Methylation Quantitative Trait Loci Associated with PTSD in a South African Population

by

Morne Du Plessis

Thesis presented in fulfilment of the requirements for the degree of MSc (Human Genetics) in the Faculty of Medicine and Health Sciences at Stellenbosch University

Supervisors: Prof Sian Hemmings and Prof Soraya Seedat

Co-Supervisors: Dr Patricia Swart, Dr Jacqueline Womersley and Prof Karoline Kuchenbaecker

December 2021

Declaration

By submitting this thesis electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

December 2021

Abstract

Posttraumatic stress disorder (PTSD) is a complex psychiatric disorder characterised by symptoms of intrusive thoughts, avoidance behaviours, hyper-arousal and negative alterations to cognition and mood. PTSD is unique among psychiatric disorders in that it is a consequence of trauma exposure. Yet, studies previously conducted in the USA have shown that although 50-85% of individuals will encounter a traumatic event in their lifetime, the prevailing prevalence of PTSD lies approximately between 1.3 and 12.2%. This discrepancy serves to highlight the existence of factors granting individuals contingent resistance or vulnerability to the development of PTSD. While the molecular mechanisms elemental to PTSD remain largely unknown, prior heritability estimates and epigenome-wide association studies have suggested that the disorder presents both genetic and epigenetic components that mediate risk and resilience to PTSD. This study aimed to integrate genomic and epigenetic data to identify methylation quantitative trait loci (mQTLs) associated with PTSD. Variants of interest were identified through a polygenic risk score (PRS) model constructed to predict PTSD case-control status through the translation of European-derived PRS to a local South African population. The PRS model was subsequently assessed to determine whether DNA methylation variation in our sample was associated with an elevated polygenic risk burden for PTSD. Positional and dosage analysis was then conducted to investigate how any risk-conferring alleles identified were associated with specific methylated regions. PRS were constructed using data pertaining to the Psychiatric Genomic Consortium's largest multi-ethnic genome-wide association study, but were not able to predict case-control status in a cohort of PTSD cases (n = 164) and trauma-exposed controls (n = 163) (p = 0.064). However, upon extracting the most predictive variants, the study was able to identify 44,614 mQTLs acting across 250 variants and 26,344 CpG probes. Moreover, the study identified evidence of substantial interconnectivity between the discovered mQTLs, wherein CpG sites were found to interact with a median of 2 different variants (IQR = 1 – 2) and each variant was found to interact with a median of 3 CpG probes (IQR = 1 – 10.5). Our results further support the hypothesis that the development of PTSD is dependent on an interconnected network of molecular interactions and highlight the need for future studies dedicated towards optimising PRS construction in multi-ethnic populations.

Opsomming

Posttraumatiese stresversteuring (PTSV) is ‘n komplekse psigiatriese versteuring wat gekenmerk word deur simptome van indringende gedagtes, vermydingsgedrag, hiper-opwinding en negatiewe veranderinge aan kognisie en gemoedstoestand. PTSV is uniek onder psigiatriese versteurings omdat dit die gevolg is van blootstelling aan trauma. Tog het studies wat voorheen in die VSA gedoen is, getoon dat alhoewel 50-85% van alle individue gedurende hul leeftyd ‘n traumatiese gebeurtenis sal ervaar, die heersende voorkoms van PTSV ongeveer tussen 1.3 en 12.2% is. Hierdie teenstrydigheid beklemtoon die bestaan van faktore wat individue voorwaardelike weerstand of kwesbaarheid vir die ontwikkeling van PTSV bied. Alhoewel die molekulêre meganismes van PTSV grotendeels onbekend is, het vorige oorerflikheidsramings en epigenoom-wye assosiasie studies voorgestel dat die versteuring beide genetiese en epigenetiese komponente bevat wat die risiko en elastisiteit vir PTSV beinvloed. Hierdie studie het ten doel gehad om genomiese en epigenetiese data te integreer om die kwantitatiewe eienskap loci van metilering (mQTLs) geassosieer met PTSV te identifiseer. Variante van belang is geïdentifiseer deur middel van ‘n poligeniese risikotelling (PRT) model wat geskep is om die PTSV geval-kontrole status te voorspel deur die toepassing van Europese-afgeleide PRT na ‘n plaaslike Suid-Afrikaanse bevolking. Die PRT-model was vervolgens ondersoek om te bepaal of DNA-metilerings variasie in ons monster geassosieer is met ‘n verhoogde poligeniese risikolas vir PTSV. Posisionele- en doseringsanalises is daarna gedoen om te ondersoek hoe enige geïdentifiseerde risiko-allele geassosieer word met spesifieke gemetileerde streke. PRT is geskep met behulp van data wat verband hou met die grootste multi-etniese genome-wye assosiasie studie van die Psigiatriese Genomiese Konsortium, maar kon nie die geval-kontrole status in ‘n groep PTSV-gevalle (n = 164) en trauma-blootgestelde kontroles (n = 163) voorspel nie (*p* = 0.064). Na die onttrekking van die mees voorspellende variante kon die studie egter 44,614 mQTLs identifiseer wat interaksie toon met 250 variante en 26,344 CpG-posisies. Daarbenewens het die studie bewyse van aansienlike interkonnektiwiteit tussen die geïdentifiseerde mQTLs waargeneem, waarin gevind is dat CpG-posisies interaksie het met ‘n mediaan van 2 verskillende variante (IKV = 1 - 2), as ook dat elke variant interaksie het met ‘n mediaan van 3 CpG-posisies (IKV = 1 – 10.5). Ons resultate ondersteun verder die hipotese dat die ontwikkeling van PTSV afhanklik is van ‘n onderling gekoppelde netwerk van molekulêre interaksies en beklemtoon die behoefte aan toekomstige studies wat daarop gemik is om PRT-konstruksie in multi-etniese bevolkings te optimaliseer.

