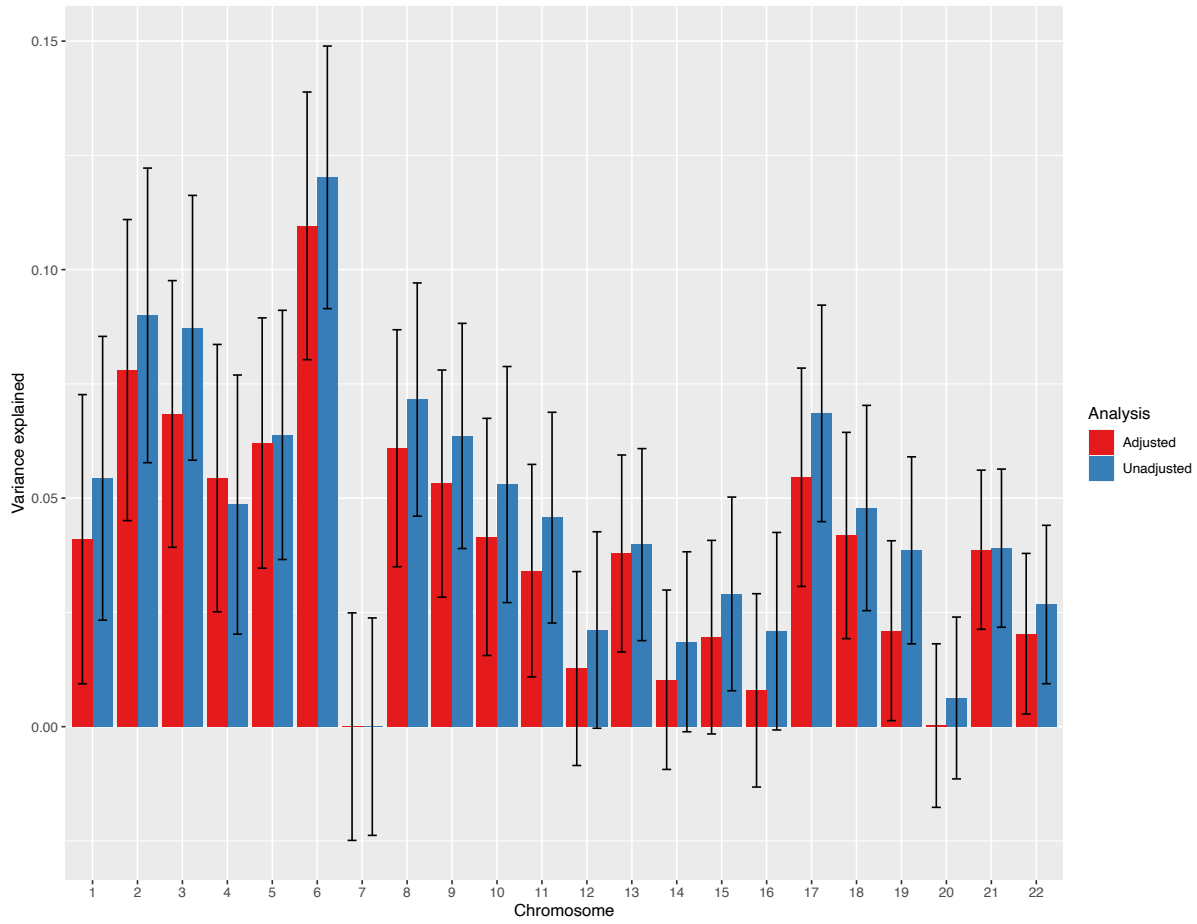


Figure 3.3 - The heritability explained by SNPs on individual chromosomes.
Adjusted analysis controls for population structure and the complex linkage disequilibrium structure in the MHC region on chromosome 6. Error bars represent the 95% confidence intervals.

Similarly, heritability varied by MAF (**Figure 3.4**). Variants on the lower end of the frequency spectrum explained more variance than those that were more common. The most common variants, with MAF between 0.4 and 0.5, explained the least variance in liability (h^2g = 0.279151 ± 0.052473, LRT = 22.6, *P* = 9.9761e-07), compared to variants in the first three MAF bins, including those with MAF less than 0.05 (h^2g = 0.551915 ± 0.081673, LRT = 38.268, *P* = 3.0840e-10) (**Appendix 6: Supplementary Table 3.3**). Similar to the analysis above, the adjusted estimates were not different from the unadjusted estimates except for MAF bin 0.3 - 0.4 where the adjusted estimates were higher than those for the unadjusted analysis.

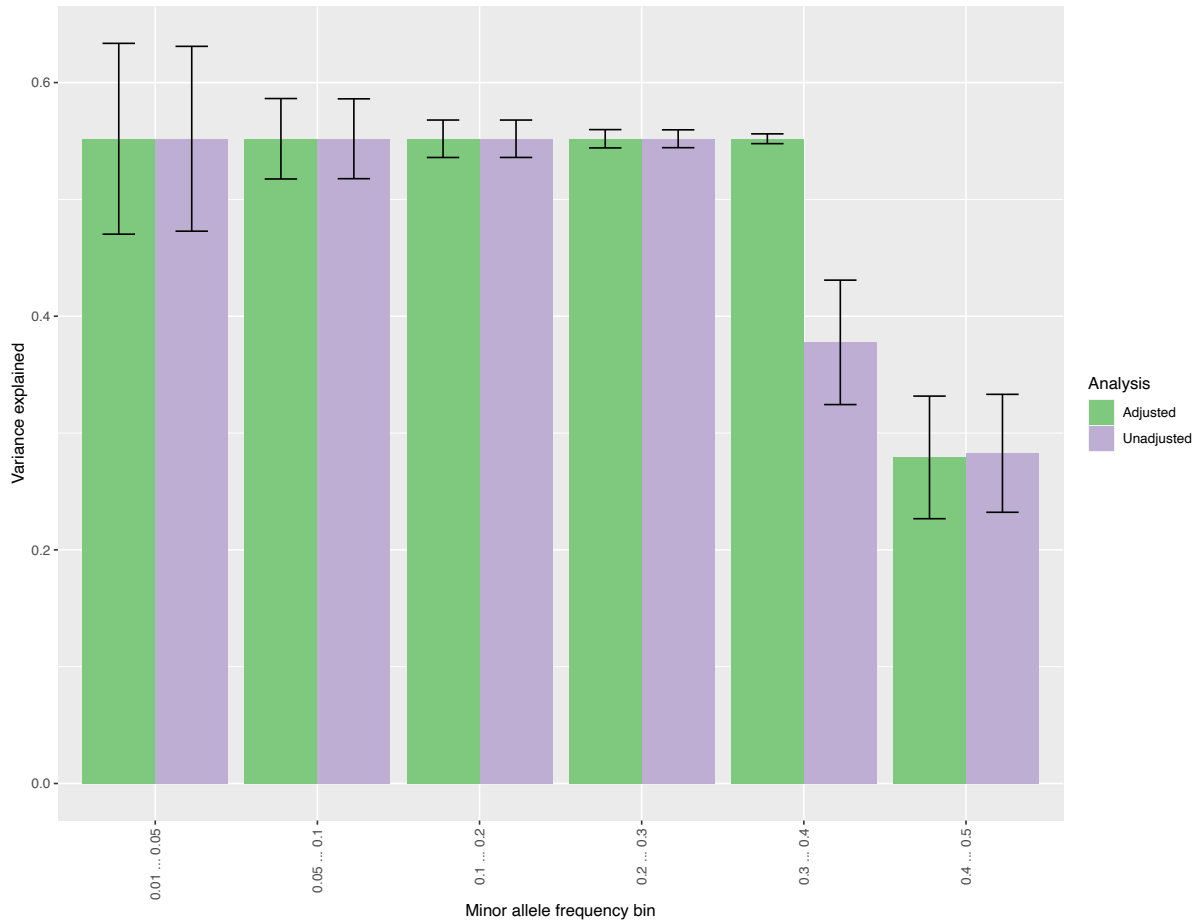


Figure 3.4 - The heritability explained by SNPs in six minor allele frequency bins. Adjusted analysis controls for population structure and the complex linkage disequilibrium structure in the MHC region on chromosome 6. Errors bars represent the 95% confidence intervals.

Table 3.3 - The heritability explained by SNPs in genes vs exons

		Adjusted	Unadjusted
Genes	h^2g (se)	0.27 (0.013)	0.27 (0.013)
	LRT	81.89	86.58
	P-value	0	0
	n SNPs	2 549 377	2 553 990
Exons	h^2g (se)	0.135 (0.027)	0.138 (0.026)
	LRT	21.38	26.94
	P-value	1.878e-06	1.046e-07
	n SNPs	537 189	540 183

When variants were partitioned by their situation in genes or in exons, the 'genes partition' explained more variance than exons (**Table 3.3**).

The heritability estimates by functional categories are shown in **Figure 3.5**, **Appendix 6: Supplementary Table 3.4**) The functional categories that explained the most variance were regions conserved in mammals, histone modifications H3Ks, digital genomic footprint, DNase I hypersensitivity sites including foetal Dnase 1 Hypersensitivity site, repressed regions, transcription binding site and transcription start sites; each of these functional categories explained about 0.55 of variance.

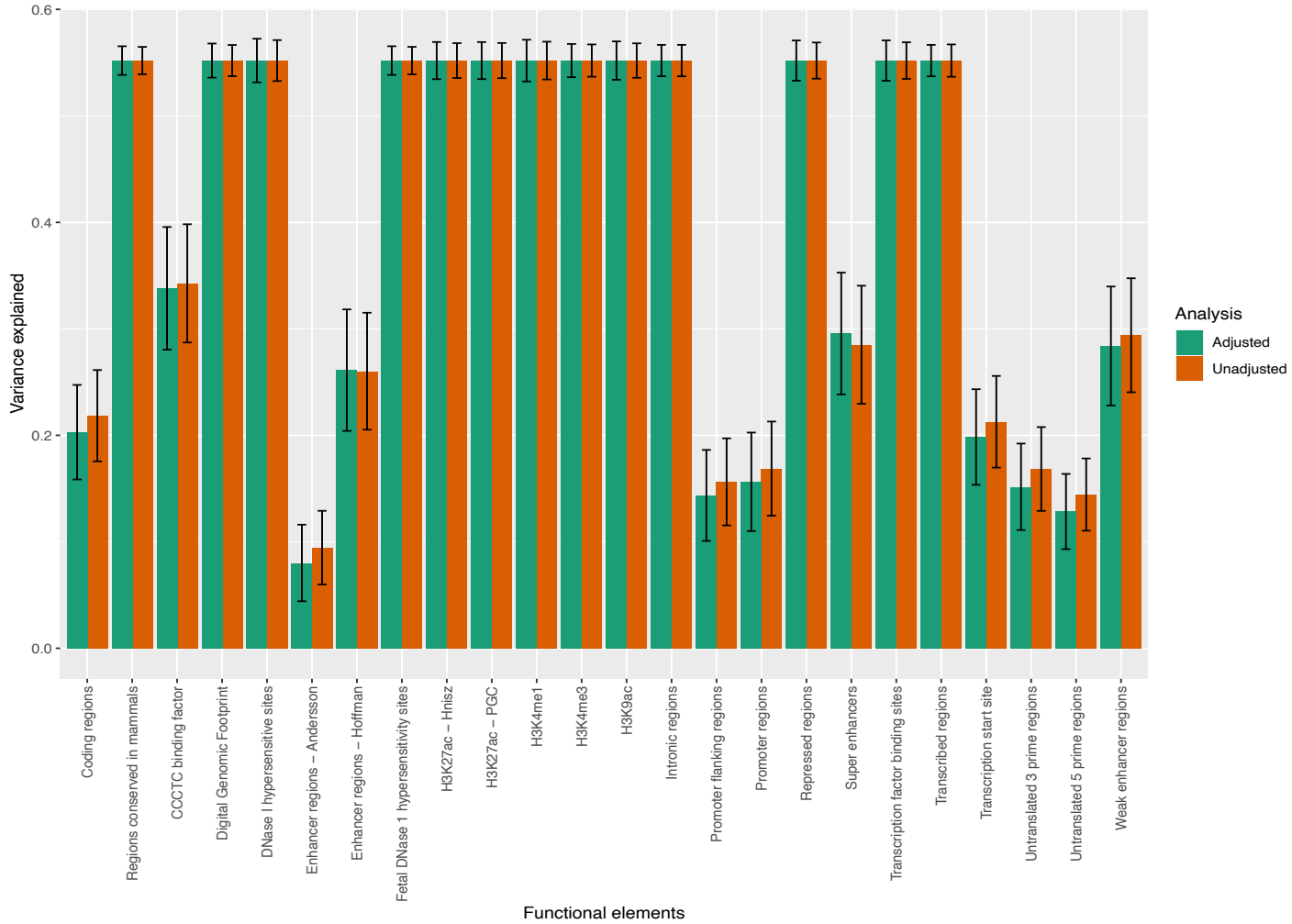


Figure 3.5 - The heritability explained by SNPs across 24 functional categories.
Adjusted analysis controls for population structure and the complex linkage disequilibrium structure in the MHC region on chromosome 6. Error bars represent the 95% confidence intervals.

3.4.2 PRS prediction accuracy

Within-ancestry PRS prediction

The GWAS from the SAX-discovery yielded no SNPs that achieved genome-wide significance. The Manhattan and QQ-plots are shown in **Appendix 6: Supplementary Figure 3.1**. After computing PRS in the SAX-target, variants with p-value < 1e-04 explained the most variance (R^2 = 2.97e-03, *P* = 0.042), with a confidence interval that included zero [CI: 6.81e-03,-6.456e-03] (**Figure 3.6**).

Trans-ancestry PRS prediction

The 1KGP EUR and EAS samples were used to clump the PGC-EUR and PGC-EAS GWAS summary statistics, respectively. After clumping PGC-EUR, 290,308 SNPs remained, and 360,573 remained after clumping PGC-EAS GWAS summary statistics. The clumped variants when then used to calculate PRS and evaluate their prediction accuracy in the SAX-target cohort. A similar maximum variance explained was achieved when using scores derived from PGC-EUR and PGC-EAS, but at different p-value thresholds. That is R^2 = 0.0057 at p-values < 1e-06 using PGC-EUR and R^2 = 0.0059 at p-values < 1e-04 using PGC-EAS (**Appendix 6: Supplementary Table 3.5**). The scores derived from the meta-analytic dataset comprising SAX-discovery, PGC-EUR and PGC-EAS, yielded maximum variance explained (R^2 = 0.0028, *P* = 4.340e-02) when all variants were considered despite their p-values (**Figure 3.6**).

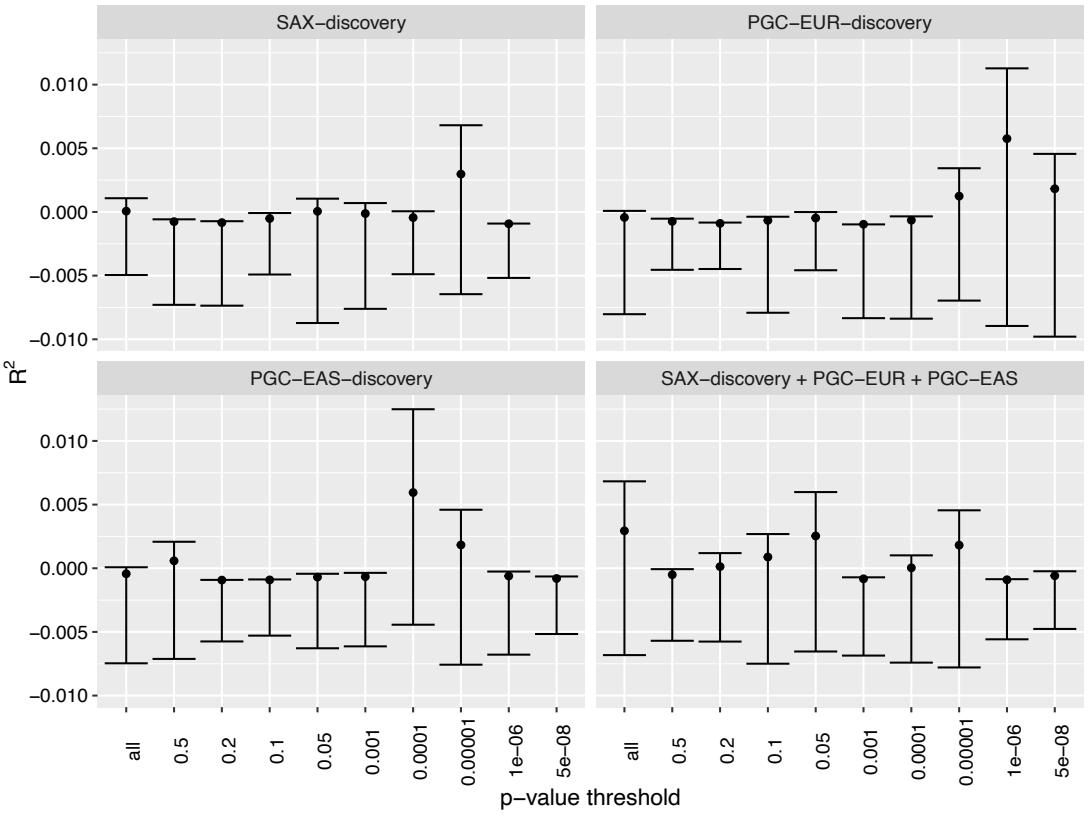


Figure 3.6 - Polygenic risk prediction accuracy within and across ancestries.
Error bars represent the 95% confidence intervals.

3.5 Discussion

In this chapter, an exploratory analysis was conducted to assess the distribution of heritability of schizophrenia in SAX, as well as evaluate the transferability of PRS within and across ancestries. The findings indicated the genome-wide SNP heritability of schizophrenia in SAX was similar to that in EUR and EAS ancestry samples. Chromosome 6 explained the most variance, compared to all the other chromosomes. When heritability was partitioned by functional elements, an enrichment was seen in genomic regions that regulate gene expression. The PRS prediction accuracy was low within and across ancestries, and was not improved after trans-ancestral meta-analysis.

Narrow-sense heritability in SAX is comparable to previous studies

The narrow-sense heritability of schizophrenia in an indigenous South African population was assessed for the first time. The heritability was slightly higher than that observed for EUR and EAS populations; 0.28 ± 0.03 in SAX, 0.24 ± 0.02 in PGC-EUR and 0.23 ± 0.03 in PGC-EAS (Lam et al., 2019; Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014). This was expected because heritability estimates from the two PGC studies were computed from GWAS summary statistics, whereas the heritability estimates for the SAX dataset in this chapter were computed from imputed genotype data. The SAX heritability estimate was however, comparable to that computed for PGC-EUR (0.274 ± 0.007) when imputed genotype data was used (Loh et al., 2015).

It is likely that estimates from genotype data are more accurate than those from GWAS summary statistics data because LD is more accurately modelled if the LD architecture of the reference panel used matches that of the population being studied. Currently, there are limited refence panels available for AFR datasets. While an AFR ancestry reference panel exists within the 1KGP3, the extensive diversity within AFR populations (as shown in Section 2.4.2) limits the practical uses of this reference panel.

Limited sample size in SAX restricts accurate estimation of heritability

The findings from the heritability estimates partitioned by chromosome and MAF were unexpected. Regarding the chromosome partition, heritability was not linearly correlated with the size of the chromosome, as previously reported (Lee et al., 2012). Additionally, the total per chromosome heritability exceeded the total genome-wide heritability. For estimation of heritability across MAF bin, the expectation is that there would be an inverse relationship between the MAF of a variant and the size of its effect (Park et al., 2011). However, variants in the bins with MAF up to 0.4 had the same heritability estimates (i.e. 0.055, **Appendix 6: Supplementary Table 3.3**).

Several explanations are possible for the observations above. First, the sample size of 2,000 is too low to fit the GCTA models, and thus limits the accurate estimation of heritability. Secondly, heritability was estimated in the chromosome and MAF partitions using the single variance component analyses that assumes that heritability is equally distributed across the genome. However, it is known that heritability is a function of both linkage disequilibrium and MAF. Thus, multiple variance component analyses using tools such as RHE-mc

(Pazokitoroudi et al., 2020), where LD and MAF can be properly modelled, would be more appropriate to estimate heritability across categories. Lastly, it is likely that populations structure is not adequately controlled for. It has been demonstrated that populations structure can inflate heritability estimates (Browning & Browning, 2011), which may explain why the total variance across the partitions exceeds the genome-wide heritability estimate.

Heritability is enriched in functional elements that regulate gene expression

The enrichment of heritability largely in functional elements, such as enhancers and promoters, that regulate gene expression is consistent with previous findings alluded to previously and discussed below, indicating that the mechanism by which schizophrenia develops may be generically related to the transcription process.

The PGC-EUR studies showed an enrichment of GWAS significant SNPs of H3K27ac enhancers in cells and tissues related to the brain (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014; Ripke et al., 2020). SNPs identified from previous studies of bipolar disorder and schizophrenia map to H3K4me3-containing regions in the anterior caudate nucleus and the frontal lobe regions of the brain (Bipolar Disorder Working Group of the Psychiatric Genomics Consortium, 2011; Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2011 ; Trynka et al., 2013). Foetal DNAse hypersensitivity sites are expressed in foetal tissues — enrichment of heritability in these sites as seen in SAX highlights that mechanisms during brain development are likely to be involved in disease aetiology. DNAse I hypersensitive sites, so named because of their hypersensitivity to nuclease cleavage or chemical modification (Gross & Garrard, 1988), are regions of the genome that are free from nucleosome and allows for binding of transcription factors to DNA.

Several transcription factors have been associated with schizophrenia — notably the transcription factor 4 (*TCF4*), which was first associated with schizophrenia in GWAS meta-analysis and remained among the top-associated SNPs from the landmark PGC-EUR (Forrest et al., 2018; Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014; Stefansson et al., 2009; Xia et al., 2018). Xia et al. (2018) showed that *TCF4* was enriched in gene sets that were preferentially expressed in the brain and were downregulated in *TCF4* knockout experiments, similar to findings from Hill et al. (2017). This study also found that of the 108 loci implicated in the landmark PGC-EUR study, 39 contained at least one *TCF4* and there was an enrichment of heritability in *TCF4* binding sites. The findings from the present study reported in this chapter are consistent with these prior findings.

It is important to interpret the findings from the heritability estimations within the context of the limitations of the size of the dataset and the models used. As previously mentioned, the SAX sample is 1,000 individuals short of meeting the GCTA requirement of at least 3,100 samples to fit models and maintain standard errors within the 0.1 range. The GCTA-REML model used assumes a Gaussian distribution, i.e. that heritability is distributed equally across the genome. By extension, when the chromosome is partitioned, it is assumed that each partition has equal weight. Other models, like those implemented in LDAK (Speed et al., 2012) and LDSC (Bulik-Sullivan et al., 2015) account for this limitation by calculating weights based on LD prior to the computing of heritability estimates. Both these software require 5,000 samples and could not be used with the SAX dataset.

Prediction accuracy of PRS is low across ancestries

In general, the prediction accuracy of PRS was low within and across ancestries. The PRS prediction accuracy for schizophrenia in EUR samples using EUR-derived summary statistics from the first application of PRS in humans conducted by the ISC was 0.032. The prediction accuracy for bipolar disorder in the same study was 0.014 (International Schizophrenia Consortium, 2009). This study showed that PRS prediction accuracy is highest when discovery and target cohorts are matched. These findings have been replicated and improved in subsequent larger PGC studies (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014; Ripke et al., 2020).

In addition to the PRS prediction accuracy being low, the computed PRS scores performed similarly when computed using summary statistics within and across ancestry. The confidence intervals around the R^2 were large and spanned zero, reflecting the sample size of SAX-target, suggesting that all true estimates could be zero. The discussion below is based on these point estimates.

P-value threshold signify the noise-to-signal ratio, with less stringent p-value thresholds likely to have more noise and include variants that are not relevant to the aetiology of the disease. Conversely, more stringent p-values have a higher proportion of variants that are likely to be causal (Martin et al., 2019). PGC-EUR summary statistics associated with SAX cases-control at p-value < 1e-06, the most stringent p-value of all three datasets, and thus likely including variants that are important to the aetiology of schizophrenia. The p-value at which variance explained is maximized is directly linked to the size of the GWAS discovery sample. In GWAS

that are underpowered to identify GWAS associated variants, it is to be expected that inclusion of all SNPs (at the least stringent p-value threshold) would explain the most variance.

Considering that the SAX-discovery dataset was smaller in size than PGC-EAS by several orders, it was surprising that the maximum variance explained p-value was more stringent (*P* < 1e-04) than that in PGC-EAS (*P* < 1e-03). It was surprising that PRS computed in PRS-EAS were different in terms of p-value at maximum variance explained, given that PGC-EAS and PGC-EUR have a high genetic correlation (Lam, Chen, et al., 2019).

These observed differences could be attributable to the differences in the LS and allele frequencies between the populations, but can also be attributed to the clumping reference panel used. The 1KGP reference panel used contains genotypes from 500 individuals from East Asia, whereas the summary statistics files was obtained from a GWAS of about 68 000 individuals, and better able to reflect the true LD structure in the sample. It is therefore likely that the clumping strategy used here was sub-optimal and eliminated variants in the GWAS summary statistics files that were not in the reference panel. This can be overcome by obtaining clumps from individual genotype data (rather than summary-level) from the PGC-EAS study.

It was surprising that meta-analysis of GWAS across ancestries did not improve the prediction accuracy of case-control status in SAX, with the most variance explained (r^2 = 0.0025) at all SNPs irrespective of their p-value. It is possible that ancestry-specific genetic effects diminish the gains acquired by the increase in sample size. The inclusion of diverse samples in future large scale studies have the potential of identifying more variants that are common among multiple ancestries; and help improve PRS prediction accuracy in general.

3.6 Conclusion

This chapter presented results from an exploratory analysis to determine the SNP heritability of schizophrenia in an indigenous South African population and suggested that the biological mechanism of schizophrenia may be similar to that reported for EUR ancestry individuals. Indeed, larger samples from AFR ancestry are required to compute more accurate point estimate calculations of heritability in AFR ancestry populations. Additionally, methodologies that are more better suited to estimate partitioned heritability, and adequately account for population stratification in non-EUR should be employed. With regards to PRS, the prediction accuracy was low across ancestries, as expected, when PRS was derived from EUR and EAS

populations to predict phenotypes in an AFR population. Larger and more diverse samples are required to improve PRS prediction accuracy in populations not currently represented in large scale GWAS studies. In next chapter, the transferability of EUR-derived PRS are assessed across AFR ancestry populations for over 34 polygenic traits.

Chapter 4: Transferability of PRS across diverse AFR populations.

4.1 Abstract

AFR populations are vastly underrepresented in genetic studies but have the most genetic variation and encounter wide-ranging environmental exposures, globally. As a result, no systematic evaluation of PRS has been conducted in diverse AFR populations. In this chapter, PRS were calculated in simulations spanning continental AFR diversity and in empirical data from South Africa, Uganda, and the UK for over 30 phenotypes including anthropometric, psychosocial and sociodemographic traits. The prediction accuracy of PRS derived from EUR samples from the UK Biobank were evaluated across several AFR populations. PRS accuracy improved with ancestry-matched discovery and target cohorts more than with ancestry-mismatched studies. Within ancestrally and ethnically diverse South Africans, the accuracy of EUR derived PRS was low for all traits and varied across groups. The differences in AFR ancestries contributed more to the variability in PRS accuracy than other large cohort differences considered between individuals in the UK versus Uganda. When PRS was computed in AFR ancestry populations using EUR only versus ancestrally diverse studies; the increased diversity produced the largest accuracy gains for haemoglobin concentration and white blood cell count, reflecting large-effect ancestry-enriched variants in genes known to influence beta-thalassemia and malaria resistance, respectively, indicating a key role for environmental factors in the prediction accuracy of PRS. The differences in PRS accuracy across AFR ancestries from diverse regions are as large as across out of Africa continental ancestries, requiring commensurate nuance.

4.2 Introduction

The results from chapter 3 showed that the prediction accuracy of PRS derived from PGC-EUR GWAS did poorly at predicting schizophrenia in SAX. This chapter expands on the evaluation of the transferability of PRS across AFR populations. Seeing that no large genomic datasets exist for psychiatric disorders in AFR populations, in this chapter, the largest existing AFR genomic datasets for over 30 non-psychiatric phenotypes were used to systematically investigate the transferability of PRS within and across diverse AFR populations

GWAS studies have yielded important biological insights into the heritable basis of many complex traits and diseases (Visscher et al., 2017). However, the vast majority of studies have been conducted in EUR populations, raising questions about their utility across diverse populations (Martin et al., 2019; Morales et al., 2018; Popejoy & Fullerton, 2016; Sirugo et al., 2019). Previous studies have evaluated the generalizability of GWAS by using PRS to compare the association between genetically-predicted versus measured phenotypes in diverse populations. These studies have found that PRS accuracy decreases with increasing genetic distance between the GWAS discovery and PRS target cohorts (Duncan et al., 2019; Martin et al., 2017; Martin et al., 2019; Scutari et al., 2016). Since the earliest applications of PRS in human genetics, and increasing numbers of large scale studies in EUR populations, it is not surprising that PRS are most accurate in EUR individuals and least accurate in AFR ancestry populations (International Schizophrenia Consortium, 2009). These study biases continue to be apparent a decade later, with several-fold differences in prediction accuracy of many traits between EUR and non-EUR ancestry populations (Martin et al., 2019).

Quantifying PRS generalizability within and among AFR populations requires considerable nuance as they represent the most genetically diverse populations globally, with more than a million more genetic variants per person than in non-AFR populations (Auton et al., 2015). Study samples collected even within the same geographic regions of Africa have complex demographic histories with complicated patterns of admixture and population structure (Choudhury et al., 2020; Choudhury et al., 2017; Pagani et al., 2015; Uren et al., 2016). Further, AFR ancestry populations experience vastly different environments within Africa, versus outside the continent. This is due to the diverse environmental exposures among communities, within countries and regions of Africa. These differences provide unique epidemiological opportunities to query the impacts of vastly differing environments on PRS accuracy (Mostafavi et al., 2020). Previous empirical analyses and theoretical work fundamentally informs how demographic history and environmental variation interplay to

produce PRS heterogeneity in traditionally underserved populations (de Vlaming et al., 2017; Wray et al., 2013; Zaidi & Mathieson, 2020).

There are also clear benefits to including AFR populations in statistical genetics efforts. Because humans originated in Africa and diversified over eons, populations from this continent have the greatest genetic diversity among global populations (Auton et al., 2015; Campbell & Tishkoff, 2008; Henn et al., 2012). For this reason, it is expected that more genotype-phenotype associations would emerge from studies in AFR populations than can be found in other populations. African Americans make up 2.8% of GWAS participants but have been shown to contribute disproportionately to GWAS findings, i.e. 7% of all trait associations (Morales et al., 2018).

Moreover, the inclusion of AFR ancestry participants in large-scale genetic studies is important for equity in medical research. They have the lowest life expectancies globally (Hero et al., 2017; Roser et al., 2013), receive the lowest access to quality medical care in the US (Health & Services, 2017), and are the most underserved by genetic technologies (Martin et al., 2018Martin et al., 2018). An understanding of PRS transferability will critically inform populations that are currently the most underserved and point to where building genetic studies and resources may have the biggest benefits globally.

For this chapter, the aim was to investigate the portability of PRS across AFR populations using simulated and empirical data, and the objectives were as follows:

- Assess PRS generalizability across AFR populations in simulations using the data from the African Genome Variation Project.
- Assess PRS generalizability in black and admixed South AFR populations using data from the Drakenstein Child Health Study.
- Assess PRS generalizability across and within AFR populations using the Ugandan General Population cohort.

4.3 Methods and Materials

4.3.1 Genetic and Phenotypic Data

Several datasets were used for the analyses in this chapter and are outlined below. The total counts of individuals by population and/or study are shown in **Table 4.1** and descriptions of each dataset given below.

1000 Genomes Project

1000 Genomes Project data from the phase 3 integrated call set was accessed and used as a reference panel and for phasing and imputation (Auton et al., 2015).

Human Genome Diversity Project (HGDP)

Genotype data on the Illumina HumanHap650K GWAS array on hg18 was publicly available for HGDP (Li et al., 2008). Genotype data was lifted over to the hg19 genome build using hail (http://hail.is).

African Genome Variation Project (AGVP)

As described previously (Gurdasani et al., 2015), the AGVP data consists of dense genotype data from 1,481 individuals from 18 ethno-linguistic groups from Eastern, Western, and Southern Africa when including the Luhya and Yoruba from the 1000 Genomes Project. When accessed from the European Genome-Phenome Archive (EGA), “Ethiopian” is the provided population label encompassing the Oromo, Amhara, and Somali groups. After collapsing these groups and counting the 1000 Genomes data separately, 1,307 individuals from 14 populations are uniquely represented in AGVP, and 2,504 individuals from 26 populations are represented in the 1000 Genomes Project data (661 individuals from 7 populations are in the AFR super population grouping).

Table 4.1 - A description of the datasets used for the analyses

Dataset name	Abbreviation	# individuals included in analyses	Summary statistics or individual-level	Description	Analysis
African Genome Variation Project	AGVP	1,307	individual-level	14 populations across continental Africa	Simulations and reference panel
1000 Genomes Project		2,504	individual-level	26 populations globally, 661 individuals from 7 AFR populations	Reference panel
Human Genome Diversity Project	HGDP	1,043	individual-level	52 populations globally, 121 individuals from 7 AFR populations	Reference panel
Drakenstein Child Health Study	DCHS	640	individual-level	A multidisciplinary longitudinal birth cohort study in South Africa study following 1,000 mother child pairs	Empirical analysis
Uganda General Population Cohort	Uganda GPC	4,778	individual-level	Cohort was set up in 1989 to examine trends in HIV prevalence and incidence, and their determinants in rural south-western Uganda	Empirical analysis
UK Biobank	UKB	Up to 500,000	individual-level	GWAS data from 500,000 people aged between 40-69 years in 2006-2010 from across the country	Empirical analysis
BioBank Japan	BBJ	Up to 159,195	summary statistics	GWAS of quantitative and disease traits in 162,255 Japanese individuals	Discovery cohort/meta-analysis
Population Architecture using Genomics and Epidemiology	PAGE	Up to 49,796	summary statistics	GWAS results from self-identified Hispanic/Latino (n=22,216), African American (n=17,299), EAS (n=4,680), Native Hawaiian (n=3,940), Native American (n=652) or Other (n=1,052)	Discovery cohort/meta-analysis

Drakenstein Children's Health Study (DCHS) in South Africa

The DCHS is an ongoing, multidisciplinary population-based birth cohort study in the Drakenstein area in Paarl, South Africa (Stein et al., 2015; Zar et al., 2015; Zar et al., 2019). After providing informed consent, pregnant women were enrolled during their second trimester (20 – 28 weeks gestation); maternal-child dyads were then followed through childbirth and longitudinally thereafter. Enrolment occurred from March 2012 to March 2015 at two primary health care clinics – TC Newman, serving a predominantly mixed ancestry population and Mbekweni which serves a predominantly Black African population). Women were eligible to participate in the DCHS if they attended one of the study clinics, were at least 18 years of age and intended to remain residing in the study area. **Table 4.2** shows the phenotypes assessed in DCHS, including the GWAS summary statistic files used to compute PRS. All GWAS summary statistics were from individuals of EUR ancestry.

Uganda General Population Cohort (GPC)

The rural Uganda GPC of MRC/UVRI & LSHTM Uganda Research Unit was set up in 1989 initially to monitor the HIV epidemic among adults, children, and adolescents, but its mandate has since expanded to include other medical conditions (Asiki et al., 2013). The 'original GPC' is located in the sub-county of Kyamulibwa in rural south-western Uganda with activities having recently been expanded to the neighbouring two peri-urban townships of Lwabenge and Lukaya. The 'original GPC' includes about 10,000 adults and about 10,000 children and adolescents. In 2011, genotype data was generated on more than 5,000 adult participants from nine ethnolinguistic groups using the Illumina HumanOmni 2.5 BeadChip at the Sanger Wellcome Trust Institute (Asiki et al., 2013; Heckerman et al., 2016). The measured phenotypes are shown in **Appendix 7: Supplementary Table 4.1**.

Table 4.2 - The phenotypes measured in mothers who participated in the Drakenstein Child Health Study, alongside the GWAS summary statistics used to compute the polygenic risk scores and their corresponding files.

Trait type	Trait name	Variable type	GWAS trait	GWAS sample size	GWAS reference	Link
Physical/ biomedical	Maternal height	continuous	Height	~360k	Neale Lab	https://www.dropbox.com/s/ou12jm89v74k55e/50_irnt.gwas.imputed_v3.both_sexes.tsv.bgz?dl=0 -O 50_irnt.gwas.imputed_v3.both_sexes.tsv.bgz
Psychosocial risk	Depression	continuous & dichotomous	Major Depression	75,607; 231,747	https://www.ncbi.nlm.nih.gov/pubmed/29700475	https://www.med.unc.edu/pgc/results-and-downloads
Psychosocial risk	Depression	continuous & dichotomous	Major Depression	135,458; 344,901	https://www.ncbi.nlm.nih.gov/pubmed/29700475	https://www.med.unc.edu/pgc/results-and-downloads
Psychosocial risk	Psychological distress	continuous & dichotomous	Subjective well-being		Turley et al, 2018	https://www.thessgac.org/data, http://ssgac.org/documents/MTAG_README.txt
Psychosocial risk	Substance use	continuous & dichotomous	Drinks per week	941,280	https://doi.org/10.1038/s41588-018-0307-5	https://conservancy.umn.edu/handle/11299/201564
Psychosocial risk	Substance use	continuous & dichotomous	Smoking initiation	1,232,091	https://doi.org/10.1038/s41588-018-0307-5	https://conservancy.umn.edu/handle/11299/201564
Sociodemographic	Ethnicity	dichotomous	N/A			
Sociodemographic	Maternal age at enrolment	continuous	N/A			

UK Biobank (UKB)

The UK Biobank (UKB) enrolled 500,000 people aged between 40-69 years in 2006-2010 from across the country, as described previously (Bycroft et al., 2018). A more detailed description of the cohort can be found at: https://www.ukbiobank.ac.uk/. Only phenotypes that overlapped with those studied in the Uganda GPC were analysed in this chapter.

4.3.2 Ancestry analysis in the UKB

The UKB consists of approximately half a million participants of primarily EUR ancestry who have thousands of measured or reported phenotypes. To assess PRS accuracy across diverse ancestries, populations of ancestral groups were identified at two levels: 1) among continental groups, and 2) among regions in Africa. To define continental ancestries, the reference data from 1KGP3 was combined with HGDP into continental ancestries according to their corresponding meta-data **Table 4.3.**

Then, PCA was run on unrelated individuals from the reference dataset. To partition individuals in the UKB based on their continental ancestry, the PC loadings from the reference dataset were used to project UKB individuals into the same PC space. A random forest classifier was trained given the continental ancestry meta-data (AFR = African, AMR = admixed American, CSA = Central/South Asian, EAS = East Asian, EUR = European, and MID = Middle Eastern) based on the top six PCs from the reference training data. Then, this random forest was applied to the projected UKB PCA data and assigned initial ancestries if the random forest probability was > 50%, otherwise individuals were dropped from further analysis.

Table 4.3 - The breakdown of datasets into continental and regional ancestries

Dataset	Continental ancestry code	Continental ancestry description	African ancestry region	Count	Discovery or target
UKB	EUR	European	N/A	351194	Discovery
UKB	AFR	African	N/A	8426	Target
UKB	AMR	Admixed American	N/A	1099	Target
UKB	CSA	Central/South Asian	N/A	10084	Target
UKB	EAS	East Asian	N/A	2753	Target
UKB	EUR	European	N/A	9947	Target
UKB	MID	Middle Eastern	N/A	1553	Target
UKB	AFR	African	Admixed	4445	Target
UKB	AFR	African	West	2778	Target
UKB	AFR	African	East	728	Target
UKB	AFR	African	Ethiopia	293	Target
UKB	AFR	African	South	182	Target
Uganda GPC	AFR	African	Uganda	2247	Target

Next, AFR ancestry individuals were further partitioned using the same random forest approach as above but without further probability thresholding using AFR ancestry reference data from AGVP, HGDP, and the 1000 Genomes Project. These reference data were partitioned into UN regional codes with an additional region for Ethiopian populations given their unique population history and collapsing in AGVP data (Admixed, Central, East, Ethiopia, South, and West Africa), as shown in **Table 4.3**. PCA with reference data at the continental and subcontinental level within Africa are shown in **Appendix 7: Supplementary Figures 4.1 - 4.2.**

4.3.3 Phasing and imputation

The Ricopili pipeline was used to conduct pre-imputation QC and perform phasing and imputation for AGVP and the Uganda GPC (Lam et al., 2019). This pipeline was also used on the DCHS data, as described previously (Duncan et al., 2018). Briefly, data was phased using Eagle 2.3.5 and variants imputed using minimac3 in chunks ≥ 3 Mb. The 1000 Genomes phase 3 haplotypes were used as the reference panel for phasing and imputation. Downstream steps for each dataset were as follows:

For the AGVP dataset, strict best guess genotypes were used, where a variant was called if it had a probability of $p > 0.8$ and a missing rate less than 0.01 and MAF > 5%. Then, variants with MAF < 0.001 were excluded from further analyses. For Uganda GPC, combined best guess genotypes were used, where a variant was called if it had a probability $p > 0.8$ or set to missing otherwise. Then, SNPs were filtered to keep sites with missingness < 0.01 and MAF > 0.05.

4.3.4 Principal component analyses

Only SNPs with high imputation quality (INFO > 0.8) were considered for PCA. The first 20 principal components were computed using PLINK v1.07b (Chang et al., 2015) with the --pca flag for autosomal SNPs MAF > 0.05 and individual missingness < 0.05.

4.3.5 Simulations

To test the PRS prediction accuracy within and across AFR populations, a quantitative trait was simulated at four heritability rates (h^2 = 0.1, 0.2, 0.4 and 0.8) as follows:

First, effect size were randomly assigned to 5, 20, 100, 2,000, 10,000 and 50,000 causal variants, respectively. The causal effect was calculated based on the relationship between effect size and MAF as shown by Schoech et al. (2019), where variants on the lower end of the MAF spectrum have larger effects than those that are more common in the population. An individual's 'true' polygenic risk was then calculated as the sum of all causal effects using the --score flag in PLINK v1.07b (Chang et al., 2015). True polygenic scores were standardized to a mean of zero and standard deviation of 1. To account for the contribution of environmental risk factors, a random environmental effect was assigned from a normal random distribution

(mean = 0; sd = 1). The quantitative phenotype was generated per heritability rate as the sum of the true polygenic risk and heritability rate, coupled with an environmental effect given in the equation below:

$$\text{phenotype} = \text{true polygenic risk} + (1 - h^2)\ \text{environmental effect}$$

GWAS was then conducted for the simulated phenotype by splitting the AGVP dataset into three groups broadly representing the three geographical areas from where samples were obtained: East (n = 589), West (n = 517) and South Africa (n = 186, **Figure 4.1A**). To allow for the quantification of PRS prediction accuracy across the geographical regions, each group was further split into discovery and target cohorts, such that size each target cohort had 186 individuals. We conducted a linear regression in PLINK v2 (Chang et al., 2015) for all the simulated traits for the East and West discovery datasets, controlling for the first 20 principal components.

In PLINK v1.07b (Chang et al., 2015), independent SNP sets were obtained for each discovery cohort by clumping SNPs from corresponding summary statistics files with an LD threshold of $r^2 > 0.1$ and a window size of 500 kb. The effect sizes from the SNP set was used as weights to compute PRS for all three of the target datasets for a range of p-values (5e-08, 1e-06, 1e-04, 1e-03, 1e-02, 0.05, 0.1, 0.2, 0.5 and all). PRS were calculated in PLINK v2 (Chang et al., 2015) using the --score and --q-score-range flag (**Figure 4.1B**).

4.3.6 Heritability estimation

For the Uganda GPC, the heritability estimates of 34 quantitative traits computed previously were used (Gurdasani et al., 2019). For UKB, heritability estimates were computed for the same traits using LD score regression with the default model (i.e. without any functional annotations) (Bulik-Sullivan et al., 2015) and using population-matched LD score references from EUR populations downloaded from the authors' website (https://data.broadinstitute.org/alkesgroup/LDSCORE/).

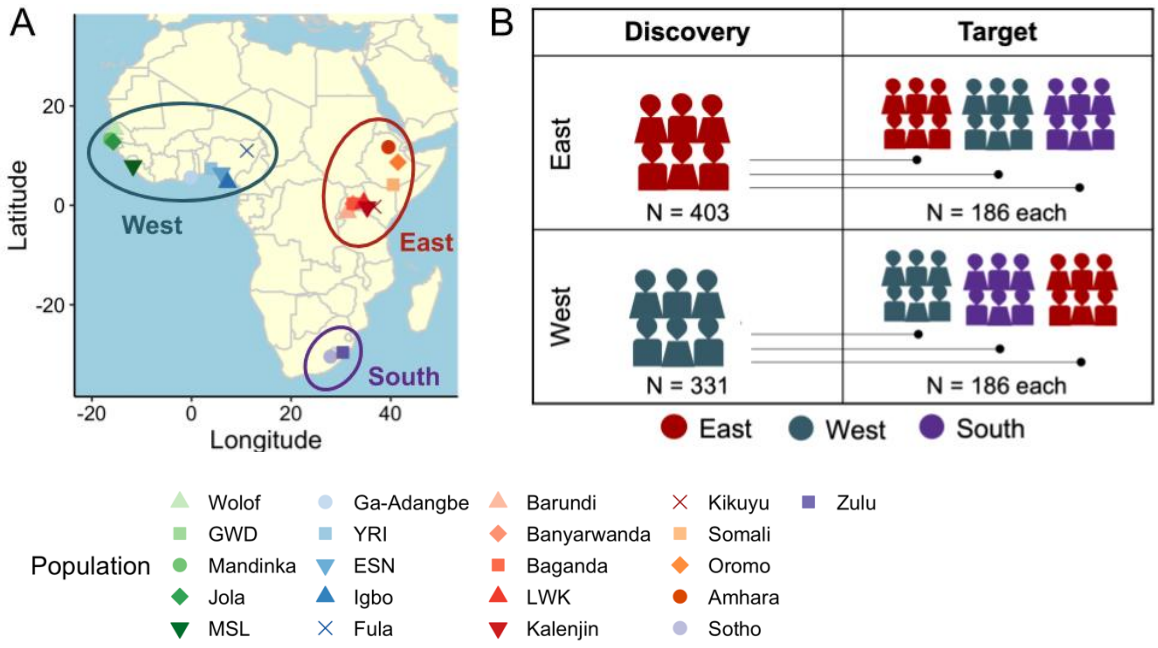


Figure 4.1 - A graphical representation of the simulated GWAS and PRS prediction accuracy across diverse regions of Africa using genetic data from the AGVP.
A) Populations were grouped into East, West, and South based on the United Nations geoscheme groupings. B) GWAS discovery cohorts included East (N = 403) and West (N=331) AFR individuals, which were independent of each target cohort (N = 186 individuals per region). South Africans were excluded from the discovery population due to the limited total sample size (2 populations and 186 individuals total)

4.3.7 PRS calculation

All PRS were calculated using a pruning and thresholding approach with an LD reference panel. All clumping was done using PLINK v2 (Chang et al., 2015) with an LD threshold of r^2 = 0.1 and a window size of 500 kb. PRS scoring calculations were conducted using the --score and --q-score-range flags in PLINK (Chang et al., 2015) for AGVP simulations and DCHS. Custom scripts in hail (http://hail.is) were used to calculate PRS in the Uganda GPC and UKB data due to the larger sample sizes. For imputed genotypes, SNP dosages in PRS calculations were used. Ten PRS were calculated for each analysis, including independent effects from the following p-value thresholds: 1, 0.5, 0.2, 0.1, 0.05, 0.01, 1e-3, 1e-4, 1e-6, 5e-8. For all analyses, the PRS that explained the most phenotypic variance was used.