Acknowledgments

The work herein was supported by the Harry Crossley Foundation, National Research Foundation and South African Research Chairs Initiative of the Department of Science and Technology. Moreover, I would like to extend my gratitude to the National Research Foundation and Professor HW Truter Private Bursary for their financial support.

To the students and staff of the Magic Lab, for their constant support and advice – both through formal channels and in spontaneous hallway conversations – which helped make this project what it is today.

To Dr Sylvanus Toikumo, for laying the foundation for the DNA methylation analysis, thank you for all your years of hard work and dedication, and for ceaselessly bringing positivity and joy to the whole Magic Lab.

To Professor Soraya Seedat, for always granting new insight on perspectives that would otherwise have been missed, thank you for helping this project grow ever better and expand upon any angles previously unnoticed.

To Professor Karoline Kuchenbaecker, for shedding light on a thoroughly unfamiliar field, thank you for all your help in navigating the translation of PRS to a South African population – a fascinating, yet often frustrating, technique upon which most of this project depends.

To Professor Sian Hemmings, for an exemplary performance as the primary supervisor of this project, thank you for all of your support and advice over the last few years – I am incredibly grateful for the opportunities for growth and betterment which the research group provides.

To Dr Jacqueline Womersley, for a near unparalleled knowledge on all things caffeinated and statistical, thank you for always having a kind word or offering of motivation in a year more difficult than most.

To Dr Patricia Swart, for answering a near endless number of questions and always listening to ramblings as to how (fairly optimistic) ideas could be incorporated, thank you for your help and guidance during this first foray into the world of bioinformatics.

Lastly, thank you to my parents, Jeanine and Rigard, without whom none of this would have been possible. From elementary school projects attempting to discern the most aerodynamic form of paper airplane to (prizewinning) cakes perfectly encapsulating the inner workings of a cell for high school biology. On a certain bookshelf adorned with photos from throughout the years, I think it's rather fitting that our first copy of 365 Simple Science Experiments still sits just a few feet away from the graduation photos it was arguably responsible for.

Table of Contents

List of Figures

List of Tables

Introduction

In 2009, the South African Stress and Health Survey (SASHS) revealed that up to 73.8% of South Africans had experienced at least one traumatic event across their lifespan (1). Comparable studies conducted in Europe and Japan reported that the nation's counterparts from the northern hemisphere encountered similar events at a prevalence of 54-64% (1). The degree to which individuals are exposed to traumatic circumstances is often reflective of the historical, cultural, and political nuances specific to the geographical region wherein they reside. Tragically, South Africa presents a prolonged history of oppression and political violence – the seeds of a federally sanctioned system of discrimination that was ousted as recently as 1994 by a lengthy liberation movement (1,2).

The progressive migration of individuals towards economic hubs and rampant inequality of both the past and present has allowed for many urban areas to be plagued by a clime of violence (2,3). Yet, any one individual is not equally likely to have experienced a traumatic event. Risk of trauma exposure is often intertwined with sociodemographic factors such as lower socioeconomic status, unemployment and impaired community support systems (4). As such, disadvantaged communities tend to be disproportionately at risk for physical, sexual, and emotional abuse – both as primary victims and as witnesses to the actions of a perpetrator within the community. Nowhere is this more clearly depicted than in the SASHS's report that the average South African citizen encounters approximately 4.3 notable traumatic incidents in their lifetime (2).

This increased rate of trauma exposure puts South Africans at risk for the development of posttraumatic stress disorder (PTSD), a complex psychiatric disorder unique in that its development is a consequence of exposure to a traumatic incident (4,5). Characterised by symptoms of intrusive thoughts, avoidance behaviours, hyper-arousal and negative alterations to cognition and mood, PTSD represents a substantially intrusive impairment of an individual's ability to function on a daily basis (4). Studies have shown that, although the severity of symptoms presented is in itself highly dependent on the nature of the initial traumatic stimulus, PTSD-associated symptoms persist, on average, for 42.3 months after the index trauma (3). Furthermore, of particular concern is the widely replicated finding that PTSD is often found to be comorbid with a plethora of other psychiatric and physical health conditions as well as suicidal thoughts and behaviours (5).

From a strictly theoretical perspective, PTSD is one of the few psychiatric disorders readily suited for the implementation of early detection and intervention measures. Both pharmacological- and behavioural-based preventative methods have shown a marked ability to reduce the likelihood of developing PTSD and at-risk individuals should ideally encounter emergency settings or personnel in the proximate aftermath of the stimulating event (6). However, existing treatment strategies fail to account for individual vulnerabilities and are often too resource-intensive to be applied in a uniform manner (7,8). As such, there is a pressing need for the development of mechanisms capable of allocating resources in a manner designed to protect those in society most vulnerable to the development of PTSD.

Yet, such efforts are complicated by the observed discrepancy between the proportion of individuals who undergo a traumatic event and those who subsequently develop PTSD. Generalizing epidemiological findings across population groups indicates that approximately 50-85% of individuals are expected to encounter at least one traumatic event in their lifetime (9). Despite this, the prevailing prevalence of PTSD is estimated to lie between 1.3 and 12.2% globally (10). Some degree of this variability may be attributed to mitigating factors, such as the individual's perception of the traumatic event (6). Existing literature suggests that heterogeneity in the development of PTSD, as well as the duration and severity of its associated symptoms, may be a product of trauma load – a loosely defined term referring to the collective impact of the type, severity and frequency of trauma experienced combined with the age at which exposure occurred (11).

However, previous attempts to control for the type of trauma encountered have alluded to the existence of an underlying mechanism capable of mediating risk and resilience to PTSD (12). Prior twin and epidemiological studies indicate that the disorder presents an estimated heritability of 40-50%, which suggests that there is at least some partial genetic component functioning in tandem with environmental influences to contribute to the risk of development (9).