PRS prediction accuracy was calculated using partial values for continuous traits computed with custom scripts in R. For AGVP simulations and DCHS (because all participants were mothers of a similar age), the first 10 PCs were included as covariates when computing the

partial R^2 or pseudo-R^2 specifically attributable to the PRS. For Uganda GPC data, age, sex, and the first 10 PCs were computed when computing partial R^2 of the PRS. For more consistency with the GWAS that were run in UKB previously (Howrigan, 2017) and here with a holdout target set, age, sex, age^2, age*sex, age^2*sex, and the first 10 PCs were included as covariates when computing PRS R^2. (The UKB EUR GWAS included 20 PCs, but fewer were used due to the particularly small sample sizes of some target ancestry groups coupled with minimal population structure observed in PCs lower than PC10, **Table 4.3**).

4.3.8 Meta-analysis

PLINK v2 (Chang et al., 2015) was used to conduct inverse variance-weighted meta-analysis across GWAS summary statistics with the --meta-analysis option.

4.3.8.1 LD reference panels and clumping

The cumulative contribution of genome-wide SNPs to polygenic risk of a range of traits and cohorts was estimated, primarily using holdout target cohorts from the UKB or other cohorts listed in **Table 4.4** based on LD clumping and p-value thresholding using PLINK v2 (Chang et al., 2015) and hail. Because not all genetic data used in the meta-analyses was available at the individual-level, the 1KGP3 data was used as a proxy LD reference panel. The ancestral representation of each population per trait was weighted to match the continental level. Individuals were matched as in **Table 4.4.**

Table 4.4 - 1KGP3 reference population used per dataset included in the meta-analysis

Cohort	1KGP3 reference data	Citation
Biobank Japan (BBJ)	East Asian (EAS)	(Nagai et al., 2017)
UK Biobank	European (EUR)	(Bycroft et al., 2018)
Ugandan Genome Resource (UGR)	African (AFR)	(Gurdasani et al., 2019)
Population Architecture using Genomics and Epidemiology (PAGE)	Proportional weighting of AFR, EAS, AMR (depending on trait, see **Appendix 7: Supplementary Table 4.2** description for more detail)	(Wojcik et al., 2019)

The maximum number of individuals was used to construct this proportional reference panel. For example, in the meta-analysis of height across the UKB, BBJ, and PAGE cohorts, UKB has the largest sample size in the discovery cohort (N = 350,353), so all EUR from 1KGP were included in the reference panel (N = 503), while a random sampling of EAS, AFR, and AMR individuals were included proportionally to the overall diversity of the discovery cohorts in the meta-analysis.

4.4 Results

4.4.1 Simulated PRS accuracy within and across diverse AFR populations

Several quantitative traits with varying numbers of causal variants (N = 5; 20; 100; 2,000; 10,000; and 50,000) and heritability rates (h^2 = 0.1, 0.2, 0.4, and 0.8) were simulated. Then independent GWAS were conducted for each number of causal variants, heritability rate and population based on geographical region (i.e. East or West Africa) combination. This resulted in a total of 48 independent GWAS. The prediction accuracy for PRS derived from the GWAS summary statistics was calculated, considering ten different p-value thresholds within and across independent target populations from East, West, and South Africa. In general, ancestry-matched results with the sparsest and most heritable genetic architectures produced the highest prediction accura**cy (Figure 4.2 - Figure 4.5**).

The prediction accuracy was highest for the most heritable trait with the sparsest genetic architecture. Specifically, the simulated trait with the heritability rate of h^2 = 0.8 and number of causal variants fewer than 100 (n = 5 and n = 20) had the highest R^2 (**Figure 4.2)**. The within-ancestry prediction at p-value threshold < 5e-08 and 5 causal variants were: R^2 = 0.86, p = 1.74 X 10^{-74} for East discovery - East target scores; R^2 = 0.85, p = 9.9e-74 for West discovery - West target scores (**Figure 4.2).** Lower prediction accuracy was observed with ancestry mismatched discovery versus target cohorts at five causal variants and p-value threshold = 1e-6 (R^2 = 0.66, p = 1.79e-42 for West discovery - West target scores, compared to R^2 = 0.53, p =1.29e-74 for East discovery - West target scores). The scores in the South target sample were comparable when using East- or West-derived summary statistics (R^2 = 0.86, p = 5.19e-84 for West-derived summary statistics, and R^2 = 0.86, p = 1.35e-83 for East-derived summary statistics). For the same trait, the negligible prediction accuracy when the number of causal variant exceeded 100 can be attributed to the small discovery cohort sample sizes leading to no variants meeting the genome-wide significance threshold in the GWAS.

The prediction accuracy declined with the decrease in the heritability rate of the trait. As expected, the least heritable trait (h^2 = 0.1) had the lowest prediction accuracy across ancestries and number of causal variants (**Figure 4.5**). This suggests that non-genetic factors contribute to the risk of this trait.

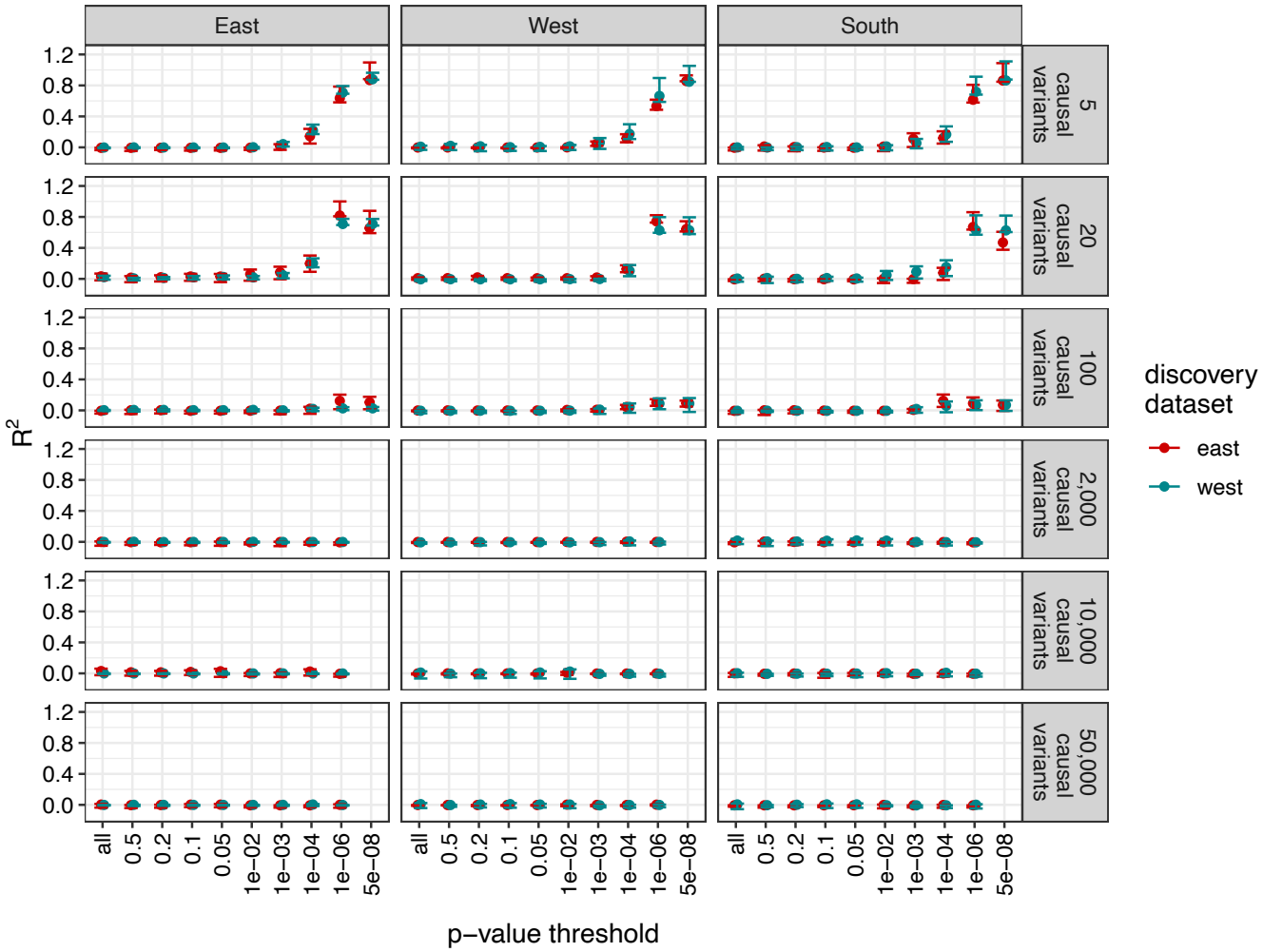


Figure 4.2 - Simulated PRS accuracy across regions of Africa, and various number of causal variants for a trait with heritability rate of 0.8.

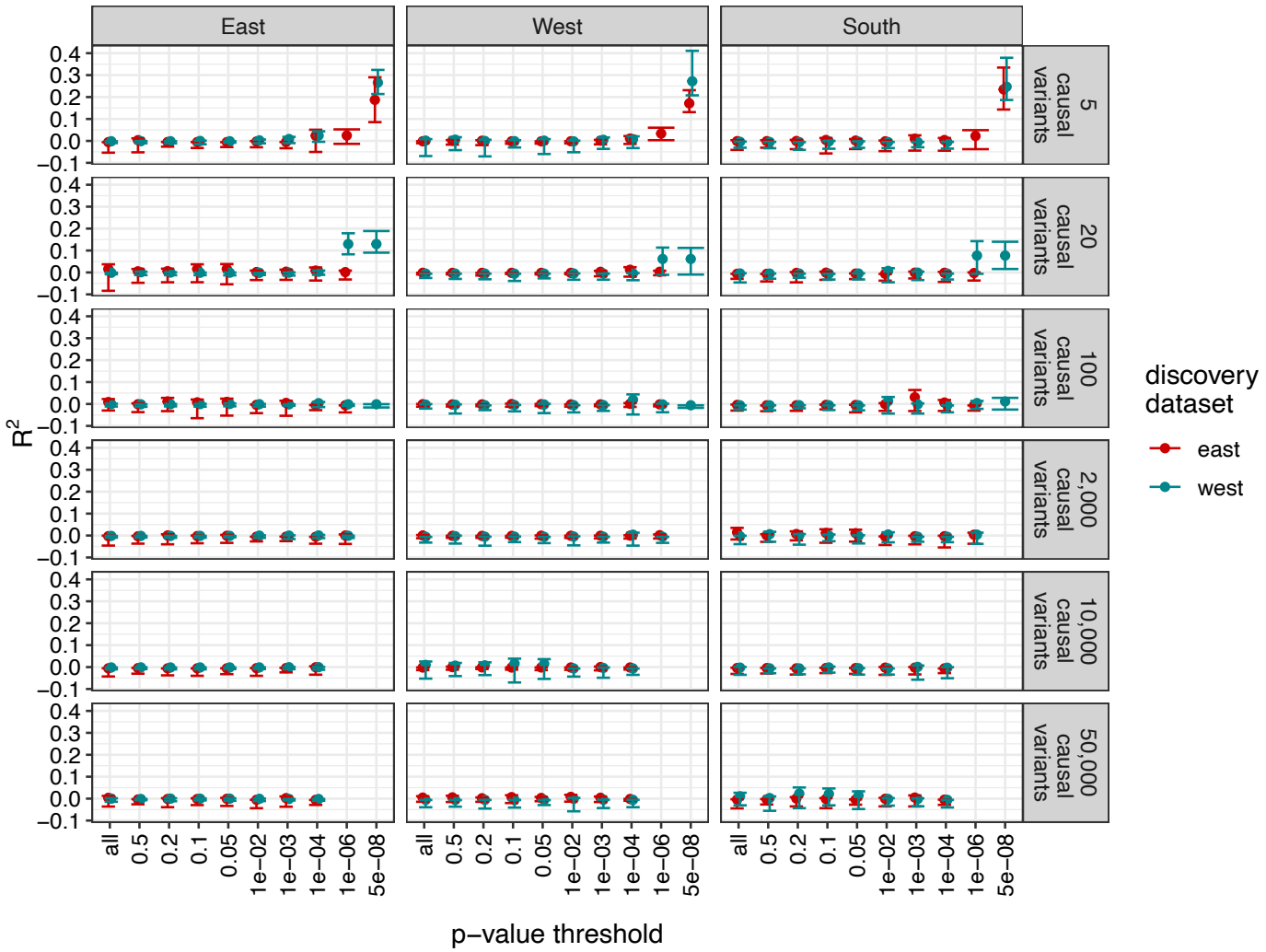


Figure 4.3 - Simulated PRS accuracy across regions of Africa, and various number of causal variants for a trait with heritability rate of 0.4

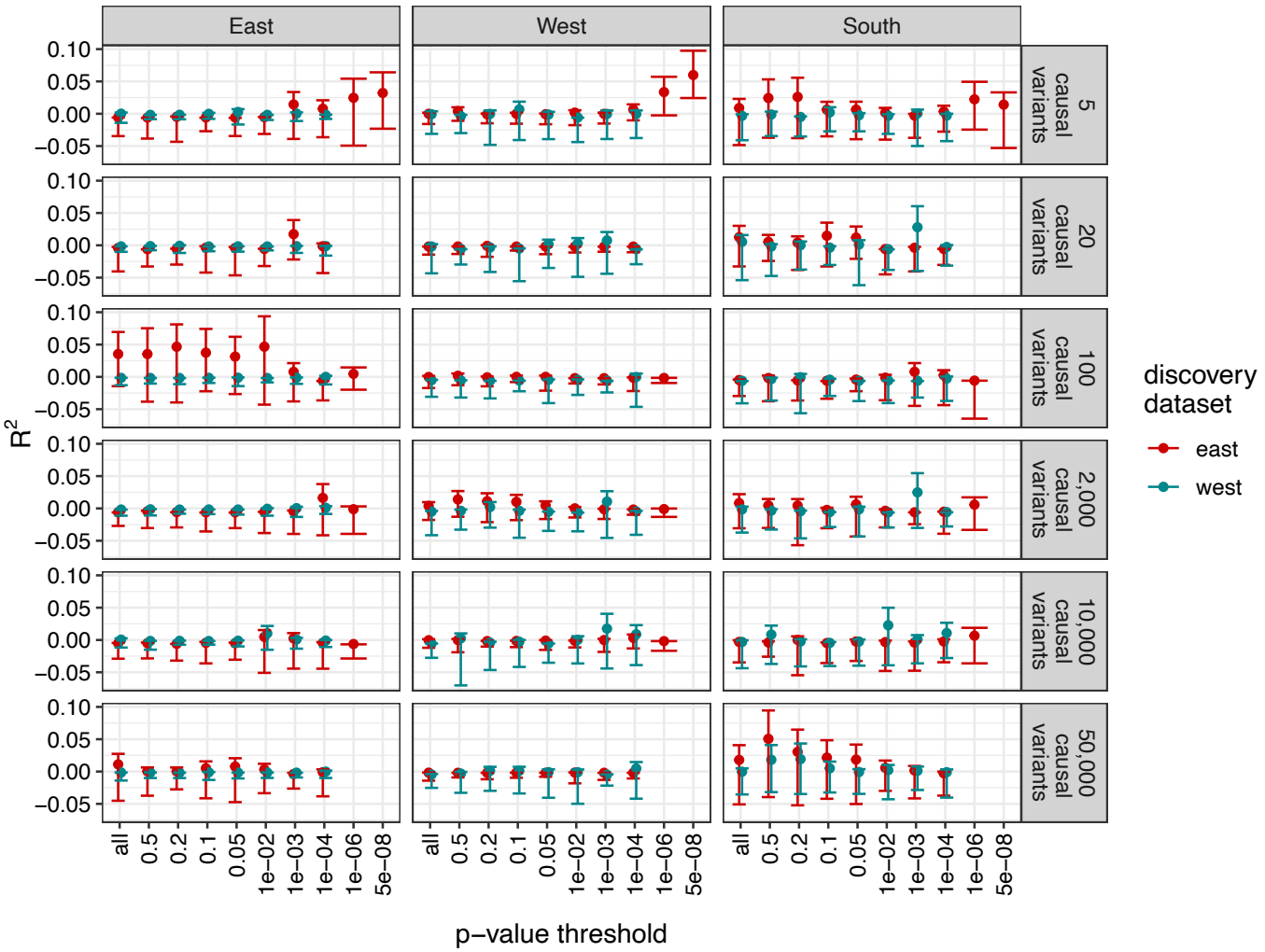


Figure 4.4 - Simulated PRS accuracy across regions of Africa, and various number of causal variants for a trait with heritability rate of 0.2

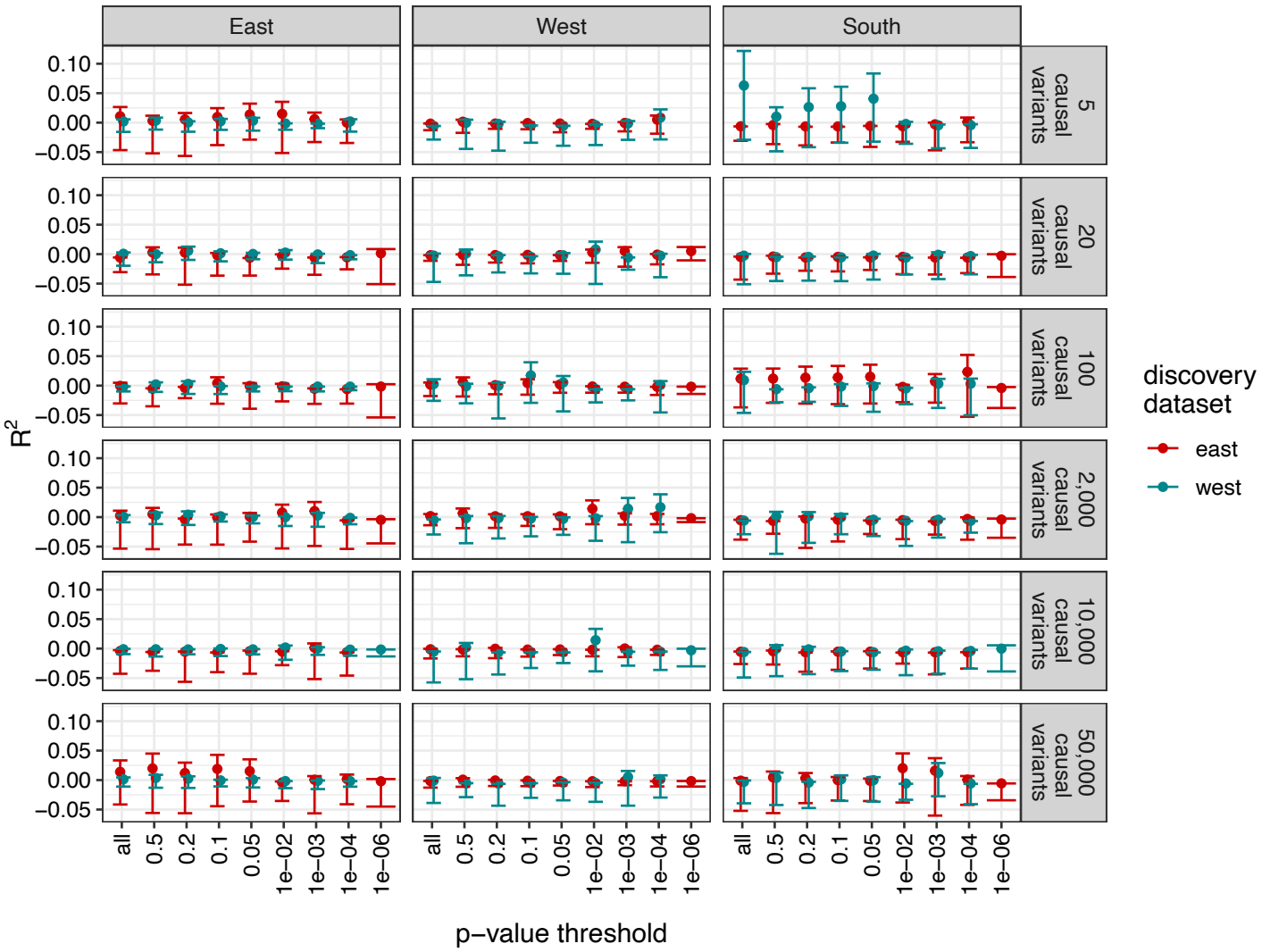


Figure 4.5 - Simulated PRS accuracy across regions of Africa, and various number of causal variants for a trait with heritability rate of 0.1

4.4.2 PRS accuracy in South African populations

While the simulations showed that PRS generalize poorly across Africa due to substantial genetic diversity and differences across the continent, there was also considerable genetic and environmental diversity within regions and countries. PRS accuracy for a range of measured phenotypes in mothers genotyped in the DCHS cohort in South Africa was quantified, including several sociodemographic, physical/biomedical, and psychosocial risk traits **(Table 4.2**). The DCHS cohort consists of participants with multiple ancestry groups that include an admixed population with ancestry from multiple continents, as well as a population with almost exclusively AFR population. These ancestry groups correlated with self-reported "Mixed" and "Black/African" ethnicities, respectively (**Figure 4.6**).

PRS were computed for maternal height, depression, psychological distress, alcohol consumption, and smoking in DCHS overall, by ethnic group, and by ancestry within the Mixed ethnic group. Across all genetically predicted phenotypes, only height was significantly predicted (**Figure 4.7**). Height was more accurately predicted on average in the ancestry clusters most correlated with the Mixed versus Black/AFR ethnic group (R^2 = 0.099, 95% bootstrapped CI = [0.012, 0.18], *P* =1.5e-7 versus R^2 = 0.021, 95% CI = [-0.031, 0.043], *P* = 5.27e-3, respectively). It was expected that PRS accuracy would increase with decreasing AFR ancestry within the Mixed ethnic group as has been shown previously in admixed AFR populations (Bitarello & Mathieson, 2020). The results showed suggestive evidence that was consistent with this trend when partitioning the Mixed group into two bins along the first principal component (PC1) (R^2 = 0.091, 95% CI = [-0.04, 0.17], *P* = 6.4e-4 in lower half of PC1 corresponding to more AFR ancestry vs R^2 = 0.12, 95% CI = [-9.0e-4, 0.21], *P* = 5.7e-5 with more out of Africa ancestry), although small sample sizes (N = 137 in each PC1 bin) limited definitive comparisons. These results are consistent with variable prediction accuracy among diverse AFR ancestry groups within South Africa, and insignificant prediction in African populations for all but the most heritable and accurately predicted traits elsewhere.

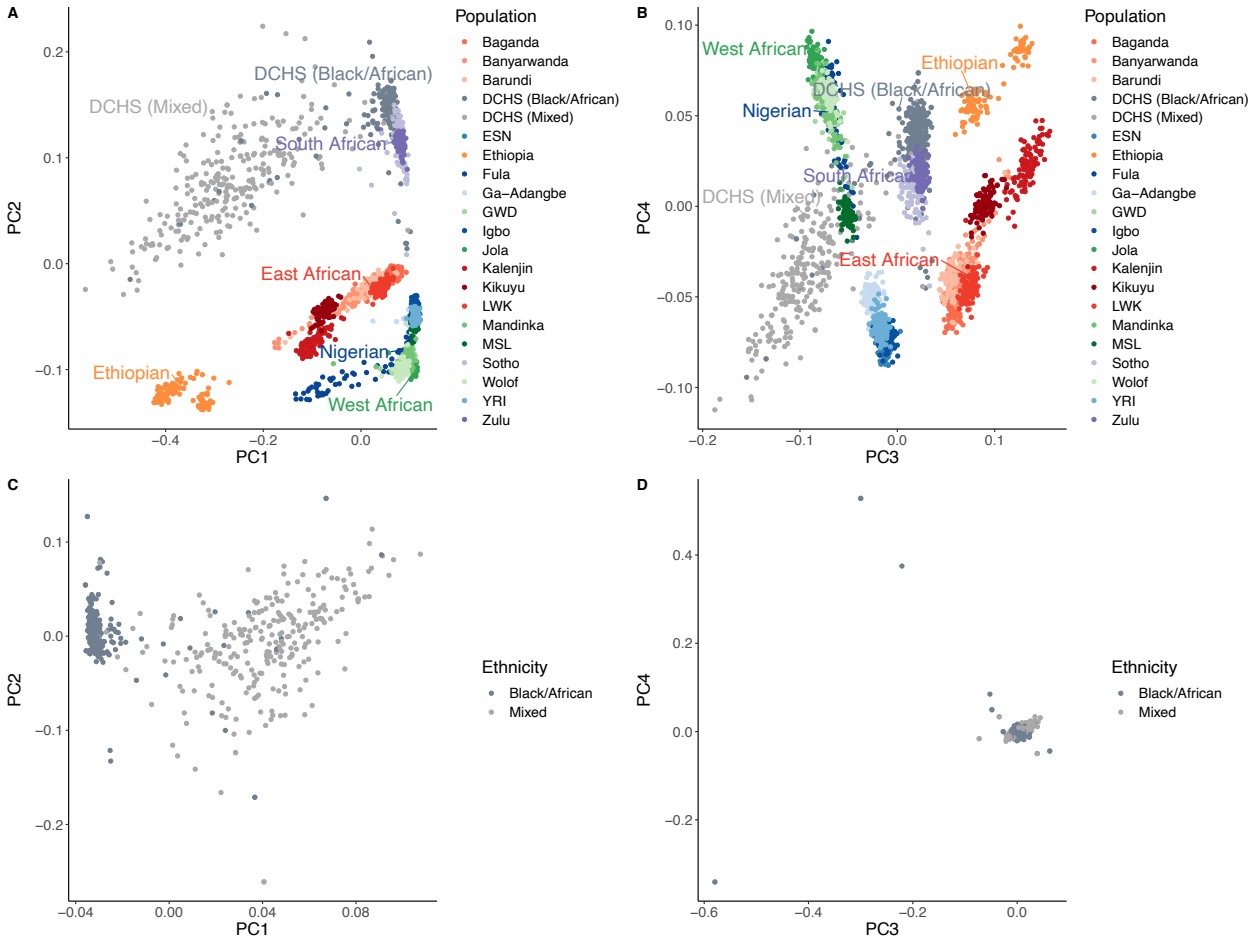


Figure 4.6 - Ancestry and ethnicity within DCHS and compared to AFR reference populations from AGVP and the 1KGP3
A-B) PCs in DCHS Black/African and Mixed ethnicities compared to reference data. C-D) PCs within DCHS without reference data. A, C) PC1 and PC2. B, D) PC3 and PC4.

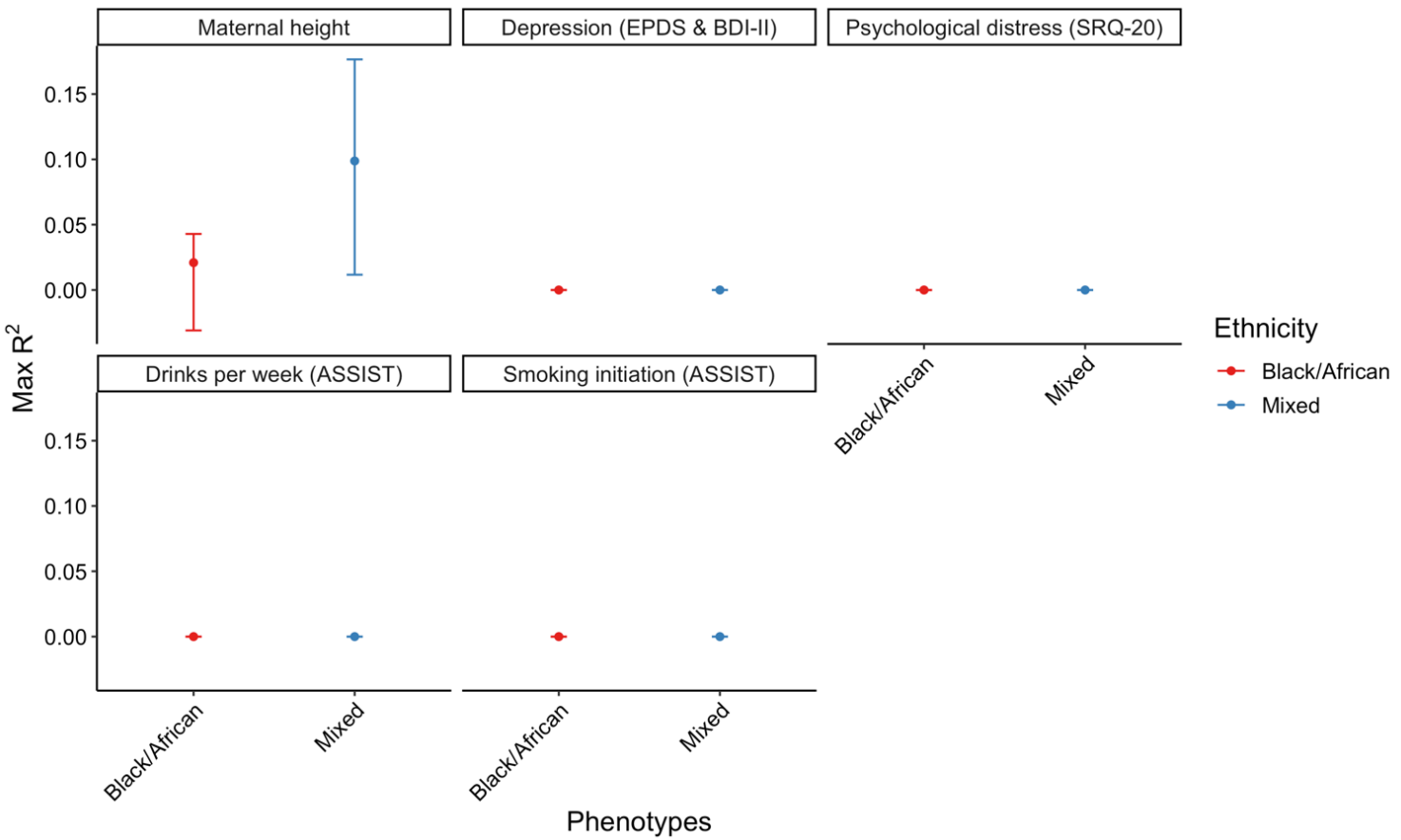


Figure 4.7 - PRS accuracy by ethnicity in the DCHS cohort for five measured phenotypes.
Max R^2 shows specifically the PRS threshold that explains the most phenotypic variation. Abbreviations are: EPDS = Edinburgh Postnatal Depression Scale, BDI = Beck Depression Index, ASSIST = Alcohol, Smoking and Substance Involvement Screening Test.

4.4.3 Phenotypic and genetic difference across the Uganda GPC and UKB

4.4.3.1 Lower phenotypic correlations in Uganda GPC suggest higher contributing environmental effects

The phenotypic similarities within and across the Uganda GPC and UKB participants were investigated. These are two of the largest cohorts with dozens of traits measured in AFR ancestry individuals. It should be noted that there were fundamental differences in the designs of these cohorts. The Uganda GPC enrolled participants using a house-to-house study design and generated genetic data on 5,000 adults from rural villages in southwestern Uganda (Asiki et al., 2013), while the UKB enrolled 500,000 people aged between 40 - 69 years in 2006 - 2010 from across the country (**Methods** (Bycroft et al., 2018)). Previous studies have reported higher rates of infectious diseases (e.g. HIV, hepatitis B and C) in the Uganda GPC than would be expected in the UKB (Asiki et al., 2013). There are many additional potential environmental explanations for mean shifts in phenotypes, such as dietary and food security differences contributing to considerable BMI differences across cohorts ($\mu = 21.3$ and $\sigma = 3.8$ in Uganda GPC versus $\mu = 27.4$ and $\sigma = 4.8$ in UKB, $p < 2.2e\text{-}16$).

To quantify comparisons while controlling for demographic differences for each of the 34 quantitative traits (**Appendix 7: Supplementary Table 4.1**) measured in both cohorts, the means for each phenotype were centered, then the effects of age and sex within each cohort were regressed out. Next, the distributions and variances of each phenotype across cohorts were compared via the Kolmogorov-Smirnov (K-S) and F-tests, respectively. Given the large sample sizes, all K-S tests were significantly different (**Appendix 7: Supplementary Table 4.2**), with several phenotypes (e.g. Bilirubin, basophil count (BASO), HbA1c, alkaline phosphatase test (ALP), eosinophil count (EOS), triglycerides (TG), and neutrophil count (NEU) showing distributional and variance differences of considerable magnitude (**Figure 4.8**).

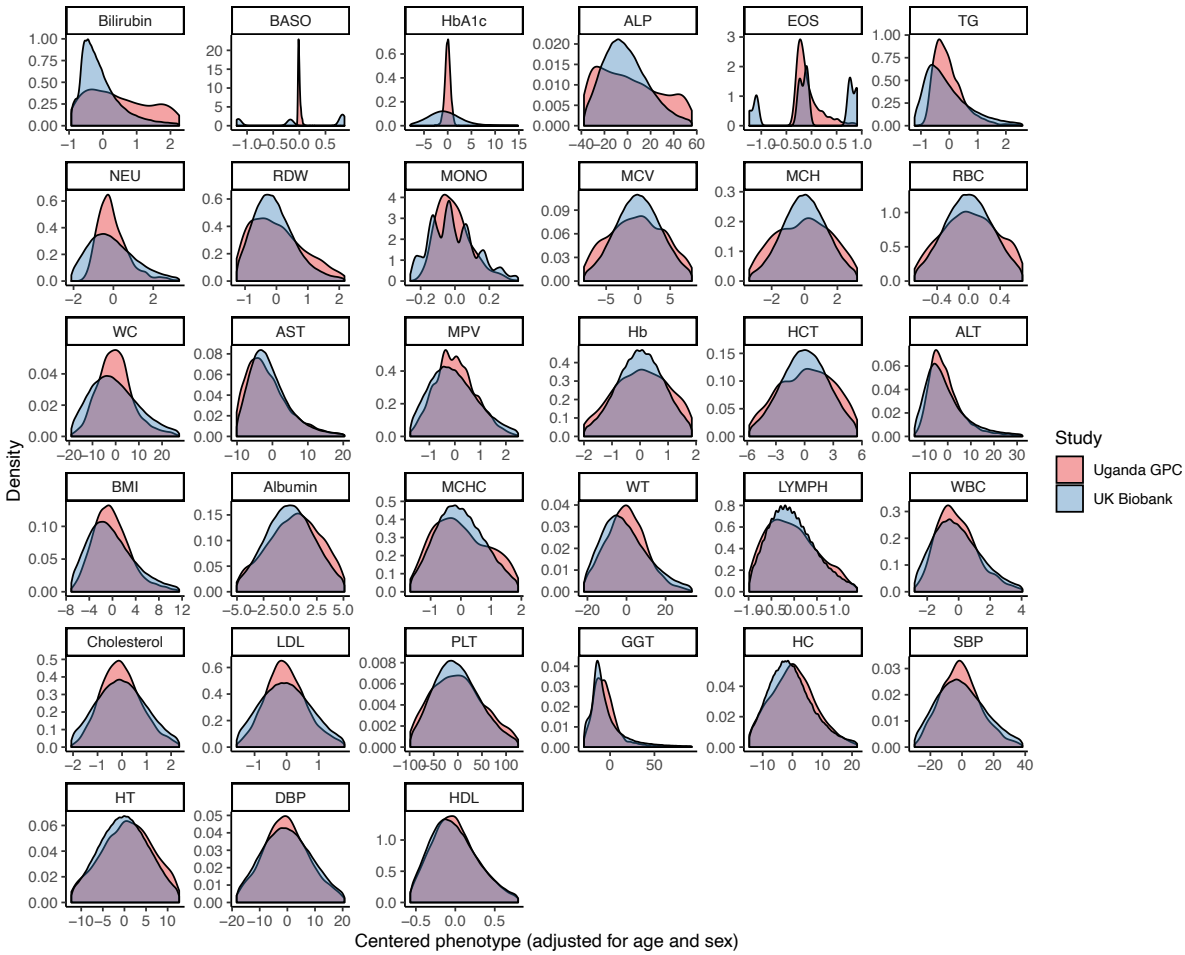


Figure 4.8 - Comparison of phenotypic distributions between the Uganda GPC and UKB cohorts
The phenotypes were mean centered and the effects of age and sex regressed out For display purposes only, the middle 95th percentile of the data is shown. Phenotypes are ordered by distributional difference as estimated by the Kolmogorov-Smirnov test statistic.

The relationships between phenotypes across the datasets were investigated, and showed similar trends overall, with distances across variance-covariance matrices for these cohorts showing evidence of significant correlation (**Figure 4.9**, Mantel test Z-statistic = 0.73, p < 1e-4). The correlations among phenotypes were slightly higher overall in the Uganda GPC than in UKB (**Figure 4.8A vs 4.8B**), both among related and unrelated individuals, as expected from a household versus volunteer-based design (**Figure 4.8**, **Appendix 7: Supplementary Figure 4.3**). These findings are expected due to the effects of shared genetics and/or shared household environments contributing to more similar phenotypes (Kong et al 2018). More specifically, there were consistent correlations among combinations of phenotypes including

systolic blood pressure (SBP) and diastolic blood pressure (DBP); red blood cell count (RBC), haemoglobin (Hb), and haematocrit (HCT); Cholesterol and low density lipoprotein (LDL); waist circumference (WC), body mass index (BMI), weight (WT), and hip circumference (HC); mean corpuscular haemoglobin concentration (MCHC), mean corpuscular haemoglobin (MCH), and mean corpuscular volume (MCV); gamma-glutamyl transpeptidase test (GGT), alanine aminotransferase test (ALT), aspartate aminotransferase test (AST), and ALP; and monocyte count (MONO), neutrophils (NEU), and white blood cell count (WBC) with high overall correlations across these datasets for these traits. Some pairs of traits, however, have significantly different correlations across datasets. The largest difference in phenotypic correlations across datasets was between ALP and WT ($\rho = 0.11$, $p < 2.2e\text{-}16$ in UKB versus $\rho = -0.36$, $p < 2.2e\text{-}16$ in Uganda GPC).

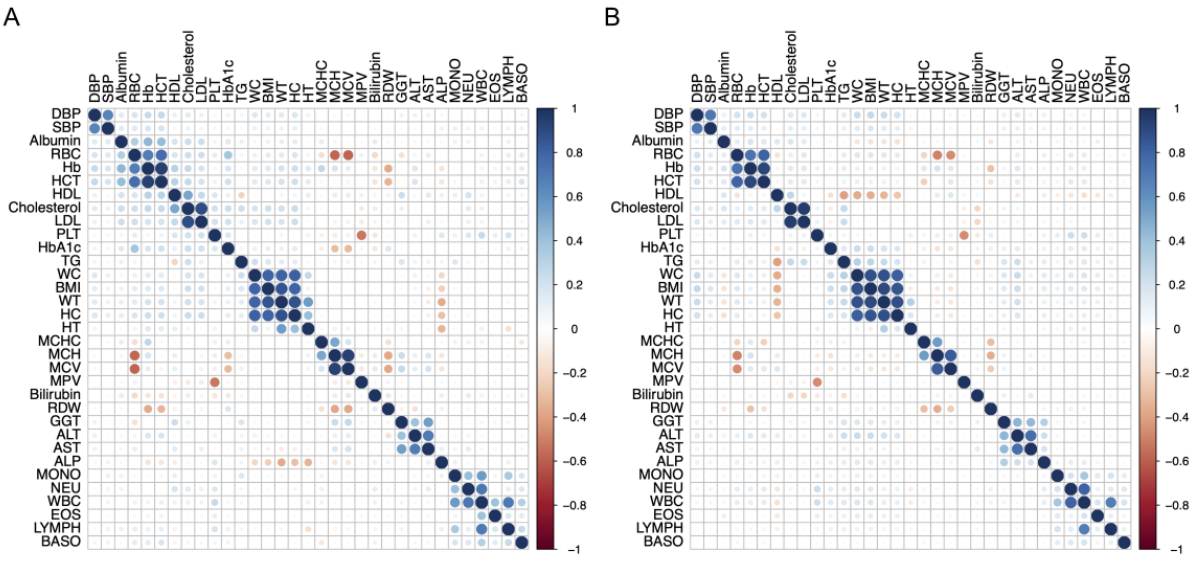


Figure 4.9 - Phenotype and genotype correlations among 33 quantitative traits measured in the Uganda GPC data and the UKB.
A) Phenotypic correlations measured in traits in the Uganda GPC among unrelated individuals. B) Phenotypic correlations in the UKB EUR ancestry unrelated individuals. A-B) Phenotypes were mean centered and adjusted for age and sex within each cohort prior to correlation analysis. The order of each phenotype correlation is determined by hierarchical clustering in the Uganda GPC.

The next goal was to compare trait heritability estimates in the UKB versus Uganda GPC data. However, the sample size and study design differences between these cohorts required the application of different methods that limit comparability. Specifically, the household design of Uganda GPC included smaller sample sizes with more relatives in which family-based heritability estimates are most appropriate, whereas the large sample size and volunteer design in UKB makes SNP-based heritability estimates from unrelated individuals most appropriate. **Figure 4.10** shows the heritability estimates across traits in the UKB versus Uganda GPC using these approaches. As expected from the differences in the methods, study designs, and sample sizes and consistent with expectation from family-based versus unrelated heritability estimates across these two studies, there was higher but noisier estimates in Uganda GPC for most traits. This is because there is a higher likelihood that a member of a family will develop a trait if the other members have the trait, than not.

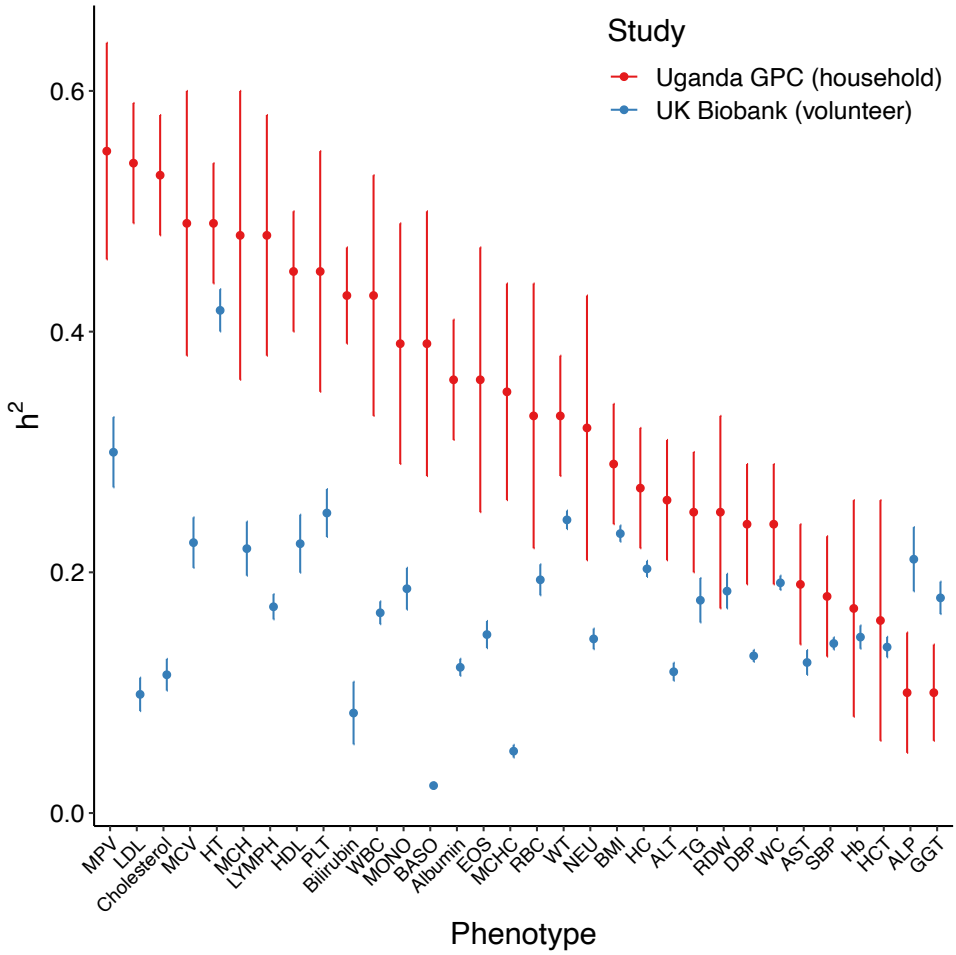


Figure 4.10 - Trait heritability comparison between UKB and the Uganda GPC
Heritability estimates in the UKB individuals was done using LD score regression from unrelated individuals, whereas a mixed model approach with multiple random effects were used previously in Uganda GPC due to the complex household and geographically diverse study design.

4.4.3.2 AFR population PRS derived from EUR ancestry GWAS data are remarkably inaccurate

To understand baseline trans-ethnic PRS accuracy using a typical approach, 32 traits in the Uganda GPC were predicted using GWAS summary statistics from the UKB EUR ancestry individuals. While several traits were significantly predicted across ancestries, prediction accuracy was low for most traits (**Figure 4.11**); the most accurate PRS was for mean platelet volume (MPV), (R^2 = 0.036, 95% CI = [0.0069, 0.063], *P* = 5.73e-7) while the average variance explained across all traits was less than 1% (mean R^2 = 0.007).