Subsequent efforts to expand upon the potential genetic architecture of PTSD have been met with limited success. While several genome-wide association studies (GWAS) have reported significant risk loci, these findings have been almost universally hindered by difficulties surpassing multiple testing correction and failure to replicate across independent cohorts (13). Such complications can largely be credited to the highly polygenic nature underlying most psychiatric disorders. As opposed to a limited collection of high-impact variants, the vast majority of variation associated with PTSD is primarily due to the cumulative effects of common variants spread throughout the genome (14). Previous GWAS conducted on more thoroughly researched psychiatric disorders, such as major depressive disorder, bipolar disorder and schizophrenia, have suggested that incredibly large sample sizes would be needed to generate sufficient statistical power for the accurate detection of variants presenting such miniscule effect sizes (13).

Consequentially, there exists a growing trend within the greater field of psychiatric genetics to promote the formation of international consortia, of which the Psychiatric Genomics Consortium (PGC) is the most notable, dedicated towards the collective pooling of resources to maximise statistical power. Yet, while such developments are encouraging for future endeavours, PTSD has historically been understudied relative to its psychiatric counterparts (9). Wherein the comparative dearth of genomic data has thus far provided limited insight as to the genetic contributions underlying suspected risk.

However, within the studies that have thus far been conducted there has emerged a series of trends alluding to the biological mechanisms through which PTSD may develop (Table 1). While relatively few of the genetic risk loci identified have been functionally characterised within the context of PTSD, several have previously been associated with neuroprotection and neurogenesis, neurotransmitter pathways, immune-related activity as well as both transcriptional and post-transcriptional gene regulation.

Table 1: Previous genome-wide association studies conducted on PTSD

Study	Discovery Sample	Discovery Case-Control Breakdown	Replication Sample	Replication Case-Control Breakdown	Association(s)	Implicated Pathway(s)
Logue *et al.*, (2012) (15)	• European (n = 491)[1]	• Case = 295 • Control = 196	• African American (n = 84)[1] • African American (n = 521)[2]	• Case = 143 • Control = 462	*RORA*	Neuroprotection
Xie *et al.*, (2013) (16)	• European (n = 1,578)[2] • African American (n = 2,766)[2]	• Case = 744 • Control = 3,600	• European (n = 1,899)[2] • African American (n = 744)[2]	• Case = 296 • Control = 2,347	*TLL1*	Neurogenesis
Guffanti *et al.*, (2013) (17)	• European (n = 45)[2] • African American (n = 342)[2] • Other Ethnicities (n = 26)[2]	• Case = 94 • Control = 319	• European (n = 2,541)[2]	• Case = 578 • Control = 1,963	*AC068718.1*	Long non-coding RNA
Wolf *et al.*, (2014) (18)	• European (n = 484)[1,2]	• Case = 292 • Control = 192	-	-	-	-
Nievergelt *et al.*, (2014) (19)	• European (n = 2,179)[1] • African American (n = 205)[1] • Hispanic and Native/Latino American (n = 640)[1] • East Asian/Other (n = 470)[1]	• Case = 940 • Control = 2,554	• European (n = 491)[1]	• Case = 313 • Control = 178	*PRTFTC1*	Tumour suppression
Almli *et al.*, (2015) (20)	• European (n = 45)[1] • African Unspecified (n = 35)[1] • Hispanic and Latino American (n = 57)[1] • Asian Unspecified (n = 6)[1] • Other Ethnicities (n = 4)[1]	• Case = 63 • Control = 84	• African American (n = 2,868)[2]	-	rs717947 (chr4:33652135)	-
Ashley-Koch *et al.*, (2015) (21)	• European (n = 759)[1] • African American (n = 949)[1]	• Case = 710 • Control = 998	-	-	-	-

* [1] *Military-based population;* [2] *Civilian-based population.*

Table 1: Previous genome-wide association studies conducted on PTSD

Study	Discovery Sample	Discovery Case-Control Breakdown	Replication Sample	Replication Case-Control Breakdown	Association(s)	Implicated Pathway(s)
Stein *et al.*, (2016) (22)	• European (n = 5,049)[1] • African American (n = 1,312)[1] • Hispanic and Latino American (n = 1,413)[1]	• Case = 3,167 • Control = 4,607	• European (n = 4,007)[1] • African American (n = 667)[1] • Hispanic and Latino American (n = 1,242)[1]	• Case = 947 • Control = 4,969	*ANKRD55*	Autoimmune and inflammatory disorders
					ZNF626	RNA transcription regulation
Duncan *et al.*, (2017) (23)	• European (n = 9,954)[1,2] • African American (n = 9,691)[1,2] • Hispanic and Latino American (n = 698)[1,2] • South African (n = 387)[1,2]	• Case = 5,239 • Control = 15,491	-	-	-	-
van der Merwe *et al.*, (2018) (24)	• European (n = 9,537)[1,2]	• Case = 2,424 • Control = 7,113	-	-	-	-
Wilker *et al.*, (2018) (11)	• Sub-Saharan African (n = 925)[2]	• Case = 195 • Control = 730	• Sub-Saharan African (n = 371)[2]	• Case = 158 • Control = 213	-	-

** [1] Military-based population; [2] Civilian-based population.*

Table 1: Previous genome-wide association studies conducted on PTSD

Study	Discovery Sample	Discovery Case-Control Breakdown	Replication Sample	Replication Case-Control Breakdown	Association(s)	Implicated Pathway(s)
Nievergelt *et al.*, (2019) (6)	• European (n = 174,659)^{1,2} • African American and African Unspecified (n = 15,339)^{1,2} • Hispanic and Native/Latino American (n = 5,703)^{1,2}	• Case = 29,556 • Control = 166,145	-	-	*HLA-B*	Immune and inflammatory response
					KAZN	Cellular processes
					PARK2	Dopaminergic pathway
					PODXL	Neuro- and synaptogenesis
					SH3RF3	Neurocognition
					ZDHHC14	Adrenergic receptor regulation
					LINC02335	Long non-coding RNA
					LINC02571	
					TUC338	
					MIR5007	MicroRNA
Shen *et al.*, (2020) (25)	• Latino American (n = 3,414)^{2}	• Case = 1,698 • Control = 1,716	-	-	-	-