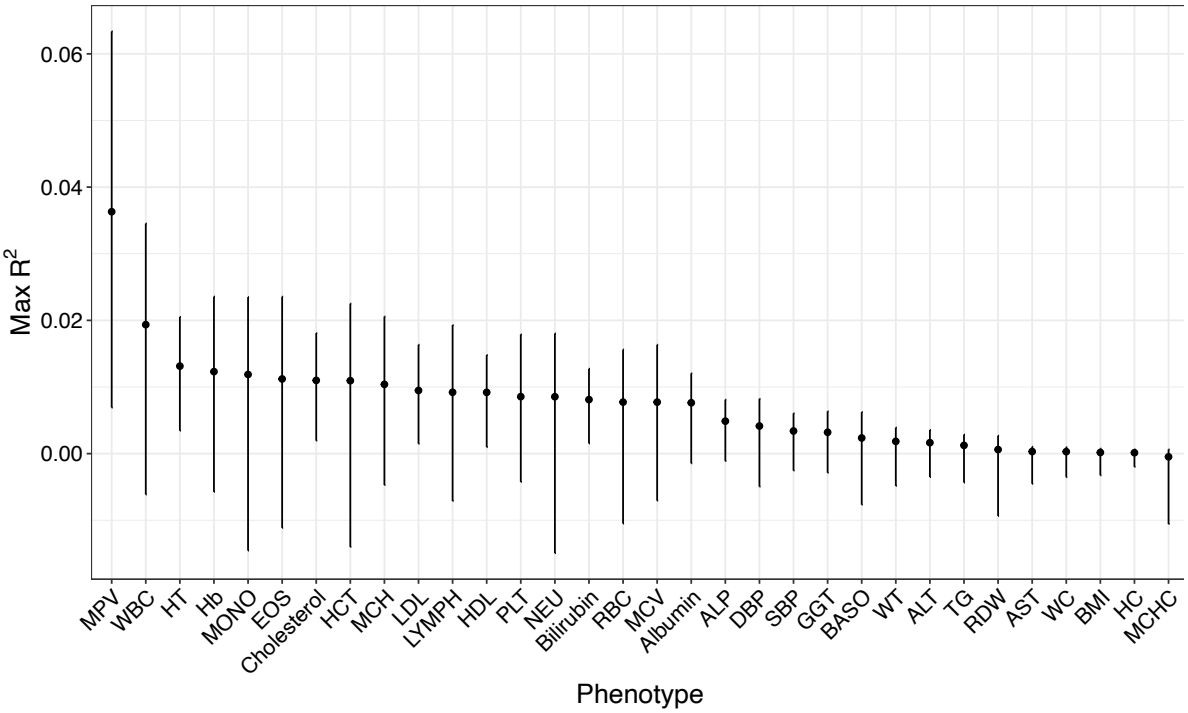


Figure 4.11 - PRS accuracy for 32 traits in unrelated Uganda GPC individuals calculated using GWAS summary statistics from UKB EUR ancestry individuals
The PRS from the p-value threshold with the highest accuracy of the 10 thresholds computed is shown

To assess the relative effects of ancestry versus cohort differences on decreases in prediction accuracy across populations, 10,000 EUR ancestry individuals from UKB were withheld for use as a EUR target cohort. Then GWAS was reran for all traits, followed by PRS computation for individuals with diverse continental ancestries in the UKB as target populations (i.e. EUR = Europeans withheld from the GWAS, AMR, MID, CSA, EAS, and AFR = African, subcontinental AFR ancestries in the UKB (Ethiopian, Admixed, South, East, West AFR ancestries), as well as the Uganda GPC. Among continental ancestries, R^2 and 95% confidence intervals were computed for each trait, then the median relative accuracy (RA) was compared to EUR and median absolute deviation (MAD) across all traits. The traits were most accurately predicted in EUR (RA = 1, MAD = 0), followed by AMR (RA = 0.784, MAD = 0.023), MID (RA = 0.643, MAD = 0.034), CSA (RA = 0.621, MAD = 0.031), EAS (RA = 0.477, MAD = 0.024), and AFR (RA = 0.219, MAD = 0.014) (**Figure 4.12**).

Next, prediction accuracy was compared within the AFR ancestry populations, but because some PRS accuracy estimates were noisy due to small sample sizes in UKB AFR (especially Ethiopian and South African individuals), the analyses was restricted to those traits predicted with a 95% confidence interval < 0.08. Among these traits, prediction accuracy was the most accurate for individuals with Ethiopian ancestry (RA = 0.511, MAD = 0.059), followed by recently admixed individuals with West AFR and EUR ancestry (RA = 0.276, MAD = 0.016), East AFR ancestry (RA = 0.193, MAD = 0.023), West AFR ancestry (RA = 0.150, MAD = 0.012), and South AFR ancestry (RA = 0.083, MAD = 0.014) (**Figure 4.11**).

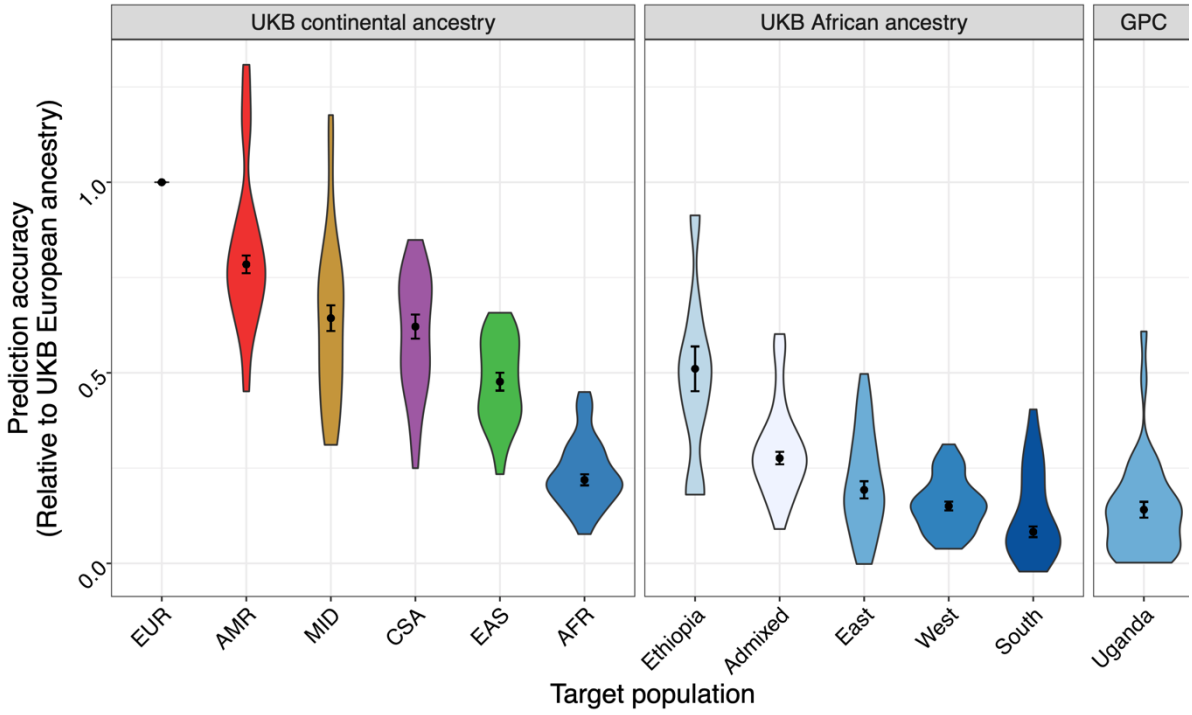


Figure 4.12 - PRS accuracy for up to 34 traits within and across diverse ancestries.
PRS prediction accuracy was compared for continental and AFR ancestry populations within the UKB, and the Uganda GPC.

4.4.3.3 Worse prediction accuracy across ancestries than across cohorts

Next, the prediction accuracy was computed among similar ancestry participants from different cohorts by computing PRS for 32 traits in UKB participants with East AFR ancestry versus Uganda GPC participants using GWAS summary statistics from UKB EUR individuals. As expected, the prediction accuracy in these populations was low across all traits in both cohorts and only slightly higher in the UK East AFR ancestry individuals than in the Uganda GPC individuals (mean R^2 = 0.017, sd = 0.013 versus mean R^2 = 0.012, sd = 0.010, respectively, **Figure 4.13**). Across traits, the differences between cohorts within the same ancestry were much smaller than the differences across ancestries within the UKB (**Figure 4.12**), indicating that ancestral diversity has a larger impact on genetic risk prediction than cross-cohort differences. Smaller effects on genetic prediction accuracy differences across cohorts may be attributable to environmental differences, such as higher rates of malnutrition and infectious diseases previously reported in Uganda and in the GPC (Asiki et al., 2013; Nalwanga et al., 2020).

4.4.3.4 Improved AFR PRS prediction accuracy with multi-ethnic GWAS summary statistics

To test the change in prediction accuracy for these phenotypes in the same target populations in the UKB using different discovery cohorts combined via meta-analysis, the prediction accuracy was computed using meta-analysed GWAS summary statistics from diverse discovery cohorts, including the UKB, BBJ, PAGE and UGR (**Table 4.4**). For each trait, discovery cohort, and target cohort combination, the PRS R^2 values were normalized from the p-value threshold that explained the maximum phenotypic variance with respect to the prediction accuracy in the EUR target cohort using UKB summary statistics only, then relative accuracies were computed as before.

The prediction accuracy improved the most across populations when using a discovery cohort consisting of the meta-analysed GWAS summary statistics across the UKB, BBJ and PAGE cohorts **(Figure 4.14**), but not the UGR data (**Appendix 7: Supplementary Figure 4.4**). Instead, meta-analysing the UGR data with UKB did not improve prediction accuracy for any population and most notably decreased accuracy in AFR ancestry target populations (discovery UKB median RA = 0.22, UGR+UKB median RA = 0.15, **Appendix 7: Supplementary Figure 4.4**). It was hypothesized that this can be explained by the relatively small sample size of UGR adding more noise than signal compared to the other relatively large discovery datasets, but another explanation could come from environmental heterogeneity. When predicting traits using the UKB, BBJ and PAGE meta-analysis as a discovery cohort, the prediction accuracy increased most for the AMR, EAS, and AFR target populations, which more closely resemble the ancestry patterns of PAGE and BBJ (**Figure 4.13**).

.

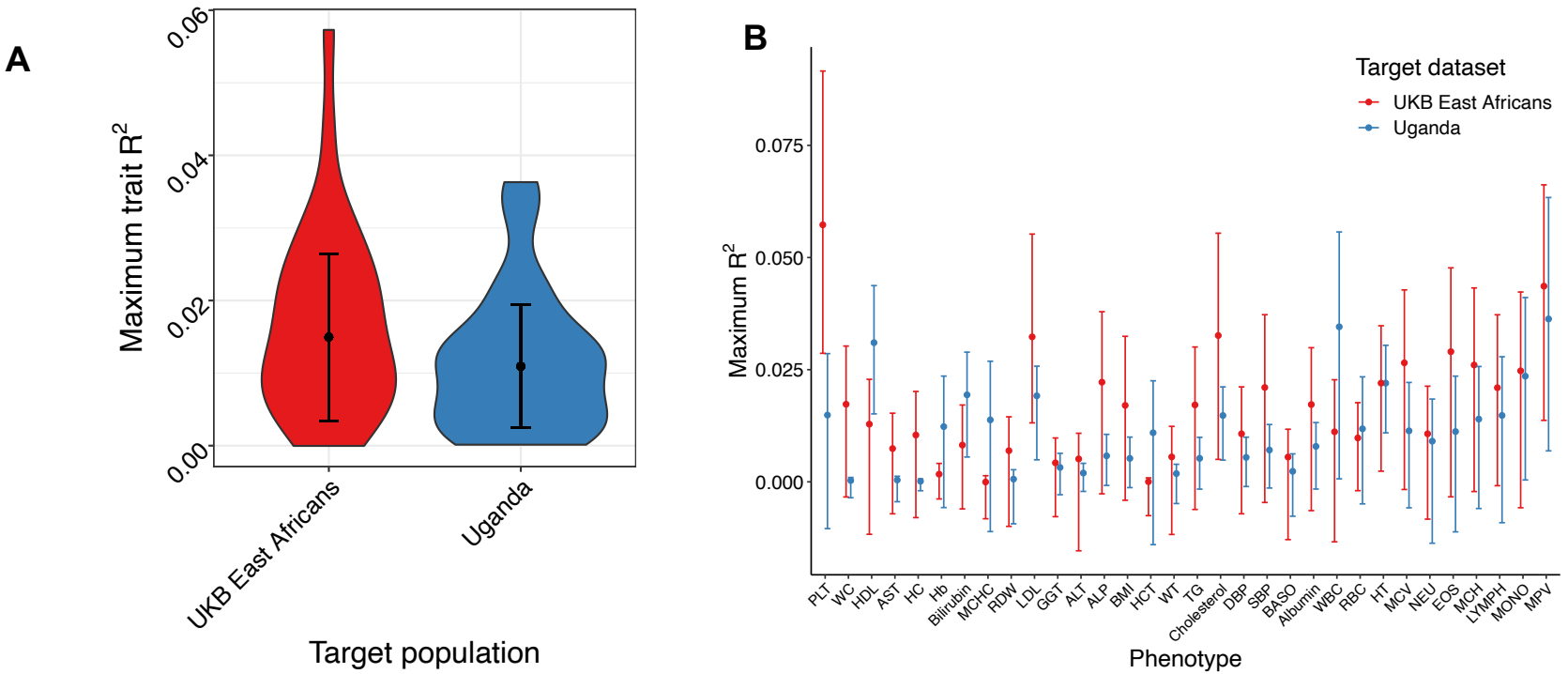


Figure 4.13 - Comparison of PRS accuracy for 32 traits across cohorts in individuals with East AFR ancestry.

Includes unrelated participants in UKB(N = 728 with East AFR ancestry) versus Uganda GPC (N =2,247), using the same summary statistics from UKB EUR ancestry individuals (N = 351,194). A) Summary of accuracy distributions across 32 traits. B) PRS accuracy in each trait. Traits are ordered by absolute difference in prediction accuracy among cohorts despite similar ancestries.

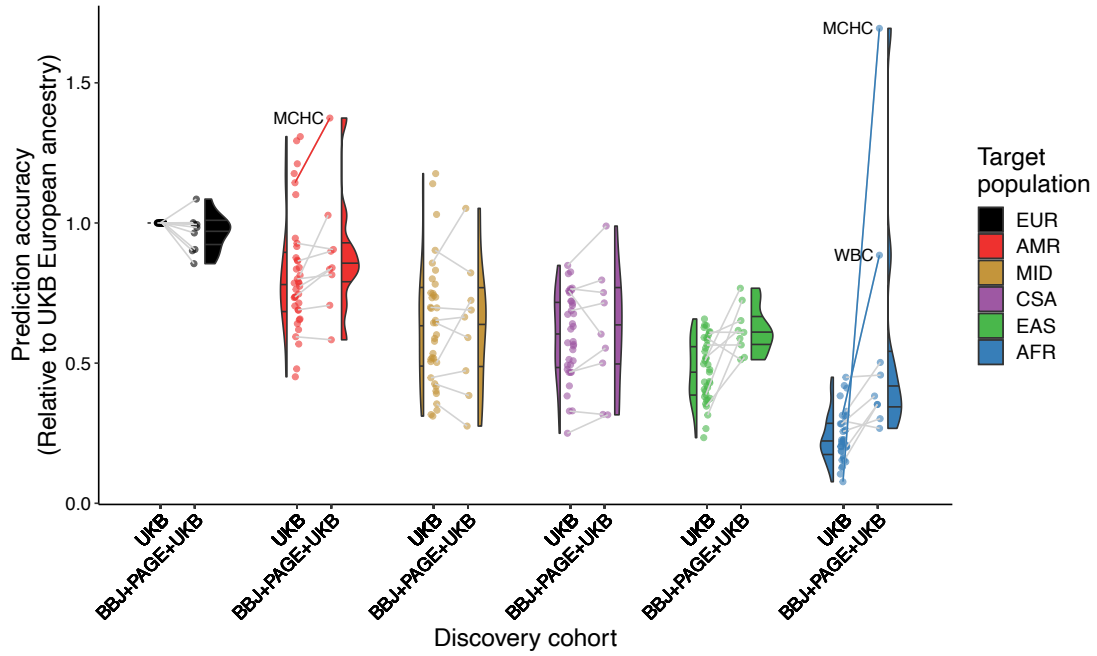


Figure 4.14 - PRS accuracy from a homogeneous versus multi-ancestry discovery dataset. GWAS discovery data consisted of summary statistics from UKB EUR ancestry data only or from the meta-analysis of UKB, BBJ and PAGE. Target populations are from the UKB. Lines connect the 10 traits available in both discovery cohorts to indicate how accuracy changed for the same trait in the UKB only versus meta-analysed discovery data, while half violin plots show the distribution across all phenotypes in each discovery cohort. When lines are missing, the trait is absent in PAGE. Trait outliers are labelled in text and with solid lines. Relative PRS accuracies are compared to the maximum for each trait in target samples withheld from discovery consisting of UKB EUR ancestry individuals. To simplify comparisons, only the PRS with the highest prediction accuracy are shown here.

4.4.3.5 Large-effect population-enriched genetic variants drive heterogeneity in PRS accuracy for blood panel traits

The accuracy improvement from increased diversity in the discovery cohorts varies across traits, and is substantially elevated in MCHC (for AMR and AFR) and WBC (for AFR) cohort (**Figure 4.14**). The specific genetic loci that could explain this pattern were investigated by comparing the significance of genetic associations in UKB alone versus the meta-analysis of UKB, BBJ and PAGE. For MCHC and WBC in particular, the genetic variants contributing to these improved PRS consist of several well-known population-enriched variants (**Figure 4.15**). For example, genetic variants that disproportionately explain population-specific risk for MCHC include variants previously associated with haemoglobin concentration, including rs9399137 upstream of *HBS1L* and *MYB* (p = 5.24e-249 and β = 0.0783 in the meta-analysis) in a study of sickle cell anaemia (Lettre et al., 2008), rs855791 in *TMPRSS6* (*P* = 3.49e-241, β = 0.0692) (Benyamin et al., 2009; Chambers et al., 2009), and rs551118 upstream of

PIEZO1 and *CDT1* (p = 5.18e-100, β = -0.0451) (Astle et al., 2016). Associations with WBC tended to show more population-enriched associations as shown in the meta-analysis (**Figure 4.14**), including rs3936197 in *MED24* (p = 5.18e-289, β = -0.0772), rs58650325 near mast cell receptor *FCER1A* (1.57e-163, β = -0.097), and rs11533993 in *CDK6* (*P* = 1.55e-84, β = -0.0799). Thus, genetic architecture and population genetic considerations are important to bear in mind when considering the generalizability of PRS.

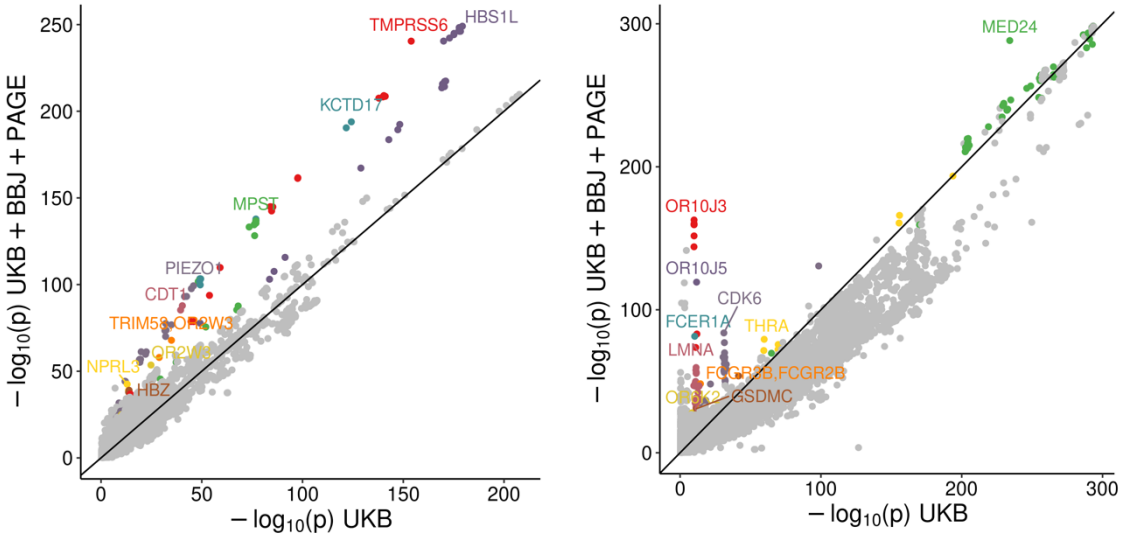


Figure 4.15 - Trait-specific genetic outlier plots
QQ-like plot showing p-values in UKB only versus multi-cohort meta-analysis of UKB, BBJ and PAGE. The ten regions that are the most significant differences between the cohorts are coloured and labelled for MCHC (left), and WBC (right). The colours correspond to the gene names on the plot that are in the same colour.

4.5 Discussion

PRS have been proposed as genetic biomarkers for use in preventative medicine, but have low prediction accuracy across global populations, especially in AFR ancestry populations. In this chapter, the prediction accuracy of PRS was investigated across diverse AFR populations for several quantitative traits, using both simulation and empirical data. The main findings were: i) PRS prediction accuracy is highest when the discovery and target populations are matched by ancestry, iii) EUR-derived PRS accuracy varies across AFR populations iii) inclusion of diverse populations in GWAS improves PRS prediction accuracy and iv) environmental factors impact on the accuracy of PRS. Each of these are discussed below.

PRS prediction accuracy is highest when discovery and target cohorts are matched

The AGVP dataset used in the simulation section which comprised 1,292 individuals was too small a sample size to make conclusive determination about the prediction of PRS. However, the simulations showed that PRS prediction accuracy would be most accurate under two circumstances: if the phenotype was highly heritable and had a sparse genetic architecture. Since PRS have been applied to phenotypes with heritable rates less than 80% (such as BMI which has a heritability rate of 40%) with some success in EUR populations, the findings from the simulation are a clear indication that the simulations were underpowered. Despite this limitation, the East-East and West-West (discovery-target) at the simulated trait with the heritability rate of 0.8 were consistent with previous observations showing that that PRS prediction accuracy is highest when the discovery and target cohorts are matched (Martin et al. 2019).

EUR-derived PRS accuracy varies across AFR populations

Although Africa is the most genetically diverse continent, the degree to which EUR-derived PRS vary across AFT populations has not been investigated before. The analysis done in this chapter is the first to demonstrate that PRS prediction accuracy varies across AFR populations. PRS prediction accuracy was most accurate in Ethiopians and the least accurate in populations with southern AFR ancestry. These results track with genetic distance from EUR and population history. The highest prediction accuracy identified in Ethiopians was expected given closer genetic proximity to EUR populations relative to other AFR populations due to the back to Africa migrations that influenced population structure in Ethiopians (Henn et al., 2012; Hodgson et al., 2014; Pagani et al., 2015). The lowest PRS prediction accuracy in populations with southern AFR ancestry, was consistent with greater genetic divergence from EUR populations and greater overall genetic diversity in this population (Busby et al., 2016; Choudhury et al., 2020; Henn et al., 2011).

PRS prediction accuracy is improved with inclusion of diverse populations in GWAS

The inclusion of GWAS from the PAGE and the Japanese (BBJ) datasets with UKB increased the PRS prediction accuracy for the AMR, EAS and AFR groups reflecting the ancestry of the populations in both BBJ and PAGE. These findings align with those from previous studies showing that ancestry-matched discovery data improves prediction accuracy in the corresponding target population (Bigdeli et al., 2019; Kuchenbaecker et al., 2019; Lam et al., 2019 ; Martin et al., 2019). Altogether, the findings are consistent with the observation that causal genetic effects tend to be similar across populations, except with LD and allele frequency differences that marginally modifying the effect size estimates (Martin et al., 2019). This is also consistent with the genetic correlation between ethnicities being very similar (Brown et al., 2016; Shi et al., 2020).

Environmental factors plays a key role in PRS prediction accuracy

In addition to reduced PRS accuracy with ancestral distance from GWAS cohorts, genetic nurture, social genetic, and environmental effects can also contribute to low portability of PRS across populations (Mostafavi et al., 2020). In this chapter, however, ancestry appears to have a larger effect on portability than cohort differences overall. An important distinction when comparing the magnitude of these and other non-genetic effects in other studies is that the traits most accurately genetically predicted here were primarily anthropometric and blood panel traits. When analysing traits with more sociodemographic influences in increasingly diverse populations, population stratification, confounding, and study design considerations introduce an additional level of complexity (Kerminen et al., 2019; Novembre & Barton, 2018; Zaidi & Mathieson, 2020). PRS accuracy comparisons across ancestrally similar but environmentally diverse populations are especially important for medically actionable traits. For example, particularly low PRS portability for TG from EUR to the Uganda GPC data and effect size heterogeneity has previously been connected to pleiotropic and gene * environment effects; specifically, most non-transferable genome-wide significant associations with TG showing pleiotropic associations with BMI in EUR (Kuchenbaecker et al., 2019).

Relevant to the traits studied in the genetic analyses in this chapter, haematological differences such as anaemia are more common in lower income countries in Africa and in AFR ancestry populations elsewhere, compared to higher income countries with primarily EUR ancestry populations. This is potentially due, in part, to genetic variation as well as the higher prevalence of infectious diseases and pathogens, poorer nutritional status, and altitude,

coinciding with the geographical distribution of AFR populations (Mugisha et al., 2013; Mugisha et al., 2016).

4.6 Conclusion

This chapter has emphasized that there is low generalizability of EUR-derived PRS across AFR populations due to high genetic and environmental diversity, and accuracy varies by proximity to EUR ancestry. The most rapid path to closing gaps in PRS transferability is to increase the inclusion of GWAS participants from populations most divergent from those already routinely studied. As empirically demonstrated here, when comparing PRS accuracy calculated from ethnically diverse cohort meta-analysis versus data from EUR only, large-scale GWAS with diverse AFR populations will most rapidly reduce portability gaps across global populations.

Chapter 5: General Discussion

Large-scale genomics studies have largely focussed on populations of EUR ancestry. This bias impacts the discovery of disease-associated and population-relevant GWAS variants; as well as the prediction accuracy of PRS in non-EUR ancestry populations. In this thesis, the genetics of schizophrenia in the South African Xhosa population was investigated, and expanded to include the extent to which EUR-ancestry derived PRS generalize among diverse African populations. An outline illustrating the approach used to analyse SAX data and the three components of the PRS aspect of the thesis is shown in **Figure 5.1A and B** below.

5.1 Summary of findings and their implications

Following a general introduction covering the epidemiological background, genetic and environmental risk factors for schizophrenia in chapter 1, chapter 2 provided insight into the genetic basis of schizophrenia in the South African Xhosa population. The intronic variants in the genomic region 17p13.2, and particularly in the gene *ZFP3* that were associated with the phenotype, despite not having previously been implicated in any GWAS of schizophrenia, suggested that *ZFP3* may be important in the aetiology of schizophrenia. This was substantiated by findings from the study by Gulsuner et al (2020), in the same cohort, where variants in the same *ZFP3* locus were implicated in a WES analysis. The convergent findings using complementary data sources, i.e. GWAS array and WES data, albeit in the same cohort suggests that *ZFP3* may be important in the aetiology of schizophrenia, at least in the SAX population, and deserve further exploration in larger cohorts.

Most genetic studies do not phenotype study participants as deeply as was done for SAX. The SAX study was designed to acquire as much information as possible about known environmental exposures including those reviewed in chapter 1 that may contribute to individuals developing schizophrenia, thus provided a unique opportunity to explore how these environmental factors impact the risk of schizophrenia. The scope of this thesis allowed for analyses investigation into the independent impact of sex and childhood trauma on schizophrenia risk.

The intensified GWAS signal in *ZFP3* after controlling for childhood trauma, as well as the positive correlation (r_g = 0.19, se = 0.066) between schizophrenia and childhood trauma suggested an interaction between the traits, i.e. along a ‘causal’ path from trauma to schizophrenia. This interaction warrants further investigation. Given the expansive literature

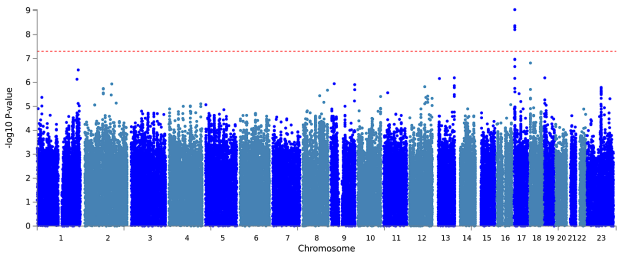
on the epidemiological link between schizophrenia and childhood trauma (as reviewed in chapter 1), it is important that future studies, where possible, collect participants' information about their childhood trauma experience, and possibly other exposures, and explore these data in the context of GWAS.

ZFP3 is ubiquitously expressed in human tissues, regulates the transcription of RNA polymerase II and may be important in regulating gene expression in the brain. Two specific findings support the gene expression regulatory role of variants in *ZFP3*. First, the variants that met the genome-wide significance threshold were intronic, and secondly, overall heritability was enriched in functional categories that are involved in the regulation of gene expression. Although the association of *ZFP3* is unique to the South African population, the plausible biological mechanism is likely that manifesting through the regulation of gene expression, and is consistent with what is known about the role of variants associated with schizophrenia in large-scale studies.

When conducting a sex-stratified GWAS, the findings from the male-only GWAS expectantly resembled those from the overall GWAS, as the sample was predominantly male. The results from the female GWAS provided little insight into how the genetic of schizophrenia might differ between males and females in SAX. While the largest and most recent GWAS study found no heterogeneity between the variants associated with schizophrenia in males and females of EUR ancestry (Ripke et al, 2020), the question of why there were higher rates of institutionalized males compared to females in the South African sample (about 4:1), specifically in the Western and Eastern Cape provinces from where the SAX samples were ascertained, still remains. This rate is 1.4:1 (male to female) in predominately EUR ancestry populations. Several hypotheses have been proposed in chapter 2 including that sociological (e.g. specific roles males and females play in their society that may be tied to cultural norms), or pathological factors (e.g. presentation and severity of disease) may influence institutionalization which warrant further investigation in larger female samples.

Schizophrenia in SAX

GWAS

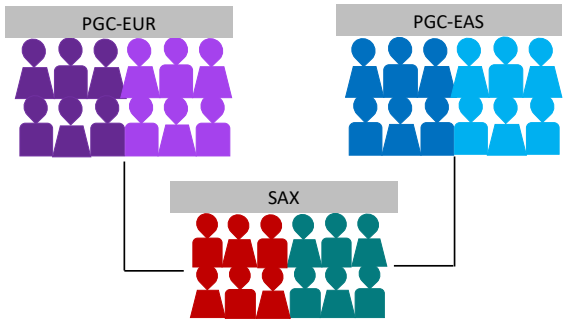


Standard GWAS

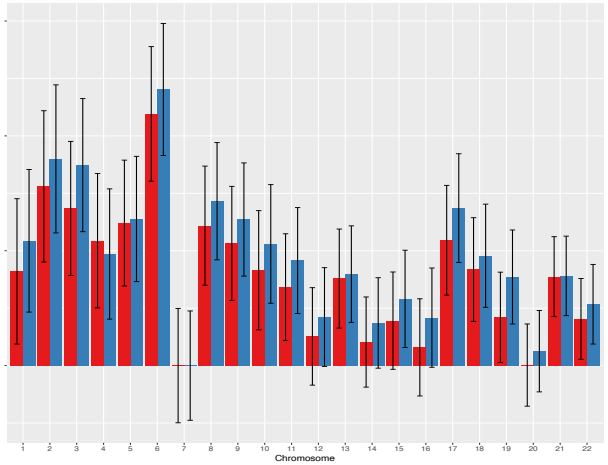
GWAS by sex

GWAS controlling for childhood trauma

PRS



Heritability enrichment



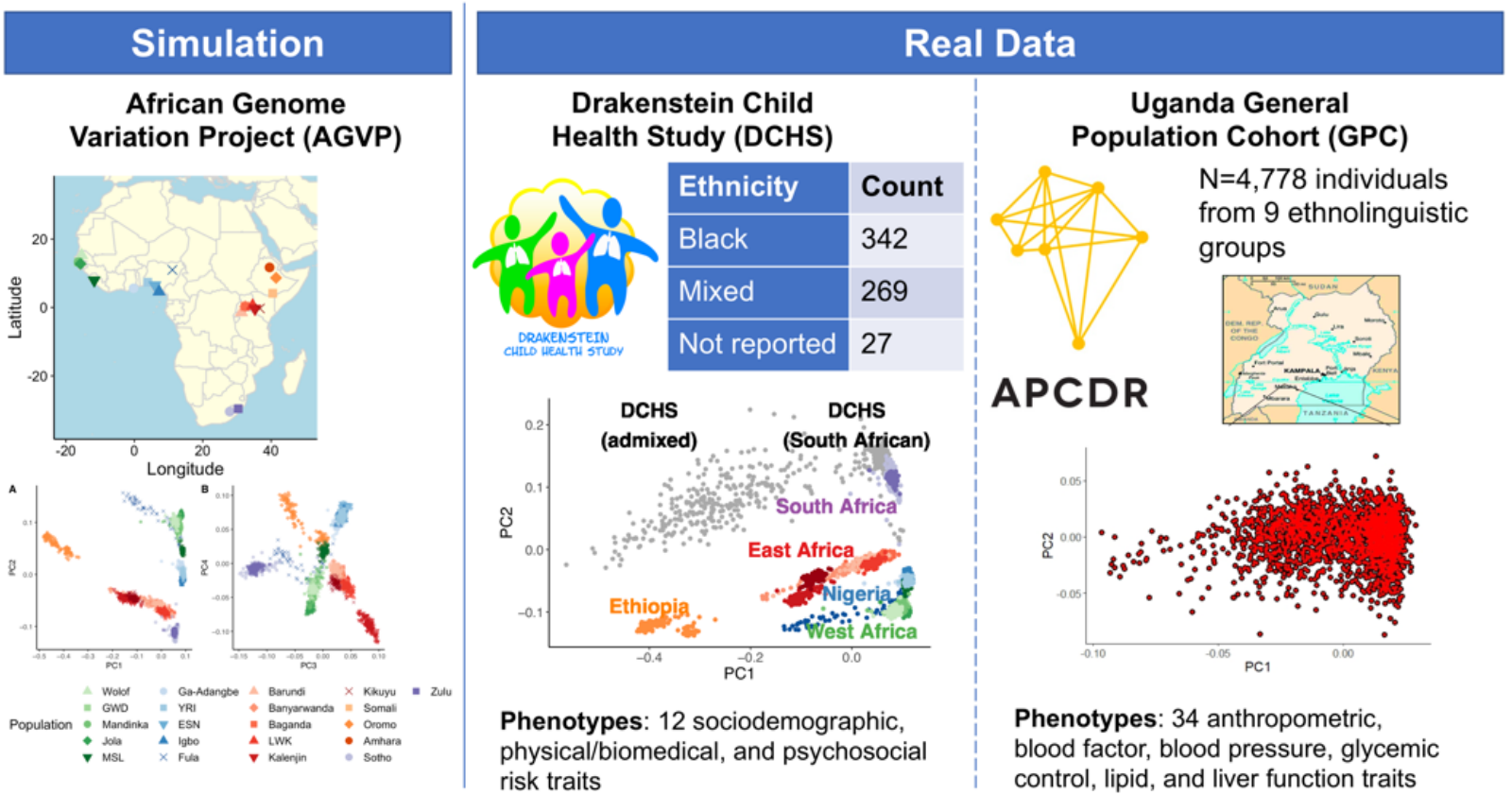


Figure 5.1 - A graphical summary of the analysis done in this thesis.
Top panel shows the analysis done for the first aim: GWAS investigation, followed by PRS and enrichment of heritability analyses. The bottom panel indicates the datasets used for the second aim: investigating the transferability of PRS across AFR populations.

In chapter 3, results from the GWAS in chapter 2 were used in two post-GWAS analyses. The first analysis investigated the enrichment of the distribution of the heritability of schizophrenia across numerous categories including chromosomes, minor allele frequencies and several functional categories. The second assessed how well PRS derived from the earlier large-scale GWAS, notable PGC studies, predict schizophrenia in SAX. From the heritability enrichment analysis, the most notable finding was that heritability was enriched in functional categories that are involved in the regulation of gene expression, consistent with findings from other studies. Regarding the transferability of PRS, the analysis showed that PRS derived from EUR and EAS populations from the PGC poorly predicted schizophrenia in SAX, consistent with the decay of PRS with increasing genetic distance from the between cohorts.

In chapter 4, the PRS work initiated in chapter 3 was expanded upon by assessing the generalizability of PRS across diverse AFR populations for non-psychiatric phenotypes using both simulated and empirical data from among the largest genetic datasets that exist of AFR ancestry individuals. For the simulations, data from the AGVP was used to assess the transferability of PRS across cohorts from different regions of Africa. The findings showed that PRS was most accurate for the most heritable traits (heritability rate = 0.8) with the sparsest genetic architecture (i.e. 5 causal variants). Further, PRS was most accurate when the discovery and target cohorts were from the same ancestral backgrounds.

Empirical data showed that prediction accuracy of PRS derived from EUR ancestry was low in DCHS for phenotypes including height, alcohol consumption, depression, psychological distress and smoking. Height was the only phenotype with a prediction accuracy higher than zero, and the accuracy was higher in mixed ancestry individuals than the Black Africans. Among the mixed ancestry individuals, prediction accuracy was the higher for those with a larger proportion of EUR ancestry than those with the low proportion.

Another empirical dataset used was the Ugandan GPC. EUR ancestry PRS from the UKB were used to predict 34 quantitative traits in GPC. Prediction accuracy was low for all traits. Further, PRS prediction accuracy was assessed for the AFR populations within UKB compared to GPC. The accuracy was highest for the Admixed populations and lowest for the southern African populations. Additionally, the prediction accuracy was lower in GPC compared to the East African individuals in the UKB cohort, which could be attributed to the differences in environment. The differences between AFR ancestries contributed to more variability in PRS than the differences between cohorts.

Lastly, the inclusion of diverse populations from PAGE and BBJ, the PRS prediction accuracy was improved particularly for the AMR, EAS and AFR populations, reflecting the ancestral backgrounds of the individuals in both PAGE and BBJ. There was disproportionate increase in PRS prediction accuracy for mean corpuscular haemoglobin concentration (MCHC) and white blood cell count (WBC), compared to all the other phenotypes. This increase was driven by the enrichment of population-specific variants relating to disorders such as sickle cell anaemia and ß-thalassemia; and allergies for WBC.

PRS have been proposed as genetic biomarkers for use in preventative medicine, but the low prediction accuracy across populations and particularly in AFR ancestry populations limit their utility (Martin et al., 2019; Sirugo et al., 2019). This study has enabled unique insights into PRS transferability within and among diverse continental African populations as well as among similar ancestry populations living in considerably different environments. It also demonstrated the looming challenges for applying current PRS in AFR ancestry populations; because relatively few genetic studies have been conducted in AFR populations coupled with uniquely deep population histories in Africa, differences in PRS accuracy across diverse AFR ancestries from varied regions can be larger than across out of Africa continents.

This is particularly problematic as widely-used algorithms that guide health decisions already have ingrained racial biases (Obermeyer et al., 2019), warning of compounding challenges with implementation. Yet, that the inclusion of ancestrally diverse populations in GWAS discovery cohorts has a positive impact that is larger in effect than a similarly sized EUR ancestry cohorts for improving PRS accuracy in underrepresented populations; this highlights the unique opportunities presented by diverse African populations to disproportionately improve PRS accuracy globally.

5.2 Study limitations

The most notable limitation of this work was that the GWAS analysis was limited by sample size. Much larger sample sizes are required to improve statistical power to identify more genome-wide significant variants and discover population specific variants, as well as variants that are known to be associated with schizophrenia across ancestries. Collaborative participation in global consortia such as the Psychiatric Genomics Consortium, who have collected tens of thousands of schizophrenia samples largely in EUR and EAS populations not only overcomes the sample size constraints but also increases diversity of studied populations, inadvertently leading to the discovery of clinically relevant schizophrenia loci.

Notwithstanding these prospects, several limitations of GWAS as a research methodology have been previously discussed by McClellan and King (McClellan & King, 2010). These limitations include: cryptic population stratification is not adequately controlled for in GWAS and has the potential to lead to spurious association; the mapping of associated variants to a gene does not mean the gene itself is biologically relevant to the disorder, the principle on which GWAS is based on, i.e. the associated variants being in LD with causal SNPs only works if there is no genetic heterogeneity between cases from different study sites.

Schizophrenia is phenomenologically diagnosed based on constantly evolving criteria. The historical basis, or foundation, of these criteria are reviewed in section **Error! Reference source not found.** of chapter 1. A broad criticism of the previous and current versions of the DSM is that it is ethnocentric, despite the efforts made by the American Psychiatric Association to be more 'culturally sensitive' in the DSM-5 (Bredström, 2019). Culture is an important consideration, as symptoms may vary across people of diverse cultural background. Additionally, it influences the understanding of disorders, and thus help-seeking behaviours, which may lead to misdiagnosis and under-treatment, respectively. For example, 'amafufunyana' and 'ukuthwasa' are terms used in the Xhosa language by traditional healers to describe schizophrenia symptoms. While 'amafufunyana' has a negative connotation as it is often believed that the affected individual was bewitched by someone jealous of them, 'ukuthwasa' refers to a divine calling by the ancestors (Campbell et al., 2017; Niehaus et al., 2004)

Another limitation of this study was that childhood trauma was retrospectively reported; this introduces recall bias — a subjective interpretation of event and the type of event (Bower, 1981; Kessler et al., 1995). Current mental health is known to influence retrospective recall of trauma, i.e. people with more severe symptoms at the time of interview are more likely to report a higher frequency and more severe trauma than those with minimal symptoms (Roemer et al., 1998; Thompson-Hollands et al., 2020).

5.3 Future considerations

An important aim of genetic research is to inform the develop of effect prevention and/or treatment therapies. However, the path from GWAS associated variants to drug discovery is complex, because the associated variants are not directly informative of the variants that cause the disease or of the mechanism by which the disease is caused. Additionally, because complex traits are polygenic, testing the functional effects of each associated SNP to determine the biology by which the disorder would be a time-consuming and costly exercise. Below, Type II diabetes (T2D) is used

as a case to demonstrate how GWAS has been applied to first identify disease-associated loci and use the information known about the biological consequence of these variants to discover new drug targets.

GWAS has identified over 100 loci that are associated with T2D (Fuchsberger et al., 2016; Morris et al., 2012). These associations have been replicated in populations of diverse ethnicity (DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium, Asian Genetic Epidemiology Network Type 2 Diabetes (AGEN-T2D) Consortium, South Asian Type 2 Diabetes (SAT2D) Consortium, Mexican American Type 2 Diabetes (MAT2D) Consortium, 2014). The more diverse samples are being genotyped, the more ethnicity-specific alleles are being found. Post-GWAS studies identified that GWAS associated loci were enrichment in genomic elements that regulate gene expression, i.e. enhancers found in pancreatic islets (Parker et al., 2013; Pasquali et al., 2014). Further expression mapping identified genes that are associated with insulin resistance and hyperlipidemia (Small et al., 2011). The integration of GWAS and candidate gene sequence data have re-assigned the GWAS signal that was initially thought to be in non-coding variants to coding regions and identified LoF function variants that impacted the secretion of insulin. LoF mutations identified in *SLC30A8* were found to be protective against T2D; this led to the development of the ZnT-8 antagonists drug which binds to a zinc transport expressed in pancreatic islet (Flannick et al., 2014).

No drugs have been developed for schizophrenia since the first antipsychotic medication from several decades ago. Part of the complexity of identifying drug targets for schizophrenia is that there are only four pathways that are hypothesized to lead to the phenotype which include dopamine, glutamate, immune modulation, calcium signalling, and nicotinic cholinergic ((Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014); many others may be relevant. Another possibility, is that the Euro-centric bias in psychiatric genomics research has limited the discovery of potentially informative disease-associated loci from diverse populations, and inadvertently the potential to discover drug targets.

The largest GWAS of schizophrenia comprises participants of EUR and EAS-ancestry, and finds schizophrenia variants that are common between the two populations. While it is expected that most risk variants will be shared across populations of different ancestry, it is still likely to find population-specific variants, as was the case for the *ZFP3* variants found in the SAX cohort, but not elsewhere. Even though the SAX cohort used in this thesis is a small fraction of the largest schizophrenia GWAS, the finding of genome-wide significant variants at the *ZFP3* locus indicate the value of studies in genetically diverse populations. It is likely that many more variants that yet

to be discovered in this and other understudied populations, that could yield insight into the aetiology of schizophrenia.

As reviewed in chapter 1, the inclusion of diverse samples in large scale studies has the following benefits: i) improved variant discovery ii) accounting for more genetic variation, more accurate PRS scores — genetically informed diagnoses iii) fine-mapping for the resolution of disease-associated loci to identify causal variants against which drugs can be developed, iv) more accurate and equitable PRS. Major efforts are underway such as the NeuroGAP Psychosis project, which aims to recruit 35,000 schizophrenia and bipolar disorder patients from Kenya, Uganda, Ethiopia and South Africa (Stevenson et al., 2019). Similarly, the H3Africa Consortium comprising investigators from across the African continent aims to improve genomics research on a broad range of diseases and disorders (H3Africa Consortium, 2014). International efforts such as All of US (Hudson et al., 2015) and PAGE (Wojcik et al., 2019), and many others mentioned by Bentley et al. (2020) are especially promising programs for rectifying missed scientific opportunities by increasing inclusion of diverse AFR participants. Because the focus of genetic studied had been EUR populations for a long time, the inclusion of diverse populations in such studies come with challenges, primarily methodological, that have been documented (Peterson et al., 2019) and discussed here.

Previously, it was standard practice to exclude individuals who are ‘population outliers’ during the QC steps that precede association test, in trying to avoid false negative signal that may result due to population stratification. Current mixed model regression models exist to allow for the joint analysis of samples from genetically diverse backgrounds, including individuals with three- or four-way admixture, whilst accounting for populations stratification. While the joint analysis is appealing because the larger sample size is maintained, it has been shown that this model does not adequately account for populations stratification, especially when there are environmental factors that are correlated with ancestry (Conomos et al., 2018; Heckerman et al., 2016; Zhang & Pan, 2015) An alternative strategy entails the stratification of individuals into homogenous ancestral groups based of PCs, followed by QC and association for each group, and then ultimately combining the summary statistics in a meta-analysis (Hellwege et al., 2017). Admixed populations are usually assigned into a group of their own, depending on their admixture background. This assignment is however not precise and leaves room for improvement (Medina-Gomez et al., 2015).

Most GWAS arrays were designed for EUR populations -- providing better coverage for variants that are common in EUR populations and not necessarily common in non-EUR populations. For example, the Affymetrix UKB array cover 80% of variants with MAF > 1% in EUR, compared to only 46% in AFR populations (Nelson et al., 2017). Many more variants are required to provide the same coverage that has been achieved for EUR populations (Barrett & Cardon, 2006). Population-

specific GWAS arrays have been designed to overcome this issue, such as the Multi-Ethnic Global Array (MEGA) (Wojcik et al., 2018), H3Africa array (Mulder et al., 2018), and the SAXv2 Affymetrix array used in chapter 2. An alternative way around the cover issue is low coverage WGS or WES. Recently, Martin et al. (2020) showed that low coverage sequencing captured genetic variants across all frequencies than GWAS arrays, at a cost comparable to that of the GWAS arrays. Low coverage should be highly considered for future studies.

Genotype imputation is an integral part of genetic studies, as variants that are not genotyped on the GWAS array or sequenced can be inferred from a reference panel. The accuracy of imputation depends on the LD reference panel used; the highest accuracy is achieved when the reference panel used has a high coverage of the variants in the genotyped or sequenced sample (Ahmad et al., 2017). Few of the existing reference panels have individuals of AFR ancestry, and far fewer Africans who live in Africa. For example, 90% of AFR ancestry individuals in the Genome Aggregation Database (gnomAD, https://gnomad.broadinstitute.org), the largest reference panel that exists, are African American or Afro-Caribbean (Martin et al., 2018). A concerted effort is being made to increase the diversity of reference panels, through projects such as the Human Variation Genome Project (HGDP, http://www.hagsc.org/hgdp/).

Cross-ancestry analyses require consistency in how the phenotype is measure across study sites. Pooling samples together is often done under the assumption that there is no heterogeneity between populations for the phenotype under investigation, and that there are no cultural biases in how phenotypes are measured. However, these assumptions are not always true, and variability in the phenotype measures may limit gene discovery and the transferability of findings between populations. Most diagnostic instruments used in psychiatry were developed and validated in high income countries (Henrich et al., 2010), but cultural differences have been noted. For example, while there is shared disease construct for major depression, cultural differences are a predictor of the level symptoms people wait to experience before seeking help (Kendler et al., 2015; Simon et al., 2002). There is a need for validation of phenotype measurement instruments in low- and middle-income countries (Mwesiga et al., 2020).

There is a wealth of data on the effect of environmental exposures on complex traits, similar to the contribution of childhood trauma to schizophrenia risk in SAX shown in this thesis, but genetic studies have largely not accounted for these. It has long been realized, and becoming increasingly evident that modelling of environmental exposures is crucial for understanding the biology of complex disorders (Dempfle et al., 2008). Additionally, studies in diverse samples may help increase the understanding of how gene-by-environment differ across social and cultural groups.

PRS are important for predicting future disease or identifying 'at-risk' populations, especially for disorders that have no valid biomarkers and are phenomenologically diagnosed, such as schizophrenia. It was demonstrated in chapter 4 of this thesis and in other studies that PRS are improved by inclusion of diverse populations in large genetics studies (Martin et al., 2019; Bigdeli et al., 2019), and also that the prediction accuracy is greatly impacted by population-enriched variants. Including such variants in future may be greatly beneficial. For example, variants enriched in *APOL1* and *G6PD* contribute especially to a high risk of chronic kidney disease and to a missed diagnosis of diabetes in AFR populations (Rotimi et al., 2017). Also, the inclusion of population-enriched variants in PRS could help counter genetic justifications for race-based medicine, which problematically reinforce implicit racial biases by overemphasizing the link between genetics and race despite the fact that there is more genetic variation within than between populations (Cerdeña et al., 2020). However, it is imperative to dispel the myth of the social 'race' construct, and to emphasise the value of ancestral/geographic origins of populations in genetic studies.

Beyond expanding on diversity by increasing the number of study participants in large-scale studies and resolving the associated methodological challenges, it is equally important to diversify researchers working on genomics studies. Currently, the vast majority of researchers in genomics studies are of EUR ancestry (Ginther et al., 2011; Hoppe et al., 2019), in line with the over-representation of EUR ancestry individuals in genomic studies. The exclusion of African researchers contributes to the disparity in research leadership and reduced scientific output from African researchers (Bentley et al., 2020). Efforts such as the NeuroGAP Global Initiative for Neuropsychiatric Genetics Education and Research (GINGER) programme (van der Merwe et al., 2018), which provides mentorship and training for early-career investigators on the African continent (particularly in Uganda, Kenya, Ethiopia and South Africa, including several of the latter publication's authors), are important in moving toward a more inclusive and representative research community.

5.4 Conclusion

This thesis has provided new insight into the genetic aetiology of schizophrenia in the South African Xhosa population, highlighting variants in the *ZFP3* gene to be of importance. It has also demonstrated that the biology of schizophrenia is shared between populations. With regards to the prediction accuracy of PRS across Africa, it was shown for the first time here that accuracy varies by African populations, and that environmental factors have a great impact on prediction accuracy. To expand on this foundation, future large-scale studies would need to include diverse samples to discovery more variants and equitably improve the prediction accuracy of PRS.

Bibliography

Aas, M., Haukvik, U. K., Djurovic, S., Bergmann, Ø., Athanasiu, L., Tesli, M. S., Hellvin, T., Steen, N. E., Agartz, I., Lorentzen, S., Sundet, K., Andreassen, O. A., & Melle, I. (2013, Oct 1). BDNF val66met modulates the association between childhood trauma, cognitive and brain abnormalities in psychoses. *Prog Neuropsychopharmacol Biol Psychiatry, 46*, 181-188. https://doi.org/10.1016/j.pnpbp.2013.07.008

Aas, M., Haukvik, U. K., Djurovic, S., Tesli, M., Athanasiu, L., Bjella, T., Hansson, L., Cattaneo, A., Agartz, I., & Andreassen, O. A. (2014). Interplay between childhood trauma and BDNF val66met variants on blood BDNF mRNA levels and on hippocampus subfields volumes in schizophrenia spectrum and bipolar disorders. *J Psychiatr Res, 59*, 14-21. https://www.sciencedirect.com/science/article/abs/pii/S0022395614002477?via%3Dihub

Abecasis, G. R., Altshuler, D., Auton, A., Brooks, L. D., Durbin, R. M., Gibbs, R. A., Hurles, M. E., & McVean, G. A. (2010, Oct 28). A map of human genome variation from population-scale sequencing. *Nature, 467*(7319), 1061-1073. https://doi.org/10.1038/nature09534

Abecasis, G. R., Auton, A., Brooks, L. D., DePristo, M. A., Durbin, R. M., Handsaker, R. E., Kang, H. M., Marth, G. T., & McVean, G. A. (2012, Nov 1). An integrated map of genetic variation from 1,092 human genomes. *Nature, 491*(7422), 56-65. https://doi.org/10.1038/nature11632

Abecasis, G. R., Burt, R. A., Hall, D., Bochum, S., Doheny, K. F., Lundy, S. L., Torrington, M., Roos, J. L., Gogos, J. A., & Karayiorgou, M. (2004, 01/28/08/11/received 11/20/accepted). Genomewide Scan in Families with Schizophrenia from the Founder Population of Afrikaners Reveals Evidence for Linkage and Uniparental Disomy on Chromosome 1. *American Journal of Human Genetics, 74*(3), 403-417. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1182255/http://ac.els-cdn.com/S000292970761859X/1-s2.0-S000292970761859X-main.pdf?_tid=99f7a5fc-e175-11e6-acd2-00000aab0f26&acdnat=1485180769_ff17cc5051be2fc04ce5855626666e15

Abel, K. M., Wicks, S., Susser, E. S., Dalman, C., Pedersen, M. G., Mortensen, P. B., & Webb, R. T. (2010). Birth weight, schizophrenia, and adult mental disorder: is risk confined to the smallest babies? *Archives of General Psychiatry, 67*(9), 923-930. https://jamanetwork.com/journals/jamapsychiatry/articlepdf/210877/yoa05020_923_930.pdf

Ahmad, M., Sinha, A., Ghosh, S., Kumar, V., Davila, S., Yajnik, C. S., & Chandak, G. R. (2017). Inclusion of population-specific reference panel from India to the 1000 genomes phase 3 panel improves imputation accuracy. *Sci Rep, 7*(1), 1-8.

Alexander, D. H., & Lange, K. (2011). Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinformatics, 12*(1), 246.

Alfimova, M. V., Kondratyev, N. V., Golov, A. K., Golubev, S. A., Galaktionova, D. Y., Nasedkina, T. V., & Golimbet, V. E. (2019, 2019/10/01). Relationship Between the rs7341475 Polymorphism and DNA Methylation in the Reelin Gene and Schizophrenia Symptoms. *Neuroscience and Behavioral Physiology, 49*(8), 1061-1066. https://doi.org/10.1007/s11055-019-00838-5

American Psychiatric Association. (1952). *Diagnostic and statistical manual mental disorders*.

American Psychiatric Association. (1968). *Diagnostic and statistical manual of mental disorders* (2d ed. ed.).

American Psychiatric Association. (1980). *Diagnostic and statistical manual of mental disorders* (3rd ed. ed.).

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders : DSM-IV-TR* (4th ed., text revision 2000. ed.).

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders : DSM-5* (Fifth edition. ed.).

Anderson, K. K., Cheng, J., Susser, E., McKenzie, K. J., & Kurdyak, P. (2015). Incidence of psychotic disorders among first-generation immigrants and refugees in Ontario. *Cmaj, 187*(9), E279-E286. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4467956/pdf/187e279.pdf

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T., Ntini, E., Arner, E., Valen, E., Li, K., Schwarzfischer, L., Glatz, D., Raithel, J., Lilje, B., Rapin, N., Bagger, F. O., Jørgensen, M., Andersen, P. R., Bertin, N., Rackham, O., Burroughs, A. M., Baillie, J. K., Ishizu, Y., Shimizu, Y., Furuhata, E., Maeda, S., Negishi, Y., Mungall, C. J., Meehan, T. F., Lassmann, T., Itoh, M., Kawaji, H., Kondo, N., Kawai, J., Lennartsson, A., Daub, C. O., Heutink, P., Hume, D. A., Jensen, T. H., Suzuki, H., Hayashizaki, Y., Müller, F., Forrest, A. R. R., Carninci, P., Rehli, M., & Sandelin, A. (2014, Mar 27). An atlas of active enhancers across human cell types and tissues. *Nature, 507*(7493), 455-461. https://doi.org/10.1038/nature12787

Andreasen, N. C. (1982). Negative symptoms in schizophrenia: Definition and reliability. *Archives of General Psychiatry, 39*(7), 784-788. https://doi.org/10.1001/archpsyc.1982.04290070020005

Andreasen, N. C., Arndt, S., Alliger, R., Miller, D., & Flaum, M. (1995). Symptoms of schizophrenia: Methods, meanings, and mechanisms. *Archives of General Psychiatry, 52*(5), 341-351. https://doi.org/10.1001/archpsyc.1995.03950170015003

Angermeyer, M. C., Carta, M. G., Matschinger, H., Millier, A., Refaï, T., Schomerus, G., & Toumi, M. (2015). Cultural differences in stigma surrounding schizophrenia: comparison between Central Europe and North Africa. *The British Journal of Psychiatry*, bjp. bp. 114.154260.

Arimondo, P. B., Barberousse, A., & Pontarotti, G. (2019, Jun). The Many Faces of EpigeneticsOxford, December 2017. *Epigenetics, 14*(6), 623-631. https://doi.org/10.1080/15592294.2019.1595298

Arnsten, A. F. (2009). Stress signalling pathways that impair prefrontal cortex structure and function. *Nature reviews neuroscience, 10*(6), 410-422.

Arseneault, L., Cannon, M., Poulton, R., Murray, R., Caspi, A., & Moffitt, T. E. (2002). Cannabis use in adolescence and risk for adult psychosis: longitudinal prospective study. *Bmj, 325*(7374), 1212-1213. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC135493/pdf/1212.pdf

Asiki, G., Murphy, G., Nakiyingi-Miiro, J., Seeley, J., Nsubuga, R. N., Karabarinde, A., Waswa, L., Biraro, S., Kasamba, I., & Pomilla, C. (2013). The general population cohort in rural south-western Uganda: a platform for communicable and non-communicable disease studies. *International journal of epidemiology, 42*(1), 129-141.

Astle, W. J., Elding, H., Jiang, T., Allen, D., Ruklisa, D., Mann, A. L., Mead, D., Bouman, H., Riveros-Mckay, F., & Kostadima, M. A. (2016). The allelic landscape of human blood cell trait variation and links to common complex disease. *Cell, 167*(5), 1415-1429. e1419.

Attademo, L., Bernardini, F., Garinella, R., & Compton, M. T. (2017). Environmental pollution and risk of psychotic disorders: A review of the science to date. *Schizophrenia Research, 181*, 55-59. https://www.sciencedirect.com/science/article/abs/pii/S0920996416304467?via%3Dihub

Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., Donnelly, P., Eichler, E. E., Flicek, P., Gabriel, S. B., Gibbs, R. A., Green, E. D., Hurles, M. E., Knoppers, B. M., Korbel, J. O., Lander, E. S., Lee, C., Lehrach, H., Mardis, E. R., Marth, G. T., McVean, G. A., Nickerson, D. A., Schmidt, J. P., Sherry, S. T., Wang, J., Wilson, R. K., Gibbs, R. A., Boerwinkle, E., Doddapaneni, H., Han, Y., Korchina, V., Kovar, C., Lee, S., Muzny, D., Reid, J. G., Zhu, Y., Wang, J., Chang, Y., Feng, Q., Fang, X., Guo, X., Jian, M., Jiang, H., Jin, X., Lan, T., Li, G., Li, J., Li, Y., Liu, S., Liu, X., Lu, Y., Ma, X., Tang, M., Wang, B., Wang, G., Wu, H., Wu, R., Xu, X., Yin, Y., Zhang, D., Zhang, W., Zhao, J., Zhao, M., Zheng, X., Lander, E. S., Altshuler, D. M., Gabriel, S. B., Gupta, N., Gharani, N., Toji, L. H., Gerry, N. P., Resch, A. M., Flicek, P., Barker, J., Clarke, L., Gil, L., Hunt, S. E., Kelman, G., Kulesha, E., Leinonen, R., McLaren, W. M., Radhakrishnan, R., Roa, A., Smirnov, D., Smith, R. E., Streeter, I., Thormann, A., Toneva, I., Vaughan, B., Zheng-Bradley, X., Bentley, D. R., Grocock, R., Humphray, S., James, T., Kingsbury, Z., Lehrach, H., Sudbrak, R., Albrecht, M. W., Amstislavskiy, V. S., Borodina, T. A., Lienhard, M., Mertes, F., Sultan, M., Timmermann, B., Yaspo, M.-L., Mardis, E. R., Wilson, R. K., Fulton, L., Fulton, R., Sherry, S. T., Ananiev, V., Belaia, Z., Beloslyudtsev, D., Bouk, N., Chen, C., Church, D., Cohen, R., Cook, C., Garner, J., Hefferon, T., Kimelman, M., Liu, C., Lopez, J., Meric, P., O'Sullivan, C., Ostapchuk, Y., Phan, L., Ponomarov, S., Schneider, V., Shekhtman, E., Sirotkin, K., Slotta, D., Zhang, H., McVean, G. A., Durbin, R. M., Balasubramaniam, S., Burton, J., Danecek, P., Keane, T. M., Kolb-Kokocinski, A., McCarthy, S., Stalker, J., Quail, M., Schmidt, J. P., Davies, C. J., Gollub, J., Webster, T., Wong, B., Zhan, Y., Auton, A., Campbell, C. L., Kong, Y., Marcketta, A., Gibbs, R. A., Yu, F., Antunes, L., Bainbridge, M., Muzny, D., Sabo, A., Huang, Z., Wang, J., Coin, L. J. M., Fang, L., Guo, X., Jin, X., Li, G., Li, Q., Li, Y., Li, Z., Lin, H., Liu, B., Luo, R., Shao, H., Xie, Y., Ye, C., Yu, C., Zhang, F., Zheng, H., Zhu, H., Alkan, C., Dal, E., Kahveci, F., Marth, G. T., Garrison, E. P., Kural, D., Lee, W.-P., Fung Leong, W., Stromberg, M., Ward, A. N., Wu, J., Zhang, M., Daly, M. J., DePristo, M. A., Handsaker, R. E., Altshuler, D. M., Banks, E., Bhatia, G., del Angel, G., Gabriel, S. B., Genovese, G., Gupta, N., Li, H., Kashin, S., Lander, E. S., McCarroll, S. A., Nemesh, J. C., Poplin, R. E., Yoon, S. C., Lihm, J., Makarov, V., Clark, A. G., Gottipati, S., Keinan, A., Rodriguez-Flores, J. L., Korbel, J. O., Rausch, T., Fritz, M. H., Stütz, A. M., Flicek, P., Beal, K., Clarke, L., Datta, A., Herrero, J., McLaren, W. M., Ritchie, G. R. S., Smith, R. E., Zerbino, D., Zheng-Bradley, X., Sabeti, P. C., Shlyakhter, I., Schaffner, S. F., Vitti, J., Cooper, D. N., Ball, E. V., Stenson, P. D., Bentley, D. R., Barnes, B., Bauer, M., Keira Cheetham, R., Cox, A., Eberle, M., Humphray, S., Kahn, S., Murray, L., Peden, J., Shaw, R., Kenny, E. E., Batzer, M. A., Konkel, M. K., Walker, J. A., MacArthur, D. G., Lek, M., Sudbrak, R., Amstislavskiy, V. S., Herwig, R., Mardis, E. R., Ding, L., Koboldt, D. C., Larson, D., Ye, K., Gravel, S., The Genomes Project, C., Corresponding, a., Steering, c., Production, g., Baylor College of, M., Shenzhen, B. G. I., Broad Institute of, M. I. T., Harvard, Coriell Institute for Medical, R., European Molecular Biology Laboratory, E. B. I., Illumina, Max Planck Institute for Molecular, G., McDonnell Genome Institute at Washington, U., Health, U. S. N. I. o., University of, O., Wellcome Trust Sanger, I., Analysis, g., Affymetrix, Albert Einstein College of, M., Bilkent, U., Boston, C., Cold Spring Harbor, L., Cornell, U., European Molecular Biology, L., Harvard, U., Human Gene Mutation, D., Icahn School of Medicine at Mount, S., Louisiana State, U., Massachusetts General, H., McGill, U., & National Eye Institute, N. I. H. (2015, 2015/10/01). A global reference for human genetic variation. *Nature, 526*(7571), 68-74. https://doi.org/10.1038/nature15393

Avramopoulos, D. (2010). Genetics of psychiatric disorders methods: molecular approaches. *Psychiatric Clinics, 33*(1), 1-13.

Awad, A. G., & Voruganti, L. N. (2016). Quality of Life and Health Costs: The Feasibility of Cost-Utility Analysis in Schizophrenia. In *Beyond Assessment of Quality of Life in Schizophrenia* (pp. 175-183). Springer.

Ayesa-Arriola, R., Pérez-Iglesias, R., Rodríguez-Sánchez, J. M., Mata, I., Gómez-Ruiz, E., García-Unzueta, M., Martínez-García, O., Tabares-Seisdedos, R., Vázquez-Barquero, J. L., & Crespo-Facorro, B. (2012). Homocysteine and cognition in first-episode psychosis patients. *Eur Arch Psychiatry Clin Neurosci, 262*(7), 557-564. https://link.springer.com/article/10.1007%2Fs00406-012-0302-2

Badano, J. L., & Katsanis, N. (2002, 2002/10/01). Beyond Mendel: an evolving view of human genetic disease transmission. *Nature Reviews Genetics, 3*(10), 779-789. https://doi.org/10.1038/nrg910

Bagot, K. S., Milin, R., & Kaminer, Y. (2015). Adolescent Initiation of Cannabis Use and Early-Onset Psychosis. *Subst Abus, 36*(4), 524-533. https://doi.org/10.1080/08897077.2014.995332

Barrett, J. C., & Cardon, L. R. (2006). Evaluating coverage of genome-wide association studies. *Nature Genetics, 38*(6), 659-662.

Behar, D. M., Villems, R., Soodyall, H., Blue-Smith, J., Pereira, L., Metspalu, E., Scozzari, R., Makkan, H., Tzur, S., & Comas, D. (2008). The dawn of human matrilineal diversity. *The American Journal of Human Genetics, 82*(5), 1130-1140.

Behr, A. A., Liu, K. Z., Liu-Fang, G., Nakka, P., & Ramachandran, S. (2016). pong: fast analysis and visualization of latent clusters in population genetic data. *Bioinformatics, 32*(18), 2817-2823. https://doi.org/10.1093/bioinformatics/btw327

Bentley, A. R., Callier, S. L., & Rotimi, C. N. (2020). Evaluating the promise of inclusion of African ancestry populations in genomics. *NPJ Genom Med, 5*, 5. https://doi.org/10.1038/s41525-019-0111-x

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., Evers, D. J., Barnes, C. L., & Bignell, H. R. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature, 456*(7218), 53-59.

Benyamin, B., Ferreira, M. A., Willemsen, G., Gordon, S., Middelberg, R. P., McEvoy, B. P., Hottenga, J.-J., Henders, A. K., Campbell, M. J., & Wallace, L. (2009). Common variants in TMPRSS6 are associated with iron status and erythrocyte volume. *Nature Genetics, 41*(11), 1173-1175.

Bernstein, D. P., Stein, J. A., Newcomb, M. D., Walker, E., Pogge, D., Ahluvalia, T., Stokes, J., Handelsman, L., Medrano, M., & Desmond, D. (2003). Development and validation of a brief screening version of the Childhood Trauma Questionnaire. *Child abuse & neglect, 27*(2), 169-190.

Bien, S. A., Wojcik, G. L., Hodonsky, C. J., Gignoux, C. R., Cheng, I., Matise, T. C., Peters, U., Kenny, E. E., & North, K. E. (2019). The future of genomic studies must be globally representative: perspectives from PAGE. *Annual review of genomics and human genetics, 20*, 181-200.

Bigdeli, T. B., Genovese, G., Georgakopoulos, P., Meyers, J. L., Peterson, R. E., Iyegbe, C. O., Medeiros, H., Valderrama, J., Achtyes, E. D., & Kotov, R. (2019). Contributions of common genetic variants to risk of schizophrenia among individuals of African and Latino ancestry. *Mol Psychiatry*, 1-13.

Bigdeli, T. B., Ripke, S., Peterson, R. E., Trzaskowski, M., Bacanu, S. A., Abdellaoui, A., Andlauer, T. F., Beekman, A. T., Berger, K., Blackwood, D. H., Boomsma, D. I., Breen, G., Buttenschon, H. N., Byrne, E. M., Cichon, S., Clarke, T. K., Couvy-Duchesne, B., Craddock, N., de Geus, E. J., Degenhardt, F., Dunn, E. C., Edwards, A. C., Fanous, A. H., Forstner, A. J., Frank, J., Gill, M., Gordon, S. D., Grabe, H. J., Hamilton, S. P., Hardiman, O., Hayward, C., Heath, A. C., Henders, A. K., Herms, S., Hickie, I. B., Hoffmann, P., Homuth, G., Hottenga, J. J., Ising, M., Jansen, R., Kloiber, S., Knowles, J. A., Lang, M., Li, Q. S., Lucae, S., MacIntyre, D. J., Madden, P. A., Martin, N. G., McGrath, P. J., McGuffin, P., McIntosh, A. M., Medland, S. E., Mehta, D., Middeldorp, C. M., Milaneschi, Y., Montgomery, G. W., Mors, O., Muller-Myhsok, B., Nauck, M., Nyholt, D. R., Nothen, M. M., Owen, M. J., Penninx, B. W., Pergadia, M. L., Perlis, R. H., Peyrot, W. J., Porteous, D. J., Potash, J. B., Rice, J. P., Rietschel, M., Riley, B. P., Rivera, M., Schoevers, R., Schulze, T. G., Shi, J., Shyn, S. I., Smit, J. H., Smoller, J. W., Streit, F., Strohmaier, J., Teumer, A., Treutlein, J., Van der Auwera, S., van Grootheest, G., van Hemert, A. M., Volzke, H., Webb, B. T., Weissman, M. M., Wellmann, J., Willemsen, G., Witt, S. H., Levinson, D. F., Lewis, C. M., Wray, N. R., Flint, J., Sullivan, P. F., & Kendler, K. S. (2017, Mar 28). Genetic effects influencing risk for major depressive disorder in China and Europe. *Transl Psychiatry, 7*(3), e1074. https://doi.org/10.1038/tp.2016.292

Bitarello, B. D., & Mathieson, I. (2020). Polygenic scores for height in admixed populations. *bioRxiv*.

Bleuler, E. (1958). Dementia praecox or the group of schizophrenias, New York (International Universities Press) 1958.

Bonoldi, I., Simeone, E., Rocchetti, M., Codjoe, L., Rossi, G., Gambi, F., Balottin, U., Caverzasi, E., Politi, P., & Fusar-Poli, P. (2013). Prevalence of self-reported childhood abuse in psychosis: a meta-analysis of retrospective studies. *Psychiatry Research, 210*(1), 8-15. https://www.sciencedirect.com/science/article/abs/pii/S0165178113002606?via%3Dihub

Bosia, M., Buonocore, M., Bechi, M., Stere, L.-M., Silvestri, M. P., Inguscio, E., Spangaro, M., Cocchi, F., Bianchi, L., & Guglielmino, C. (2019). Schizophrenia, cannabis use and Catechol-O-Methyltransferase (COMT): Modeling the interplay on cognition. *Progress in Neuro-Psychopharmacology and Biological Psychiatry, 92*, 363-368. https://www.sciencedirect.com/science/article/abs/pii/S0278584618307206?via%3Dihub

Bourque, F., van der Ven, E., & Malla, A. (2011). A meta-analysis of the risk for psychotic disorders among first-and second-generation immigrants. *Psychological Medicine, 41*(5), 897. https://www.cambridge.org/core/journals/psychological-medicine/article/abs/metaanalysis-of-the-risk-for-psychotic-disorders-among-first-and-secondgeneration-immigrants/7427585927F99E8EF88C4F1AA6546C02

Bower, G. H. (1981, Feb). Mood and memory. *Am Psychol, 36*(2), 129-148. https://doi.org/10.1037//0003-066x.36.2.129

Boyle, A. P., Hong, E. L., Hariharan, M., Cheng, Y., Schaub, M. A., Kasowski, M., Karczewski, K. J., Park, J., Hitz, B. C., & Weng, S. (2012). Annotation of functional variation in personal genomes using RegulomeDB. *Genome Research, 22*(9), 1790-1797.

Bredström, A. (2019, 2019/09/01). Culture and Context in Mental Health Diagnosing: Scrutinizing the DSM-5 Revision. *Journal of Medical Humanities, 40*(3), 347-363. https://doi.org/10.1007/s10912-017-9501-1

Brohan, E., Clement, S., Rose, D., Sartorius, N., Slade, M., & Thornicroft, G. (2013). Development and psychometric evaluation of the Discrimination and Stigma Scale (DISC). *Psychiatry Research, 208*(1), 33-40.

Brown, A. S. (2011). The environment and susceptibility to schizophrenia. *Progress in neurobiology, 93*(1), 23-58. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3521525/pdf/nihms-426417.pdf

Brown, A. S., Cohen, P., Harkavy-Friedman, J., Babulas, V., Malaspina, D., Gorman, J. M., & Susser, E. S. (2001). Prenatal rubella, premorbid abnormalities, and adult schizophrenia. *Biological psychiatry, 49*(6), 473-486. https://www.biologicalpsychiatryjournal.com/article/S0006-3223(01)01068-X/fulltext

Brown, Brielin C., Ye, Chun J., Price, Alkes L., & Zaitlen, N. (2016, 2016/07/07/). Transethnic Genetic-Correlation Estimates from Summary Statistics. *The American Journal of Human Genetics, 99*(1), 76-88. https://doi.org/https://doi.org/10.1016/j.ajhg.2016.05.001

Browning, S. R., & Browning, B. L. (2011, Jul 15). Population structure can inflate SNP-based heritability estimates. *Am J Hum Genet, 89*(1), 191-193; author reply 193-195. https://doi.org/10.1016/j.ajhg.2011.05.025

Bulat, V., Rast, M., & Pielage, J. (2014). Presynaptic CK2 promotes synapse organization and stability by targeting Ankyrin2. *Journal of Cell Biology, 204*(1), 77-94.

Bulik-Sullivan, B. K., Loh, P. R., Finucane, H. K., Ripke, S., Yang, J., Patterson, N., Daly, M. J., Price, A. L., & Neale, B. M. (2015, Mar). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet, 47*(3), 291-295. https://doi.org/10.1038/ng.3211

Burgdorf, K. S., Trabjerg, B. B., Pedersen, M. G., Nissen, J., Banasik, K., Pedersen, O. B., Sørensen, E., Nielsen, K. R., Larsen, M. H., & Erikstrup, C. (2019). Large-scale study of Toxoplasma and Cytomegalovirus shows an association between infection and serious psychiatric disorders. *Brain, behavior, and immunity, 79*, 152-158.

Busby, G. B., Band, G., Le, Q. S., Jallow, M., Bougama, E., Mangano, V. D., Amenga-Etego, L. N., Enimil, A., Apinjoh, T., & Ndila, C. M. (2016). Admixture into and within sub-Saharan Africa. *Elife, 5*, e15266.

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., & O'Connell, J. (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature, 562*(7726), 203-209.

Campbell, M. C., & Tishkoff, S. A. (2008). African genetic diversity: Implications for human demographic history, modern human origins, and complex disease mapping. In *Annual review of genomics and human genetics* (Vol. 9, pp. 403-433). https://doi.org/10.1146/annurev.genom.9.081307.164258

Campbell, M. M., Sibeko, G., Mall, S., Baldinger, A., Nagdee, M., Susser, E., & Stein, D. J. (2017, Jan 24). The content of delusions in a sample of South African Xhosa people with schizophrenia. *BMC Psychiatry, 17*(1), 41. https://doi.org/10.1186/s12888-017-1196-3

Cannon, T. D., Kaprio, J., Lonnqvist, J., Huttunen, M., & Koskenvuo, M. (1998, Jan). The genetic epidemiology of schizophrenia in a Finnish twin cohort. A population-based modeling study. *Arch Gen Psychiatry, 55*(1), 67-74.

Cantor-Graae, E., Pedersen, C. B., Mcneil, T. F., & Mortensen, P. B. (2003). Migration as a risk factor for schizophrenia: a Danish population-based cohort study. *The British Journal of Psychiatry, 182*(2), 117-122.

Cardno, A. G., Marshall, E. J., Coid, B., Macdonald, A. M., Ribchester, T. R., Davies, N. J., Venturi, P., Jones, L. A., Lewis, S. W., & Sham, P. C. (1999). Heritability estimates for psychotic disorders: the Maudsley twin psychosis series. *Archives of General Psychiatry, 56*(2), 162-168.

Carter, C., & Evans, K. (1969). Inheritance of congenital pyloric stenosis. *Journal of Medical Genetics, 6*(3), 233.

Caspi, A., Moffitt, T. E., Cannon, M., McClay, J., Murray, R., Harrington, H., Taylor, A., Arseneault, L., Williams, B., & Braithwaite, A. (2005). Moderation of the effect of adolescent-onset cannabis use on adult psychosis by a functional polymorphism in the catechol-O-methyltransferase gene: longitudinal evidence of a gene X environment interaction. *Biological psychiatry, 57*(10), 1117-1127. https://www.biologicalpsychiatryjournal.com/article/S0006-3223(05)00103-4/fulltext

Cerdeña, J. P., Plaisime, M. V., & Tsai, J. (2020). From race-based to race-conscious medicine: how anti-racist uprisings call us to act. *The Lancet, 396*(10257), 1125-1128.

Chambers, J. C., Zhang, W., Li, Y., Sehmi, J., Wass, M. N., Zabaneh, D., Hoggart, C., Bayele, H., McCarthy, M. I., & Peltonen, L. (2009). Genome-wide association study identifies variants in TMPRSS6 associated with hemoglobin levels. *Nature Genetics, 41*(11), 1170-1172.

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience, 4*(1), s13742-13015-10047-13748.

Chimusa, E. R., Meintjies, A., Tchanga, M., Mulder, N., Seoighe, C., Soodyall, H., & Ramesar, R. (2015, Mar). A genomic portrait of haplotype diversity and signatures of selection in indigenous southern African populations. *PLoS Genet, 11*(3), e1005052. https://doi.org/10.1371/journal.pgen.1005052

Choudhury, A., Aron, S., Botigué, L. R., Sengupta, D., Botha, G., Bensellak, T., Wells, G., Kumuthini, J., Shriner, D., & Fakim, Y. J. (2020). High-depth African genomes inform human migration and health. *Nature, 586*(7831), 741-748.

Choudhury, A., Ramsay, M., Hazelhurst, S., Aron, S., Bardien, S., Botha, G., Chimusa, E. R., Christoffels, A., Gamieldien, J., Sefid-Dashti, M. J., Joubert, F., Meintjes, A., Mulder, N., Ramesar, R., Rees, J., Scholtz, K., Sengupta, D., Soodyall, H., Venter, P., Warnich, L., & Pepper, M. S. (2017, 2017/12/12). Whole-genome sequencing for an enhanced understanding of genetic variation among South Africans. *Nature Communications, 8*(1), 2062. https://doi.org/10.1038/s41467-017-00663-9

Clouston, T. S. (1904). Clinical Lectures on Mental Diseases. London. *UK: J&A Churchill*.

Cloutier, M., Aigbogun, M. S., Guerin, A., Nitulescu, R., Ramanakumar, A. V., Kamat, S. A., DeLucia, M., Duffy, R., Legacy, S. N., Henderson, C., Francois, C., & Wu, E. (2016, Jun). The Economic Burden of Schizophrenia in the United States in 2013. *J Clin Psychiatry, 77*(6), 764-771. https://doi.org/10.4088/JCP.15m10278

Coleman, J. R. I., Euesden, J., Patel, H., Folarin, A. A., Newhouse, S., & Breen, G. (2015). Quality control, imputation and analysis of genome-wide genotyping data from the Illumina HumanCoreExome microarray. *Briefings in Functional Genomics, 15*(4), 298-304. https://doi.org/10.1093/bfgp/elv037

Comer, A. L., Jinadasa, T., Sriram, B., Phadke, R. A., Kretsge, L. N., Nguyen, T. P., Antognetti, G., Gilbert, J. P., Lee, J., & Newmark, E. R. (2020). Increased expression of schizophrenia-associated gene C4 leads to hypoconnectivity of prefrontal cortex and reduced social interaction. *PLoS Biology, 18*(1), e3000604.

Conomos, M. P., Reiner, A. P., McPeek, M. S., & Thornton, T. A. (2018). Genome-Wide Control of Population Structure and Relatedness in Genetic Association Studies via Linear Mixed Models with Orthogonally Partitioned Structure. *bioRxiv*, 409953.

Cornuet, J. M., & Luikart, G. (1996). Description and power analysis of two tests for detecting recent population bottlenecks from allele frequency data. *Genetics, 144*(4), 2001-2014.

Corrigan, P. W., Druss, B. G., & Perlick, D. A. (2014). The impact of mental illness stigma on seeking and participating in mental health care. *Psychological Science in the Public Interest, 15*(2), 37-70.

Costas, J., Sanjuán, J., Ramos-Ríos, R., Paz, E., Agra, S., Tolosa, A., Páramo, M., Brenlla, J., & Arrojo, M. (2011). Interaction between COMT haplotypes and cannabis in schizophrenia: a case-only study in two samples from Spain. *Schizophrenia Research, 127*(1-3), 22-27. https://www.sciencedirect.com/science/article/abs/pii/S092099641100048X?via%3Dihub

Creese, B., Vassos, E., Bergh, S., Athanasiu, L., Johar, I., Rongve, A., Medbøen, I. T., Vasconcelos Da Silva, M., Aakhus, E., Andersen, F., Bettella, F., Braekhus, A., Djurovic, S., Paroni, G., Proitsi, P., Saltvedt, I., Seripa, D., Stordal, E., Fladby, T., Aarsland, D., Andreassen, O. A., Ballard, C., Selbaek, G., on behalf of the AddNeuroMed, c., & the Alzheimer's Disease Neuroimaging, I. (2019, 2019/10/22). Examining the association between genetic liability for schizophrenia and psychotic symptoms in Alzheimer's disease. *Transl Psychiatry, 9*(1), 273. https://doi.org/10.1038/s41398-019-0592-5

Crisp, A. H., Gelder, M. G., Rix, S., Meltzer, H. I., & Rowlands, O. J. (2000). Stigmatisation of people with mental illnesses. *British Journal of Psychiatry, 177*(1), 4-7. https://doi.org/10.1192/bjp.177.1.4

Cutajar, M. C., Mullen, P. E., Ogloff, J. R., Thomas, S. D., Wells, D. L., & Spataro, J. (2010). Schizophrenia and other psychotic disorders in a cohort of sexually abused children. *Archives of General Psychiatry, 67*(11), 1114-1119.

de Castro-Catala, M., van Nierop, M., Barrantes-Vidal, N., Cristóbal-Narváez, P., Sheinbaum, T., Kwapil, T. R., Peña, E., Jacobs, N., Derom, C., & Thiery, E. (2016). Childhood trauma, BDNF Val66Met and subclinical psychotic experiences. Attempt at replication in two independent samples. *J Psychiatr Res, 83*, 121-129. http://diposit.ub.edu/dspace/bitstream/2445/124981/1/663788.pdf

de Leeuw, C. A., Mooij, J. M., Heskes, T., & Posthuma, D. (2015). MAGMA: generalized gene-set analysis of GWAS data. *PLoS Comput Biol, 11*(4), e1004219.

de Vlaming, R., Okbay, A., Rietveld, C. A., Johannesson, M., Magnusson, P. K., Uitterlinden, A. G., van Rooij, F. J., Hofman, A., Groenen, P. J., & Thurik, A. R. (2017). Meta-GWAS Accuracy and Power (MetaGAP) calculator shows that hiding heritability is partially due to imperfect genetic correlations across studies. *Plos Genetics, 13*(1), e1006495.

Dempfle, A., Scherag, A., Hein, R., Beckmann, L., Chang-Claude, J., & Schäfer, H. (2008). Gene–environment interactions for complex traits: definitions, methodological requirements and challenges. *European Journal of Human Genetics, 16*(10), 1164-1172.

Denomme, M. M., Haywood, M. E., Parks, J. C., Schoolcraft, W. B., & Katz-Jaffe, M. G. (2020). The inherited methylome landscape is directly altered with paternal aging and associated with offspring neurodevelopmental disorders. *Aging cell, 19*(8), e13178. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7431824/pdf/ACEL-19-e13178.pdf

Derks, E. M., Vorstman, J. A., Ripke, S., Kahn, R. S., & Ophoff, R. A. (2012). Investigation of the genetic association between quantitative measures of psychosis and schizophrenia: a polygenic risk score analysis. *Plos One, 7*(6), e37852. https://doi.org/10.1371/journal.pone.0037852

DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium, A. G. E. N. T. D. A.-T. D. C., South Asian Type 2 Diabetes (SAT2D) Consortium, Mexican American Type 2 Diabetes (MAT2D) Consortium. (2014). Type 2 Diabetes Genetic Exploration by Next-generation Sequencing in Multi-Ethnic Samples *Nat Genet, 46*(3), 234-244.

Dickerson, F., Schroeder, J. R., Nimgaonkar, V., Gold, J., & Yolken, R. (2020). The association between exposure to herpes simplex virus type 1 (HSV-1) and cognitive functioning in schizophrenia: A meta-analysis. *Psychiatry Research, 291*, 113157. https://www.sciencedirect.com/science/article/abs/pii/S0165178119325338?via%3Dihub

Dohrenwend, B. P., Levav, I., Shrout, P. E., Schwartz, S., Naveh, G., Link, B. G., Skodol, A. E., & Stueve, A. (1992, Feb 21). Socioeconomic status and psychiatric disorders: the causation-selection issue. *Science, 255*(5047), 946-952. https://doi.org/10.1126/science.1546291

Dorrington, S., Zammit, S., Asher, L., Evans, J., Heron, J., & Lewis, G. (2014). Perinatal maternal life events and psychotic experiences in children at twelve years in a birth cohort study. *Schizophrenia Research, 152*(1), 158-163. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3906533/pdf/main.pdf

Driver, D. I., Gogtay, N., & Rapoport, J. L. (2013, 06/18). Childhood Onset Schizophrenia and Early Onset Schizophrenia spectrum disorders. *Child and adolescent psychiatric clinics of North America, 22*(4), 539-555. https://doi.org/10.1016/j.chc.2013.04.001

Duncan, L., Shen, H., Gelaye, B., Meijsen, J., Ressler, K., Feldman, M., Peterson, R., & Domingue, B. (2019). Analysis of polygenic risk score usage and performance in diverse human populations. *Nature Communications, 10*(1), 1-9.

Duncan, L. E., & Keller, M. C. (2011). A critical review of the first 10 years of candidate gene-by-environment interaction research in psychiatry. *American Journal of Psychiatry, 168*(10), 1041-1049.

Duncan, L. E., Ratanatharathorn, A., Aiello, A. E., Almli, L. M., Amstadter, A. B., Ashley-Koch, A. E., Baker, D. G., Beckham, J. C., Bierut, L. J., & Bisson, J. (2018). Largest GWAS of PTSD (N= 20 070) yields genetic overlap with schizophrenia and sex differences in heritability. *Mol Psychiatry, 23*(3), 666-673.

Dunham, I., Birney, E., Lajoie, B. R., Sanyal, A., Dong, X., Greven, M., Lin, X., Wang, J., Whitfield, T. W., & Zhuang, J. (2012). An integrated encyclopedia of DNA elements in the human genome.

Dupont, C., Armant, D. R., & Brenner, C. A. (2009, Sep). Epigenetics: definition, mechanisms and clinical perspective. *Semin Reprod Med, 27*(5), 351-357. https://doi.org/10.1055/s-0029-1237423

Ellersgaard, D., Jessica Plessen, K., Richardt Jepsen, J., Soeborg Spang, K., Hemager, N., Klee Burton, B., Jerlang Christiani, C., Gregersen, M., Søndergaard, A., Uddin, M. J., Poulsen, G., Greve, A., Gantriis, D., Mors, O., Nordentoft, M., & Elgaard Thorup, A. A. (2018, Jun). Psychopathology in 7-year-old children with familial high risk of developing schizophrenia spectrum psychosis or bipolar disorder - The Danish High Risk and Resilience Study - VIA 7, a population-based cohort study. *World Psychiatry, 17*(2), 210-219. https://doi.org/10.1002/wps.20527

Ellingson, S. R., & Fardo, D. W. (2016). Automated quality control for genome wide association studies. *F1000Research, 5*.

Ellman, L. M., Murphy, S. K., Maxwell, S. D., Calvo, E. M., Cooper, T., Schaefer, C. A., Bresnahan, M. A., Susser, E. S., & Brown, A. S. (2019). Maternal cortisol during pregnancy and offspring schizophrenia: Influence of fetal sex and timing of exposure. *Schizophrenia Research, 213*, 15-22. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7074891/pdf/nihms-1558224.pdf

ENCODE Project, Consortium. (2012, Sep 6). An integrated encyclopedia of DNA elements in the human genome. *Nature, 489*(7414), 57-74. https://doi.org/10.1038/nature11247

Ermis, A., Erkiran, M., Dasdemir, S., Turkcan, A. S., Ceylan, M. E., Bireller, E. S., & Cakmakoglu, B. (2015). The relationship between catechol-O-methyltransferase gene Val158Met (COMT) polymorphism and premorbid cannabis use in Turkish male patients with schizophrenia. *in vivo, 29*(1), 129-132. https://iv.iiarjournals.org/content/invivo/29/1/129.full.pdf

Ernst, J., & Kellis, M. (2012). ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods, 9*(3), 215-216.

Escott-Price, V., Smith, D. J., Kendall, K., Ward, J., Kirov, G., Owen, M. J., Walters, J., & O'Donovan, M. C. (2018). Polygenic risk for schizophrenia and season of birth within the UK Biobank cohort. *Psychological Medicine*.

Esposito, G., Azhari, A., & Borelli, J. L. (2018, 2018-October-26). Gene × Environment Interaction in Developmental Disorders: Where Do We Stand and What's Next? [Hypothesis and Theory]. *Frontiers in Psychology, 9*(2036). https://doi.org/10.3389/fpsyg.2018.02036

Esshili, A., Manitz, M.-P., Freund, N., & Juckel, G. (2020). Induction of inducible nitric oxide synthase expression in activated microglia and astrocytes following pre-and postnatal immune challenge in an animal model of schizophrenia. *European Neuropsychopharmacology*.

Estimates, G. H. (2016). Deaths by cause, age, sex, by country and by region, 2000–2015.

Estrada, G., Fatjó-Vilas, M., Munoz, M., Pulido, G., Minano, M., Toledo, E., Illa, J., Martin, M., Miralles, M., & Miret, S. (2011). Cannabis use and age at onset of psychosis: further evidence of interaction with

COMT Val158Met polymorphism. *Acta Psychiatr Scand, 123*(6), 485-492. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1600-0447.2010.01665.x

Eyles, D. W., Trzaskowski, M., Vinkhuyzen, A. A., Mattheisen, M., Meier, S., Gooch, H., Anggono, V., Cui, X., Tan, M. C., & Burne, T. H. (2018). The association between neonatal vitamin D status and risk of schizophrenia. *Sci Rep, 8*(1), 1-8.

Falkenburger, B. H., Jensen, J. B., Dickson, E. J., Suh, B. C., & Hille, B. (2010). Symposium Review: Phosphoinositides: lipid regulators of membrane proteins. *The Journal of physiology, 588*(17), 3179-3185.

Fernández-Rhodes, L., Gong, J., Haessler, J., Franceschini, N., Graff, M., Nishimura, K. K., Wang, Y., Highland, H. M., Yoneyama, S., Bush, W. S., Goodloe, R., Ritchie, M. D., Crawford, D., Gross, M., Fornage, M., Buzkova, P., Tao, R., Isasi, C., Avilés-Santa, L., Daviglus, M., Mackey, R. H., Houston, D., Gu, C. C., Ehret, G., Nguyen, K.-D. H., Lewis, C. E., Leppert, M., Irvin, M. R., Lim, U., Haiman, C. A., Le Marchand, L., Schumacher, F., Wilkens, L., Lu, Y., Bottinger, E. P., Loos, R. J. L., Sheu, W. H. H., Guo, X., Lee, W.-J., Hai, Y., Hung, Y.-J., Absher, D., Wu, I. C., Taylor, K. D., Lee, I. T., Liu, Y., Wang, T.-D., Quertermous, T., Juang, J.-M. J., Rotter, J. I., Assimes, T., Hsiung, C. A., Chen, Y.-D. I., Prentice, R., Kuller, L. H., Manson, J. E., Kooperberg, C., Smokowski, P., Robinson, W. R., Gordon-Larsen, P., Li, R., Hindorff, L., Buyske, S., Matise, T. C., Peters, U., & North, K. E. (2017, 2017/06/01). Trans-ethnic fine-mapping of genetic loci for body mass index in the diverse ancestral populations of the Population Architecture using Genomics and Epidemiology (PAGE) Study reveals evidence for multiple signals at established loci. *Human Genetics, 136*(6), 771-800. https://doi.org/10.1007/s00439-017-1787-6

First, M. B. (1997). Structured clinical interview for DSM-IV axis I disorders. *Biometrics Research Department*

Flannick, J., Thorleifsson, G., Beer, N. L., Jacobs, S. B., Grarup, N., Burtt, N. P., Mahajan, A., Fuchsberger, C., Atzmon, G., & Benediktsson, R. (2014). Loss-of-function mutations in SLC30A8 protect against type 2 diabetes. *Nature Genetics, 46*(4), 357-363.