* ^{1} *Military-based population;* ^{2} *Civilian-based population.*

Logue *et al.,* (2012) reported one of the first PTSD-associated loci in the form of the nuclear hormone receptor *RAR related orphan receptor A* (*RORA*) (15). *RORA* is widely expressed in neuronal tissue and is thought to play a neuroprotective role in hindering the harmful effects of oxidative stress and proinflammatory cytokines (15). This finding is particularly pertinent to the development of PTSD as it aligns with previous thoughts as to the mechanism through which trauma exposure is suspected of influencing the brain – wherein growing consensus indicates that trauma-mediated increases in both oxidative stress and proinflammatory cytokines are capable of introducing functional and structural alterations to neural tissue (26). Interestingly, *RORA* perfectly highlights how genetic susceptibility can contribute to the risk of developing PTSD, in that a pre-existing vulnerability to certain biological processes is further compromised by the aberrant conditions stimulated upon exposure to a traumatic event. Furthermore, *RORA* provided some of the first evidence that PTSD possessed similar genetic underpinnings to that shared by several other psychiatric disorders, including attention deficit hyperactivity disorder, bipolar disorder, autism and major depressive disorder (15).

Findings by Xie *et al.,* (2013) and Nievergelt *et al.,* (2019) further implicated neurocentric mechanisms through *tolloid like 1* (*TLL1*) and *podocalyxin like* (*PODXL*), respectively (6,16). While the respective variants have yet to be functionally verified in humans, mice- and cell culture-based models have previously alluded that the genes play a role in neuro-and synaptogenesis. Tamura *et al.,* (2005) first demonstrated that mice presenting increased levels of *TLL1* expression displayed greater neurogenesis (27). Once paired with subsequent reports that glucocorticoids, integral components of the greater stress response, are capable of reducing *in vitro TLL1* expression, *TLL1* serves as a similar example to *RORA* in how pre-existing vulnerabilities may hinder one's ability to respond to a traumatic event (27). Additionally, previous studies by Vitureria *et al.,* (2010) have shown that *PODXL* aids neural adhesion proteins in promoting neuronal growth and plasticity, noting that simulating decreased *PODXL* expression *in vitro* resulted in a lower degree of synapse formation across central nervous system derived neural samples (28).

Nievergelt *et al.,* (2019) also identified two genes, *parkin RBR E3 ubiquitin protein ligase* (*PARK2*) and *zinc finger DHHC-type palmitoyltransferase 14* (*ZDHHC14*), which potentially implicate aberrant neurotransmitter regulation in the development of PTSD (6). *PARK2* is closely interlinked with the molecular machinery responsible for regulating dopamine transport across cellular membranes (29). Previous studies have shown that fear extinction is primarily driven by an overriding stimulus from the dopaminergic system, wherein activation of the reward pathway serves to indicate that the fear-inducing stimulus has been nullified (30). Therefore, impairment in the degree to which dopamine is successfully transported can alter synaptic dopamine concentration and may prevent adequate fear extinction, thus simulating a critical component of the PTSD phenotype. Moreover, *ZDHHC14* has been associated with the functioning of beta-adrenergic receptors – neurotransmitter pathways which are suspected of directly contributing to both the intrusive thoughts and hyper-arousal symptom clusters under PTSD (31,32).

In addition to uncovering innate neurocentric mechanisms, previous PTSD GWAS have also highlighted the potential role played by disparate immune functioning. While its function has yet to be identified, *ankyrin repeat domain 55* (*ANKRD55*), as implicated by Stein *et al.,* (2016), has previously been found to be associated with a number of autoimmune and inflammation-based disorders (e.g. rheumatoid arthritis and type 2 diabetes) (22). Interestingly, this finding aligns with previous epidemiological reports that both PTSD and schizophrenia are independently comorbid for a similar range of immune disorders – a fact which hints at the existence of a shared immunological pathophysiology underlying distinct psychiatric disorders (22). Additionally, the involvement of *major histocompatibility complex B class 1* (*HLA-B*), as reported by Nievergelt *et al.,* (2019), further supports previous findings that PTSD presents increased levels of proinflammatory cytokines accompanied by widespread immune dysregulation (33,34).

Functional implications of the various microRNA (*MIR5007*) and long non-coding RNA (*AC068718.1*, *LINC02335*, *LINC02571* & *TUC338*) identified are substantially less evident. Barring subsequent functional analysis, current thoughts are limited to the theoretical capabilities of each molecular class – wherein microRNAs and long non-coding RNAs are widely thought to influence both transcriptional and post-transcriptional gene regulation (35,36). Moreover, this regulatory ability appears to be further advanced by Stein *et al.,* (2016)'s suggested involvement of *zinc finger protein 626* (*ZNF626*), which encodes a small protein structure notable for its ability to regulate transcriptional activity (22).

Yet, despite their findings, the existing PTSD GWAS have been accompanied by a number of notable caveats. The vast majority of studies have primarily been conducted utilising military-based individuals sourced from European and African American ancestry groups. Such limited sampling raises questions as to the portability of these findings both to civilian populations, where trauma load differs drastically from the unique conditions of active combat, and to ancestrally diverse populations from other geographical locations. These concerns are further exacerbated by the manner in which several studies used a combination of singular- and trans-ethnic cohort groupings in an attempt to replicate any findings across all recruited individuals. Specifically, significant associations tended to be found in one particular ancestry subset, with minimal, or often no, indications of replication present across the other ethnic groupings. This would appear to suggest that PTSD may present ancestry-mediated differences in causal variants, a fact which further emphasises the need for greater participant diversity in future attempts to unravel the genetic mechanisms underlying the disorder.