Fond, G., Godin, O., Boyer, L., Llorca, P.-M., Andrianarisoa, M., Brunel, L., Aouizerate, B., Berna, F., Capdevielle, D., & D'Amato, T. (2017). Advanced paternal age is associated with earlier schizophrenia onset in offspring. Results from the national multicentric FACE-SZ cohort. *Psychiatry Research, 254*, 218-223. https://www.sciencedirect.com/science/article/abs/pii/S0165178116318327?via%3Dihub

Forrest, M. P., Hill, M. J., Kavanagh, D. H., Tansey, K. E., Waite, A. J., & Blake, D. J. (2018, Aug 20). The Psychiatric Risk Gene Transcription Factor 4 (TCF4) Regulates Neurodevelopmental Pathways Associated With Schizophrenia, Autism, and Intellectual Disability. *Schizophr Bull, 44*(5), 1100-1110. https://doi.org/10.1093/schbul/sbx164

Forster, P., & Matsumura, S. (2005, May 13). Evolution. Did early humans go north or south? *Science, 308*(5724), 965-966. https://doi.org/10.1126/science.1113261

Forster, P., & Romano, V. (2007). Timing of a back-migration into Africa. *Science, 316*(5821), 50-53.

Fountoulakis, K. N., Gonda, X., Siamouli, M., Panagiotidis, P., Moutou, K., Nimatoudis, I., & Kasper, S. (2018). Paternal and maternal age as risk factors for schizophrenia: a case–control study. *International journal of psychiatry in clinical practice, 22*(3), 170-176. https://www.tandfonline.com/doi/full/10.1080/13651501.2017.1391292

Franceschini, N., Carty, C. L., Lu, Y., Tao, R., Sung, Y. J., Manichaikul, A., Haessler, J., Fornage, M., Schwander, K., & Zubair, N. (2016). Variant discovery and fine mapping of genetic loci associated with blood pressure traits in Hispanics and African Americans. *Plos One, 11*(10), e0164132.

Franzek, E., & Beckmann, H. (1998, Jan). Different genetic background of schizophrenia spectrum psychoses: a twin study. *Am J Psychiatry, 155*(1), 76-83. https://doi.org/10.1176/ajp.155.1.76

Fromer, M., Pocklington, A. J., Kavanagh, D. H., Williams, H. J., Dwyer, S., Gormley, P., Georgieva, L., Rees, E., Palta, P., Ruderfer, D. M., Carrera, N., Humphreys, I., Johnson, J. S., Roussos, P., Barker, D. D., Banks, E., Milanova, V., Grant, S. G., Hannon, E., Rose, S. A., Chambert, K., Mahajan, M., Scolnick, E. M., Moran, J. L., Kirov, G., Palotie, A., McCarroll, S. A., Holmans, P., Sklar, P., Owen, M. J., Purcell, S. M., & O'Donovan, M. C. (2014, Feb 13). De novo mutations in schizophrenia implicate synaptic networks. *Nature, 506*(7487), 179-184. https://doi.org/10.1038/nature12929

Fuchsberger, C., Flannick, J., Teslovich, T. M., Mahajan, A., Agarwala, V., Gaulton, K. J., Ma, C., Fontanillas, P., Moutsianas, L., & McCarthy, D. J. (2016). The genetic architecture of type 2 diabetes. *Nature, 536*(7614), 41-47.

Gage, S. H., Hickman, M., & Zammit, S. (2016, Apr 1). Association Between Cannabis and Psychosis: Epidemiologic Evidence. *Biol Psychiatry, 79*(7), 549-556. https://doi.org/10.1016/j.biopsych.2015.08.001

Garcia, M., Montalvo, I., Creus, M., Cabezas, Á., Solé, M., Algora, M. J., Moreno, I., Gutiérrez-Zotes, A., & Labad, J. (2016, 2016/07/01/). Sex differences in the effect of childhood trauma on the clinical expression of early psychosis. *Comprehensive Psychiatry, 68*, 86-96. https://doi.org/https://doi.org/10.1016/j.comppsych.2016.04.004

Ge, T., Chen, C.-Y., Neale, B. M., Sabuncu, M. R., & Smoller, J. W. (2017). Phenome-wide heritability analysis of the UK Biobank. *Plos Genetics, 13*(4), e1006711.

Genovese, G., Fromer, M., Stahl, E. A., Ruderfer, D. M., Chambert, K., Landén, M., Moran, J. L., Purcell, S. M., Sklar, P., Sullivan, P. F., Hultman, C. M., & McCarroll, S. A. (2016, Nov). Increased burden of ultra-rare protein-altering variants among 4,877 individuals with schizophrenia. *Nat Neurosci, 19*(11), 1433-1441. https://doi.org/10.1038/nn.4402

Gerlinger, G., Hauser, M., De Hert, M., Lacluyse, K., Wampers, M., & Correll, C. U. (2013, Jun). Personal stigma in schizophrenia spectrum disorders: a systematic review of prevalence rates, correlates, impact and interventions. *World Psychiatry, 12*(2), 155-164. https://doi.org/10.1002/wps.20040

Ginther, D. K., Schaffer, W. T., Schnell, J., Masimore, B., Liu, F., Haak, L. L., & Kington, R. (2011). Race, ethnicity, and NIH research awards. *Science, 333*(6045), 1015-1019.

Gottesman, I. I. (1991). *Schizophrenia genesis: The origins of madness*. WH Freeman/Times Books/Henry Holt & Co.

Gratten, J., Wray, N. R., Peyrot, W. J., McGrath, J. J., Visscher, P. M., & Goddard, M. E. (2016, 2016/07/01). Risk of psychiatric illness from advanced paternal age is not predominantly from de novo mutations. *Nature Genetics, 48*(7), 718-724. https://doi.org/10.1038/ng.3577

Green, M. J., Chia, T. Y., Cairns, M. J., Wu, J., Tooney, P. A., Scott, R. J., & Carr, V. J. (2014, Feb). Catechol-O-methyltransferase (COMT) genotype moderates the effects of childhood trauma on cognition and symptoms in schizophrenia. *J Psychiatr Res, 49*, 43-50. https://doi.org/10.1016/j.jpsychires.2013.10.018

Green, M. J., Raudino, A., Cairns, M. J., Wu, J., Tooney, P. A., Scott, R. J., & Carr, V. J. (2015, Nov). Do common genotypes of FK506 binding protein 5 (FKBP5) moderate the effects of childhood maltreatment on cognition in schizophrenia and healthy controls? *J Psychiatr Res, 70*, 9-17. https://doi.org/10.1016/j.jpsychires.2015.07.019

Gresham, D., Dunham, M. J., & Botstein, D. (2008, Apr). Comparing whole genomes using DNA microarrays. *Nat Rev Genet, 9*(4), 291-302. https://doi.org/10.1038/nrg2335

Gronholm, P. C., Henderson, C., Deb, T., & Thornicroft, G. (2017). Interventions to reduce discrimination and stigma: the state of the art. *Social psychiatry and psychiatric epidemiology, 52*(3), 249-258.

Gross, D. S., & Garrard, W. T. (1988). Nuclease hypersensitive sites in chromatin. *Annu Rev Biochem, 57*, 159-197. https://doi.org/10.1146/annurev.bi.57.070188.001111

Grossman, L. S., Harrow, M., Rosen, C., Faull, R., & Strauss, G. P. (2008). Sex differences in schizophrenia and other psychotic disorders: a 20-year longitudinal study of psychosis and recovery. *Comprehensive Psychiatry, 49*(6), 523-529.

Gulsuner, S., Stein, D. J., Susser, E. S., Sibeko, G., Pretorius, A., Walsh, T., Majara, L., Mndini, M. M., Mqulwana, S. G., Ntola, O. A., Casadei, S., Ngqengelele, L. L., Korchina, V., van der Merwe, C., Malan, M., Fader, K. M., Feng, M., Willoughby, E., Muzny, D., Baldinger, A., Andrews, H. F., Gur, R. C., Gibbs, R. A., Zingela, Z., Nagdee, M., Ramesar, R. S., King, M. C., & McClellan, J. M. (2020, Jan 31). Genetics of schizophrenia in the South African Xhosa. *Science, 367*(6477), 569-573. https://doi.org/10.1126/science.aay8833

Gulsuner, S., Walsh, T., Watts, A. C., Lee, M. K., Thornton, A. M., Casadei, S., Rippey, C., Shahin, H., Consortium on the Genetics of, S., Group, P. S., Nimgaonkar, V. L., Go, R. C. P., Savage, R. M., Swerdlow, N. R., Gur, R. E., Braff, D. L., King, M.-C., & McClellan, J. M. (2013). Spatial and temporal mapping of de novo mutations in schizophrenia to a fetal prefrontal cortical network. *Cell, 154*(3), 518-529. https://doi.org/10.1016/j.cell.2013.06.049

Guna, A., Butcher, N. J., & Bassett, A. S. (2015, 2015/07/01). Comparative mapping of the 22q11.2 deletion region and the potential of simple model organisms. *Journal of Neurodevelopmental Disorders, 7*(1), 18. https://doi.org/10.1186/s11689-015-9113-x

Gurdasani, D., Carstensen, T., Fatumo, S., Chen, G., Franklin, C. S., Prado-Martinez, J., Bouman, H., Abascal, F., Haber, M., & Tachmazidou, I. (2019). Uganda genome resource enables insights into population history and genomic discovery in Africa. *Cell, 179*(4), 984-1002. e1036.

Gurdasani, D., Carstensen, T., Tekola-Ayele, F., Pagani, L., Tachmazidou, I., Hatzikotoulas, K., Karthikeyan, S., Iles, L., Pollard, M. O., & Choudhury, A. (2015). The African genome variation project shapes medical genetics in Africa. *Nature, 517*(7534), 327-332.

Gutiérrez, B., Rivera, M., Obel, L., McKenney, K., Martínez-Leal, R., Molina, E., Dolz, M., Ochoa, S., Usall, J., & Haro, J. M. (2009). Variability of the COMT gene and modification of the risk of schizophrenia conferred by cannabis consumption. *Revista de Psiquiatría y Salud Mental (English Edition), 2*(2), 89-94.

H3Africa Consortium, (2014). Enabling the genomic revolution in Africa. *Science, 344*(6190), 1346-1348.

Hakulinen, C., Webb, R. T., Pedersen, C. B., Agerbo, E., & Mok, P. L. H. (2020, Jan 1). Association Between Parental Income During Childhood and Risk of Schizophrenia Later in Life. *JAMA Psychiatry, 77*(1), 17-24. https://doi.org/10.1001/jamapsychiatry.2019.2299

Halvorsen, M., Huh, R., Oskolkov, N., Wen, J., Netotea, S., Giusti-Rodriguez, P., Karlsson, R., Bryois, J., Nystedt, B., Ameur, A., Kähler, A. K., Ancalade, N., Farrell, M., Crowley, J. J., Li, Y., Magnusson, P. K. E., Gyllensten, U., Hultman, C. M., Sullivan, P. F., & Szatkiewicz, J. P. (2020, Apr 15). Increased burden of ultra-rare structural variants localizing to boundaries of topologically associated domains in schizophrenia. *Nat Commun, 11*(1), 1842. https://doi.org/10.1038/s41467-020-15707-w

Hatzimanolis, A., Stefanatou, P., Kattoulas, E., Ralli, I., Dimitrakopoulos, S., Foteli, S., Kosteletos, I., Mantonakis, L., Selakovic, M., Soldatos, R. F., Vlachos, I., Xenaki, L. A., Smyrnis, N., & Stefanis, N. C. (2020, Apr 29). Familial and socioeconomic contributions to premorbid functioning in psychosis: Impact on age at onset and treatment response. *Eur Psychiatry, 63*(1), e44. https://doi.org/10.1192/j.eurpsy.2020.41

Hayes, J. F., Marston, L., Walters, K., King, M. B., & Osborn, D. P. J. (2017). Mortality gap for people with bipolar disorder and schizophrenia: UK-based cohort study 2000–2014. *British Journal of Psychiatry, 211*(3), 175-181. https://doi.org/10.1192/bjp.bp.117.202606

Hayes, L., Hawthorne, G., Farhall, J., O'hanlon, B., & Harvey, C. (2015). Quality of life and social isolation among caregivers of adults with schizophrenia: Policy and outcomes. *Community mental health journal, 51*(5), 591.

He, P., Chen, G., Guo, C., Wen, X., Song, X., & Zheng, X. (2018). Long-term effect of prenatal exposure to malnutrition on risk of schizophrenia in adulthood: evidence from the Chinese famine of 1959–1961. *European Psychiatry, 51*, 42-47. https://www.sciencedirect.com/science/article/abs/pii/S0924933818300038?via%3Dihub

Healthcare Research and Quality Agency. (2016). 2015 National healthcare quality and disparities report and 5th anniversary update on the national quality strategy. *Pub. no. 16-0015*

Heckerman, D., Gurdasani, D., Kadie, C., Pomilla, C., Carstensen, T., Martin, H., Ekoru, K., Nsubuga, R. N., Ssenyomo, G., & Kamali, A. (2016). Linear mixed model for heritability estimation that explicitly addresses environmental variation. *Proceedings of the National Academy of Sciences, 113*(27), 7377-7382.

Helgason, A., Pálsson, S., Thorleifsson, G., Grant, S. F., Emilsson, V., Gunnarsdottir, S., Adeyemo, A., Chen, Y., Chen, G., Reynisdottir, I., Benediktsson, R., Hinney, A., Hansen, T., Andersen, G., Borch-Johnsen, K., Jorgensen, T., Schäfer, H., Faruque, M., Doumatey, A., Zhou, J., Wilensky, R. L., Reilly, M. P., Rader, D. J., Bagger, Y., Christiansen, C., Sigurdsson, G., Hebebrand, J., Pedersen, O., Thorsteinsdottir, U., Gulcher, J. R., Kong, A., Rotimi, C., & Stefánsson, K. (2007, Feb). Refining the impact of TCF7L2 gene variants on type 2 diabetes and adaptive evolution. *Nat Genet, 39*(2), 218-225. https://doi.org/10.1038/ng1960

Hellwege, J. N., Keaton, J. M., Giri, A., Gao, X., Velez Edwards, D. R., & Edwards, T. L. (2017). Population stratification in genetic association studies. *Current Protocols in Human Genetics, 95*(1), 1.22. 21-21.22. 23.

Henn, B. M., Botigué, L. R., Gravel, S., Wang, W., Brisbin, A., Byrnes, J. K., Fadhlaoui-Zid, K., Zalloua, P. A., Moreno-Estrada, A., & Bertranpetit, J. (2012). Genomic ancestry of North Africans supports back-to-Africa migrations. *PLoS Genet, 8*(1), e1002397.

Henn, B. M., Gignoux, C. R., Jobin, M., Granka, J. M., Macpherson, J., Kidd, J. M., Rodríguez-Botigué, L., Ramachandran, S., Hon, L., & Brisbin, A. (2011). Hunter-gatherer genomic diversity suggests a southern African origin for modern humans. *Proceedings of the National Academy of Sciences, 108*(13), 5154-5162.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and brain sciences, 33*(2-3), 61-83.

Herman, A. A., Stein, D. J., Seedat, S., Heeringa, S. G., Moomal, H., & Williams, D. R. (2009). The South African Stress and Health (SASH) study: 12-month and lifetime prevalence of common mental disorders. *South African Medical Journal, 99*(5).

Hero, J. O., Zaslavsky, A. M., & Blendon, R. J. (2017). The United States leads other nations in differences by income in perceptions of health and health care. *Health Affairs, 36*(6), 1032-1040.

Heston, L. L. (1966, Aug). Psychiatric disorders in foster home reared children of schizophrenic mothers. *Br J Psychiatry, 112*(489), 819-825.

Higgins, J., Gore, R., Gutkind, D., Mednick, S. A., Parnas, J., Schulsinger, F., & Cannon, T. D. (1997, Nov). Effects of child-rearing by schizophrenic mothers: a 25-year follow-up. *Acta Psychiatr Scand, 96*(5), 402-404.

Hill, M. J., Killick, R., Navarrete, K., Maruszak, A., McLaughlin, G. M., Williams, B. P., & Bray, N. J. (2017, May). Knockdown of the schizophrenia susceptibility gene TCF4 alters gene expression and proliferation of progenitor cells from the developing human neocortex. *J Psychiatry Neurosci, 42*(3), 181-188. https://doi.org/10.1503/jpn.160073

Hirschhorn, J. N., & Daly, M. J. (2005, 2005/02/01). Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics, 6*(2), 95-108. https://doi.org/10.1038/nrg1521

Hitzeroth, A., Niehaus, D. J., Koen, L., Botes, W. C., Deleuze, J., & Warnich, L. (2007). Association between the MnSOD Ala-9Val polymorphism and development of schizophrenia and abnormal involuntary movements in the Xhosa population. *Progress in Neuro-Psychopharmacology and Biological Psychiatry, 31*(3), 664-672.

Hnisz, D., Abraham, B. J., Lee, T. I., Lau, A., Saint-André, V., Sigova, A. A., Hoke, H. A., & Young, R. A. (2013, Nov 7). Super-enhancers in the control of cell identity and disease. *Cell, 155*(4), 934-947. https://doi.org/10.1016/j.cell.2013.09.053

Hodgson, J. A., Mulligan, C. J., Al-Meeri, A., & Raaum, R. L. (2014). Early back-to-Africa migration into the Horn of Africa. *PLoS Genet, 10*(6), e1004393.

Hoffman, M. M., Ernst, J., Wilder, S. P., Kundaje, A., Harris, R. S., Libbrecht, M., Giardine, B., Ellenbogen, P. M., Bilmes, J. A., Birney, E., Hardison, R. C., Dunham, I., Kellis, M., & Noble, W. S. (2013, Jan). Integrative annotation of chromatin elements from ENCODE data. *Nucleic Acids Res, 41*(2), 827-841. https://doi.org/10.1093/nar/gks1284

Hoppe, T. A., Litovitz, A., Willis, K. A., Meseroll, R. A., Perkins, M. J., Hutchins, B. I., Davis, A. F., Lauer, M. S., Valantine, H. A., & Anderson, J. M. (2019). Topic choice contributes to the lower rate of NIH awards to African-American/black scientists. *Science advances, 5*(10), eaaw7238.

Hovatta, I., Varilo, T., Suvisaari, J., Terwilliger, J. D., Ollikainen, V., Arajarvi, R., Juvonen, H., Kokko-Sahin, M. L., Vaisanen, L., Mannila, H., Lonnqvist, J., & Peltonen, L. (1999, Oct). A genomewide screen for schizophrenia genes in an isolated Finnish subpopulation, suggesting multiple susceptibility loci. *Am J Hum Genet, 65*(4), 1114-1124. https://doi.org/10.1086/302567

Howrigan, D. (2017). Details and Considerations of the UK Biobank GWAS

Huang, L., Ohi, K., Chang, H., Yu, H., Wu, L., Yue, W., Zhang, D., Gao, L. and Li, M. (2016). A comprehensive meta-analysis of ZNF804A SNPs in the risk of schizophrenia among Asian populations. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics, 171*(3), 437-446.

Hudson, K., Lifton, R., & Patrick-Lake, B. (2015). The precision medicine initiative cohort program—Building a Research Foundation for 21st Century Medicine. *Precision Medicine Initiative (PMI) Working Group Report to the Advisory Committee to the Director, ed*.

Ikeda, M., Aleksic, B., Kinoshita, Y., Okochi, T., Kawashima, K., Kushima, I., Ito, Y., Nakamura, Y., Kishi, T., Okumura, T., Fukuo, Y., Williams, H. J., Hamshere, M. L., Ivanov, D., Inada, T., Suzuki, M., Hashimoto, R., Ujike, H., Takeda, M., Craddock, N., Kaibuchi, K., Owen, M. J., Ozaki, N., O'Donovan, M. C., & Iwata, N. (2011, Mar 1). Genome-wide association study of schizophrenia in a Japanese population. *Biol Psychiatry, 69*(5), 472-478. https://doi.org/10.1016/j.biopsych.2010.07.010

Ikeda, M., Takahashi, A., Kamatani, Y., Momozawa, Y., Saito, T., Kondo, K., Shimasaki, A., Kawase, K., Sakusabe, T., Iwayama, Y., Toyota, T., Wakuda, T., Kikuchi, M., Kanahara, N., Yamamori, H., Yasuda, Y., Watanabe, Y., Hoya, S., Aleksic, B., Kushima, I., Arai, H., Takaki, M., Hattori, K., Kunugi, H., Okahisa, Y., Ohnuma, T., Ozaki, N., Someya, T., Hashimoto, R., Yoshikawa, T., Kubo, M., & Iwata, N. (2019, Jun 18). Genome-Wide Association Study Detected Novel Susceptibility Genes for Schizophrenia and Shared Trans-Populations/Diseases Genetic Effect. *Schizophr Bull, 45*(4), 824-834. https://doi.org/10.1093/schbul/sby140

International Schizophrenia Consortium. (2009, 08/06/print). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder [10.1038/nature08185]. *Nature, 460*(7256), 748-752. https://doi.org/http://www.nature.com/nature/journal/v460/n7256/suppinfo/nature08185_S1.html

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS medicine, 2*(8), e124.

Jansen, I. E., Savage, J. E., Watanabe, K., Bryois, J., Williams, D. M., Steinberg, S., Sealock, J., Karlsson, I. K., Hägg, S., Athanasiu, L., Voyle, N., Proitsi, P., Witoelar, A., Stringer, S., Aarsland, D., Almdahl, I. S., Andersen, F., Bergh, S., Bettella, F., Bjornsson, S., Brækhus, A., Bråthen, G., de Leeuw, C., Desikan, R. S., Djurovic, S., Dumitrescu, L., Fladby, T., Hohman, T. J., Jonsson, P. V., Kiddle, S. J., Rongve, A., Saltvedt, I., Sando, S. B., Selbæk, G., Shoai, M., Skene, N. G., Snaedal, J., Stordal, E., Ulstein, I. D., Wang, Y., White, L. R., Hardy, J., Hjerling-Leffler, J., Sullivan, P. F., van der Flier, W. M., Dobson, R., Davis, L. K., Stefansson, H., Stefansson, K., Pedersen, N. L., Ripke, S., Andreassen, O. A., & Posthuma, D. (2019, Mar). Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. *Nat Genet, 51*(3), 404-413. https://doi.org/10.1038/s41588-018-0311-9

Jayatilleke, N., Hayes, R. D., Dutta, R., Shetty, H., Hotopf, M., Chang, C. K., & Stewart, R. (2017, Jun). Contributions of specific causes of death to lost life expectancy in severe mental illness. *Eur Psychiatry, 43*, 109-115. https://doi.org/10.1016/j.eurpsy.2017.02.487

Jensen, S. K. G., Berens, A. E., & Nelson, C. A., 3rd. (2017, Nov). Effects of poverty on interacting biological systems underlying child development. *Lancet Child Adolesc Health, 1*(3), 225-239. https://doi.org/10.1016/s2352-4642(17)30024-x

Jeste, D. V., Palmer, B. W., Appelbaum, P. S., Golshan, S., Glorioso, D., Dunn, L. B., Kim, K., Meeks, T., & Kraemer, H. C. (2007). A new brief instrument for assessing decisional capacity for clinical research. *Archives of General Psychiatry, 64*(8), 966-974.

Kahlbaum, K. L. (1874). Clinische abhandlungen einige psychische krankheiten. I. Katatonia oder das spannungsirresein.

Kantrowitz, J. T., Nolan, K. A., Sen, S., Simen, A. A., Lachman, H. M., & Bowers, M. B. (2009). Adolescent cannabis use, psychosis and catechol-O-methyltransferase genotype in African Americans and Caucasians. *Psychiatric quarterly, 80*(4), 213-218. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2888694/pdf/nihms-138485.pdf

Karayiorgou, M., Morris, M. A., Morrow, B., Shprintzen, R. J., Goldberg, R., Borrow, J., Gos, A., Nestadt, G., Wolyniec, P. S., & Lasseter, V. K. (1995). Schizophrenia susceptibility associated with interstitial deletions of chromosome 22q11. *Proceedings of the National Academy of Sciences, 92*(17), 7612-7616. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC41195/pdf/pnas01495-0015.pdf

Karlsson, H., Dal, H., Gardner, R. M., Torrey, E. F., & Dalman, C. (2019, 2019/09/01/). Birth month and later diagnosis of schizophrenia. A population-based cohort study in Sweden. *J Psychiatr Res, 116*, 1-6. https://doi.org/https://doi.org/10.1016/j.jpsychires.2019.05.025

Kay, S. R., Fiszbein, A., & Opler, L. A. (1987). The positive and negative syndrome scale (PANSS) for schizophrenia. *Schizophr Bull, 13*(2), 261-276.

Kendler, K., Aggen, S., Li, Y., Lewis, C., Breen, G., Boomsma, D., Bot, M., Penninx, B., & Flint, J. (2015). The similarity of the structure of DSM-IV criteria for major depression in depressed women from China, the United States and Europe. *Psychological Medicine, 45*(9), 1945-1954.

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., & Haussler, D. (2002, Jun). The human genome browser at UCSC. *Genome Res, 12*(6), 996-1006. https://doi.org/10.1101/gr.229102

Kępińska, A. P., Iyegbe, C. O., Vernon, A. C., Yolken, R., Murray, R. M., & Pollak, T. A. (2020). Schizophrenia and influenza at the centenary of the 1918-1919 Spanish influenza pandemic: mechanisms of psychosis risk. *Frontiers in Psychiatry, 11*, 72.

Kerminen, S., Martin, A. R., Koskela, J., Ruotsalainen, S. E., Havulinna, A. S., Surakka, I., Palotie, A., Perola, M., Salomaa, V., Daly, M. J., Ripatti, S., & Pirinen, M. (2019, Jun 6). Geographic Variation and Bias in the Polygenic Scores of Complex Diseases and Traits in Finland. *Am J Hum Genet, 104*(6), 1169-1181. https://doi.org/10.1016/j.ajhg.2019.05.001

Kessler, R. C., Sonnega, A., Bromet, E., Hughes, M., & Nelson, C. B. (1995, Dec). Posttraumatic stress disorder in the National Comorbidity Survey. *Arch Gen Psychiatry, 52*(12), 1048-1060. https://doi.org/10.1001/archpsyc.1995.03950240066012

Khashan, A. S., Abel, K. M., McNamee, R., Pedersen, M. G., Webb, R. T., Baker, P. N., Kenny, L. C., & Mortensen, P. B. (2008). Higher Risk of Offspring Schizophrenia Following Antenatal Maternal Exposure to Severe Adverse Life Events. *Archives of General Psychiatry, 65*(2), 146-152. https://doi.org/10.1001/archgenpsychiatry.2007.20

Kilian, S., Burns, J., Seedat, S., Asmal, L., Chiliza, B., Du Plessis, S., Olivier, M., Kidd, M., & Emsley, R. (2017). Factors moderating the relationship between childhood trauma and premorbid adjustment in first-episode schizophrenia. *Plos One, 12*(1), e0170178. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5249082/pdf/pone.0170178.pdf

Kim, J., Park, C., Choi, J., Park, E., Tchoe, H., Choi, M., Suh, J., Kim, Y., Won, S., & Chung, Y. (2017). The association between season of birth, age at onset, and clozapine use in schizophrenia. *Acta Psychiatr Scand, 136*(5), 445-454. https://onlinelibrary.wiley.com/doi/abs/10.1111/acps.12776

Kim, T. H., & Moon, S. W. (2011). Serum homocysteine and folate levels in korean schizophrenic patients. *Psychiatry investigation, 8*(2), 134. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3149108/pdf/pi-8-134.pdf

Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M., & Shendure, J. (2014, Mar). A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet, 46*(3), 310-315. https://doi.org/10.1038/ng.2892

Kirkbride, J. B., Errazuriz, A., Croudace, T. J., Morgan, C., Jackson, D., Boydell, J., Murray, R. M., & Jones, P. B. (2012). Incidence of schizophrenia and other psychoses in England, 1950–2009: a systematic review and meta-analyses. *Plos One, 7*(3), e31660. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3310436/pdf/pone.0031660.pdf

Kirov, G., Rees, E., Walters, J. T., Escott-Price, V., Georgieva, L., Richards, A. L., Chambert, K. D., Davies, G., Legge, S. E., Moran, J. L., McCarroll, S. A., O'Donovan, M. C., & Owen, M. J. (2014, Mar 1). The penetrance of copy number variations for schizophrenia and developmental delay. *Biol Psychiatry, 75*(5), 378-385. https://doi.org/10.1016/j.biopsych.2013.07.022

Kitsios, G., & Zintzaras, E. (2009, 07/29). Genome-Wide Association Studies: hypothesis-"free" or "engaged"? *Translational research : the journal of laboratory and clinical medicine, 154*(4), 161-164. https://doi.org/10.1016/j.trsl.2009.07.001

Klei, L., Bacanu, S. A., Myles-Worsley, M., Galke, B., Xie, W., Tiobech, J., Otto, C., Roeder, K., Devlin, B., & Byerley, W. (2005, Aug). Linkage analysis of a completely ascertained sample of familial schizophrenics and bipolars from Palau, Micronesia. *Hum Genet, 117*(4), 349-356. https://doi.org/10.1007/s00439-005-1320-1

Klemm, S. L., Shipony, Z., & Greenleaf, W. J. (2019, 2019/04/01). Chromatin accessibility and the regulatory epigenome. *Nature Reviews Genetics, 20*(4), 207-220. https://doi.org/10.1038/s41576-018-0089-8

Koen, L., Niehaus, D. J., Wright, G., Warnich, L., De Jong, G., Emsley, R. A., & Mall, S. (2012, Feb 23). Chromosome 22q11 in a Xhosa schizophrenia population. *S Afr Med J, 102*(3 Pt 1), 165-166. https://core.ac.uk/download/37349150.pdf

Konrath, L., Beckius, D., & Tran, U. S. (2016). Season of birth and population schizotypy: results from a large sample of the adult general population. *Psychiatry Research, 242*, 245-250. https://www.sciencedirect.com/science/article/abs/pii/S016517811530161X?via%3Dihub

Kouzarides, T. (2007, 2007/02/23/). Chromatin Modifications and Their Function. *Cell, 128*(4), 693-705. https://doi.org/https://doi.org/10.1016/j.cell.2007.02.005

Kraepelin, E. (1971). *Dementia praecox and paraphrenia*. Krieger Publishing Company.

Kraft, P., Zeggini, E., & Ioannidis, J. P. (2009). Replication in genome-wide association studies. *Statistical Science: A review journal of the Institute of Mathematical Statistics, 24*(4), 561.

Kuchenbaecker, K., Reiker, T., Gilly, A., Gurdasani, D., Prins, B., Suveges, D., Southam, L., Asiki, G., Seeley, J., & Kamali, A. (2019). Associations of polygenic scores with lipid biomarkers in diverse populations. European Journal of Human Genetics,

Kugathasan, P., Stubbs, B., Aagaard, J., Jensen, S. E., Munk Laursen, T., & Nielsen, R. E. (2019). Increased mortality from somatic multimorbidity in patients with schizophrenia: a Danish nationwide cohort study. *Acta Psychiatr Scand, 140*(4), 340-348.

Labonté, B., Suderman, M., Maussion, G., Navaro, L., Yerko, V., Mahar, I., Bureau, A., Mechawar, N., Szyf, M., Meaney, M. J., & Turecki, G. (2012, Jul). Genome-wide epigenetic regulation by early-life trauma. *Arch Gen Psychiatry, 69*(7), 722-731. https://doi.org/10.1001/archgenpsychiatry.2011.2287

Lam, M., Awasthi, S., Watson, H. J., Goldstein, J., Panagiotaropoulou, G., Trubetskoy, V., Karlsson, R., Frei, O., Fan, C.-C., De Witte, W., Mota, N. R., Mullins, N., Brügger, K., Lee, S. H., Wray, N. R., Skarabis, N., Huang, H., Neale, B., Daly, M. J., Mattheisen, M., Walters, R., & Ripke, S. (2019). RICOPILI: Rapid Imputation for COnsortias PIpeLIne. *Bioinformatics, 36*(3), 930-933. https://doi.org/10.1093/bioinformatics/btz633

Lam, M., Chen, C.-Y., Li, Z., Martin, A. R., Bryois, J., Ma, X., Gaspar, H., Ikeda, M., Benyamin, B., Brown, B. C., Liu, R., Zhou, W., Guan, L., Kamatani, Y., Kim, S.-W., Kubo, M., Kusumawardhani, A. A. A. A., Liu, C.-M., Ma, H., Periyasamy, S., Takahashi, A., Xu, Z., Yu, H., Zhu, F., Schizophrenia Working Group of the Psychiatric Genomics, C., Indonesia Schizophrenia, C., Genetic, R. o. s. n.-C., the, N., Chen, W. J., Faraone, S., Glatt, S. J., He, L., Hyman, S. E., Hwu, H.-G., McCarroll, S. A., Neale, B. M., Sklar, P., Wildenauer, D. B., Yu, X., Zhang, D., Mowry, B. J., Lee, J., Holmans, P., Xu, S., Sullivan, P. F., Ripke, S., O'Donovan, M. C., Daly, M. J., Qin, S., Sham, P., Iwata, N., Hong, K. S., Schwab, S. G., Yue, W., Tsuang, M., Liu, J., Ma, X., Kahn, R. S., Shi, Y., & Huang, H. (2019). Comparative genetic architectures of schizophrenia in East Asian and European populations. *Nature Genetics, 51*(12), 1670-1678. https://doi.org/10.1038/s41588-019-0512-x

Lan, K.-C., Chiang, H.-J., Huang, T.-L., Chiou, Y.-J., Hsu, T.-Y., Ou, Y.-C., & Yang, Y.-H. (2020). Association between paternal age and risk of schizophrenia: a nationwide population–based study. *Journal of assisted reproduction and genetics*, 1-9.

Larsson, S., Andreassen, O. A., Aas, M., Røssberg, J. I., Mork, E., Steen, N. E., Barrett, E. A., Lagerberg, T. V., Peleikis, D., & Agartz, I. (2013). High prevalence of childhood trauma in patients with schizophrenia spectrum and affective disorder. *Comprehensive Psychiatry, 54*(2), 123-127. https://www.sciencedirect.com/science/article/abs/pii/S0010440X12001216?via%3Dihub

Laurent, C., Niehaus, D., Bauché, S., Levinson, D. F., Soubigou, S., Pimstone, S., Hayden, M., Mbanga, I., Emsley, R., & Deleuze, J. F. (2003). CAG repeat polymorphisms in KCNN3 (HSKCa3) and PPP2R2B show no association or linkage to schizophrenia. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics, 116*(1), 45-50.

Lee, K. W., San Woon, P., Teo, Y. Y., & Sim, K. (2012). Genome wide association studies (GWAS) and copy number variation (CNV) studies of the major psychoses: what have we learnt? *Neuroscience & Biobehavioral Reviews, 36*(1), 556-571.

Lee, S., Yang, J., Goddard, M., Visscher, P., & Wray, N. (2012). Estimation of pleiotropy between complex diseases using SNP-derived genomic relationships and restricted maximum likelihood. *Bioinformatics, 28*(19), 2540-2542.

Lee, S. H., DeCandia, T. R., Ripke, S., Yang, J., Sullivan, P. F., Goddard, M. E., Keller, M. C., Visscher, P. M., & Wray, N. R. (2012, Feb 19). Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. *Nat Genet, 44*(3), 247-250. https://doi.org/10.1038/ng.1108

Lee, S. H., Ripke, S., Neale, B. M., Faraone, S. V., Purcell, S. M., Perlis, R. H., Mowry, B. J., Thapar, A., Goddard, M. E., Witte, J. S., Absher, D., Agartz, I., Akil, H., Amin, F., Andreassen, O. A., Anjorin, A., Anney, R., Anttila, V., Arking, D. E., Asherson, P., Azevedo, M. H., Backlund, L., Badner, J. A., Bailey, A. J., Banaschewski, T., Barchas, J. D., Barnes, M. R., Barrett, T. B., Bass, N., Battaglia, A., Bauer, M., Bayes, M., Bellivier, F., Bergen, S. E., Berrettini, W., Betancur, C., Bettecken, T., Biederman, J., Binder, E. B., Black, D. W., Blackwood, D. H., Bloss, C. S., Boehnke, M., Boomsma, D. I., Breen, G., Breuer, R., Bruggeman, R., Cormican, P., Buccola, N. G., Buitelaar, J. K., Bunney, W. E., Buxbaum, J. D., Byerley, W. F., Byrne, E. M., Caesar, S., Cahn, W., Cantor, R. M., Casas, M., Chakravarti, A., Chambert, K., Choudhury, K., Cichon, S., Cloninger, C. R., Collier, D. A., Cook, E. H., Coon, H., Cormand, B., Corvin, A., Coryell, W. H., Craig, D. W., Craig, I. W., Crosbie, J., Cuccaro, M. L., Curtis, D., Czamara, D., Datta, S., Dawson, G., Day, R., De Geus, E. J., Degenhardt, F., Djurovic, S., Donohoe, G. J., Doyle, A. E., Duan, J., Dudbridge, F., Duketis, E., Ebstein, R. P., Edenberg, H. J., Elia, J., Ennis, S., Etain, B., Fanous, A., Farmer, A. E., Ferrier, I. N., Flickinger, M., Fombonne, E., Foroud, T., Frank, J., Franke, B., Fraser, C., Freedman, R., Freimer, N. B., Freitag, C. M., Friedl, M., Frisen, L., Gallagher, L., Gejman, P. V., Georgieva, L., Gershon, E. S., Geschwind, D. H., Giegling, I., Gill, M., Gordon, S. D., Gordon-Smith, K., Green, E. K., Greenwood, T. A., Grice, D. E., Gross, M., Grozeva, D., Guan, W., Gurling, H., De Haan, L., Haines, J. L., Hakonarson, H., Hallmayer, J., Hamilton, S. P., Hamshere, M. L., Hansen, T. F., Hartmann, A. M., Hautzinger, M., Heath, A. C., Henders, A. K., Herms, S., Hickie, I. B., Hipolito, M., Hoefels, S., Holmans, P. A., Holsboer, F., Hoogendijk, W. J., Hottenga, J. J., Hultman, C. M., Hus, V., Ingason, A., Ising, M., Jamain, S., Jones, E. G., Jones, I., Jones, L., Tzeng, J. Y., Kahler, A. K., Kahn, R. S., Kandaswamy, R., Keller, M. C., Kennedy, J. L., Kenny, E., Kent, L., Kim, Y., Kirov, G. K., Klauck, S. M., Klei, L., Knowles, J. A., Kohli, M. A., Koller, D. L., Konte, B., Korszun, A., Krabbendam, L., Krasucki, R., Kuntsi, J., Kwan, P., Landen, M., Langstrom, N., Lathrop, M., Lawrence, J., Lawson, W. B., Leboyer, M., Ledbetter, D. H., Lee, P. H., Lencz, T., Lesch, K. P., Levinson, D. F., Lewis, C. M., Li, J., Lichtenstein, P., Lieberman, J. A., Lin, D. Y., Linszen, D. H., Liu, C., Lohoff, F. W., Loo, S. K., Lord, C., Lowe, J. K., Lucae, S., MacIntyre, D. J., Madden, P. A., Maestrini, E., Magnusson, P. K., Mahon, P. B., Maier, W., Malhotra, A. K., Mane, S. M., Martin, C. L., Martin, N. G., Mattheisen, M., Matthews, K., Mattingsdal, M., McCarroll, S. A., McGhee, K. A., McGough, J. J., McGrath, P. J., McGuffin, P., McInnis, M. G., McIntosh, A., McKinney, R., McLean, A. W., McMahon, F. J., McMahon, W. M., McQuillin, A., Medeiros, H., Medland, S. E., Meier, S., Melle, I., Meng, F., Meyer, J., Middeldorp, C. M., Middleton, L., Milanova, V., Miranda, A., Monaco, A. P., Montgomery, G. W., Moran, J. L., Moreno-De-Luca, D.,

Morken, G., Morris, D. W., Morrow, E. M., Moskvina, V., Muglia, P., Muhleisen, T. W., Muir, W. J., Muller-Myhsok, B., Murtha, M., Myers, R. M., Myin-Germeys, I., Neale, M. C., Nelson, S. F., Nievergelt, C. M., Nikolov, I., Nimgaonkar, V., Nolen, W. A., Nothen, M. M., Nurnberger, J. I., Nwulia, E. A., Nyholt, D. R., O'Dushlaine, C., Oades, R. D., Olincy, A., Oliveira, G., Olsen, L., Ophoff, R. A., Osby, U., Owen, M. J., Palotie, A., Parr, J. R., Paterson, A. D., Pato, C. N., Pato, M. T., Penninx, B. W., Pergadia, M. L., Pericak-Vance, M. A., Pickard, B. S., Pimm, J., Piven, J., Posthuma, D., Potash, J. B., Poustka, F., Propping, P., Puri, V., Quested, D. J., Quinn, E. M., Ramos-Quiroga, J. A., Rasmussen, H. B., Raychaudhuri, S., Rehnstrom, K., Reif, A., Ribases, M., Rice, J. P., Rietschel, M., Roeder, K., Roeyers, H., Rossin, L., Rothenberger, A., Rouleau, G., Ruderfer, D., Rujescu, D., Sanders, A. R., Sanders, S. J., Santangelo, S. L., Sergeant, J. A., Schachar, R., Schalling, M., Schatzberg, A. F., Scheftner, W. A., Schellenberg, G. D., Scherer, S. W., Schork, N. J., Schulze, T. G., Schumacher, J., Schwarz, M., Scolnick, E., Scott, L. J., Shi, J., Shilling, P. D., Shyn, S. I., Silverman, J. M., Slager, S. L., Smalley, S. L., Smit, J. H., Smith, E. N., Sonuga-Barke, E. J., St Clair, D., State, M., Steffens, M., Steinhausen, H. C., Strauss, J. S., Strohmaier, J., Stroup, T. S., Sutcliffe, J. S., Szatmari, P., Szelinger, S., Thirumalai, S., Thompson, R. C., Todorov, A. A., Tozzi, F., Treutlein, J., Uhr, M., van den Oord, E. J., Van Grootheest, G., Van Os, J., Vicente, A. M., Vieland, V. J., Vincent, J. B., Visscher, P. M., Walsh, C. A., Wassink, T. H., Watson, S. J., Weissman, M. M., Werge, T., Wienker, T. F., Wijsman, E. M., Willemsen, G., Williams, N., Willsey, A. J., Witt, S. H., Xu, W., Young, A. H., Yu, T. W., Zammit, S., Zandi, P. P., Zhang, P., Zitman, F. G., Zollner, S., Devlin, B., Kelsoe, J. R., Sklar, P., Daly, M. J., O'Donovan, M. C., Craddock, N., Sullivan, P. F., Smoller, J. W., Kendler, K. S., & Wray, N. R. (2013, Sep). Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nat Genet, 45*(9), 984-994. https://doi.org/10.1038/ng.2711

Lee, S. H., Wray, N. R., Goddard, M. E., & Visscher, P. M. (2011, Mar 11). Estimating missing heritability for disease from genome-wide association studies. *Am J Hum Genet, 88*(3), 294-305. https://doi.org/10.1016/j.ajhg.2011.02.002

Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., O'Donnell-Luria, A. H., Ware, J. S., Hill, A. J., Cummings, B. B., Tukiainen, T., Birnbaum, D. P., Kosmicki, J. A., Duncan, L. E., Estrada, K., Zhao, F., Zou, J., Pierce-Hoffman, E., Berghout, J., Cooper, D. N., Deflaux, N., DePristo, M., Do, R., Flannick, J., Fromer, M., Gauthier, L., Goldstein, J., Gupta, N., Howrigan, D., Kiezun, A., Kurki, M. I., Moonshine, A. L., Natarajan, P., Orozco, L., Peloso, G. M., Poplin, R., Rivas, M. A., Ruano-Rubio, V., Rose, S. A., Ruderfer, D. M., Shakir, K., Stenson, P. D., Stevens, C., Thomas, B. P., Tiao, G., Tusie-Luna, M. T., Weisburd, B., Won, H. H., Yu, D., Altshuler, D. M., Ardissino, D., Boehnke, M., Danesh, J., Donnelly, S., Elosua, R., Florez, J. C., Gabriel, S. B., Getz, G., Glatt, S. J., Hultman, C. M., Kathiresan, S., Laakso, M., McCarroll, S., McCarthy, M. I., McGovern, D., McPherson, R., Neale, B. M., Palotie, A., Purcell, S. M., Saleheen, D., Scharf, J. M., Sklar, P., Sullivan, P. F., Tuomilehto, J., Tsuang, M. T., Watkins, H. C., Wilson, J. G., Daly, M. J., & MacArthur, D. G. (2016, Aug 18). Analysis of protein-coding genetic variation in 60,706 humans. *Nature, 536*(7616), 285-291. https://doi.org/10.1038/nature19057

Lesh, T. A., Careaga, M., Rose, D. R., McAllister, A. K., Van de Water, J., Carter, C. S., & Ashwood, P. (2018, 2018/05/26). Cytokine alterations in first-episode schizophrenia and bipolar disorder: relationships to brain structure and symptoms. *Journal of Neuroinflammation, 15*(1), 165. https://doi.org/10.1186/s12974-018-1197-2

Lettre, G., Sankaran, V. G., Bezerra, M. A. C., Araújo, A. S., Uda, M., Sanna, S., Cao, A., Schlessinger, D., Costa, F. F., & Hirschhorn, J. N. (2008). DNA polymorphisms at the BCL11A, HBS1L-MYB, and β-globin loci associate with fetal hemoglobin levels and pain crises in sickle cell disease. *Proceedings of the National Academy of Sciences, 105*(33), 11869-11874.

Li, H., & Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. *Nature, 475*(7357), 493-496.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009, Aug 15). The Sequence Alignment/Map format and SAMtools. *Bioinformatics, 25*(16), 2078-2079. https://doi.org/10.1093/bioinformatics/btp352

Li, J. Z., Absher, D. M., Tang, H., Southwick, A. M., Casto, A. M., Ramachandran, S., Cann, H. M., Barsh, G. S., Feldman, M., & Cavalli-Sforza, L. L. (2008). Worldwide human relationships inferred from genome-wide patterns of variation. *Science, 319*(5866), 1100-1104.

Li, X.-B., Bo, Q.-J., Zhang, G.-P., Zheng, W., Wang, Z.-M., Li, A.-N., Tian, Q., Liu, J.-T., Tang, Y.-L., & Wang, C.-Y. (2017). Effect of childhood trauma on cognitive functions in a sample of Chinese patients with schizophrenia. *Comprehensive Psychiatry, 76*, 147-152. https://www.sciencedirect.com/science/article/abs/pii/S0010440X16307064?via%3Dihub

Li, Z., Chen, J., Yu, H., He, L., Xu, Y., Zhang, D., Yi, Q., Li, C., Li, X., Shen, J., Song, Z., Ji, W., Wang, M., Zhou, J., Chen, B., Liu, Y., Wang, J., Wang, P., Yang, P., Wang, Q., Feng, G., Liu, B., Sun, W., Li, B., He, G., Li, W., Wan, C., Xu, Q., Li, W., Wen, Z., Liu, K., Huang, F., Ji, J., Ripke, S., Yue, W., Sullivan, P. F., O'Donovan, M. C., & Shi, Y. (2017, Nov). Genome-wide association analysis identifies 30 new susceptibility loci for schizophrenia. *Nat Genet, 49*(11), 1576-1583. https://doi.org/10.1038/ng.3973

Li, Z., Shen, T., Xin, R., Liang, B., Jiang, J., Ling, W., Wei, B., & Su, L. (2017, Apr). Association of NKAPL, TSPAN18, and MPC2 gene variants with schizophrenia based on new data and a meta-analysis in Han Chinese. *Acta Neuropsychiatr, 29*(2), 87-94. https://doi.org/10.1017/neu.2016.36

Li, Z., Xiang, Y., Chen, J., Li, Q., Shen, J., Liu, Y., Li, W., Xing, Q., Wang, Q., Wang, L., Feng, G., He, L., Zhao, X., & Shi, Y. (2015, Dec). Loci with genome-wide associations with schizophrenia in the Han Chinese population. *Br J Psychiatry, 207*(6), 490-494. https://doi.org/10.1192/bjp.bp.114.150490

Lindblad-Toh, K., Garber, M., Zuk, O., Lin, M. F., Parker, B. J., Washietl, S., Kheradpour, P., Ernst, J., Jordan, G., Mauceli, E., Ward, L. D., Lowe, C. B., Holloway, A. K., Clamp, M., Gnerre, S., Alföldi, J., Beal, K., Chang, J., Clawson, H., Cuff, J., Di Palma, F., Fitzgerald, S., Flicek, P., Guttman, M., Hubisz, M. J., Jaffe, D. B., Jungreis, I., Kent, W. J., Kostka, D., Lara, M., Martins, A. L., Massingham, T., Moltke, I., Raney, B. J., Rasmussen, M. D., Robinson, J., Stark, A., Vilella, A. J., Wen, J., Xie, X., Zody, M. C., Baldwin, J., Bloom, T., Chin, C. W., Heiman, D., Nicol, R., Nusbaum, C., Young, S., Wilkinson, J., Worley, K. C., Kovar, C. L., Muzny, D. M., Gibbs, R. A., Cree, A., Dihn, H. H., Fowler, G., Jhangiani, S., Joshi, V., Lee, S., Lewis, L. R., Nazareth, L. V., Okwuonu, G., Santibanez, J., Warren, W. C., Mardis, E. R., Weinstock, G. M., Wilson, R. K., Delehaunty, K., Dooling, D., Fronik, C., Fulton, L., Fulton, B., Graves, T., Minx, P., Sodergren, E., Birney, E., Margulies, E. H., Herrero, J., Green, E. D., Haussler, D., Siepel, A., Goldman, N., Pollard, K. S., Pedersen, J. S., Lander, E. S., & Kellis, M. (2011, Oct 12). A high-resolution map of human evolutionary constraint using 29 mammals. *Nature, 478*(7370), 476-482. https://doi.org/10.1038/nature10530

Loh, P.-R., Danecek, P., Palamara, P. F., Fuchsberger, C., Reshef, Y. A., Finucane, H. K., Schoenherr, S., Forer, L., McCarthy, S., & Abecasis, G. R. (2016). Reference-based phasing using the Haplotype Reference Consortium panel. *Nature Genetics, 48*(11), 1443-1448.

Loh, P. R., Bhatia, G., Gusev, A., Finucane, H. K., Bulik-Sullivan, B. K., Pollack, S. J., de Candia, T. R., Lee, S. H., Wray, N. R., Kendler, K. S., O'Donovan, M. C., Neale, B. M., Patterson, N., & Price, A. L. (2015, Dec). Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. *Nat Genet, 47*(12), 1385-1392. https://doi.org/10.1038/ng.3431

Lupiáñez, D. G., Kraft, K., Heinrich, V., Krawitz, P., Brancati, F., Klopocki, E., Horn, D., Kayserili, H., Opitz, J. M., & Laxova, R. (2015). Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell, 161*(5), 1012-1025.

Maeng, S.-R., Kim, W.-H., Kim, J.-H., Bae, J.-N., Lee, J.-S., & Kim, C.-E. (2016). Factors Affecting Quality of Life and Family Burden among the Families of Patients with Schizophrenia. *Korean Journal of Schizophrenia Research, 19*(2), 78-88.

Mall, S., Platt, J. M., Temmingh, H., Musenge, E., Campbell, M., Susser, E., & Stein, D. J. (2019, Aug 7). The relationship between childhood trauma and schizophrenia in the Genomics of Schizophrenia in the Xhosa people (SAX) study in South Africa. *Psychol Med*, 1-8. https://doi.org/10.1017/s0033291719001703

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., & Chakravarti, A. (2009). Finding the missing heritability of complex diseases. *Nature, 461*(7265), 747-753.

Marcellusi, A., Fabiano, G., Viti, R., Francesa Morel, P. C., Nicolò, G., Siracusano, A., & Mennini, F. S. (2018, Feb 8). Economic burden of schizophrenia in Italy: a probabilistic cost of illness analysis. *Bmj Open, 8*(2), e018359. https://doi.org/10.1136/bmjopen-2017-018359

Marshall, C. R., Howrigan, D. P., Merico, D., Thiruvahindrapuram, B., Wu, W., Greer, D. S., Antaki, D., Shetty, A., Holmans, P. A., Pinto, D., Gujral, M., Brandler, W. M., Malhotra, D., Wang, Z., Fajarado, K. V. F., Maile, M. S., Ripke, S., Agartz, I., Albus, M., Alexander, M., Amin, F., Atkins, J., Bacanu, S. A., Belliveau, R. A., Jr., Bergen, S. E., Bertalan, M., Bevilacqua, E., Bigdeli, T. B., Black, D. W., Bruggeman, R., Buccola, N. G., Buckner, R. L., Bulik-Sullivan, B., Byerley, W., Cahn, W., Cai, G., Cairns, M. J., Campion, D., Cantor, R. M., Carr, V. J., Carrera, N., Catts, S. V., Chambert, K. D., Cheng, W., Cloninger, C. R., Cohen, D., Cormican, P., Craddock, N., Crespo-Facorro, B., Crowley, J. J., Curtis, D., Davidson, M., Davis, K. L., Degenhardt, F., Del Favero, J., DeLisi, L. E., Dikeos, D., Dinan, T., Djurovic, S., Donohoe, G., Drapeau, E., Duan, J., Dudbridge, F., Eichhammer, P., Eriksson, J., Escott-Price, V., Essioux, L., Fanous, A. H., Farh, K. H., Farrell, M. S., Frank, J., Franke, L., Freedman, R., Freimer, N. B., Friedman, J. I., Forstner, A. J., Fromer, M., Genovese, G., Georgieva, L., Gershon, E. S., Giegling, I., Giusti-Rodriguez, P., Godard, S., Goldstein, J. I., Gratten, J., de Haan, L., Hamshere, M. L., Hansen, M., Hansen, T., Haroutunian, V., Hartmann, A. M., Henskens, F. A., Herms, S., Hirschhorn, J. N., Hoffmann, P., Hofman, A., Huang, H., Ikeda, M., Joa, I., Kahler, A. K., Kahn, R. S., Kalaydjieva, L., Karjalainen, J., Kavanagh, D., Keller, M. C., Kelly, B. J., Kennedy, J. L., Kim, Y., Knowles, J. A., Konte, B., Laurent, C., Lee, P., Lee, S. H., Legge, S. E., Lerer, B., Levy, D. L., Liang, K. Y., Lieberman, J., Lonnqvist, J., Loughland, C. M., Magnusson, P. K. E., Maher, B. S., Maier, W., Mallet, J., Mattheisen, M., Mattingsdal, M., McCarley, R. W., McDonald, C., McIntosh, A. M., Meier, S., Meijer, C. J., Melle, I., Mesholam-Gately, R. I., Metspalu, A., Michie, P. T., Milani, L., Milanova, V., Mokrab, Y., Morris, D. W., Muller-Myhsok, B., Murphy, K. C., Murray, R. M., Myin-Germeys, I., Nenadic, I., Nertney, D. A., Nestadt, G., Nicodemus, K. K., Nisenbaum, L., Nordin, A., O'Callaghan, E., O'Dushlaine, C., Oh, S. Y., Olincy, A., Olsen, L., O'Neill, F. A., Van Os, J., Pantelis, C., Papadimitriou, G. N., Parkhomenko, E., Pato, M. T., Paunio, T., Perkins, D. O., Pers, T. H., Pietilainen, O., Pimm, J., Pocklington, A. J., Powell, J., Price, A., Pulver, A. E., Purcell, S. M., Quested, D., Rasmussen, H. B., Reichenberg, A., Reimers, M. A., Richards, A. L., Roffman, J. L., Roussos, P., Ruderfer, D. M., Salomaa, V., Sanders, A. R., Savitz, A., Schall, U., Schulze, T. G., Schwab, S. G., Scolnick, E. M., Scott, R. J., Seidman, L. J., Shi, J., Silverman, J. M., Smoller, J. W., Soderman, E., Spencer, C. C. A., Stahl, E. A., Strengman, E., Strohmaier, J., Stroup, T. S., Suvisaari, J., Svrakic, D. M., Szatkiewicz, J. P., Thirumalai, S., Tooney, P. A., Veijola, J., Visscher, P. M., Waddington, J., Walsh, D., Webb, B. T., Weiser, M., Wildenauer, D. B., Williams, N. M., Williams, S., Witt, S. H., Wolen, A. R., Wormley, B. K., Wray, N. R., Wu, J. Q., Zai, C. C., Adolfsson, R., Andreassen, O. A., Blackwood, D. H. R., Bramon, E., Buxbaum, J. D., Cichon, S., Collier, D. A., Corvin, A., Daly, M. J., Darvasi, A., Domenici, E., Esko, T., Gejman, P. V., Gill, M., Gurling, H., Hultman, C. M., Iwata, N., Jablensky, A. V., Jonsson, E. G., Kendler, K. S., Kirov, G., Knight, J., Levinson, D. F., Li, Q. S., McCarroll, S. A., McQuillin, A., Moran, J. L., Mowry, B. J., Nothen, M. M., Ophoff, R. A., Owen, M. J., Palotie, A., Pato, C. N., Petryshen, T. L., Posthuma, D., Rietschel, M., Riley, B. P., Rujescu, D., Sklar, P., St Clair, D., Walters, J. T. R., Werge, T., Sullivan, P. F., O'Donovan, M. C., Scherer, S. W., Neale, B. M., & Sebat, J. (2017, Jan). Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects. *Nat Genet, 49*(1), 27-35. https://doi.org/10.1038/ng.3725

Martin, A. R., Atkinson, E. G., Chapman, S. B., Stevenson, A., Stroud, R. E., Abebe, T., Akena, D., Alemayehu, M., Ashaba, F. K., & Atwoli, L. (2020). Low-coverage sequencing cost-effectively detects known and novel variation in underrepresented populations. *bioRxiv*.

Martin, A. R., Daly, M. J., Robinson, E. B., Hyman, S. E., & Neale, B. M. (2019, Jul 15). Predicting Polygenic Risk of Psychiatric Disorders. *Biol Psychiatry, 86*(2), 97-109. https://doi.org/10.1016/j.biopsych.2018.12.015

Martin, A. R., Gignoux, C. R., Walters, R. K., Wojcik, G. L., Neale, B. M., Gravel, S., Daly, M. J., Bustamante, C. D., & Kenny, E. E. (2017, Apr 6). Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. *Am J Hum Genet, 100*(4), 635-649. https://doi.org/10.1016/j.ajhg.2017.03.004

Martin, A. R., Kanai, M., Kamatani, Y., Okada, Y., Neale, B. M., & Daly, M. J. (2019, Apr). Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat Genet, 51*(4), 584-591. https://doi.org/10.1038/s41588-019-0379-x

Martin, A. R., Teferra, S., Moller, M., Hoal, E. G., & Daly, M. J. (2018, Dec). The critical needs and challenges for genetic architecture studies in Africa. *Curr Opin Genet Dev, 53*, 113-120. https://doi.org/10.1016/j.gde.2018.08.005

McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A. R., Teumer, A., Kang, H. M., Fuchsberger, C., Danecek, P., Sharp, K., Luo, Y., Sidore, C., Kwong, A., Timpson, N., Koskinen, S., Vrieze, S., Scott, L. J., Zhang, H., Mahajan, A., Veldink, J., Peters, U., Pato, C., van Duijn, C. M., Gillies, C. E., Gandin, I., Mezzavilla, M., Gilly, A., Cocca, M., Traglia, M., Angius, A., Barrett, J. C., Boomsma, D., Branham, K., Breen, G., Brummett, C. M., Busonero, F., Campbell, H., Chan, A., Chen, S., Chew, E., Collins, F. S., Corbin, L. J., Smith, G. D., Dedoussis, G., Dorr, M., Farmaki, A. E., Ferrucci, L., Forer, L., Fraser, R. M., Gabriel, S., Levy, S., Groop, L., Harrison, T., Hattersley, A., Holmen, O. L., Hveem, K., Kretzler, M., Lee, J. C., McGue, M., Meitinger, T., Melzer, D., Min, J. L., Mohlke, K. L., Vincent, J. B., Nauck, M., Nickerson, D., Palotie, A., Pato, M., Pirastu, N., McInnis, M., Richards, J. B., Sala, C., Salomaa, V., Schlessinger, D., Schoenherr, S., Slagboom, P. E., Small, K., Spector, T., Stambolian, D., Tuke, M., Tuomilehto, J., Van den Berg, L. H., Van Rheenen, W., Volker, U., Wijmenga, C., Toniolo, D., Zeggini, E., Gasparini, P., Sampson, M. G., Wilson, J. F., Frayling, T., de Bakker, P. I., Swertz, M. A., McCarroll, S., Kooperberg, C., Dekker, A., Altshuler, D., Willer, C., Iacono, W., Ripatti, S., Soranzo, N., Walter, K., Swaroop, A., Cucca, F., Anderson, C. A., Myers, R. M., Boehnke, M., McCarthy, M. I., & Durbin, R. (2016, Oct). A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet, 48*(10), 1279-1283. https://doi.org/10.1038/ng.3643

McClellan, J., & King, M.-C. (2010, 2010/04/16/). Genetic Heterogeneity in Human Disease. *Cell, 141*(2), 210-217. https://doi.org/https://doi.org/10.1016/j.cell.2010.03.032

McGrath, J. J., Petersen, L., Agerbo, E., Mors, O., Mortensen, P. B., & Pedersen, C. B. (2014, Mar). A comprehensive assessment of parental age and psychiatric disorders. *JAMA Psychiatry, 71*(3), 301-309. https://doi.org/10.1001/jamapsychiatry.2013.4081

McRae, A. F. (2017). Analysis of Genome-Wide Association Data. In J. M. Keith (Ed.), *Bioinformatics: Volume II: Structure, Function, and Applications* (pp. 161-173). Springer New York. https://doi.org/10.1007/978-1-4939-6613-4_9

Medina-Gomez, C., Felix, J. F., Estrada, K., Peters, M. J., Herrera, L., Kruithof, C. J., Duijts, L., Hofman, A., van Duijn, C. M., & Uitterlinden, A. G. (2015). Challenges in conducting genome-wide association studies in highly admixed multi-ethnic populations: the Generation R Study. *European journal of epidemiology, 30*(4), 317-330.

Merico, D., Costain, G., Butcher, N. J., Warnica, W., Ogura, L., Alfred, S. E., Brzustowicz, L. M., & Bassett, A. S. (2014). MicroRNA dysregulation, gene networks, and risk for schizophrenia in 22q11. 2 deletion syndrome. *Frontiers in neurology, 5*, 238.

Miller, S. A., Dykes, D. D., & Polesky, H. F. (1988). A simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic Acids Research, 16*(3), 1215. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC334765/

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC334765/pdf/nar00145-0424.pdf

Moises, H. W., Yang, L., Kristbjarnarson, H., Wiese, C., Byerley, W., Macciardi, F., Arolt, V., Blackwood, D., Liu, X., Sjogren, B., & et al. (1995, Nov). An international two-stage genome-wide search for schizophrenia susceptibility genes. *Nat Genet, 11*(3), 321-324. https://doi.org/10.1038/ng1195-321

Morales, J., Welter, D., Bowler, E. H., Cerezo, M., Harris, L. W., McMahon, A. C., Hall, P., Junkins, H. A., Milano, A., & Hastings, E. (2018). A standardized framework for representation of ancestry data in genomics studies, with application to the NHGRI-EBI GWAS Catalog. *Genome Biol, 19*(1), 21.

Morel, B.-A. (1860). *Traité des maladies mentales*. Victor Masson.

Morris, A. P., Voight, B. F., Teslovich, T. M., Ferreira, T., Segre, A. V., Steinthorsdottir, V., Strawbridge, R. J., Khan, H., Grallert, H., & Mahajan, A. (2012). Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nature Genetics, 44*(9), 981.

Mortensen, P., & Kyvik, K. (1996). Increased occurrence of schizophrenia and other psychiatric illnesses among twins. *The British Journal of Psychiatry, 168*(6), 688-692.

Mortensen, P. B., Pedersen, M. G., & Pedersen, C. B. (2010, Feb). Psychiatric family history and schizophrenia risk in Denmark: which mental disorders are relevant? *Psychol Med, 40*(2), 201-210. https://doi.org/10.1017/s0033291709990419

Mostafavi, H., Harpak, A., Agarwal, I., Conley, D., Pritchard, J. K., & Przeworski, M. (2020). Variable prediction accuracy of polygenic scores within an ancestry group. *Elife, 9*, e48376.

Moustafa, A. A., Hewedi, D. H., Eissa, A. M., Frydecka, D., & Misiak, B. (2014). Homocysteine levels in schizophrenia and affective disorders—focus on cognition. *Frontiers in behavioral neuroscience, 8*, 343. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4186289/pdf/fnbeh-08-00343.pdf

Mugisha, J. O., Baisley, K., Asiki, G., Seeley, J., & Kuper, H. (2013). Prevalence, types, risk factors and clinical correlates of anaemia in older people in a rural Ugandan population. *Plos One, 8*(10), e78394.

Mugisha, J. O., Schatz, E. J., Randell, M., Kuteesa, M., Kowal, P., Negin, J., & Seeley, J. (2016). Chronic disease, risk factors and disability in adults aged 50 and above living with and without HIV: findings from the Wellbeing of Older People Study in Uganda. *Global health action, 9*(1), 31098.

Mulder, N., Abimiku, A., Adebamowo, S. N., de Vries, J., Matimba, A., Olowoyo, P., Ramsay, M., Skelton, M., & Stein, D. J. (2018). H3Africa: current perspectives. *Pharmgenomics Pers Med, 11*, 59-66. https://doi.org/10.2147/pgpm.S141546

Mustonen, A., Niemelä, S., Nordström, T., Murray, G. K., Mäki, P., Jääskeläinen, E., & Miettunen, J. (2018). Adolescent cannabis use, baseline prodromal symptoms and the risk of psychosis. *The British Journal of Psychiatry, 212*(4), 227-233. https://www.cambridge.org/core/services/aop-cambridge-core/content/view/D5CAA12A5F424146DABB9C6A6AB4CB56/S0007125017000526a.pdf/div-

class-title-adolescent-cannabis-use-baseline-prodromal-symptoms-and-the-risk-of-psychosis-div.pdf

Mwesiga, E. K., Akena, D., Koen, N., Senono, R., Obuku, E. A., Gumikiriza, J. L., Robbins, R. N., Nakasujja, N., & Stein, D. J. (2020, 2020/12/01/). A systematic review of research on neuropsychological measures in psychotic disorders from low and middle-income countries: The question of clinical utility. *Schizophrenia Research: Cognition, 22*, 100187. https://doi.org/https://doi.org/10.1016/j.scog.2020.100187

Nagai, A., Hirata, M., Kamatani, Y., Muto, K., Matsuda, K., Kiyohara, Y., Ninomiya, T., Tamakoshi, A., Yamagata, Z., & Mushiroda, T. (2017). Overview of the BioBank Japan Project: study design and profile. *Journal of epidemiology, 27*(Supplement_III), S2-S8.

Nalwanga, D., Musiime, V., Kizito, S., Kiggundu, J. B., Batte, A., Musoke, P., & Tumwine, J. K. (2020). Mortality among children under five years admitted for routine care of severe acute malnutrition: a prospective cohort study from Kampala, Uganda. *BMC Pediatrics, 20*, 1-11.

Nelson, S. C., Romm, J. M., Doheny, K. F., Pugh, E. W., & Laurie, C. C. (2017). Imputation-based genomic coverage assessments of current genotyping arrays: Illumina HumanCore, OmniExpress, Multi-Ethnic global array and sub-arrays, Global Screening Array, Omni2. 5M, Omni5M, and Affymetrix UK Biobank. *bioRxiv*, 150219.

Ng, S. B., Turner, E. H., Robertson, P. D., Flygare, S. D., Bigham, A. W., Lee, C., Shaffer, T., Wong, M., Bhattacharjee, A., & Eichler, E. E. (2009). Targeted capture and massively parallel sequencing of 12 human exomes. *Nature, 461*(7261), 272-276.

Nica, A. C., Montgomery, S. B., Dimas, A. S., Stranger, B. E., Beazley, C., Barroso, I., & Dermitzakis, E. T. (2010, Apr 1). Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. *PLoS Genet, 6*(4), e1000895. https://doi.org/10.1371/journal.pgen.1000895

Nicolae, D. L., Gamazon, E., Zhang, W., Duan, S., Dolan, M. E., & Cox, N. J. (2010, Apr 1). Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet, 6*(4), e1000888. https://doi.org/10.1371/journal.pgen.1000888

Niehaus, D. J., Oosthuizen, P., Lochner, C., Emsley, R. A., Jordaan, E., Mbanga, N. I., Keyter, N., Laurent, C., Deleuze, J. F., & Stein, D. J. (2004, Mar-Apr). A culture-bound syndrome 'amafufunyana' and a culture-specific event 'ukuthwasa': differentiated by a family history of schizophrenia and other psychiatric disorders. *Psychopathology, 37*(2), 59-63. https://doi.org/10.1159/000077579

Nieman, D. H., Dragt, S., van Duin, E. D., Denneman, N., Overbeek, J. M., de Haan, L., Rietdijk, J., Ising, H. K., Klaassen, R. M., & van Amelsvoort, T. (2016). COMT Val158Met genotype and cannabis use in people with an At Risk Mental State for psychosis: Exploring Gene x Environment interactions. *Schizophrenia Research, 174*(1-3), 24-28. https://www.sciencedirect.com/science/article/abs/pii/S0920996416301086?via%3Dihub

Novembre, J., & Barton, N. H. (2018). Tread lightly interpreting polygenic tests of selection. *Genetics, 208*(4), 1351.

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science, 366*(6464), 447-453.

Oloniniyi, I. O., Akinsulore, A., Aloba, O. O., Mapayi, B. M., Oginni, O. A., & Makanjuola, R. (2019, Jan-Mar). Economic Cost of Schizophrenia in a Nigerian Teaching Hospital. *J Neurosci Rural Pract, 10*(1), 39-47. https://doi.org/10.4103/jnrp.jnrp_210_18

Pagani, L., Schiffels, S., Gurdasani, D., Danecek, P., Scally, A., Chen, Y., Xue, Y., Haber, M., Ekong, R., & Oljira, T. (2015). Tracing the route of modern humans out of Africa by using 225 human genome sequences from Ethiopians and Egyptians. *The American Journal of Human Genetics, 96*(6), 986-991.

Pardiñas, A. F., Holmans, P., Pocklington, A. J., Escott-Price, V., Ripke, S., Carrera, N., Legge, S. E., Bishop, S., Cameron, D., Hamshere, M. L., Han, J., Hubbard, L., Lynham, A., Mantripragada, K., Rees, E., MacCabe, J. H., McCarroll, S. A., Baune, B. T., Breen, G., Byrne, E. M., Dannlowski, U., Eley, T. C., Hayward, C., Martin, N. G., McIntosh, A. M., Plomin, R., Porteous, D. J., Wray, N. R., Caballero, A., Geschwind, D. H., Huckins, L. M., Ruderfer, D. M., Santiago, E., Sklar, P., Stahl, E. A., Won, H., Agerbo, E., Als, T. D., Andreassen, O. A., Bækvad-Hansen, M., Mortensen, P. B., Pedersen, C. B., Børglum, A. D., Bybjerg-Grauholm, J., Djurovic, S., Durmishi, N., Pedersen, M. G., Golimbet, V., Grove, J., Hougaard, D. M., Mattheisen, M., Molden, E., Mors, O., Nordentoft, M., Pejovic-Milovancevic, M., Sigurdsson, E., Silagadze, T., Hansen, C. S., Stefansson, K., Stefansson, H., Steinberg, S., Tosato, S., Werge, T., Harold, D., Sims, R., Gerrish, A., Chapman, J., Escott-Price, V., Abraham, R., Hollingworth, P., Pahwa, J., Denning, N., Thomas, C., Taylor, S., Powell, J., Proitsi, P., Lupton, M., Lovestone, S., Passmore, P., Craig, D., McGuinness, B., Johnston, J., Todd, S., Maier, W., Jessen, F., Heun, R., Schurmann, B., Ramirez, A., Becker, T., Herold, C., Lacour, A., Drichel, D., Nothen, M., Goate, A., Cruchaga, C., Nowotny, P., Morris, J. C., Mayo, K., Holmans, P., O'Donovan, M., Owen, M., Williams,

J., Achilla, E., Agerbo, E., Barr, C. L., Böttger, T. W., Breen, G., Cohen, D., Collier, D. A., Curran, S., Dempster, E., Dima, D., Sabes-Figuera, R., Flanagan, R. J., Frangou, S., Frank, J., Gasse, C., Gaughran, F., Giegling, I., Grove, J., Hannon, E., Hartmann, A. M., Heißerer, B., Helthuis, M., Horsdal, H. T., Ingimarsson, O., Jollie, K., Kennedy, J. L., Köhler, O., Konte, B., Lang, M., Legge, S. E., Lewis, C., MacCabe, J., Malhotra, A. K., McCrone, P., Meier, S. M., Mill, J., Mors, O., Mortensen, P. B., Nöthen, M. M., O'Donovan, M. C., Owen, M. J., Pardiñas, A. F., Pedersen, C. B., Rietschel, M., Rujescu, D., Schwalber, A., Sigurdsson, E., Sørensen, H. J., Spencer, B., Stefansson, H., Støvring, H., Strohmaier, J., Sullivan, P., Vassos, E., Verbelen, M., Walters, J. T. R., Werge, T., Collier, D. A., Rujescu, D., Kirov, G., Owen, M. J., O'Donovan, M. C., Walters, J. T. R., Consortium, G., & Consortium, C. (2018, 2018/03/01). Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. *Nature Genetics, 50*(3), 381-389. https://doi.org/10.1038/s41588-018-0059-2

Park, J.-H., Gail, M. H., Weinberg, C. R., Carroll, R. J., Chung, C. C., Wang, Z., Chanock, S. J., Fraumeni, J. F., & Chatterjee, N. (2011). Distribution of allele frequencies and effect sizes and their interrelationships for common genetic susceptibility variants. *Proceedings of the National Academy of Sciences, 108*(44), 18026-18031.

Parker, S. C., Stitzel, M. L., Taylor, D. L., Orozco, J. M., Erdos, M. R., Akiyama, J. A., van Bueren, K. L., Chines, P. S., Narisu, N., & Black, B. L. (2013). Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. *Proceedings of the National Academy of Sciences, 110*(44), 17921-17926.

Pasquali, L., Gaulton, K. J., Rodríguez-Seguí, S. A., Mularoni, L., Miguel-Escalada, I., Akerman, I., Tena, J. J., Morán, I., Gómez-Marín, C., & Van De Bunt, M. (2014). Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants. *Nature Genetics, 46*(2), 136-143.

Paunio, T., Ekelund, J., Varilo, T., Parker, A., Hovatta, I., Turunen, J. A., Rinard, K., Foti, A., Terwilliger, J. D., Juvonen, H., Suvisaari, J., Arajärvi, R., Suokas, J., Partonen, T., Lönnqvist, J., Meyer, J., & Peltonen, L. (2001, Dec 15). Genome-wide scan in a nationwide study sample of schizophrenia families in Finland reveals susceptibility loci on chromosomes 2q and 5q. *Hum Mol Genet, 10*(26), 3037-3048. https://doi.org/10.1093/hmg/10.26.3037

Pazokitoroudi, A., Wu, Y., Burch, K. S., Hou, K., Zhou, A., Pasaniuc, B., & Sankararaman, S. (2020, 2020/08/11). Efficient variance components analysis across millions of genomes. *Nature Communications, 11*(1), 4020. https://doi.org/10.1038/s41467-020-17576-9

Pellet, J., Golay, P., Nguyen, A., Suter, C., Ismailaj, A., Bonsack, C., & Favrod, J. (2019, May). The relationship between self-stigma and depression among people with schizophrenia-spectrum disorders: A longitudinal study. *Psychiatry Res, 275*, 115-119. https://doi.org/10.1016/j.psychres.2019.03.022

Perälä, J., Suvisaari, J., Saarni, S. I., Kuoppasalmi, K., Isometsä, E., Pirkola, S., Partonen, T., Tuulio-Henriksson, A., Hintikka, J., Kieseppä, T., Härkänen, T., Koskinen, S., & Lönnqvist, J. (2007, Jan). Lifetime prevalence of psychotic and bipolar I disorders in a general population. *Arch Gen Psychiatry, 64*(1), 19-28. https://doi.org/10.1001/archpsyc.64.1.19

Petersen, L., & Sørensen, T. I. (2011). The Danish adoption register. *Scandinavian Journal of Public Health, 39*(7_suppl), 83-86.

Petersen, L., & Sørensen, T. I. A. (2011). Studies based on the Danish Adoption Register: Schizophrenia, BMI, smoking, and mortality in perspective [Review]. *Scandinavian Journal of Public Health, 39*(7), 191-195. https://doi.org/10.1177/1403494810396560

Peterson, R. E., Kuchenbaecker, K., Walters, R. K., Chen, C. Y., Popejoy, A. B., Periyasamy, S., Lam, M., Iyegbe, C., Strawbridge, R. J., Brick, L., Carey, C. E., Martin, A. R., Meyers, J. L., Su, J., Chen, J., Edwards, A. C., Kalungi, A., Koen, N., Majara, L., Schwarz, E., Smoller, J. W., Stahl, E. A., Sullivan, P. F., Vassos, E., Mowry, B., Prieto, M. L., Cuellar-Barboza, A., Bigdeli, T. B., Edenberg, H. J., Huang, H., & Duncan, L. E. (2019, Oct 7). Genome-wide Association Studies in Ancestrally Diverse Populations: Opportunities, Methods, Pitfalls, and Recommendations. *Cell*. https://doi.org/10.1016/j.cell.2019.08.051

Petrovski, S., Gussow, A. B., Wang, Q., Halvorsen, M., Han, Y., Weir, W. H., Allen, A. S., & Goldstein, D. B. (2015). The intolerance of regulatory sequence to genetic variation predicts gene dosage sensitivity. *PLoS Genet, 11*(9), e1005492.

Phillipson, D. W. (2005). *African archaeology*. Cambridge University Press.

Popejoy, A. B., & Fullerton, S. M. (2016, Oct 13). Genomics is failing on diversity. *Nature, 538*(7624), 161-164. https://doi.org/10.1038/538161a

Psychiatric, G. C. B. D. W. G. (2011, Sep 18). Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. *Nat Genet, 43*(10), 977-983. https://doi.org/10.1038/ng.943

Psychiatric Genomics Consortium, S. (2011, Sep 18). Genome-wide association study identifies five new schizophrenia loci. *Nat Genet, 43*(10), 969-976. https://doi.org/10.1038/ng.940

Pugliese, V., Bruni, A., Carbone, E. A., Calabrò, G., Cerminara, G., Sampogna, G., Luciano, M., Steardo Jr, L., Fiorillo, A., & Garcia, C. S. (2019). Maternal stress, prenatal medical illnesses and obstetric complications: risk factors for schizophrenia spectrum disorder, bipolar disorder and major depressive disorder. *Psychiatry Research, 271*, 23-30. https://www.sciencedirect.com/science/article/abs/pii/S0165178118306012?via%3Dihub

Purcell, S., Cherny, S. S., & Sham, P. C. (2003, Jan). Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits. *Bioinformatics, 19*(1), 149-150.

Purcell, S. M., Moran, J. L., Fromer, M., Ruderfer, D., Solovieff, N., Roussos, P., O'Dushlaine, C., Chambert, K., Bergen, S. E., Kähler, A., Duncan, L., Stahl, E., Genovese, G., Fernández, E., Collins, M. O., Komiyama, N. H., Choudhary, J. S., Magnusson, P. K. E., Banks, E., Shakir, K., Garimella, K., Fennell, T., DePristo, M., Grant, S. G. N., Haggarty, S. J., Gabriel, S., Scolnick, E. M., Lander, E. S., Hultman, C. M., Sullivan, P. F., McCarroll, S. A., & Sklar, P. (2014, 2014/02/01). A polygenic burden of rare disruptive mutations in schizophrenia. *Nature, 506*(7487), 185-190. https://doi.org/10.1038/nature12975

Quintana-Murci, L., Semino, O., Bandelt, H. J., Passarino, G., McElreavey, K., & Santachiara-Benerecetti, A. S. (1999, Dec). Genetic evidence of an early exit of Homo sapiens sapiens from Africa through eastern Africa. *Nat Genet, 23*(4), 437-441. https://doi.org/10.1038/70550

Rajman, M., & Schratt, G. (2017). MicroRNAs in neural development: from master regulators to fine-tuners. *Development, 144*(13), 2310-2322.

Rasic, D., Hajek, T., Alda, M., & Uher, R. (2014, Jan). Risk of mental illness in offspring of parents with schizophrenia, bipolar disorder, and major depressive disorder: a meta-analysis of family high-risk studies. *Schizophr Bull, 40*(1), 28-38. https://doi.org/10.1093/schbul/sbt114

Raznahan, A., Parikshak, N. N., Chandran, V., Blumenthal, J. D., Clasen, L. S., Alexander-Bloch, A. F., Zinn, A. R., Wangsa, D., Wise, J., & Murphy, D. G. (2018). Sex-chromosome dosage effects on gene expression in humans. *Proceedings of the National Academy of Sciences, 115*(28), 7398-7403.

Reed, F. A., & Tishkoff, S. A. (2006, Dec). African human diversity, origins and migrations. *Curr Opin Genet Dev, 16*(6), 597-605. https://doi.org/10.1016/j.gde.2006.10.008

Rees, E., Han, J., Morgan, J., Carrera, N., Escott-Price, V., Pocklington, A. J., Duffield, M., Hall, L. S., Legge, S. E., Pardiñas, A. F., Richards, A. L., Roth, J., Lezheiko, T., Kondratyev, N., Kaleda, V., Golimbet, V., Parellada, M., González-Peñas, J., Arango, C., Alizadeh, B. Z., van Amelsvoort, T., Bruggeman, R., Cahn, W., de Haan, L., Luykx, J. J., Rutten, B. P. F., van Os, J., van Winkel, R., Gawlik, M., Kirov, G., Walters, J. T. R., Holmans, P., O'Donovan, M. C., Owen, M. J., & Investigators, G. (2020, 2020/02/01). De novo mutations identified by exome sequencing implicate rare missense variants in SLC6A1 in schizophrenia. *Nat Neurosci, 23*(2), 179-184. https://doi.org/10.1038/s41593-019-0565-2

Rees, E., Kirov, G., O'Donovan, M. C., & Owen, M. J. (2012, May). De novo mutation in schizophrenia. *Schizophr Bull, 38*(3), 377-381. https://doi.org/10.1093/schbul/sbs047

Rees, E., Kirov, G., Sanders, A., Walters, J. T. R., Chambert, K. D., Shi, J., Szatkiewicz, J., O'Dushlaine, C., Richards, A. L., Green, E. K., Jones, I., Davies, G., Legge, S. E., Moran, J. L., Pato, C., Pato, M., Genovese, G., Levinson, D., Duan, J., Moy, W., Göring, H. H. H., Morris, D., Cormican, P., Kendler, K. S., O'Neill, F. A., Riley, B., Gill, M., Corvin, A., Craddock, N., Sklar, P., Hultman, C., Sullivan, P. F., Gejman, P. V., McCarroll, S. A., O'Donovan, M. C., & Owen, M. J. (2014). Evidence that duplications of 22q11.2 protect against schizophrenia [Article]. *Mol Psychiatry, 19*(1), 37-40. https://doi.org/10.1038/mp.2013.156

Rees, E., Walters, J. T. R., Georgieva, L., Isles, A. R., Chambert, K. D., Richards, A. L., Mahoney-Davies, G., Legge, S. E., Moran, J. L., McCarroll, S. A., O'Donovan, M. C., Owen, M. J., & Kirov, G. (2014). Analysis of copy number variations at 15 schizophrenia-associated loci. *Br J Psychiatry, 204*(2), 108-114. https://doi.org/10.1192/bjp.bp.113.131052

Reich, D. E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P. C., Richter, D. J., Lavery, T., Kouyoumjian, R., Farhadian, S. F., Ward, R., & Lander, E. S. (2001, 05/10/print). Linkage disequilibrium in the human genome [10.1038/35075590]. *Nature, 411*(6834), 199-204. http://dx.doi.org/10.1038/35075590

http://www.nature.com/nature/journal/v411/n6834/pdf/411199a0.pdf

Riley, B., Mogudi-Carter, M., Jenkins, T., & Williamson, R. (1996, Nov 22). No evidence for linkage of chromosome 22 markers to schizophrenia in southern African Bantu-speaking families. *Am J Med Genet, 67*(6), 515-522. https://doi.org/10.1002/(sici)1096-8628(19961122)67:6<515::aid-ajmg2>3.0.co;2-g

Riley, B. P., Tahir, E., Rajagopalan, S., Mogudi-Carter, M., Faure, S., Weissenbach, J., Jenkins, T., & Williamson, R. (1997, Summer). A linkage study of the N-methyl-D-aspartate receptor subunit gene loci and schizophrenia in southern African Bantu-speaking families. *Psychiatr Genet, 7*(2), 57-74.

Ripke, S., O'Dushlaine, C., Chambert, K., Moran, J. L., Kahler, A. K., Akterin, S., Bergen, S. E., Collins, A. L., Crowley, J. J., Fromer, M., Kim, Y., Lee, S. H., Magnusson, P. K., Sanchez, N., Stahl, E. A., Williams, S., Wray, N. R., Xia, K., Bettella, F., Borglum, A. D., Bulik-Sullivan, B. K., Cormican, P., Craddock, N., de Leeuw, C., Durmishi, N., Gill, M., Golimbet, V., Hamshere, M. L., Holmans, P., Hougaard, D. M., Kendler, K. S., Lin, K., Morris, D. W., Mors, O., Mortensen, P. B., Neale, B. M., O'Neill, F. A., Owen, M. J., Milovancevic, M. P., Posthuma, D., Powell, J., Richards, A. L., Riley, B. P., Ruderfer, D., Rujescu, D., Sigurdsson, E., Silagadze, T., Smit, A. B., Stefansson, H., Steinberg, S., Suvisaari, J., Tosato, S., Verhage, M., Walters, J. T., Levinson, D. F., Gejman, P. V., Kendler, K. S., Laurent, C., Mowry, B. J., O'Donovan, M. C., Owen, M. J., Pulver, A. E., Riley, B. P., Schwab, S. G., Wildenauer, D. B., Dudbridge, F., Holmans, P., Shi, J., Albus, M., Alexander, M., Campion, D., Cohen, D., Dikeos, D., Duan, J., Eichhammer, P., Godard, S., Hansen, M., Lerer, F. B., Liang, K. Y., Maier, W., Mallet, J., Nertney, D. A., Nestadt, G., Norton, N., O'Neill, F. A., Papadimitriou, G. N., Ribble, R., Sanders, A. R., Silverman, J. M., Walsh, D., Williams, N. M., Wormley, B., Arranz, M. J., Bakker, S., Bender, S., Bramon, E., Collier, D., Crespo-Facorro, B., Hall, J., Iyegbe, C., Jablensky, A., Kahn, R. S., Kalaydjieva, L., Lawrie, S., Lewis, C. M., Lin, K., Linszen, D. H., Mata, I., McIntosh, A., Murray, R. M., Ophoff, R. A., Powell, J., Rujescu, D., Van Os, J., Walshe, M., Weisbrod, M., Wiersma, D., Donnelly, P., Barroso, I., Blackwell, J. M., Bramon, E., Brown, M. A., Casas, J. P., Corvin, A. P., Deloukas, P., Duncanson, A., Jankowski, J., Markus, H. S., Mathew, C. G., Palmer, C. N., Plomin, R., Rautanen, A., Sawcer, S. J., Trembath, R. C., Viswanathan, A. C., Wood, N. W., Spencer, C. C., Band, G., Bellenguez, C., Freeman, C., Hellenthal, G., Giannoulatou, E., Pirinen, M., Pearson, R. D., Strange, A., Su, Z., Vukcevic, D., Donnelly, P., Langford, C., Hunt, S. E., Edkins, S., Gwilliam, R., Blackburn, H., Bumpstead, S. J., Dronov, S., Gillman, M., Gray, E., Hammond, N., Jayakumar, A., McCann, O. T., Liddle, J., Potter, S. C., Ravindrarajah, R., Ricketts, M., Tashakkori-Ghanbaria, A., Waller, M. J., Weston, P., Widaa, S., Whittaker, P., Barroso, I., Deloukas, P., Mathew, C. G., Blackwell, J. M., Brown, M. A., Corvin, A. P., McCarthy, M. I., Spencer, C. C., Bramon, E., Corvin, A. P., O'Donovan, M. C., Stefansson, K., Scolnick, E., Purcell, S., McCarroll, S. A., Sklar, P., Hultman, C. M., & Sullivan, P. F. (2013, Oct). Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nat Genet, 45*(10), 1150-1159. https://doi.org/10.1038/ng.2742

Ripke, S., Walters, J. T., O'Donovan, M. C., & Consortium, S. W. G. o. t. P. G. (2020). Mapping genomic loci prioritises genes and implicates synaptic biology in schizophrenia. *medRxiv*.

Roemer, L., Litz, B. T., Orsillo, S. M., Ehlich, P. J., & Friedman, M. J. (1998, Jul). Increases in retrospective accounts of war-zone exposure over time: the role of PTSD symptom severity. *J Trauma Stress, 11*(3), 597-605. https://doi.org/10.1023/a:1024469116047

Roser, M., Ortiz-Ospina, E., & Ritchie, H. (2013). Life expectancy. *Our World in Data*.

Rotimi, C. N., Bentley, A. R., Doumatey, A. P., Chen, G., Shriner, D., & Adeyemo, A. (2017). The genomic landscape of African populations in health and disease. *Human Molecular Genetics, 26*(R2), R225-R236.

Rubinacci, S., Delaneau, O., & Marchini, J. (2020). Genotype imputation using the Positional Burrows Wheeler Transform. *bioRxiv*, 797944. https://doi.org/10.1101/797944

Ryckman, K., & Williams, S. M. (2008). Calculation and use of the Hardy-Weinberg model in association studies. *Current Protocols in Human Genetics, 57*(1), 1.18. 11-11.18. 11.

Saha, S., Chant, D., & McGrath, J. (2007). A systematic review of mortality in schizophrenia: is the differential mortality gap worsening over time? *Archives of General Psychiatry, 64*(10), 1123-1131.

Salas, A., Richards, M., De la Fe, T., Lareu, M.-V., Sobrino, B., Sánchez-Diz, P., Macaulay, V., & Carracedo, Á. (2002). The making of the African mtDNA landscape. *The American Journal of Human Genetics, 71*(5), 1082-1111.

Sanchez-Gistau, V., Romero, S., Moreno, D., de la Serna, E., Baeza, I., Sugranyes, G., Moreno, C., Sanchez-Gutierrez, T., Rodriguez-Toscano, E., & Castro-Fornieles, J. (2015, Oct). Psychiatric disorders in child and adolescent offspring of patients with schizophrenia and bipolar disorder: A controlled study. *Schizophr Res, 168*(1-2), 197-203. https://doi.org/10.1016/j.schres.2015.08.034

Sánchez-Quinto, F., Botigué, L. R., Civit, S., Arenas, C., Ávila-Arcos, M. C., Bustamante, C. D., Comas, D., & Lalueza-Fox, C. (2012). North African populations carry the signature of admixture with Neandertals. *Plos One, 7*(10), e47765.

Sanders, S. J., Ercan-Sencicek, A. G., Hus, V., Luo, R., Murtha, M. T., Moreno-De-Luca, D., Chu, S. H., Moreau, M. P., Gupta, A. R., Thomson, S. A., Mason, C. E., Bilguvar, K., Celestino-Soper, P. B., Choi, M., Crawford, E. L., Davis, L., Wright, N. R., Dhodapkar, R. M., DiCola, M., DiLullo, N. M., Fernandez, T. V., Fielding-Singh, V., Fishman, D. O., Frahm, S., Garagaloyan, R., Goh, G. S., Kammela, S., Klei, L., Lowe, J. K., Lund, S. C., McGrew, A. D., Meyer, K. A., Moffat, W. J., Murdoch, J. D., O'Roak, B. J., Ober, G. T., Pottenger, R. S., Raubeson, M. J., Song, Y., Wang, Q., Yaspan, B. L., Yu, T. W., Yurkiewicz, I. R., Beaudet, A. L., Cantor, R. M., Curland, M., Grice, D. E., Gunel, M., Lifton, R. P., Mane, S. M., Martin, D. M., Shaw, C. A., Sheldon, M., Tischfield, J. A., Walsh, C. A., Morrow, E. M., Ledbetter, D. H., Fombonne, E., Lord, C., Martin, C. L., Brooks, A. I., Sutcliffe, J. S., Cook, E. H., Jr., Geschwind, D., Roeder, K., Devlin, B., & State, M. W. (2011, Jun 9). Multiple recurrent de novo CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. *Neuron, 70*(5), 863-885. https://doi.org/10.1016/j.neuron.2011.05.002

Sartorius, N., & Janca, A. (1996). Psychiatric assessment instruments developed by the World Health Organization. *Social psychiatry and psychiatric epidemiology, 31*(2), 55-69.

Schizophrenia Working Group of the Psychiatric Genomics, C. (2014, Jul 24). Biological insights from 108 schizophrenia-associated genetic loci. *Nature, 511*(7510), 421-427. https://doi.org/10.1038/nature13595

Schlebusch, C. M., Skoglund, P., Sjodin, P., Gattepaille, L. M., Hernandez, D., Jay, F., Li, S., De Jongh, M., Singleton, A., Blum, M. G., Soodyall, H., & Jakobsson, M. (2012, Oct 19). Genomic variation in seven Khoe-San groups reveals adaptation and complex African history. *Science, 338*(6105), 374-379. https://doi.org/10.1126/science.1227721

Schneider, K. (1959). Clinical psychopathology.(Trans. by MW Hamilton).