Nevertheless, the potential benefits of a consortia-based approach are clearly evident when comparing the findings of Nievergelt *et al.,* (2019) against that of the field. The PGC was able to assemble a combined sample orders of magnitude larger than that typically possible for a single academic institution or group. Moreover, increasing statistical power allowed the study to identify several significant associations – wherein each of the previously implicated pathways all featured in a single analytical run. As such, striving towards the assembly of ever-greater sample sizes may be critical to uncovering the broader spectrum of genetic contributions mediating risk and resilience to PTSD.

However, increasing interest in the idea that one's risk of developing PTSD is the product of interactions between intrinsic and environmental factors has drawn attention to the potential role played by epigenetic modifications. Briefly, epigenetic changes refer to the introduction of heritable structural alterations capable of influencing gene expression without physically changing the basal genetic code (37). One such adjustment, DNA methylation, has been found to be particularly susceptible to external forces – wherein a range of toxins, stressors and physical health conditions have been shown to replicate the methyl alterations typically reserved for innate biological and developmental mechanisms (38,39).

When considered within the context of previous evidence demonstrating how trauma exposure is associated with the presentation of differential DNA methylation patterns, the mediatory effects that environmental influences hold over DNA methylation are framed as the most likely method through which external stimuli are translated to tangible effects on the development of PTSD (40,41). Subsequent epigenome-wide association studies (EWAS) have further supported such discourse with reports linking PTSD status to differential DNA methylation among several pathways previously implicated in PTSD GWAS (Table 2). Specifically, pathways pertaining to neuronal development as well as the respective immune- and stress-responses appear to present the greatest degree of concordance with existing genomic findings.

Table 2: Previous epigenome-wide association studies conducted on PTSD

Study	Discovery Sample	Discovery Case-Control Breakdown	Association(s)	Implicated Pathway(s)
Smith *et al.*, (2011) (42)	• African American (n = 110)[2]	• Case = 50 • Control = 60	*TPR*	Stress response
			ANXA2	Immune and inflammatory response
			APC5	
			CLEC9A	
			TLR8	
Rutten *et al.*, (2017) (43)	• European (n = 191)[1]	• Case = 67 • Control = 124	*COL1A2*	Collagen formation
			DUSP22	Neurogenesis
			NINJ2	
			HIST1H2APS2	Histone structure
			HOOK2	Mitosis-related processes
			MYT1L	Neurocognition
			PAX8	Neurogenesis & Endocrine regulation
			SDK1	Immune response
Kuan *et al.*, (2017) (44)	• European (n = 304)[2] • Other Ethnicities (n = 69)[2]	• Case = 171 • Control = 202	-	-
Mehta *et al.*, (2017) (45)	• European (n = 96)[1]	• Case = 48 • Control = 48	*BRSK1*	Neurotransmitter release
			DOCK2	Inflammatory response
			LCN8	Epididymis-specific expression
			NGF	Neuro- and synaptogenesis

* [1] *Military-based population;* [2] *Civilian-based population.*

Table 2: Previous epigenome-wide association studies conducted on PTSD

Study	Discovery Sample	Discovery Case-Control Breakdown	Association(s)	Implicated Pathway(s)
Uddin *et al.*, (2018) (46)	- European (n = 164)[2] - African American (n = 343)[2] - Other Ethnicities (n = 38)[2]	- Case = 196 - Control = 349	*HGS*	Immune and inflammatory response
			NRG1	Neuro- and synaptogenesis & Stress response
Mehta *et al.*, (2019) (47)	- European (n = 38)[1]	- Case = 16 - Control = 22	*CCDC88C*	Spinal-related disorders
Snijders *et al.*, (2020) (48)	- European (n = 211)[1] - African American (n = 10)[1] - Other Ethnicities (n = 45)[1]	- Case = 123 - Control = 143	*HEXDC*	Autoimmune and inflammatory disorders
			MAD1C1	Mitosis-related processes
			SPRY4	Stress response
Logue *et al.*, (2020) (49)	- Ancestry not Reported (n = 513)[1]	- Case = 378 - Control = 135	*G0S2*	Cellular signalling
Smith *et al.*, (2020) (12)	- Ancestry not Reported (n = 1,896)[1,2]	- Case = 758 - Control = 1,138	*AHRR*	Smoking-related
Katrinli *et al.*, (2021) (50)	- African American (n = 521)[2] - Other Ethnicities (n = 33)[2]	- Case = 187 - Control = 367	-	-

* [1] *Military-based population;* [2] *Civilian-based population.*

Rutten *et al.*, (2017) (43), Mehta *et al.*, (2017) (45) and Uddin *et al.*, (2018) (46) all identified epigenome-wide significant associations across multiple genes that had previously been shown to be capable of regulating neurogenesis: *dual specificity phosphatase 22* (*DUSP22*) (51), *ninjurin 2* (*NINJ2*) (52), *paired box 8* (*PAX8*) (53), *nerve growth factor* (*NGF*) (54) and *neuregulin 1* (*NRG1*). Moreover, *NGF*, *PAX8* and *NRG1* were linked to supplementary neurocentric mechanisms in the form of aided synaptogenesis, the ability to regulate endocrine functioning, and altered hypothalamus-pituitary-adrenal axis activity, respectively (46,53,54). The potential effects that *PAX8* may hold over the disparate production of thyroid hormones is a particularly intriguing concept, as the aberrant sleeping patterns often observed under hyperthyroidism are also replicated in a lesser fashion under PTSD (55).

Furthermore, existing EWAS indicate that the abnormal immune- and inflammation-based processes thought to underlie PTSD may in fact be the product of a both widespread and multifaceted dysregulation in immune response. Indeed, significant epigenetic changes are routinely observed as affecting several distinct components of traditional immunity. *Annexin A2* (*ANXA2*) and *hepatocyte growth factor-regulated tyrosine kinase substrate* (*HGS*), as implicated by Smith *et al.*, (2011) and Uddin *et al.*, (2018), have previously been linked with an increased proinflammatory response specifically within peripheral systems (42,46,56). However, Smith *et al.*, (2011) also observed *anaphase promoting complex subunit 5* (*APC5*) and *c-type lectin domain containing 9A* (*CLEC9a*) – two genes which, in like manner to Mehta *et al.*, (2017)'s finding of *dedicator of cytokinesis 2* (*DOCK2*), were reported as being exclusively associated with neuroinflammation (42,45,57). Moreover, both *toll like receptor 8* (*TLR8*) and *sidekick cell adhesion molecule 1* (*SDK1*), as implicated by Smith *et al.*, (2011) and Rutten *et al.*, (2017), have previously been documented as playing a role in innate pathogen detection (42,43,58,59). Lastly, Snijders *et al.*, (2020) observed *hexosaminidase D* (*HEXDC*) – a gene which, while it's functional mechanisms have yet to be ascertained, is linked with the presentation of the inflammatory disorder rheumatoid arthritis (48).

In addition to the findings reflective of altered neurocentric and immune-based mechanisms, previous PTSD EWAS have also provided nominal evidence of epigenetic-mediated dysregulation in the stress response. Specifically, aberrations in *translocated promoter region, nuclear basket protein* (*TPR*) and *sprouty RTK signalling antagonist 4* (*SPRY4*), which pertain to glucocorticoid receptor assembly and cellular signalling within the stress response pathway, respectively (42,48).

Suggestive associations notwithstanding, epigenetic studies are often plagued by similar issues to their genetic counterparts, where existing limitations in sample sizes have greatly hampered attempts to replicate findings across differing cohorts (5). As such, the molecular mechanisms elemental to PTSD remain largely unknown (60). This has greatly hindered subsequent efforts to facilitate improvement in existing prevention and mitigation strategies as well as the identification of biologically relevant pharmacological targets (6,11). Furthermore, the disproportionate degree to which studies have been conducted in both European and military-derived populations has amplified an extant lack of understanding as to how current findings translate to ancestrally diverse populations and within general society (13).

However, within the limitations imposed by restricted resources there exists an opportunity to further expand upon current research outside of large-scale consortiums. While studies have historically committed to investigating the various biological domains as independent entities, there is mounting consensus that PTSD may in fact be the product of an extensively integrated network of molecular interactions (61). Such arguments for a systems biology, or multi-omics, approach are largely corroborated by the observation that, while relatively few findings consistently replicate across independent studies, several routinely implicate similar pathways or associated functions (13).

Two particularly enlightening examples thereof pertain to reports regarding the genes *G0/G1 switch 2* (*G0S2*) and *histone deacetylase 4* (*HDAC4*). *G0S2*, which encodes a protein associated with regulating lipid metabolism, was initially identified by Logue *et al.*, (2020) as an epigenome-wide significant association within a cohort of 513 military veterans (49,62). Interestingly, the same association had previously been reported in two distinct mediums – firstly, Daskalakis *et al.*, (2014) observed a reduced expression of *G0S2* transcripts in amygdala- and hippocampus- derived brain tissue obtained during a predator-scent-stress PTSD model in female rats (63). Thereafter, Bam *et al.*, (2016) discovered a similar reduction in gene expression in whole blood samples originally sourced from a small cohort of 10 military veterans (64).

Moreover, while investigating how circulating estrogen levels affected the development of PTSD in a subset of women recruited under the Grady Trauma Project, Maddox *et al.*, (2018) noted an epigenome-wide significant finding associated with increased methylation at *HDAC4* (65). Serving as one of the antitheses of the histone acetyltransferase enzymes, *HDAC4* functions as an histone deacetylase in regulating gene expression through the physical condensing of chromatin (66). In addition to the CpG site of interest, the authors also reported a proximally located single-nucleotide polymorphism (SNP) (rs7570903) that was consistently associated with similar changes in DNA methylation as well as the reduced expression of transcripts encoding *HDAC4*. While the SNP itself was found not to be associated with PTSD status, subsequent analysis determined that the corresponding genotype was significantly linked to increased fear-related reactivity on the fear-potentiated startle task (65).

Considering the suggested benefits of such a multi-omics approach, this study aims to further elucidate the molecular mechanisms underlying PTSD by creating a unified bioinformatics pipeline that will integrate genetic and epigenetic data to identify methylation quantitative trait loci (mQTLs) associated with PTSD. Briefly, mQTLs are genomic loci where specific genetic variants are capable of directly influencing DNA methylation patterns (67). In attempting to identify mQTLs, one would be examining the interconnected relationship between genetic and epigenetic data within the context of PTSD. Traditionally, mQTL mapping would be carried out through dosage-analysis models designed to create matrices comparing large GWAS and EWAS datasets at the individual SNP and CpG-site levels. However, this poses a problem in that the available sample is too small to identify novel genetic variants associated with PTSD in a GWAS. One solution to this would be to instead implement polygenic risk score (PRS) analysis to predict PTSD case-control status within our sample, and subsequently isolate SNPs associated with elevated polygenic risk.

Briefly, PRS are statistical models that provide probabilistic estimates indicative of an individual's genetic predisposition to a particular trait or disorder (68). The technique relies on the inherent polygenicity of complex disorders, wherein increasingly larger GWAS have suggested that the genetic architecture underlying traditionally complex traits is not predicated on the powerful effects of singular variants, but rather stems from the cumulative contributions of hundreds to thousands of variants spread throughout the genome (69). Considering this, one could theoretically quantify genetic risk by calculating the sum of an individual's genotypes across predetermined loci weighted by the effect size for risk as determined by how strongly said variant had previously been associated with the outcome under consideration (70). PRS have become a particularly enticing premise for small-scale genomic studies, as their two-dataset system – in which models are first trained in publicly available GWAS before being applied in smaller host samples – allows one to generate an analytical proxy capable of detecting finer effects than if association testing were conducted on the smaller dataset alone (70).