Schoech, A. P., Jordan, D. M., Loh, P.-R., Gazal, S., O'Connor, L. J., Balick, D. J., Palamara, P. F., Finucane, H. K., Sunyaev, S. R., & Price, A. L. (2019). Quantification of frequency-dependent genetic architectures in 25 UK Biobank traits reveals action of negative selection. *Nature Communications, 10*(1), 1-10.

Schomerus, G., Schwahn, C., Holzinger, A., Corrigan, P. W., Grabe, H. J., Carta, M. G., & Angermeyer, M. C. (2012, Jun). Evolution of public attitudes about mental illness: a systematic review and meta-analysis. *Acta Psychiatr Scand, 125*(6), 440-452. https://doi.org/10.1111/j.1600-0447.2012.01826.x

Schurz, H., Müller, S. J., van Helden, P. D., Tromp, G., Hoal, E. G., Kinnear, C. J., & Möller, M. (2019, 2019-February-05). Evaluating the Accuracy of Imputation Methods in a Five-Way Admixed Population [Methods]. *Front Genet, 10*(34). https://doi.org/10.3389/fgene.2019.00034

Scutari, M., Mackay, I., & Balding, D. (2016). Using genetic distance to infer the accuracy of genomic prediction. *Plos Genetics, 12*(9), e1006288.

Sekar, A., Bialas, A. R., de Rivera, H., Davis, A., Hammond, T. R., Kamitaki, N., Tooley, K., Presumey, J., Baum, M., & Van Doren, V. (2016). Schizophrenia risk from complex variation of complement component 4. *Nature, 530*(7589), 177-183.

Sekar, A., Bialas, A. R., de Rivera, H., Davis, A., Hammond, T. R., Kamitaki, N., Tooley, K., Presumey, J., Baum, M., Van Doren, V., Genovese, G., Rose, S. A., Handsaker, R. E., Schizophrenia Working Group of the Psychiatric Genomics, C., Daly, M. J., Carroll, M. C., Stevens, B., & McCarroll, S. A. (2016, 02/11/print). Schizophrenia risk from complex variation of complement component 4 [Article]. *Nature, 530*(7589), 177-183. https://doi.org/10.1038/nature16549

http://www.nature.com/nature/journal/v530/n7589/abs/nature16549.html#supplementary-information

Shannon, C., Douse, K., McCusker, C., Feeney, L., Barrett, S., & Mulholland, C. (2011). The association between childhood trauma and memory functioning in schizophrenia. *Schizophrenia bulletin, 37*(3), 531-537. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3080678/pdf/sbp096.pdf

Shi, H., Burch, K. S., Johnson, R., Freund, M. K., Kichaev, G., Mancuso, N., Manuel, A. M., Dong, N., & Pasaniuc, B. (2020, 2020/06/04/). Localizing Components of Shared Transethnic Genetic Architecture of Complex Traits from GWAS Summary Data. *The American Journal of Human Genetics, 106*(6), 805-817. https://doi.org/https://doi.org/10.1016/j.ajhg.2020.04.012

Shi, H., Kichaev, G., & Pasaniuc, B. (2016, Jul 7). Contrasting the Genetic Architecture of 30 Complex Traits from Summary Association Data. *Am J Hum Genet, 99*(1), 139-153. https://doi.org/10.1016/j.ajhg.2016.05.013

Shi, Y., Li, Z., Xu, Q., Wang, T., Li, T., Shen, J., Zhang, F., Chen, J., Zhou, G., Ji, W., Li, B., Xu, Y., Liu, D., Wang, P., Yang, P., Liu, B., Sun, W., Wan, C., Qin, S., He, G., Steinberg, S., Cichon, S., Werge, T., Sigurdsson, E., Tosato, S., Palotie, A., Nothen, M. M., Rietschel, M., Ophoff, R. A., Collier, D. A., Rujescu, D., Clair, D. S., Stefansson, H., Stefansson, K., Ji, J., Wang, Q., Li, W., Zheng, L., Zhang, H., Feng, G., & He, L.

(2011, Oct 30). Common variants on 8p12 and 1q24.2 confer risk of schizophrenia. *Nat Genet, 43*(12), 1224-1227. https://doi.org/10.1038/ng.980

Shifman, S., Johannesson, M., Bronstein, M., Chen, S. X., Collier, D. A., Craddock, N. J., Kendler, K. S., Li, T., O'Donovan, M., O'Neill, F. A., Owen, M. J., Walsh, D., Weinberger, D. R., Sun, C., Flint, J., & Darvasi, A. (2008, Feb). Genome-wide association identifies a common variant in the reelin gene that increases the risk of schizophrenia only in women. *PLoS Genet, 4*(2), e28. https://doi.org/10.1371/journal.pgen.0040028

Simon, G. E., VON KORFF, M., & T B, Ü. (2002). Understanding cross-national differences in depression prevalence. *Psychological Medicine, 32*(4), 585.

Singh, T., Kurki, M. I., Curtis, D., Purcell, S. M., Crooks, L., McRae, J., Suvisaari, J., Chheda, H., Blackwood, D., & Breen, G. (2016). Rare loss-of-function variants in SETD1A are associated with schizophrenia and developmental disorders. *Nat Neurosci, 19*(4), 571-577.

Singh, T., Poterba, T., Curtis, D., Akil, H., Al Eissa, M., Barchas, J. D., Bass, N., Bigdeli, T. B., Breen, G., & Bromet, E. J. (2020). Exome sequencing identifies rare coding variants in 10 genes which confer substantial risk for schizophrenia. *medRxiv*.

Singh, T., Walters, J. T., Johnstone, M., Curtis, D., Suvisaari, J., Torniainen, M., Rees, E., Iyegbe, C., Blackwood, D., & McIntosh, A. M. (2017). The contribution of rare variants to risk of schizophrenia in individuals with and without intellectual disability. *Nature Genetics, 49*(8), 1167.

Sirugo, G., Williams, S. M., & Tishkoff, S. A. (2019). The missing diversity in human genetic studies. *Cell, 177*(1), 26-31.

Small, K. S., Hedman, A. K., Grundberg, E., Nica, A., Thorleifsson, G., Kong, A., Thorsteindottir, U., Shin, S.-Y., Richards, H. B., & Soranzo, N. (2011). Identification of an imprinted master trans regulator at the KLF14 locus related to multiple metabolic phenotypes. *Nature Genetics, 43*(6), 561-564.

[Record #1821 is using a reference type undefined in this output style.]

Sozuguzel, M. D., Sazci, A., & Yildiz, M. (2019, 2019/06/01). Female gender specific association of the Reelin (RELN) gene rs7341475 variant with schizophrenia. *Molecular Biology Reports, 46*(3), 3411-3416. https://doi.org/10.1007/s11033-019-04803-w

Speed, D., Hemani, G., Johnson, M. R., & Balding, D. J. (2012). Improved heritability estimation from genome-wide SNPs. *The American Journal of Human Genetics, 91*(6), 1011-1021.

Spencer, C. C., Su, Z., Donnelly, P., & Marchini, J. (2009). Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. *Plos Genetics, 5*(5), e1000477.

Stefansson, H., Ophoff, R. A., Steinberg, S., Andreassen, O. A., Cichon, S., Rujescu, D., Werge, T., Pietilainen, O. P., Mors, O., Mortensen, P. B., Sigurdsson, E., Gustafsson, O., Nyegaard, M., Tuulio-Henriksson, A., Ingason, A., Hansen, T., Suvisaari, J., Lonnqvist, J., Paunio, T., Borglum, A. D., Hartmann, A., Fink-Jensen, A., Nordentoft, M., Hougaard, D., Norgaard-Pedersen, B., Bottcher, Y., Olesen, J., Breuer, R., Moller, H. J., Giegling, I., Rasmussen, H. B., Timm, S., Mattheisen, M., Bitter, I., Rethelyi, J. M., Magnusdottir, B. B., Sigmundsson, T., Olason, P., Masson, G., Gulcher, J. R., Haraldsson, M., Fossdal, R., Thorgeirsson, T. E., Thorsteinsdottir, U., Ruggeri, M., Tosato, S., Franke, B., Strengman, E., Kiemeney, L. A., Melle, I., Djurovic, S., Abramova, L., Kaleda, V., Sanjuan, J., de Frutos, R., Bramon, E., Vassos, E., Fraser, G., Ettinger, U., Picchioni, M., Walker, N., Toulopoulou, T., Need, A. C., Ge, D., Yoon, J. L., Shianna, K. V., Freimer, N. B., Cantor, R. M., Murray, R., Kong, A., Golimbet, V., Carracedo, A., Arango, C., Costas, J., Jonsson, E. G., Terenius, L., Agartz, I., Petursson, H., Nothen, M. M., Rietschel, M., Matthews, P. M., Muglia, P., Peltonen, L., St Clair, D., Goldstein, D. B., Stefansson, K., & Collier, D. A. (2009, Aug 6). Common variants conferring risk of schizophrenia. *Nature, 460*(7256), 744-747. https://doi.org/10.1038/nature08186

Stein, D. J., Koen, N., Donald, K., Adnams, C. M., Koopowitz, S., Lund, C., Marais, A., Myers, B., Roos, A., & Sorsdahl, K. (2015). Investigating the psychosocial determinants of child health in Africa: The Drakenstein Child Health Study. *Journal of neuroscience methods, 252*, 27-35.

Steinberg, S., Gudmundsdottir, S., Sveinbjornsson, G., Suvisaari, J., Paunio, T., Torniainen-Holm, M., Frigge, M. L., Jonsdottir, G. A., Huttenlocher, J., & Arnarsdottir, S. (2017). Truncating mutations in RBM12 are associated with psychosis. *Nature Genetics, 49*(8), 1251-1254.

Stevenson, A., Akena, D., Stroud, R. E., Atwoli, L., Campbell, M. M., Chibnik, L. B., Kwobah, E., Kariuki, S. M., Martin, A. R., de Menil, V., Newton, C., Sibeko, G., Stein, D. J., Teferra, S., Zingela, Z., & Koenen, K. C. (2019, Feb 19). Neuropsychiatric Genetics of African Populations-Psychosis (NeuroGAP-Psychosis): a case-control study protocol and GWAS in Ethiopia, Kenya, South Africa and Uganda. *Bmj Open, 9*(2), e025469. https://doi.org/10.1136/bmjopen-2018-025469

Stilo, S. A., & Murray, R. M. (2019, 2019/09/14). Non-Genetic Factors in Schizophrenia. *Current Psychiatry Reports, 21*(10), 100. https://doi.org/10.1007/s11920-019-1091-3

Suarez, B. K., Duan, J., Sanders, A. R., Hinrichs, A. L., Jin, C. H., Hou, C., Buccola, N. G., Hale, N., Weilbaecher, A. N., Nertney, D. A., Olincy, A., Green, S., Schaffer, A. W., Smith, C. J., Hannah, D. E., Rice, J. P., Cox, N. J., Martinez, M., Mowry, B. J., Amin, F., Silverman, J. M., Black, D. W., Byerley, W. F., Crowe, R. R., Freedman, R., Cloninger, C. R., Levinson, D. F., & Gejman, P. V. (2006, Feb). Genomewide linkage scan of 409 European-ancestry and African American families with schizophrenia: suggestive evidence of linkage at 8p23.3-p21.2 and 11p13.1-q14.1 in the combined sample. *Am J Hum Genet, 78*(2), 315-333. https://doi.org/10.1086/500272

Sullivan, P. F., Kendler, K. S., & Neale, M. C. (2003). Schizophrenia as a complex trait: Evidence from a meta-analysis of twin studies. *Archives of General Psychiatry, 60*(12), 1187-1192. https://doi.org/10.1001/archpsyc.60.12.1187

Szatkiewicz, J. P., O'Dushlaine, C., Chen, G., Chambert, K., Moran, J. L., Neale, B. M., Fromer, M., Ruderfer, D., Akterin, S., Bergen, S. E., Kahler, A., Magnusson, P. K., Kim, Y., Crowley, J. J., Rees, E., Kirov, G., O'Donovan, M. C., Owen, M. J., Walters, J., Scolnick, E., Sklar, P., Purcell, S., Hultman, C. M., McCarroll, S. A., & Sullivan, P. F. (2014, Jul). Copy number variation in schizophrenia in Sweden. *Mol Psychiatry, 19*(7), 762-773. https://doi.org/10.1038/mp.2014.40

Teoh, S. L., Chong, H. Y., Abdul Aziz, S., Chemi, N., Othman, A. R., Md Zaki, N., Vanichkulpitak, P., & Chaiyakunapruk, N. (2017). The economic burden of schizophrenia in Malaysia. *Neuropsychiatr Dis Treat, 13*, 1979-1987. https://doi.org/10.2147/ndt.S137140

Thompson Ray, M., Weickert, C. S., Wyatt, E., & Webster, M. J. (2011, May). Decreased BDNF, trkB-TK+ and GAD67 mRNA expression in the hippocampus of individuals with schizophrenia and mood disorders. *J Psychiatry Neurosci, 36*(3), 195-203. https://doi.org/10.1503/jpn.100048

Thompson-Hollands, J., Marx, B. P., Lee, D. J., & Sloan, D. M. (2020). Longitudinal change in self-reported peritraumatic dissociation during and after a course of posttraumatic stress disorder treatment: Contributions of symptom severity and time. *Psychological Trauma: Theory, Research, Practice, and Policy*.

Thurman, R. E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M. T., Haugen, E., Sheffield, N. C., Stergachis, A. B., Wang, H., Vernot, B., Garg, K., John, S., Sandstrom, R., Bates, D., Boatman, L., Canfield, T. K., Diegel, M., Dunn, D., Ebersol, A. K., Frum, T., Giste, E., Johnson, A. K., Johnson, E. M., Kutyavin, T., Lajoie, B., Lee, B. K., Lee, K., London, D., Lotakis, D., Neph, S., Neri, F., Nguyen, E. D., Qu, H., Reynolds, A. P., Roach, V., Safi, A., Sanchez, M. E., Sanyal, A., Shafer, A., Simon, J. M., Song, L., Vong, S., Weaver, M., Yan, Y., Zhang, Z., Zhang, Z., Lenhard, B., Tewari, M., Dorschner, M. O., Hansen, R. S., Navas, P. A., Stamatoyannopoulos, G., Iyer, V. R., Lieb, J. D., Sunyaev, S. R., Akey, J. M., Sabo, P. J., Kaul, R., Furey, T. S., Dekker, J., Crawford, G. E., & Stamatoyannopoulos, J. A. (2012, Sep 6). The accessible chromatin landscape of the human genome. *Nature, 489*(7414), 75-82. https://doi.org/10.1038/nature11232

Tienari, P. (1991, Nov). Interaction between genetic vulnerability and family environment: the Finnish adoptive family study of schizophrenia. *Acta Psychiatr Scand, 84*(5), 460-465.

Tienari, P., Wynne, L. C., Moring, J., Lahti, I., Naarala, M., Sorri, A., Wahlberg, K. E., Saarento, O., Seitamaa, M., Kaleva, M., & et al. (1994, Apr). The Finnish adoptive family study of schizophrenia. Implications for family research. *Br J Psychiatry Suppl*(23), 20-26.

Tortelli, A., Morgan, C., Szoke, A., Nascimento, A., Skurnik, N., de Caussade, E. M., Fain-Donabedian, E., Fridja, F., Henry, M., & Ezembe, F. (2014). Different rates of first admissions for psychosis in migrant groups in Paris. *Social psychiatry and psychiatric epidemiology, 49*(7), 1103-1109. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4283097/pdf/emss-61597.pdf

Traglia, M., Bseiso, D., Gusev, A., Adviento, B., Park, D. S., Mefford, J. A., Zaitlen, N., & Weiss, L. A. (2017). Genetic mechanisms leading to sex differences across common diseases and anthropometric traits. *Genetics, 205*(2), 979-992.

Trynka, G., Sandor, C., Han, B., Xu, H., Stranger, B. E., Liu, X. S., & Raychaudhuri, S. (2013, Feb). Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nat Genet, 45*(2), 124-130. https://doi.org/10.1038/ng.2504

Trześniowska-Drukała, B., Kalinowska, S., Safranow, K., Kłoda, K., Misiak, B., & Samochowiec, J. (2019). Evaluation of hyperhomocysteinemia prevalence and its influence on the selected cognitive functions in patients with schizophrenia. *Progress in Neuro-Psychopharmacology and Biological Psychiatry, 95*, 109679.

Tsujita, T., Okazaki, Y., Fujimaru, K., Minami, Y., Mutoh, Y., Maeda, H., Fukazawa, T., Yamashita, H., & Nakane, Y. (1992). Twin concordance rate of DSM-III-R schizophrenia in a new Japanese sample. Abstracts Seventh International Congress on Twin Studies, Tokyo, Japan,

Uren, C., Kim, M., Martin, A. R., Bobo, D., Gignoux, C. R., van Helden, P. D., Moller, M., Hoal, E. G., & Henn, B. M. (2016, Sep). Fine-Scale Human Population Structure in Southern Africa Reflects Ecogeographic Boundaries. *Genetics, 204*(1), 303-314. https://doi.org/10.1534/genetics.116.187369

Van den Bergh, B. R., van den Heuvel, M. I., Lahti, M., Braeken, M., de Rooij, S. R., Entringer, S., Hoyer, D., Roseboom, T., Räikkönen, K., & King, S. (2017). Prenatal developmental origins of behavior and mental health: The influence of maternal stress in pregnancy. *Neuroscience & Biobehavioral Reviews*.

van der Merwe, C., Mwesiga, E. K., McGregor, N. W., Ejigu, A., Tilahun, A. W., Kalungi, A., Akimana, B., Dubale, B. W., Omari, F., & Mmochi, J. (2018). Advancing neuropsychiatric genetics training and collaboration in Africa. *The Lancet Global Health, 6*(3), e246-e247.

van Os, J., Kenis, G., & Rutten, B. P. F. (2010, 2010/11/01). The environment and schizophrenia. *Nature, 468*(7321), 203-212. https://doi.org/10.1038/nature09563

van Winkel, R. (2011). Family-based analysis of genetic variation underlying psychosis-inducing effects of cannabis: sibling analysis and proband follow-up. *Archives of General Psychiatry, 68*(2), 148-157. https://jamanetwork.com/journals/jamapsychiatry/articlepdf/211074/yoa05072_148_157.pdf

Varese, F., Smeets, F., Drukker, M., Lieverse, R., Lataster, T., Viechtbauer, W., Read, J., van Os, J., & Bentall, R. P. (2012). Childhood adversities increase the risk of psychosis: a meta-analysis of patient-control, prospective-and cross-sectional cohort studies. *Schizophrenia bulletin, 38*(4), 661-671. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3406538/pdf/sbs050.pdf

Vassos, E., Pedersen, C. B., Murray, R. M., Collier, D. A., & Lewis, C. M. (2012). Meta-analysis of the association of urbanicity with schizophrenia. *Schizophrenia bulletin, 38*(6), 1118-1123. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3494055/pdf/sbs096.pdf

Vaucher, J., Keating, B. J., Lasserre, A. M., Gan, W., Lyall, D. M., Ward, J., Smith, D. J., Pell, J. P., Sattar, N., Paré, G., & Holmes, M. V. (2018, 2018/05/01). Cannabis use and risk of schizophrenia: a Mendelian randomization study. *Mol Psychiatry, 23*(5), 1287-1292. https://doi.org/10.1038/mp.2016.252

Veling, W., Susser, E., Van Os, J., Mackenbach, J. P., Selten, J.-P., & Hoek, H. W. (2008). Ethnic density of neighborhoods and incidence of psychotic disorders among immigrants. *American Journal of Psychiatry, 165*(1), 66-73.

Vieland, V. J., Walters, K. A., Lehner, T., Azaro, M., Tobin, K., Huang, Y., & Brzustowicz, L. M. (2014, Mar). Revisiting schizophrenia linkage data in the NIMH Repository: reanalysis of regularized data across multiple studies. *Am J Psychiatry, 171*(3), 350-359. https://doi.org/10.1176/appi.ajp.2013.11121766

Vink, J. M., Bartels, M., Van Beijsterveldt, T. C., Van Dongen, J., Van Beek, J. H., Distel, M. A., De Moor, M. H., Smit, D. J., Minica, C. C., & Ligthart, L. (2012). Sex differences in genetic architecture of complex phenotypes? *Plos One, 7*(12), e47371.

Vinkers, C. H., Van Gastel, W. A., Schubart, C. D., Van Eijk, K. R., Luykx, J. J., Van Winkel, R., Joëls, M., Ophoff, R. A., Boks, M. P., & Bruggeman, R. (2013). The effect of childhood maltreatment and cannabis use on adult psychotic symptoms is modified by the COMT Val158Met polymorphism. *Schizophrenia Research, 150*(1), 303-311.

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., & Yang, J. (2017). 10 years of GWAS discovery: biology, function, and translation. *The American Journal of Human Genetics, 101*(1), 5-22.

Vos, T., Lim, S. S., Abbafati, C., Abbas, K. M., Abbasi, M., Abbasifard, M., Abbasi-Kangevari, M., Abbastabar, H., Abd-Allah, F., & Abdelalim, A. (2020). Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. *The Lancet, 396*(10258), 1204-1222.

Wainschtein, P., Jain, D. P., Yengo, L., Zheng, Z., Cupples, L. A., Shadyab, A. H., McKnight, B., Shoemaker, B. M., Mitchell, B. D., & Psaty, B. M. (2019). Recovery of trait heritability from whole genome sequence data. *bioRxiv*, 588020.

Walker, E. F., & Diforio, D. (1997, Oct). Schizophrenia: a neural diathesis-stress model. *Psychol Rev, 104*(4), 667-685. https://doi.org/10.1037/0033-295x.104.4.667

Wang, C., & Zhang, Y. (2017). Season of birth and schizophrenia: Evidence from China. *Psychiatry Research, 253*, 189-196. https://www.sciencedirect.com/science/article/abs/pii/S016517811631931X?via%3Dihub

Wang, K., Li, M., & Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research, 38*(16), e164-e164.

Wang, Y., Wei, Y., Edmiston, E. K., Womer, F. Y., Zhang, X., Duan, J., Zhu, Y., Zhang, R., Yin, Z., & Zhang, Y. (2020). Altered structural connectivity and cytokine levels in Schizophrenia and Genetic high-risk individuals: Associations with disease states and vulnerability. *Schizophrenia Research*.

Watanabe, K., Taskesen, E., van Bochoven, A., & Posthuma, D. (2017, 2017/11/28). Functional mapping and annotation of genetic associations with FUMA. *Nature Communications, 8*(1), 1826. https://doi.org/10.1038/s41467-017-01261-5

Weinstein, Y., Levav, I., Gelkopf, M., Roe, D., Yoffe, R., Pugachova, I., & Levine, S. Z. (2018). Association of maternal exposure to terror attacks during pregnancy and the risk of schizophrenia in the offspring: a population-based study. *Schizophrenia Research, 199*, 163-167. https://www.sciencedirect.com/science/article/abs/pii/S0920996418302378?via%3Dihub

Weisstein, E. W. (2004). Bonferroni correction. *https://mathworld. wolfram. com/*.

Wender, P. H., Rosenthal, D., Kety, S. S., Schulsinger, F., & Welner, J. (1974, Jan). Crossfostering. A research strategy for clarifying the role of genetic and experiential factors in the etiology of schizophrenia. *Arch Gen Psychiatry, 30*(1), 121-128.

Whiteford, H. A., Ferrari, A. J., Degenhardt, L., Feigin, V., & Vos, T. (2015). The global burden of mental, neurological and substance use disorders: an analysis from the Global Burden of Disease Study 2010. *Plos One, 10*(2), e0116820. https://doi.org/10.1371/journal.pone.0116820

Wicks, S., Hjern, A., Gunnell, D., Lewis, G., & Dalman, C. (2005, Sep). Social adversity in childhood and the risk of developing psychosis: a national cohort study. *Am J Psychiatry, 162*(9), 1652-1657. https://doi.org/10.1176/appi.ajp.162.9.1652

Wiehahn, G. J., Bosch, G. P., du Preez, R. R., Pretorius, H. W., Karayiorgou, M., & Roos, J. L. (2004, Aug 15). Assessment of the frequency of the 22q11 deletion in Afrikaner schizophrenic patients. *Am J Med Genet B Neuropsychiatr Genet, 129b*(1), 20-22. https://doi.org/10.1002/ajmg.b.20168

Wilcox, M. A., Faraone, S. V., Su, J., Van Eerdewegh, P., & Tsuang, M. T. (2002, Nov 1). Genome scan of three quantitative traits in schizophrenia pedigrees. *Biol Psychiatry, 52*(9), 847-854. https://doi.org/10.1016/s0006-3223(02)01465-8

Wojcik, G. L., Fuchsberger, C., Taliun, D., Welch, R., Martin, A. R., Shringarpure, S., Carlson, C. S., Abecasis, G., Kang, H. M., Boehnke, M., Bustamante, C. D., Gignoux, C. R., & Kenny, E. E. (2018, Oct 3). Imputation-Aware Tag SNP Selection To Improve Power for Large-Scale, Multi-ethnic Association Studies. *G3 (Bethesda), 8*(10), 3255-3267. https://doi.org/10.1534/g3.118.200502

Wojcik, G. L., Graff, M., Nishimura, K. K., Tao, R., Haessler, J., Gignoux, C. R., Highland, H. M., Patel, Y. M., Sorokin, E. P., & Avery, C. L. (2019). Genetic analyses of diverse populations improves discovery for complex traits. *Nature, 570*(7762), 514-518.

Wong, E. H., So, H. C., Li, M., Wang, Q., Butler, A. W., Paul, B., Wu, H. M., Hui, T. C., Choi, S. C., So, M. T., Garcia-Barcelo, M. M., McAlonan, G. M., Chen, E. Y., Cheung, E. F., Chan, R. C., Purcell, S. M., Cherny, S. S., Chen, R. R., Li, T., & Sham, P. C. (2014, Jul). Common variants on Xq28 conferring risk of schizophrenia in Han Chinese. *Schizophr Bull, 40*(4), 777-786. https://doi.org/10.1093/schbul/sbt104

Wray, N. R., Yang, J., Hayes, B. J., Price, A. L., Goddard, M. E., & Visscher, P. M. (2013). Pitfalls of predicting complex traits from SNPs. *Nature Reviews Genetics, 14*(7), 507-515.

Xia, H., Jahr, F. M., Kim, N. K., Xie, L., Shabalin, A. A., Bryois, J., Sweet, D. H., Kronfol, M. M., Palasuberniam, P., McRae, M., Riley, B. P., Sullivan, P. F., van den Oord, E. J., & McClay, J. L. (2018, Sep 15). Building a schizophrenia genetic network: transcription factor 4 regulates genes involved in neuronal development and schizophrenia risk. *Hum Mol Genet, 27*(18), 3246-3256. https://doi.org/10.1093/hmg/ddy222

Xiao, X., Luo, X. J., Chang, H., Liu, Z., & Li, M. (2017, Aug). Evaluation of European Schizophrenia GWAS Loci in Asian Populations via Comprehensive Meta-Analyses. *Mol Neurobiol, 54*(6), 4071-4080. https://doi.org/10.1007/s12035-016-9990-3

Xie, P., Wu, K., Zheng, Y., Guo, Y., Yang, Y., He, J., Ding, Y., & Peng, H. (2018, 2018/03/01/). Prevalence of childhood trauma and correlations between childhood trauma, suicidal ideation, and social support in patients with depression, bipolar disorder, and schizophrenia in southern China. *Journal of Affective Disorders, 228*, 41-48. https://doi.org/https://doi.org/10.1016/j.jad.2017.11.011

Xu, L., Xu, T., Tan, W., Yan, B., Wang, D., Li, H., Lin, Y., Li, K., Wen, H., Qin, X., Sun, X., Guan, L., Bass, J. K., Ma, H., & Yu, X. (2019, Dec 16). Household economic burden and outcomes of patients with schizophrenia after being unlocked and treated in rural China. *Epidemiol Psychiatr Sci, 29*, e81. https://doi.org/10.1017/s2045796019000775

Yang, J., Lee, S. H., Goddard, M. E., & Visscher, P. M. (2011, Jan 07). GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet, 88*(1), 76-82. https://doi.org/10.1016/j.ajhg.2010.11.011

Yoneyama, S., Yao, J., Guo, X., Fernandez-Rhodes, L., Lim, U., Boston, J., Buzková, P., Carlson, C. S., Cheng, I., Cochran, B., Cooper, R., Ehret, G., Fornage, M., Gong, J., Gross, M., Gu, C. C., Haessler, J., Haiman, C. A., Henderson, B., Hindorff, L. A., Houston, D., Irvin, M. R., Jackson, R., Kuller, L., Leppert, M., Lewis, C. E., Li, R., Le Marchand, L., Matise, T. C., Nguyen, K. D., Chakravarti, A., Pankow, J. S., Pankratz, N., Pooler, L., Ritchie, M. D., Bien, S. A., Wassel, C. L., Chen, Y. D., Taylor, K. D., Allison, M., Rotter, J. I., Schreiner, P. J., Schumacher, F., Wilkens, L., Boerwinkle, E., Kooperberg, C., Peters, U., Buyske, S., Graff, M., North, K. E., & Consortium, P. (2017, 2017/02/01). Generalization and fine mapping of European ancestry-based central adiposity variants in African ancestry populations. *International Journal of Obesity, 41*(2), 324-331. https://doi.org/10.1038/ijo.2016.207

Yu, H., Yan, H., Li, J., Li, Z., Zhang, X., Ma, Y., Mei, L., Liu, C., Cai, L., Wang, Q., Zhang, F., Iwata, N., Ikeda, M., Wang, L., Lu, T., Li, M., Xu, H., Wu, X., Liu, B., Yang, J., Li, K., Lv, L., Ma, X., Wang, C., Li, L., Yang, F., Jiang, T., Shi, Y., Li, T., Zhang, D., & Yue, W. (2017, Jul). Common variants on 2p16.1, 6p22.1 and 10q24.32 are associated with schizophrenia in Han Chinese population. *Mol Psychiatry, 22*(7), 954-960. https://doi.org/10.1038/mp.2016.212

Yue, W. H., Wang, H. F., Sun, L. D., Tang, F. L., Liu, Z. H., Zhang, H. X., Li, W. Q., Zhang, Y. L., Zhang, Y., Ma, C. C., Du, B., Wang, L. F., Ren, Y. Q., Yang, Y. F., Hu, X. F., Wang, Y., Deng, W., Tan, L. W., Tan, Y. L., Chen, Q., Xu, G. M., Yang, G. G., Zuo, X. B., Yan, H., Ruan, Y. Y., Lu, T. L., Han, X., Ma, X. H., Wang, Y., Cai, L. W., Jin, C., Zhang, H. Y., Yan, J., Mi, W. F., Yin, X. Y., Ma, W. B., Liu, Q., Kang, L., Sun, W., Pan, C. Y., Shuang, M., Yang, F. D., Wang, C. Y., Yang, J. L., Li, K. Q., Ma, X., Li, L. J., Yu, X., Li, Q. Z., Huang, X., Lv, L. X., Li, T., Zhao, G. P., Huang, W., Zhang, X. J., & Zhang, D. (2011, Oct 30). Genome-wide

association study identifies a susceptibility locus for schizophrenia in Han Chinese at 11p11.2. *Nat Genet, 43*(12), 1228-1231. https://doi.org/10.1038/ng.979

Zaidi, A. A., & Mathieson, I. (2020). Demographic history mediates the effect of stratification on polygenic scores. *Elife, 9*, e61548.

Zammit, S., Spurlock, G., Williams, H., Norton, N., Williams, N., O'Donovan, M. C., & Owen, M. J. (2007). Genotype effects of CHRNA7, CNR1 and COMT in schizophrenia: interactions with tobacco and cannabis use. *The British Journal of Psychiatry, 191*(5), 402-407.

Zar, H., Barnett, W., Myer, L., Stein, D., & Nicol, M. (2015). Investigating the early-life determinants of illness in Africa: the Drakenstein Child Health Study. *Thorax, 70*(6), 592-594.

Zar, H. J., Pellowski, J. A., Cohen, S., Barnett, W., Vanker, A., Koen, N., & Stein, D. J. (2019). Maternal health and birth outcomes in a South African birth cohort study. *Plos One, 14*(11), e0222399.

Zäske, H., Linden, M., Degner, D., Jockers-Scherübl, M., Klingberg, S., Klosterkötter, J., Maier, W., Möller, H. J., Sauer, H., Schmitt, A., & Gaebel, W. (2019, Jun). Stigma experiences and perceived stigma in patients with first-episode schizophrenia in the course of 1 year after their first in-patient treatment. *Eur Arch Psychiatry Clin Neurosci, 269*(4), 459-468. https://doi.org/10.1007/s00406-018-0892-4

Zein, J. G., & Erzurum, S. C. (2015, Jun). Asthma is Different in Women. *Curr Allergy Asthma Rep, 15*(6), 28. https://doi.org/10.1007/s11882-015-0528-y

Zhang, Y., & Pan, W. (2015). Principal component regression and linear mixed model in association analysis of structured samples: competitors or complements? *Genetic epidemiology, 39*(3), 149-155.

Zubair, N., Graff, M., Luis Ambite, J., Bush, W. S., Kichaev, G., Lu, Y., Manichaikul, A., Sheu, W. H., Absher, D., Assimes, T. L., Bielinski, S. J., Bottinger, E. P., Buzkova, P., Chuang, L. M., Chung, R. H., Cochran, B., Dumitrescu, L., Gottesman, O., Haessler, J. W., Haiman, C., Heiss, G., Hsiung, C. A., Hung, Y. J., Hwu, C. M., Juang, J. J., Le Marchand, L., Lee, I. T., Lee, W. J., Lin, L. A., Lin, D., Lin, S. Y., Mackey, R. H., Martin, L. W., Pasaniuc, B., Peters, U., Predazzi, I., Quertermous, T., Reiner, A. P., Robinson, J., Rotter, J. I., Ryckman, K. K., Schreiner, P. J., Stahl, E., Tao, R., Tsai, M. Y., Waite, L. L., Wang, T. D., Buyske, S., Ida Chen, Y. D., Cheng, I., Crawford, D. C., Loos, R. J. F., Rich, S. S., Fornage, M., North, K.

E., Kooperberg, C., & Carty, C. L. (2016, Dec 15). Fine-mapping of lipid regions in global populations discovers ethnic-specific signals and refines previously identified lipid loci. *Hum Mol Genet, 25*(24), 5500-5512. https://doi.org/10.1093/hmg/ddw358

Appendices

Appendix 1 — UBACC Assessment of capacity to consent

Declaration of Competence to Give Informed Consent

I, _____, am a registered nurse / medical doctor (delete which is non-applicable). I do hereby declare that, on the date and time indicated below, I have interviewed: I have found that they are competent and therefore capable to give informed consent regarding participating in study "The Genomics of Schizophrenia in the South African Xhosa People"

Signed: _____

Registration number: _____

Signed on the _____ (day) of _____ (month) 20_____ at _____ (time) and at

_____ _____ (place)

UBACC Questions and Answers	1	2	3	4
1. Yintoni injongo yoluphondo lwenziwayo ndigqiba ukukucacisela ngalo?				
Response: **Two key concepts (1) Genetics in the Xhosa People (2) Causes of Schizophrenia**	Score	Score	Score	Score
Neither concept	0	0	0	0
One concept – either (1) drawing the participant's blood OR (2) look at causes of Schizophrenia	1	1	1	1
Both concepts - 1) drawing the participant's blood IN ORDER TO (2) look at causes of Schizophrenia	2	2	2	2
2. Yintoni ekwenze ukuba ufune ukuthabatha inxaxheba koluphando?				
Response: **Two key concepts: (1) Personal benefit and (2) Helping humanity**	Score	Score	Score	Score
Neither concept or Give my blood for money	0	0	0	0
One concept – either (1) Personal benefit OR (2) Helping humanity	1	1	1	1
Both concepts - 1) Personal benefit AND (2) Helping humanity	2	2	2	2
3. Ingaba ucinga ukuba olu luphando okanye lunyango?				
Response:	Score	Score	Score	Score
Treatment; I don't know; both;	0	0	0	0
Research	2	2	2	2
4. Ucinga ukuba kunyanzelekile na ukuba ube koluphando nokuba awufuni?				
Response:	Score	Score	Score	Score
Yes; I don't know	0	0	0	0
No	2	2	2	2
5. Ucinga ukuba uye wayeka ukuthabatha inxaxheba koluphando ungakwazi ukufumana unyango lwakho njengesiqhelo?				
Response:	Score	Score	Score	Score
No; I don't know	0	0	0	0
Yes	2	2	2	2
6. Ukuba uthe wathatha inxaxheba koluphando zeziphi ezinye zezinto ozakucelwa uzenze?				
Response: **Four tasks: (1) answer questions about myself; (2) do a**	Score	Score	Score	Score

computerized task to test my thinking; (3) give 2 or 3 tubes of blood; (4) HIV test,			
None of the above	0	0	0
One or two tasks	1	1	1
Three or all four tasks	2	2	2
7. Ndicela uchaze ubungozi okanye ubunzima onokubufumana ukuba uthe wathatha inxaxheba koluphando?			
Response: **Five key concepts: (1) Taking blood is uncomfortable, small chance of infection; (2) HIV test could be positive; (3) Small chance of being identified through DNA; (4) Cell information stored in the US and could be accessed by US government; (5) No control over research on stem cells in the future.**	Score	Score	Score
None of the above	0	0	0
One or two concepts	1	1	1
Three to five concepts	2	2	2
8. Ndicela uchaze inzuzo/amanye amancedo anokufumaneka koluphando?			
Response: **Two key concepts: (1) HIV-test (2) talk about experiences with mental health professional (3) in mother-tongue**	Score	Score	Score
Neither concept: I will be paid to participate	0	0	0
One concept – either (1) HIV testing OR (2) talk about experience to professional in mother tongue	1	1	1
Both concepts - 1) HIV testing AND (2) talk about experience to professional in mother tongue	2	2	2
9. Ingaba igenzeka into yokuba oluphando lungangabi luncedo kuwe?			
Response:	Score	Score	Score
No, I don't know (0)	0	0	0
Yes (2)	2	2	2
10. Ingaba kunyanzelekile ukuba imisebe yakho yegazi iyokugcinwa?			
Response:	Score	Score	Score
Yes; I don't know	0	0	0
No	2	2	2

Appendix 2 — SAX consent form

The Genomics of Schizophrenia in the Xhosa People of South Africa

Primary Investigators: Profs D. Stein, O. Alonso Betancourt, R. Ramesar. E. Susser, R. Gur, R. Gur, M.C. King
University of Cape Town Human Research Ethics Committee number: 049/2013
Walter Sisulu University Research Ethics and Bio-safety Committee Number: 003/2013
Columbia University Internal Review Board number: UCT IRB of record
University of Washington IRB number: 29501

Participant Consent Form – Donation of DNA OR Stem Cells

Consent for DNA to be stored and available for later use

I agree that my DNA will be stored for 15 years and that it may be used by other researchers. I agree to my DNA being made available on an online database for use by other researchers. The DNA will be used for:

Research into schizophrenia and related disorders.

OR

Any medical research.

I understand that my identity will be kept secret.

Signature_____ Date _____

OR

Consent for cell immortalisation

I agree to have some cells from my blood treated so that they can be stored for 15 years. I agree to my DNA being made available on an online database. I understand that my identity will be kept secret. These cells and my DNA will be used for:

Research into schizophrenia and related disorders.

OR

Any medical research.

Signature_____ Date _____

Appendix 3 — Recipes for buffer solutions

Red Blood Cell lysis buffer (RBC Lysis Buffer)

- NH4Cl(Ammonium chloride): 8.28g
- NH4HCO3 (Ammonium bicarbonate powder): 0.79g
- EDTA (0.5M,pH 7.4): 0.2ml
- Make up volume to 1L with ddH2O (sterile filter)

White blood cell Lysis Buffer

- Tris-HCl (1M, pH 7.5): 25ml
- NaCl (3M): 16.7ml
- EDTA (0.5M):1ml
- Make up volume to 500ml with dH_2O (sterile water)

20mg/ml Proteinase K

0.02g Proteinase K and add 1000ul of filter-sterilized 1mM CaOAc at pH8

20% Sodium dodecyl sulfate

2g SDS in 10 ml distilled H_2O

6M NaCl

Dissolve 35.064g of NaCl in100ml dH_2O

1M Tris-HCl (pH 7.5)

Dissolve 12.1g Tris/Trizma base in 100ml dH_2O and adjust pH to 7.5

Appendix 4 — Population codes and descriptions used in PCA and admixture analyses

Superpopulation	Code	Subpopulation	Description
East Asian	CHB	Han Chinese	Han Chinese in Beijing, China
	JPT	Japanese	Japanese in Tokyo, Japan
	CHS	Southern Han Chinese	Han Chinese South
	CDX	Dai Chinese	Chinese Dai in Xishuangbanna, China
	KHV	Kinh Vietnamese	Kinh in Ho Chi Minh City, Vietnam
	CHD	Denver Chinese	Chinese in Denver, Colorado (pilot 3 only)
European	CEU	CEPH	Utah residents (CEPH) with Northern and Western European ancestry
	TSI	Tuscan	Toscani in Italia
	GBR	British	British in England and Scotland
	FIN	Finnish	Finnish in Finland
	IBS	Spanish	Iberian populations in Spain
African	**Central Niger Bantu**		
	Kaba	Kaba	Kaba from Central African Republic
	Fang	Fang	Fang from Southern Gabon and Cameroon
	Kongo	Kongo	Kongo from Democratic Republic of Congo
	Hausa	Hausa	Hausa from Niger Bulala
	Bulala	Bulala	Bulala from Malawi
	Mada	Mada	Mada from Cameroon
	Bamoun	Bamoun	Bamoun from Cameroon
	Fulani	Fulani	Fulani from Niger
	Eastern Bantu		
	MKK	MKK	Maasai in Kinyawe, Kenya
	LWK	LWK	Luhya in Webuye, Kenya
	SAW	SAW	Sandawe from Tanzania
	BAN	BAN	Bantu-Kenya

	Southern Bantu		
	ZUL	ZUL	Zulu-South-Africa
	HER	HER	Herero , South Africa-Namibia
	STS	STS	Sotho-Tswana,South Africa
	XHS	XHS	Xhosa-South-Africa
	XHS2	XHS2	Xhosa from Southeastern South Africa
	XHOSA_CASE	XHOSA_CASE	SAX-schizophrenia cases
	XHOSA_CONTROL	XHOSA_CONTROL	SAX-schizophrenia controls
	Khoi-San		
	SAN	SAN	Namibia Khoi-San
	KHS	KHS	Namibia Khoi-San
	BUS	BUS	Bushmen
	West Niger Bantu		
	Brong	Brong	Brong from mid-western Ghana
	Igbo	Igbo	Igbo from southeastern Nigeria
	MAN	MAN	Mandenka from Senegal
	YOR	YOR	Yoruba in Ibadan, Nigeria
	ESN	Esan	Esan in Nigeria
	YRI	YRI	Yoruba in Ibadan, Nigeria
	GWD	Gambian	Gambian in Western Division, The Gambia
	MSL	Mende	Mende in Sierra Leone
American	ASW	African-American SW	African Ancestry in Southwest US
	ACB	African-Caribbean	African Caribbean in Barbados
	MXL	Mexican-American	Mexican Ancestry in Los Angeles, California
	PUR	Puerto Rican	Puerto Rican in Puerto Rico
	CLM	Colombian	Colombian in Medellin, Colombia
	PEL	Peruvian	Peruvian in Lima, Peru
South Asian	GIH	Gujarati	Gujarati Indian in Houston, TX
	PJL	Punjabi	Punjabi in Lahore, Pakistan

	BEB	Bengali	Bengali in Bangladesh
	STU	Sri Lankan	Sri Lankan Tamil in the UK
	ITU	Indian	Indian Telugu in the UK

Appendix 5 — Supplementary information for chapter 2

Supplementary Table 2.1 - Biological data repositories used in FUMA

Category	Name	Description	Link
Reference Variants	dbSNP 146	Map rsID of input files to dbSNP build 146	ftp://ftp.ncbi.nlm.nih.gov/snp/organisms/human_9606_b146_grch137p13/database/organism_data/RsMergeArch.bcp.gz
Reference Genome	1000 Genomes Project Phase3	Compute MAF and r^2 for each available population	ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/
Functional annotations of SNPs	CADD v1.3	Deleteriousness score of variants	http://cadd.gs.washington.edu/download
	RegulomeDB	Score of regulatory variants	http://www.regulomedb.org/downloads
	15-core chromatin state	Chromatin states of genomic region in 127 tissue/cell types	http://egg2.wustl.edu/roadmap/data/byFileType/chromhmmSegmentations/ChmmModels/coreMarks/jointModel/final/
	GWAS catalog	Known trait associated variants	https://www.ebi.ac.uk/gwas/
eQTLs	GTEx v6	cis-eQTLs of 44 tissue types	http://www.gtexportal.org/home/
	Blood eQTL Browser	cis-eQTLs of blood cell	http://genenetwork.nl/bloodeqtlbrowser/
	BIOS QTL Browser	cis-eQTLs of blood cell	http://genenetwork.nl/biosqtlbrowser/
	BRAINEAC	cis-eQTLs of 10 brain regions	http://www.braineac.org/
HiC	GSE87112	HiC data for 14 tissue types and 7 cell lines	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE87112
Regulatory elements	Roadmapc epigenomics project	Enhancer, promoter and dyadic enhancer/promoter regions in 111 epigenomes	http://egg2.wustl.edu/roadmap/data/byDataType/dnase/
Gene score	pLI	Probability of being loss-of-function intolerance	ftp://ftp.broadinstitute.org/pub/ExAC_release/release0.3.1/functional_gene_constraint
	ncRVIS	Non-coding residual variation intolerance score	http://journals.plos.org/plosgenetics/article/file?type=supplementary&id=info:doi/10.1371/journal.pgen.1005492.s011

Gene expression	GTEx v6	Normalized gene expression (RPKM: Read Per Kilo base per Million) for 53 tissue types	http://www.gtexportal.org/home/
Gene sets	MsigDB v5.2	Curated pathways and gene sets	http://software.broadinstitute.org/gsea/msigdb/
	WikiPathways	Curated pathways	http://wikipathways.org/index.php/WikiPathways
Tools	ANNOVAR	Variant annotation tool	http://annovar.openbioinformatics.org/en/latest/
	MAGMA v6.0	Software for gene-based test and gene-set analyses of GWAS	https://ctg.cncr.nl/software/magma

Table modified from (https://fuma.ctglab.nl/links)

Appendix 6 — Supplementary information for chapter 3

Supplementary Table 3.1 - Functional elements used for the heritability enrichment analyses