PRS analysis is a rapidly growing field, and while its predictive ability still needs to improve to qualify for routine use in clinical settings, the associated data produced remains highly biologically relevant in its ability to shed light on the underlying nature of complex diseases. However, much in the same way that PTSD remains relatively understudied as compared to its psychiatric counterparts, there has yet to be substantial research assessing the feasibility of PRS-mediated risk prediction within the disorder. Moreover, the existing literature is further limited in that extant differences in both the analytical approach utilised and the manner in which results are reported renders it difficult to compare inter-study predictive performance.

As such, we have collated a brief summary of previous PTSD-PRS in Table 3 – wherein reported findings were largely restricted to the component of PTSD assessed and whether the generated model was capable of significantly differentiating between the conditions tested.

Table 3: Previous attempts to construct polygenic risk scores for PTSD

Study	Discovery Sample	Discovery Case-Control Breakdown	Reference Dataset	Component Assessed	Predictive Success
Nievergelt *et al.*, (2019) (6)	• European (n = 174,659)[1,2]	• Case = 23,212 • Control = 151,447	PGC-PTSD GWAS (European)	PTSD Status	Significant
Misganaw *et al.*, (2019) (14)	• European (n = 77)[1] • African American (n = 64)[1] • Hispanic and Latino American (n = 85)[1] • Asian (n = 12)[1] • Other Ethnicities (n = 6)[1]	• Case = 128 • Control = 116	PGC-PTSD GWAS (European)	PTSD Status	Significant
				PTSD Symptom Severity	Significant
Schur *et al.*, (2019) (71)	• European (n = 516)[1]	-	PGC-PTSD GWAS (European) & Independent GWAS** (European)	PTSD Symptom Development	Non-Significant
Waszczuk *et al.*, (2020) (72)	• European (n = 1,490)[2]	• Case = 355 • Control = 1,135	PGC-PTSD GWAS (European)	PTSD Status	Non-Significant
Shen *et al.*, (2020) (25)	• Latino American (n = 3,414)[2]	• Case = 1,698 • Control = 1,716	PGC-PTSD GWAS (Multi-ethnic)	PTSD Status	Non-Significant

* [1] *Military-based population;* [2] *Civilian-based population.*

*** Study generated a combined reference using summary statistics derived from the PGC and the Army Study to Assess Risk and Resilience in Servicemembers (STARRS)* (22).

From a performance perspective, extant PTSD-PRS have fared variably – with only two of the five studies identified presenting PRS models capable of significantly predicting the PTSD component under study. Moreover, upon further disassembling the PTSD phenotype into a series of interdependent components, neither Schur *et al.*, (2012) (71) nor Waszczuk *et al.*, (2020) (72) were able to attain significant prediction. Schur *et al.,* (2019) conducted their study in a longitudinal manner, wherein a cohort of military veterans were repeatedly assessed for PTSD symptom development at a series of interviews over five years. The primary benefit of such a strategy lies in that it allows one to better characterise overall risk by identifying subthreshold individuals who may not have been detected had the cohort been assessed at a singular moment. However, Schur *et al.,* (2019) reported that no PRS model achieved predictive significance at any of the timepoints tested (71). Moreover, upon conducting their initial assessment Waszczuk *et al.*, (2020) took the opportunity to reinterview their recruited participants so as to further characterise each individual with respect to specific PTSD symptoms – wherein PRS were subsequently recalculated to assess predictive performance against re-experiencing, avoidance, numbing and hyper-arousal symptoms (72). Yet, despite narrowing the manner in which outcomes were defined, PTSD-PRS were still not able to significantly predict the trait tested (72).

Additionally, there is a distinct lack of PTSD-PRS studies centred around both non-European and multi-ethnic populations. The majority of studies conducted thus far have utilised European summary statistics to predict PTSD status in individuals of European ancestry. Subsequent attempts to rectify this have been met with little success. Upon repeating their initial analysis, Misganaw *et al.*, (2019) failed to generate successful PRS when applying both European and African summary statistics to the African American subset of their study cohort (14). Furthermore, the Peruvian cohort employed by Shen *et al.,* (2020) represents the first instance of PTSD-PRS being applied in a multi-ethnic population (25). However, despite utilising multi-ethnic summary statistics, the study found that PTSD-PRS was not capable of distinguishing between PTSD status in its sample population. Most importantly to this thesis, there has thus far been no studies assessing the feasibility of PTSD-PRS in African populations.

Clearly, there exists a strong need for both increased efforts in assessing the overall validity of PTSD-PRS as well as increased contributions to studies implementing PTSD-PRS in non-European populations. As such, this thesis has two broad aims which it intends to accomplish: (i) the utilization of PRS modelling to construct a variable capable of predicting PTSD case-control status within a South African population; and (ii) the integration of PRS variants and DNA methylation data to identify mQTLs associated with PTSD.