Code	Category name	Category description	Reference
Coding_UCSC	Coding regions	Coding annotations from Refseq gene model obtained from University of California Santa Cruz	(Kent et al., 2002)
Conserved_LindbladToh	Regions conserved in mammals	Regions conserved in mammals	(Lindblad-Toh et al., 2011)
CTCF_Hoffman	CCCTC binding factor	Conserved zing finger protein annotations obtained from Hoffman et al.	(Hoffman et al., 2013)
DGF_ENCODE	Digital Genomic Footprint	Digital Genomic Footprint annotations obtained from ENCODE	(ENCODE Project, 2012)
DHS_Trynka	DNase I hypersensitive sites	DNase I hypersensitive sites of all cell types obtained from ENCODE and Roadmap; processed by Trynka	(Trynka et al., 2013)
Enhancer_Andersson	Enhancer regions - Andersson	Enhancer region annotations obtained from Andersson et al.	(Andersson et al., 2014)
Enhancer_Hoffman	Enhancer regions - Hoffman	Enhancer region annotations obtained from Hoffman et al.	(Hoffman et al., 2013)

FetalDHS_Trynka	Fetal DNase 1 hypersensitivity sites	DNase I hypersensitive sites of the fetal cell type obtained from Trynka et al.	(Trynka et al., 2013)
H3K27ac_Hnisz	H3K27ac - Hnisz	Acetylation of the 27th lysine residue on histone 3 epigenetic modification annotations obtained from Hnisz et al.	(Hnisz et al., 2013)
H3K27ac_PGC2	H3K27ac - PGC	Acetylation of the 27th lysine residue on histone 3 epigenetic modification annotations obtained from Roadmap and processed by PGC2	(Schizophrenia Working Group of the Psychiatric Genomics, 2014)
H3K4me1_Trynka	H3K4me1	Mono-methylation of 4th lysine residue on histone 3 epigenetic modification annotations obtained from Trynka	(Trynka et al., 2013)
H3K4me3_Trynka	H3K4me3	Tri-methylation of 4th lysine residue on histone 3 epigenetic modification annotations obtained from Trynka	(Trynka et al., 2013)
H3K9ac_Trynka	H3K9ac	Acetylation of 9th lysine residue on histone 3 epigenetic modification to Histone 3 annotations obtained from Trynka	(Trynka et al., 2013)
Intron_UCSC	Intronic regions	Intron annotations from Refseq gene model obtained from University of California Santa Cruz	(Kent et al., 2002)
PromoterFlanking_Hoffman	Promoter flanking regions	Promoter flanking regions	(Hoffman et al., 2013)
Promoter_UCSC	Promoter regions	Promoter region annotations from Refseq gene model obtained from University of California Santa Cruz	(Kent et al., 2002)

Repressed_Hoffman	Repressed regions	Repressed region annotations obtained from Hoffman	(Hoffman et al., 2013)
SuperEnhancer_Hnisz	Super enhancers	Super-enhancers obtained from Hnisz et al.	(Hnisz et al., 2013)
TFBS_ENCODE	Transcription factor binding sites	Transcription factor binding site annotations obtained from the Encyclopaedia of DNA elements	(ENCODE Project, 2012)
Transcribed_Hoffman	Transcribed regions	Transcribed region annotations obtained from Hoffman et al	(Hoffman et al., 2013)
TSS_Hoffman	Transcription start site	Transcription start site annotations obtained from Hoffman et al.	(Hoffman et al., 2013)
UTR_3_UCSC	Untranslated 3' regions	Untranslated 3' region annotations from Refseq gene model obtained from University of California Santa Cruz	(Kent et al., 2002)
UTR_5_UCSC	Untranslated 5' regions	Untranslated 5' region annotations from Refseq gene model obtained from University of California Santa Cruz	(Kent et al., 2002)
WeakEnhancer_Hoffman	Weak enhancer regions	Weak enhancer region annotations obtained from Hoffman et al.	(Hoffman et al., 2013)

Supplementary Table 3.2 - The heritability explained by SNPs across the autosomes

Chromosome	Number of SNPs	h^2g	SE	*P*-value	Description
1	1130065	0.054348	0.031043	3.6758e-02	Unadjusted
2	1241199	0.089984	0.032231	1.9002e-03	Unadjusted
3	1037441	0.087269	0.028972	7.6779e-04	Unadjusted
4	1050907	0.048587	0.028355	4.4511e-02	Unadjusted
5	948377	0.063835	0.027265	7.4972e-03	Unadjusted
6	924497	0.120193	0.028725	7.7307e-06	Unadjusted
7	852440	0.000001	0.023791	5.0000e-01	Unadjusted
8	830828	0.071570	0.025516	1.0851e-03	Unadjusted
9	615006	0.063606	0.024637	1.6617e-03	Unadjusted
10	717165	0.052971	0.025829	1.9280e-02	Unadjusted
11	716778	0.045732	0.023056	1.7443e-02	Unadjusted
12	671591	0.021147	0.021496	1.5129e-01	Unadjusted
13	525253	0.039831	0.021008	1.7371e-02	Unadjusted
14	472529	0.018569	0.019688	1.6359e-01	Unadjusted
15	421853	0.029044	0.021188	8.1273e-02	Unadjusted
16	465485	0.020884	0.021610	1.6177e-01	Unadjusted
17	395569	0.068545	0.023683	1.0200e-03	Unadjusted
18	400976	0.047841	0.022455	1.5312e-02	Unadjusted
19	330185	0.038580	0.020460	2.4815e-02	Unadjusted
20	318592	0.006273	0.017705	3.6324e-01	Unadjusted
21	200449	0.039053	0.017301	6.0213e-03	Unadjusted
22	197883	0.026722	0.017329	5.0703e-02	Unadjusted

1	1130065	0.041018	0.031634	9.6045e-02	Adjusted
2	1241199	0.078020	0.032939	8.0890e-03	Adjusted
3	1037441	0.068419	0.029180	8.0598e-03	Adjusted
4	1050907	0.054379	0.029248	3.1558e-02	Adjusted
5	948377	0.062059	0.027382	7.9530e-03	Adjusted
6	914829	0.109578	0.029283	7.0097e-05	Adjusted
7	852440	0.000001	0.024894	5.0000e-01	Adjusted
8	830828	0.060916	0.025927	6.0730e-03	Adjusted
9	615006	0.053175	0.024853	8.5468e-03	Adjusted
10	717165	0.041515	0.025944	5.5811e-02	Adjusted
11	716778	0.034132	0.023251	6.5618e-02	Adjusted
12	671591	0.012713	0.021216	2.6478e-01	Adjusted
13	525253	0.037892	0.021549	2.7559e-02	Adjusted
14	472529	0.010262	0.019626	2.9634e-01	Adjusted
15	421853	0.019579	0.021178	1.7773e-01	Adjusted
16	465485	0.007938	0.021154	3.5203e-01	Adjusted
17	395569	0.054573	0.023883	1.0020e-02	Adjusted
18	400976	0.041826	0.022568	2.9751e-02	Adjusted
19	330185	0.020998	0.019679	1.4502e-01	Adjusted
20	318592	0.000234	0.017896	4.9496e-01	Adjusted
21	200449	0.038713	0.017408	6.0795e-03	Adjusted
22	197883	0.020334	0.01757	1.2038e-01	Adjusted

h^2g, heritability; SE, standard ; SNPs from the MHC region on chromosome six were not removed for the unadjusted analysis, whereas they were removed for the adjusted analysis, as well as the inclusion of principal components

Supplementary Table 3.3 - The heritability explained by SNPs across MAF bins

MAF-bin	Number of SNPs	h^2g	SE	*P*-value	MAF frequency range	Description
1	6015609	0.551915	0.079105	9.1538e-14	"0.01 – 0.05"	Unadjusted
2	2203236	0.551915	0.034172	0	"0.05 – 0.1"	Unadjusted
3	2025937	0.551915	0.016004	0	"0.1 – 0.2"	Unadjusted
4	1161526	0.551915	0.007640	0	"0.2 – 0.3"	Unadjusted
5	859609	0.377630	0.053267	2.0874e-09	"0.3 – 0.4"	Unadjusted
6	754218	0.282653	0.050478	7.6875e-08	"0.4 – 0.5"	Unadjusted
1	6014696	0.551915	0.081673	3.0840e-10	"0.01 – 0.05"	Adjusted
2	2201179	0.551915	0.034430	3.8858e-16	"0.05 – 0.1"	Adjusted
3	2023203	0.551915	0.016024	0	"0.1 – 0.2"	Adjusted
4	1160036	0.551915	0.007827	5.5511e-17	"0.2 – 0.3"	Adjusted
5	858275	0.551915	0.004191	8.8818e-16	"0.3 – 0.4"	Adjusted
6	753093	0.279151	0.052473	9.9761e-07	"0.4 – 0.5"	Adjusted

h^2g, heritability; SE, standard error; SNPs from the MHC region on chromosome six were not removed for the unadjusted analysis, whereas they were removed for the adjusted analysis, as well as the inclusion of principal components

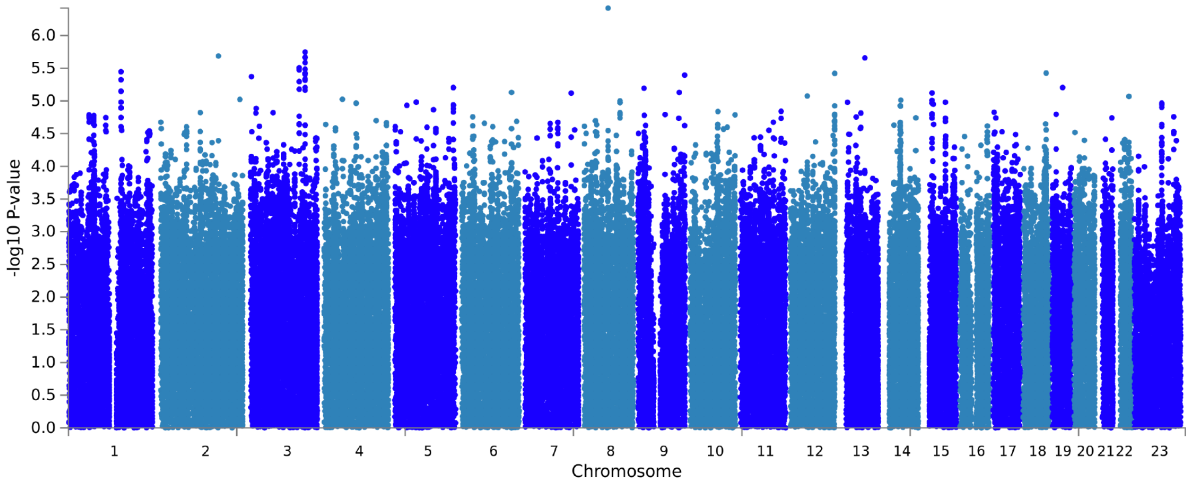
Supplementary Table 3.4 - The heritability explained by 24 functional categories

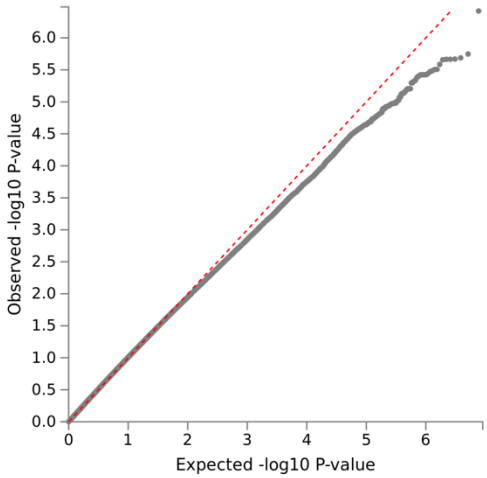
	Number of SNPs	h^2g	SE	*P*-value	Analysis	Code	Category name
1	64599	0.218416	0.042874	1.1395e-07	Unadjusted	Coding_UCSC	Coding regions
2	118847	0.551915	0.012840	0	Unadjusted	Conserved_LindbladToh	Regions conserved in mammals
3	108752	0.342713	0.055547	1.9907e-08	Unadjusted	CTCF_Hoffman	CCCTC binding factor
4	643970	0.551915	0.014637	0	Unadjusted	DGF_ENCODE	Digital Genomic Footprint
5	805942	0.551915	0.019233	0	Unadjusted	DHS_Trynka	DNase I hypersensitive sites
6	19810	0.094527	0.034576	2.0000e-03	Unadjusted	Enhancer_Andersson	Enhancer regions - Andersson
7	195758	0.260210	0.054922	3.2275e-06	Unadjusted	Enhancer_Hoffman	Enhancer regions - Hoffman
8	404056	0.551915	0.012913	0	Unadjusted	FetalDHS_Trynka	Fetal DNase 1 hypersensitivity sites
9	1809010	0.551915	0.016366	0	Unadjusted	H3K27ac_Hnisz	H3K27ac - Hnisz
10	1244067	0.551915	0.016524	0	Unadjusted	H3K27ac_PGC2	H3K27ac - PGC
11	2020569	0.551915	0.017740	0	Unadjusted	H3K4me1_Trynka	H3K4me1
12	612860	0.551915	0.015170	2.2204e-16	Unadjusted	H3K4me3_Trynka	H3K4me3
13	582923	0.551915	0.016180	0	Unadjusted	H3K9ac_Trynka	H3K9ac
14	1809993	0.551915	0.014785	6.3977e-13	Unadjusted	Intron_UCSC	Intronic regions
15	38511	0.156280	0.040868	5.1878e-05	Unadjusted	PromoterFlanking_Hoffman	Promoter flanking regions
16	134438	0.168747	0.044196	4.0348e-05	Unadjusted	Promoter_UCSC	Promoter regions
17	2141208	0.551915	0.016971	0	Unadjusted	Repressed_Hoffman	Repressed regions
18	766061	0.285066	0.055426	1.4856e-06	Unadjusted	SuperEnhancer_Hnisz	Super enhancers

19	615823	0.551915	0.017164	0	Unadjusted	TFBS_ENCODE	Transcription factor binding sites
20	1606468	0.551915	0.015179	0	Unadjusted	Transcribed_Hoffman	Transcribed regions
21	79520	0.212722	0.043037	3.2219e-07	Unadjusted	TSS_Hoffman	Transcription start site
22	50931	0.168377	0.039369	6.1557e-06	Unadjusted	UTR_3_UCSC	Untranslated 3 prime regions
23	23959	0.144357	0.033924	5.3663e-06	Unadjusted	UTR_5_UCSC	Untranslated 5 prime regions
24	98376	0.293961	0.053534	1.1116e-07	Unadjusted	WeakEnhancer_Hoffman	Weak enhancer regions
1	64107	0.202933	0.044394	2.9689e-06	Adjusted	Coding_UCSC	Coding regions
2	118588	0.551915	0.013456	0	Adjusted	Conserved_LindbladToh	Regions conserved in mammals
3	108493	0.338004	0.057618	3.4474e-07	Adjusted	CTCF_Hoffman	CCCTC binding factor
4	641690	0.551915	0.015993	0	Adjusted	DGF_ENCODE	Digital Genomic Footprint
5	803690	0.551915	0.020493	0	Adjusted	DHS_Trynka	DNase I hypersensitive sites
6	19742	0.080119	0.035960	1.1295e-02	Adjusted	Enhancer_Andersson	Enhancer regions - Andersson
7	194938	0.261185	0.057056	1.5098e-05	Adjusted	Enhancer_Hoffman	Enhancer regions - Hoffman
8	402725	0.551915	0.013479	0	Adjusted	FetalDHS_Trynka	Fetal DNase 1 hypersensitivity sites
9	1803420	0.551915	0.017473	0	Adjusted	H3K27ac_Hnisz	H3K27ac - Hnisz
10	1239903	0.551915	0.017418	0	Adjusted	H3K27ac_PGC2	H3K27ac - PGC
11	2014927	0.551915	0.019593	0	Adjusted	H3K4me1_Trynka	H3K4me1
12	609687	0.551915	0.015594	1.7208e-15	Adjusted	H3K4me3_Trynka	H3K4me3

13	580198	0.551915	0.018018	1.6653e-16	Adjusted	H3K9ac_Trynka	H3K9ac
14	1806413	0.551915	0.014737	3.1265e-12	Adjusted	Intron_UCSC	Intronic regions
15	38134	0.143543	0.042730	5.3226e-04	Adjusted	PromoterFlanking_Hoffman	Promoter flanking regions
16	133102	0.156322	0.046282	3.8456e-04	Adjusted	Promoter_UCSC	Promoter regions
17	2138736	0.551915	0.018897	0	Adjusted	Repressed_Hoffman	Repressed regions
18	762870	0.295578	0.057192	4.1703e-06	Adjusted	SuperEnhancer_Hnisz	Super enhancers
19	613009	0.551915	0.018948	0	Adjusted	TFBS_ENCODE	Transcription factor binding sites
20	1603094	0.551915	0.014751	0	Adjusted	Transcribed_Hoffman	Transcribed regions
21	78404	0.198379	0.044908	8.3552e-06	Adjusted	TSS_Hoffman	Transcription start site
22	50563	0.151648	0.040614	9.0531e-05	Adjusted	UTR_3_UCSC	Untranslated 3 prime regions
23	23734	0.128428	0.035321	1.5844e-04	Adjusted	UTR_5_UCSC	Untranslated 5 prime regions
24	98018	0.283909	0.055832	2.5575e-06	Adjusted	WeakEnhancer_Hoffman	Weak enhancer regions

h2g, heritability; SE, standard error; SNPs from the MHC region on chromosome six were not removed for the unadjusted analysis, whereas they were removed for the adjusted analysis, as well as the inclusion of principal components





Supplementary Figure 3.1 - Manhattan and QQ-plot of SAX-discovery GWAS

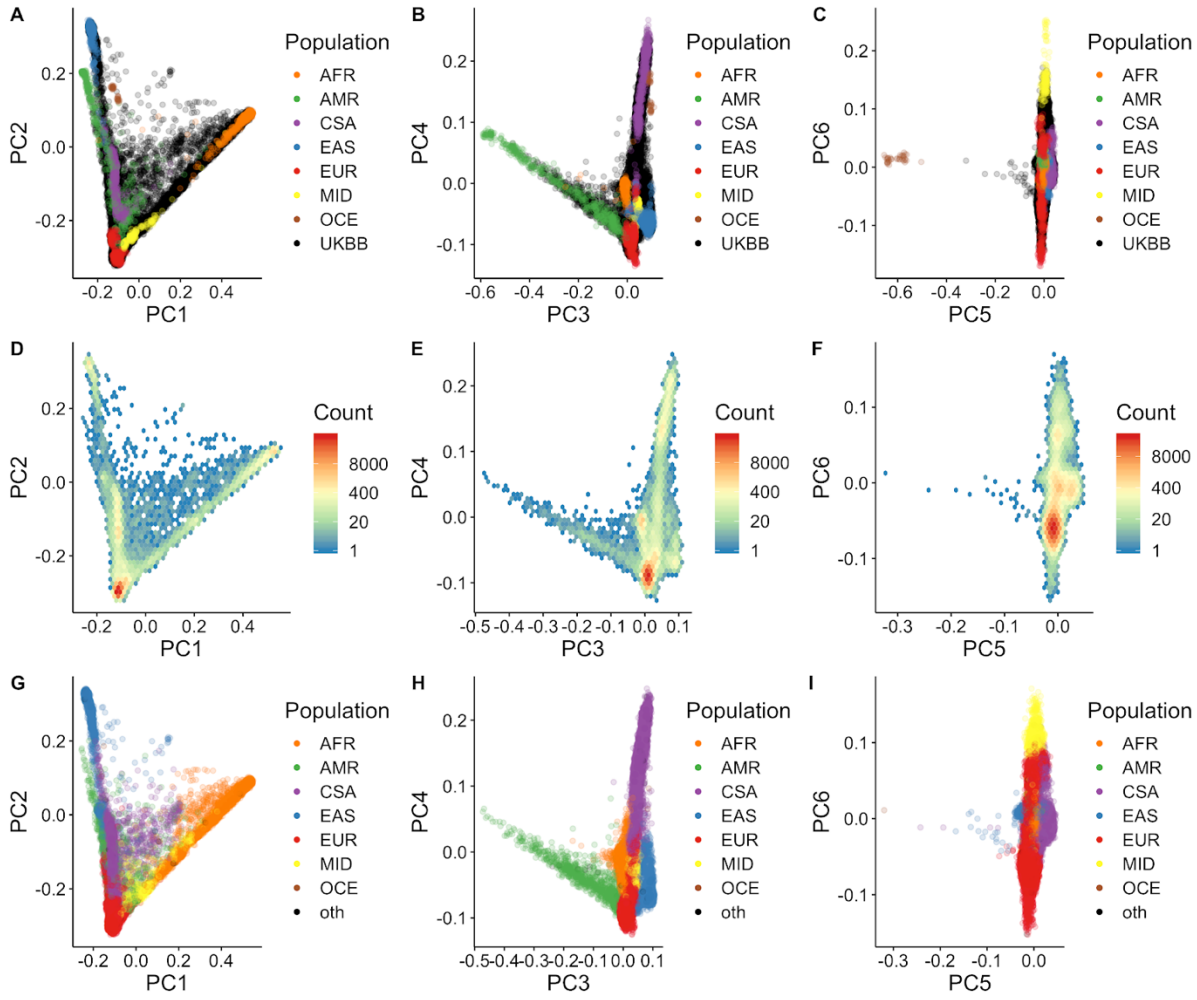
Supplementary Table 3.5 - R^2 values for polygenic risk scores computed within and across ancestries

R^2	CI2.5	CI97.5	*P*-value	*P*-threshold	discovery cohort	target cohort	n target
0.000063	0.001079	-0.0050	0.3020	all	SAX discovery	SAX target	1044
-0.000762	-0.0005795	-0.0073	0.6528	0.5	SAX discovery	SAX target	1044
-0.000838	-0.0007264	-0.0074	0.7261	0.2	SAX discovery	SAX target	1044
-0.000514	-0.0000836	-0.0049	0.4970	0.1	SAX discovery	SAX target	1044
0.000051	0.0010439	-0.0087	0.3049	0.05	SAX discovery	SAX target	1044
-0.000125	0.0006990	-0.0076	0.3513	0.001	SAX discovery	SAX target	1044
-0.000442	0.0000544	-0.0049	0.4637*	0.0001	SAX discovery	SAX target	1044
0.002973	0.0068102	-0.0065	0.0425	0.00001	SAX discovery	SAX target	1044
-0.000928	-0.0009113	-0.0052	0.8650	0.000001	SAX discovery	SAX target	1044
-0.000431	0.0000840	-0.0080	0.4565	all	PGC-EUR discovery	SAX entire cohort	2087
-0.000741	-0.0005268	-0.0045	0.6281	0.5	PGC-EUR discovery	SAX entire cohort	2087
-0.000897	-0.0008325	-0.0045	0.7846	0.2	PGC-EUR discovery	SAX entire cohort	2087
-0.000664	-0.0003747	-0.0079	0.5747	0.1	PGC-EUR discovery	SAX entire cohort	2087
-0.000479	-0.0000065	-0.0046	0.4773	0.05	PGC-EUR discovery	SAX entire cohort	2087
-0.000969	-0.0009768	-0.0083	0.9815	0.001	PGC-EUR discovery	SAX entire cohort	2087
-0.000648	-0.0003472	-0.0084	0.5651	0.0001	PGC-EUR discovery	SAX entire cohort	2087
0.001242	0.0034343	-0.0070	0.1310	0.00001	PGC-EUR discovery	SAX entire cohort	2087
0.005758	0.0112782	-0.0090	0.0084*	0.000001	PGC-EUR discovery	SAX entire cohort	2087
0.001823	0.0045612	-0.0098	0.0897	0.00000005	PGC-EUR discovery	SAX entire cohort	2087
-0.000435	0.0000788	-0.0075	0.4580	all	PGC-EAS discovery	SAX entire cohort	2087
0.000583	0.0020811	-0.0071	0.2059	0.5	PGC-EAS discovery	SAX entire cohort	2087
-0.000931	-0.0009134	-0.0057	0.8436	0.2	PGC-EAS discovery	SAX entire cohort	2087

-0.000916	-0.0008780	-0.0053	0.8155	0.1	PGC-EAS discovery	SAX entire cohort	2087
-0.000694	-0.0004340	-0.0063	0.5942	0.05	PGC-EAS discovery	SAX entire cohort	2087
-0.000660	-0.0003619	-0.0061	0.5724	0.001	PGC-EAS discovery	SAX entire cohort	2087
0.005946	0.0124894	-0.0044	0.0075*	0.0001	PGC-EAS discovery	SAX entire cohort	2087
0.001832	0.0045952	-0.0076	0.0891	0.00001	PGC-EAS discovery	SAX entire cohort	2087
-0.000608	-0.0002578	-0.0068	0.5418	0.000001	PGC-EAS discovery	SAX entire cohort	2087
-0.000805	-0.0006448	-0.0052	0.6805	0.00000005	PGC-EAS discovery	SAX entire cohort	2087
0.002942	0.0068268	-0.0068	0.0434*	all	SAX-discovery + PGC-EUR + PGC-EAS	SAX target	1044
-0.000506	-0.0000699	-0.0057	0.4932	0.5	SAX-discovery + PGC-EUR + PGC-EAS	SAX target	1044
0.000128	0.0011898	-0.0058	0.2871	0.2	SAX-discovery + PGC-EUR + PGC-EAS	SAX target	1044
0.000883	0.0026858	-0.0075	0.1655	0.1	SAX-discovery + PGC-EUR + PGC-EAS	SAX target	1044
0.002536	0.0059829	-0.0065	0.0559	0.05	SAX-discovery + PGC-EUR + PGC-EAS	SAX target	1044
-0.000829	-0.0007102	-0.0069	0.7155	0.001	SAX-discovery + PGC-EUR + PGC-EAS	SAX target	1044
0.000036	0.0010137	-0.0074	0.3087	0.0001	SAX-discovery + PGC-EUR + PGC-EAS	SAX target	1044
0.001813	0.0045513	-0.0078	0.0887	0.00001	SAX-discovery + PGC-EUR + PGC-EAS	SAX target	1044
-0.000902	-0.0008573	-0.0056	0.8121	0.000001	SAX-discovery + PGC-EUR + PGC-EAS	SAX target	1044
-0.000585	-0.0002349	-0.0048	0.5336	0.00000005	SAX-discovery + PGC-EUR + PGC-EAS	SAX target	1044

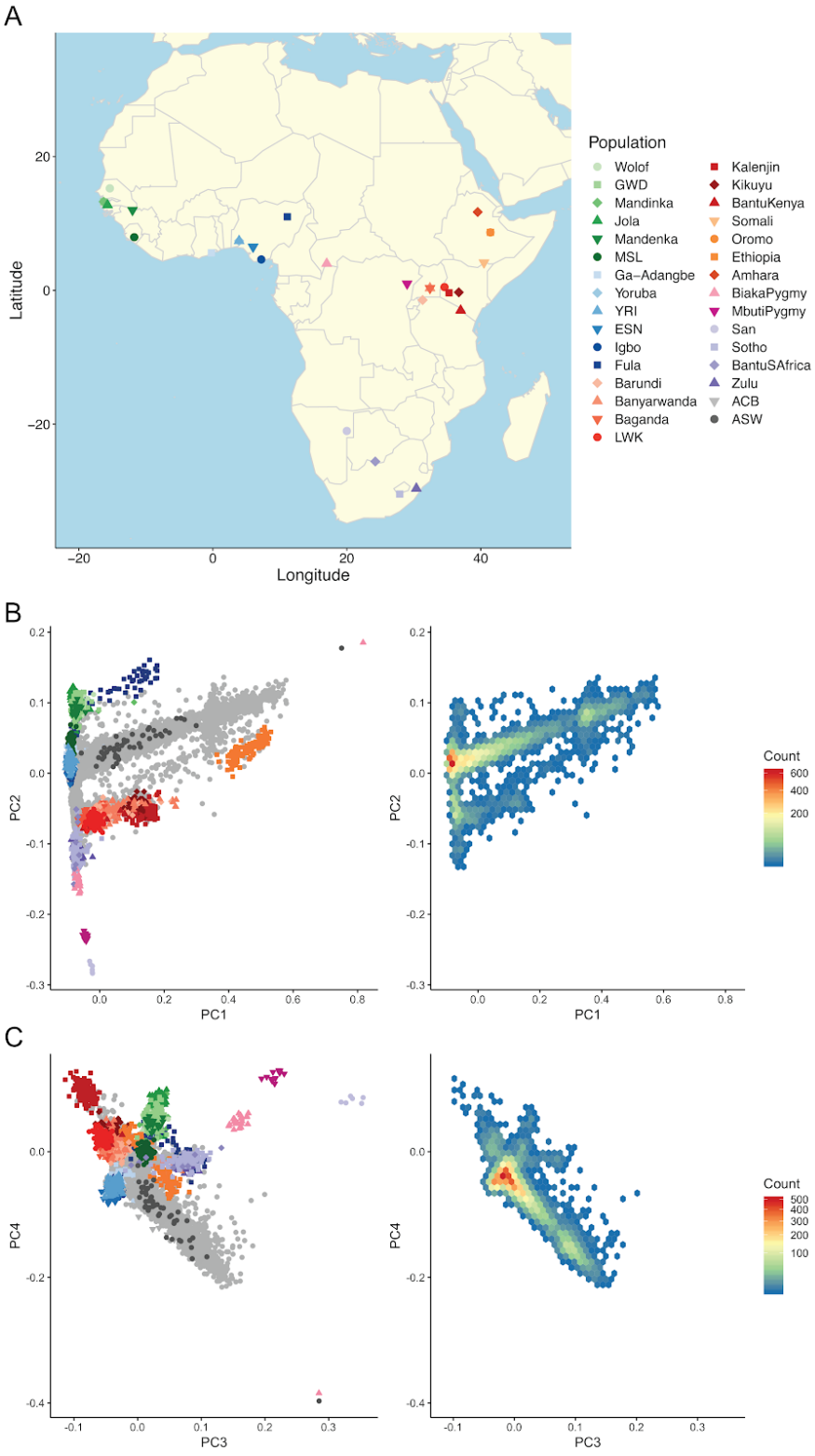
R^2, Nagelkerke's R^2; CI2.5, lower bound of the 95% confidence interval; CI97.5, upper bound of the 95% confidence interval; significance, the level of significance for the p-value * indicates p-value less than 0.05; n target, the number of samples in the target cohort.

Appendix 6 — Supplementary information for chapter 4

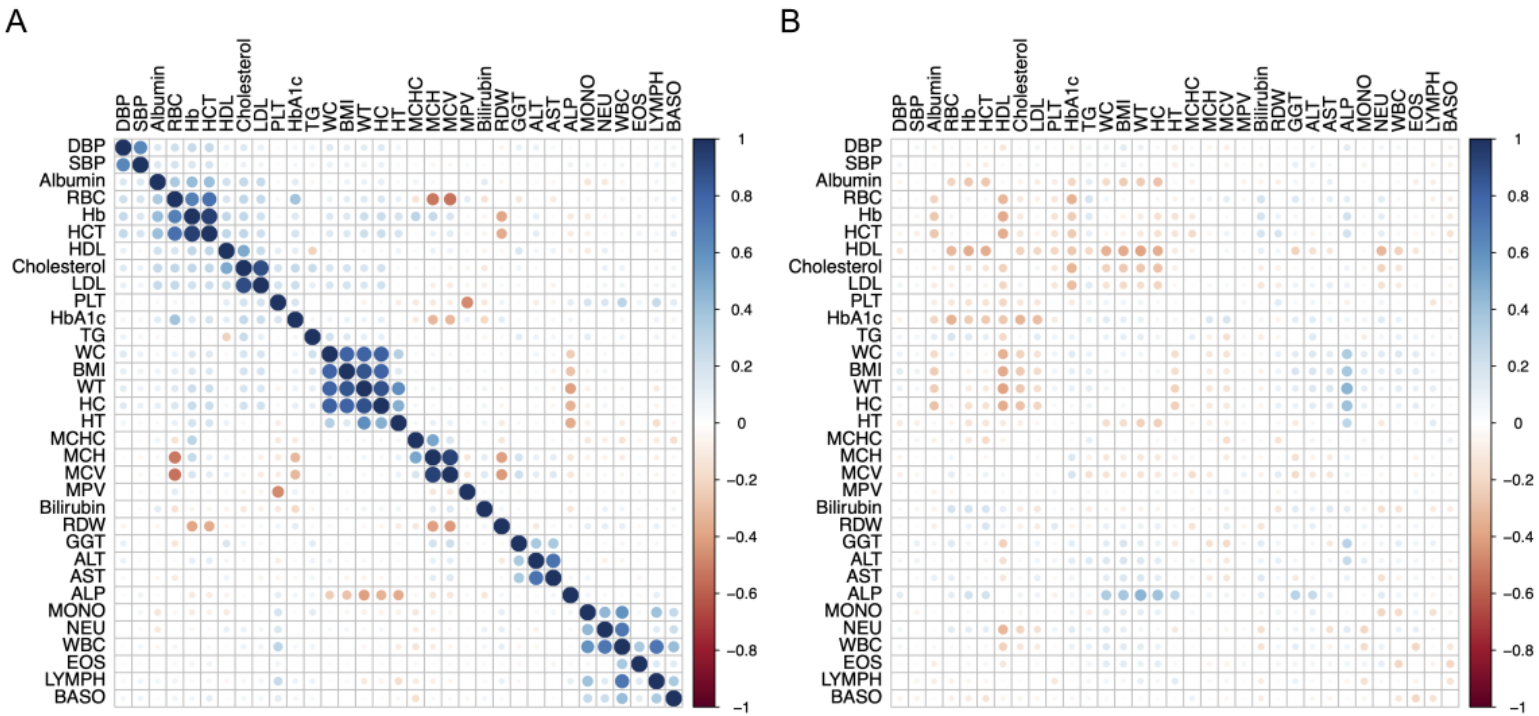


Supplementary Figure 4.1. - Continental ancestries in the UKB data with reference data from the 1000 Genomes Project and Human Genome Diversity Panel (HGDP)

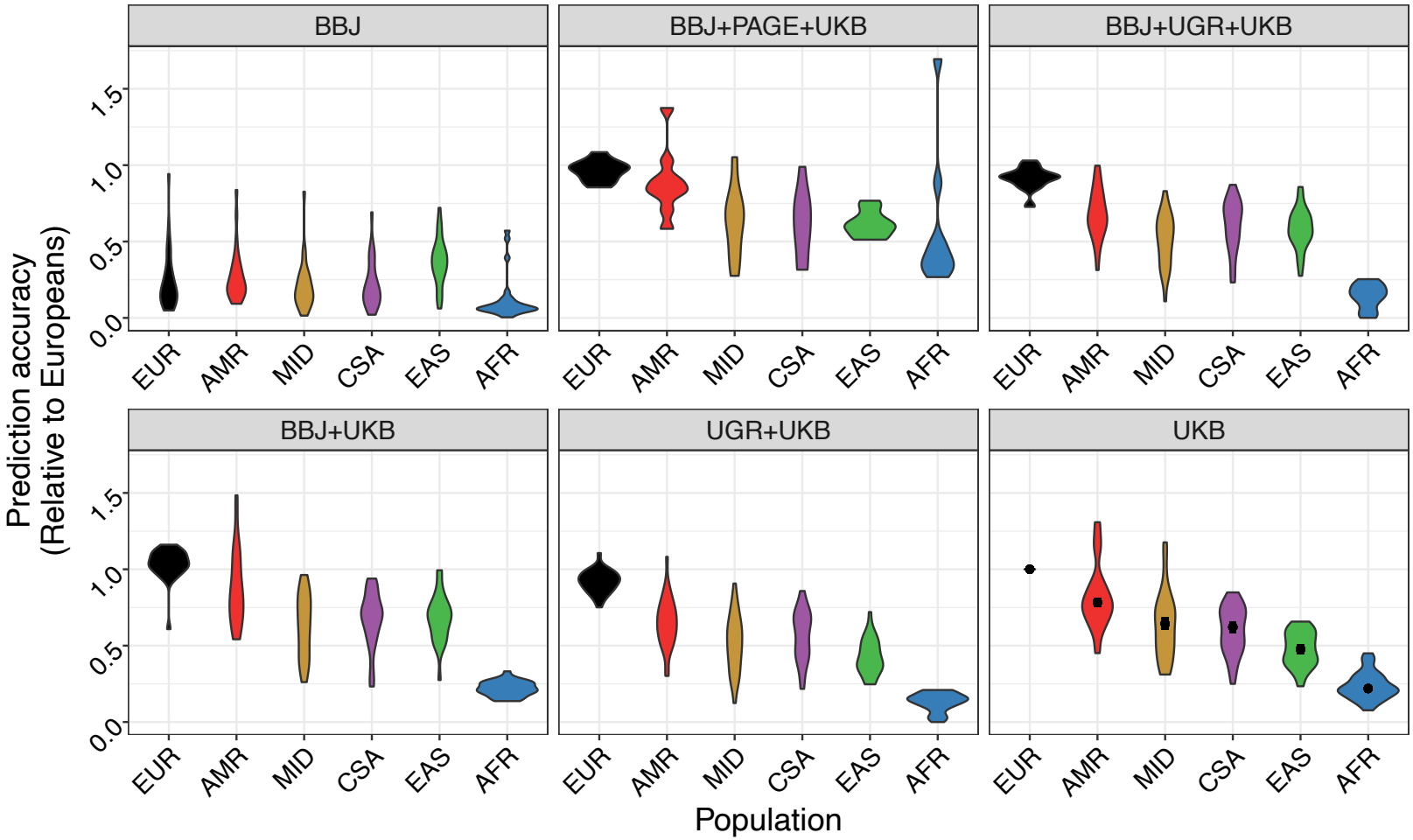
A-C) Principal components analysis biplots (PCs 1-6) with loadings defined by reference data from 1000 Genomes and HGDP with colours corresponding to continental ancestry meta-data from these projects. UKB data (UKBB) is projected into the same PC space and coloured in black. D-F) Density of UKB data (excluding reference panels) in PC1-6. G-I) Continental ancestry assignments in the UKB using a random forest trained on meta-data from 1000 Genomes and HGDP (excluding reference panels). “Oth” are individuals whose ancestry was not confidently assigned to any ancestry group.



Supplementary Figure 4.2 - African subcontinental ancestry in the UKB
A) Coordinates of African reference panel data from the African Genome Variation Project, 1000 Genomes Project, and Human Genome Diversity Panel. B) PC1-2 of UKB and/or reference data. C) PC3-4 of UKB and/or reference data.
B-C) Colors and shapes correspond to A), while grey points in the left plots are UKB projected PC coordinates. Right plots show density of AFR ancestry UKB individuals.



Supplementary Figure 4.3 - Phenotype correlations among 34 quantitative traits measured in the Uganda GPC data versus UKB data.
Analysis conducted as in **Figure 4.8**. A) Phenotypic correlation matrix when relatives are included in the Uganda GPC. B) Differences between phenotypic correlation matrices among unrelated individuals in the UKB - Uganda GPC.



Supplementary Figure 4.4 - PRS accuracy from a homogeneous versus multi-ancestry discovery dataset
All relative comparisons are with respect to accuracy in withheld EUR when predicting with UKB GWAS summary statistics

Supplementary Table 4.1 - Description of the phenotypes assessed in Uganda GPC and UKB

Trait type	Phenotype code	Description
Liver function	Albumin	Serum albumin test
Liver function	ALP	Alkaline phosphatase test
Liver function	ALT	Alanine aminotransferase test
Liver function	AST	Aspartate aminotransferase test
Blood factor	BASO	Basophil count
Liver function	Bilirubin	Bilirubin
Anthropometric index	BMI	Body mass index
Lipid test	Cholesterol	Total cholesterol
Blood pressure	DBP	Diastolic blood pressure
Blood factor	EOS	Eosinophil count
Liver function	GGT	Gamma-glutamyl transpeptidase test
Blood factor	Hb	Hemoglobin
Glycemic control	HbA1c	HbA1c2
Anthropometric index	HC	Hip circumference
Blood factor	HCT	Hematocrit
Lipid test	HDL	High-density lipoprotein
Anthropometric index	HT	Height
Lipid test	LDL	Low-density lipoprotein
Blood factor	LYMPH	Lymphocyte count
Blood factor	MCH	Mean corpuscular hemoglobin
Blood factor	MCHC	Mean corpuscular hemoglobin concentration
Blood factor	MCV	Mean corpuscular volume
Blood factor	MONO	Monocyte count
Blood factor	MPV	Mean platelet volume
Blood factor	NEU	Neutrophil count
Blood factor	PLT	Platelet count

Blood factor	RBC	Red blood cell count
Blood factor	RDW	Red blood cell distribution width
Blood pressure	SBP	Systolic blood pressure
Lipid test	TG	Triglycerides
Blood factor	WBC	White blood cell count
Anthropometric index	WC	Waist circumference
Anthropometric index	WT	Weight

Supplementary Table 4.2 - Meta-analysis discovery cohort summaries across phenotypes.

Phenotype code	N (UGR)	N (UKB)	N (BBJ)	N (PAGE)	N (PAGE AFR)	N (PAGE AMR)	N (PAGE EAS)
Albumin	13125	306557	102223	N/A	N/A	N/A	N/A
ALP	9322	334766	105030	N/A	N/A	N/A	N/A
ALT	9401	334622	134182	N/A	N/A	N/A	N/A
AST	8995	333492	134154	N/A	N/A	N/A	N/A
BASO	2681	340121	62076	N/A	N/A	N/A	N/A
Bilirubin	9326	284860	110207	N/A	N/A	N/A	N/A
BMI	13976	349957	158284	49335	17127	22600	4647
Cholesterol	13116	334752	128305	33185	10137	18406	2387
DBP	13618	330693	136615	N/A	N/A	N/A	N/A
EOS	2671	340121	62076	N/A	N/A	N/A	N/A
GGT	8995	334586	118309	N/A	N/A	N/A	N/A
Hb	2741	340718	108769	N/A	N/A	N/A	N/A
HbA1c	6116	334658	42790	11178	559	10412	92
HC	13966	350485	N/A	N/A	N/A	N/A	N/A
HCT	2744	340719	108757	N/A	N/A	N/A	N/A
HDL	13114	306424	70657	33063	10085	18355	2378
HT	14126	350353	159195	49796	17286	22839	4680
LDL	13086	334112	72866	32221	9720	17964	2316
LYMPH	2681	340121	62076	N/A	N/A	N/A	N/A
MCH	2742	340716	108054	N/A	N/A	N/A	N/A
MCHC	2744	340712	108728	19803	3750	15522	128
MCV	2742	340717	108256	N/A	N/A	N/A	N/A
MONO	2681	340121	62076	N/A	N/A	N/A	N/A
MPV	N/A	340714	N/A	N/A	N/A	N/A	N/A
NEU	2671	340121	62076	N/A	N/A	N/A	N/A
PLT	2723	340718	108208	29328	8850	19552	541
RBC	2744	340719	108794	N/A	N/A	N/A	N/A
RDW	2744	340717	N/A	N/A	N/A	N/A	N/A
SBP	13613	330690	136597	N/A	N/A	N/A	N/A

TG	13115	334471	105597	33096	9980	18460	2381
WBC	2741	340714	107964	28608	8825	18857	543
WC	13963	350529	N/A	N/A	N/A	N/A	N/A
WT	14005	350088	N/A	N/A	N/A	N/A	N/A

Supplementary Table 4.3 - Kolmogorov-Smirnov (K-S) and F-tests used to compare overall phenotypic distributions and variances.

Trait type	Phenotype code	Description	K-S test statistic	K-S p-value	F-test statistic	F-test CI 2.5	F-test CI 97.5	F-test p-value
Liver function	Albumin	Serum albumin test	0.09	0	0.45	0.43	0.47	0
Liver function	ALP	Alkaline phosphatase test	0.39	0	0.08	0.08	0.09	0
Liver function	ALT	Alanine aminotransferase test	0.10	0	1.16	1.12	1.21	8.01E-13
Liver function	AST	Aspartate aminotransferase test	0.11	0	0.18	0.18	0.19	0
Blood factor	BASO	Basophil count	0.54	0	518.37	480.89	557.24	0
Liver function	Bilirubin	Bilirubin	0.59	0	0.01	0.01	0.01	0
Anthropometric index	BMI	Body mass index	0.09	0	1.71	1.64	1.78	0
Lipid test	Cholesterol	Total cholesterol	0.08	0	1.55	1.49	1.62	0
Blood pressure	DBP	Diastolic blood pressure	0.03	0.0001216667441	1.05	1.01	1.09	0.01773934371
Blood factor	EOS	Eosinophil count	0.31	0	3.16	2.93	3.40	0
Liver function	GGT	Gamma-glutamyl transpeptidase test	0.07	0	0.33	0.32	0.34	0
Blood factor	Hb	Hemoglobin	0.11	1.89E-15	0.43	0.40	0.46	9.87E-155
Glycemic control	HbA1c	HbA1c2	0.42	0	87.05	83.57	90.61	0
Anthropometric index	HC	Hip circumference	0.06	4.44E-15	1.11	1.06	1.15	2.46E-06
Blood factor	HCT	Hematocrit	0.11	8.66E-15	0.46	0.42	0.49	4.60E-133
Lipid test	HDL	High-density lipoprotein	0.02	0.008592505313	0.80	0.77	0.84	5.35E-28

Anthropometric index	HT	Height	0.06	5.00E-14	0.59	0.57	0.61	2.13E-172
Lipid test	LDL	Low-density lipoprotein	0.07	0	1.46	1.40	1.51	0
Blood factor	LYMPH	Lymphocyte count	0.08	4.18E-08	1.81	1.68	1.94	0
Blood factor	MCH	Mean corpuscular hemoglobin	0.12	0	0.41	0.38	0.44	4.15E-181
Blood factor	MCHC	Mean corpuscular hemoglobin concentration	0.08	3.97E-09	0.80	0.75	0.86	7.32E-10
Blood factor	MCV	Mean corpuscular volume	0.12	0	0.34	0.32	0.37	9.98E-271
Blood factor	MONO	Monocyte count	0.13	0	3.02	2.80	3.24	0
Blood factor	MPV	Mean platelet volume	0.11	1.78E-15	1.74	1.62	1.87	0
Blood factor	NEU	Neutrophil count	0.19	0	2.52	2.34	2.71	0
Blood factor	PLT	Platelet count	0.07	1.56E-07	0.57	0.53	0.61	2.68E-64
Blood factor	RBC	Red blood cell count	0.12	0	0.37	0.34	0.39	5.07E-234
Blood factor	RDW	Red blood cell distribution width	0.14	0	0.49	0.45	0.52	1.13E-109
Blood pressure	SBP	Systolic blood pressure	0.06	3.54E-14	1.31	1.25	1.36	0
Lipid test	TG	Triglycerides	0.21	0	2.75	2.64	2.86	0
Blood factor	WBC	White blood cell count	0.08	4.40E-08	1.81	1.69	1.95	0
Anthropometric index	WC	Waist circumference	0.12	0	2.05	1.96	2.13	0
Anthropometric index	WT	Weight	0.08	0	1.63	1.56	1.70	0