Methods

1. Participants

Our sample consisted of 327 individuals (164 PTSD cases & 163 trauma-exposed controls (TEC)) recruited as part of a larger project investigating the potential commonality of genomic, neural, cellular and environmental features observed across neuropsychiatric disorders and cardiovascular disease risk (*Understanding the SHARED ROOTS of Neuropsychiatric Disorders and Modifiable Risk Factors for Cardiovascular Disease*). Participants were enrolled via subjective sampling measures applied to one of three potential recruitment avenues: (i) hospitals and community clinics located in the greater area surrounding Tygerberg Hospital in the Western Cape; (ii) print, radio and web-based advertisements; and (iii) referrals from the Mental Health Information Centre at Stellenbosch University's Department of Psychiatry (Faculty of Medicine and Health Sciences, Tygerberg, South Africa). According to Shared Roots inclusion criteria, all participants were at least 18 years of age and self-identified as members of the South African Mixed Ancestry (South African Coloured) ethnic group, hereafter referred to as "SAC". Recruitment efforts were restricted to a single ethnic group so as to limit undue confounding effects due to population structure. Furthermore, inclusion required that participants could proficiently read and write in either English or Afrikaans and were committed to attending subsequent follow-up meetings. All recruited subjects were matched for age, gender, education and socioeconomic status. Ethics approval for the parent study was granted by the Stellenbosch University Health Research Ethics Committee institutional review board (Ethics Approval Number: N13/08/115). The recruited sample lacked sufficient statistical power for the detection of PTSD-associated variants through GWAS. Preliminary analysis using notable variants identified in an international meta-analysis of PTSD GWAS indicated that the study presented approximately 42% power for the detection of an appropriate candidate SNP under a genome-wide significance threshold of 5e-8 (minor allele frequency = 34%; odds ratio = 1.12) (6).

2. Demographic and clinical assessments

All participants underwent assessments in the form of diagnostic interviews and a series of clinical questionnaires administered by qualified medical personnel in the Department of Psychiatry, Stellenbosch University. PTSD was diagnosed using the Clinician Administered Posttraumatic Stress Disorder Scale for DSM-5 (CAPS-5) (73). A CAPS-5 total severity score of 23 or higher was used to separate participants into the PTSD cohort. Childhood trauma exposure was evaluated through the Childhood Trauma Questionnaire (CTQ) (74). Extant literature has suggested that CTQ scores can be divided into a series of representative intervals, wherein the proposed cut-offs are reflective of none to minimal trauma exposure (25 – 36), low to moderate trauma exposure (41 – 51), moderate to severe trauma exposure (56 – 68) and severe to extreme trauma exposure (73 – 125) (75,76). For this study, a score of 41 or higher on the CTQ was used as a screening cut-off value to categorically identify participants that met the basal requirement for having had experienced low childhood trauma. Current major depressive disorder (MDD) was identified using the Mini-International Neuropsychiatric Interview (M.I.N.I v6.0) (77). While phenotypically varied, increasing evidence suggests that common psychiatric disorders share overlapping genetic architecture (78). Thus, to avoid potential confounding, any study participant presenting evidence of any other major psychiatric disorder (e.g. schizophrenia or bipolar disorder) as per the M.I.N.I v6.0 was excluded from the current study.

Demographic information pertaining to individual age, gender and medical history was also obtained from each participant. Both smoking status and alcohol use were assessed according to individual responses given to the study-specific medical questionnaire – wherein any lifetime intake of nicotine or alcohol was regarded as presenting a history of smoking or alcohol use. Participants were also screened for Metabolic Syndrome (MetS) according to internationally agreed-upon practices established by the efforts of Alberti *et al.,* (2009) as the joint interim statement of the International Diabetes Federation Task Force on Epidemiology and Prevention; National Heart, Lung, and Blood Institute; American Heart Association; World Heart Federation; International Atherosclerosis Society and International Association for the Study of Obesity (79).

MetS was thus defined as any participant who possessed three or more of the following five criteria: (i) fasting glucose greater than 100 mg/dl or receiving treatment for elevated glucose levels; (ii) systolic blood pressure greater than 130 mm Hg, diastolic blood pressure greater than 85 mm Hg or receiving treatment for chronic hypertension; (iii) high density lipoprotein (HDL-C) levels less than 40 mg/dl in males and 50 mg/dl in females, or receiving treatment for low HDL-C; (iv) triglyceride levels greater than 150 mg/dL or receiving treatment for elevated triglyceride levels; and (v) a waist circumference greater than the population-specific threshold. A waist circumference cut-off of 90 cm was employed in both males and females, as per the recommendations of Matsha *et al.,* (2013) for the SAC population (80). The appropriate physical measurements were obtained in accordance with criteria established by the World Health Organisation STEPwise Approach to Surveillance (STEPS) instrument (81).

3. Biological measures

Blood samples were drawn by a qualified nurse and DNA extraction was performed through traditional phenol-chloroform techniques. Blood samples were obtained immediately upon participant recruitment and stored for no longer than two days at 4° C before performing DNA extraction. Extraction was conducted in such a manner that the resulting genomic product aligned with the standard concentration of ~ 200 ng/µl. Spectrophotometric analysis was subsequently conducted using the NanoDrop 2000c Spectrophotometer (Thermo Fisher Scientific, Waltham, MA) as a method of quantification and quality assessment. After preparation, DNA samples were stored at -20° C for long-term storage. Genotyping data was obtained using the Multi-Ethnic Genotyping Array (Illumina) in collaboration with the PGC. Array selection prioritised accurately capturing the diverse ancestry of our sample population, where the Multi-Ethnic Genotyping array tailors to exploratory mapping across multiple ethnicities (82). Moreover, epigenome-wide methylation data was attained using the EPIC 850K Array (Illumina) due to the array's improved coverage of regulatory regions and historical technical replicability (83). Genotyping array and DNA methylation data was available for 327 (164 PTSD cases & 163 TEC) and 120 (61 PTSD cases & 59 TEC) participants, respectively.

4. Quality control of genotyping data

Genotyping data was processed using PLINK v1.9 through an iterative quality control procedure previously described by Anderson *et al.*, (2010) (84), Coleman *et al.*, (2016) (85) and Schurz *et al.*, (2019) (86). The quality control process can be differentiated into two broad actions: filtering at the individual SNP level followed by the filtering of participants. These two phases are traditionally separated so as to minimize the potential loss of any genetic data – wherein studies would rather remove singular SNPs than whole individuals from downstream analysis (84). In the following paragraphs we will provide a brief description of the various steps, as well as their accompanying exclusion thresholds, constituting the quality control pipeline for genotyping data. A graphical representation of this protocol has been provided in Figure 1, below.

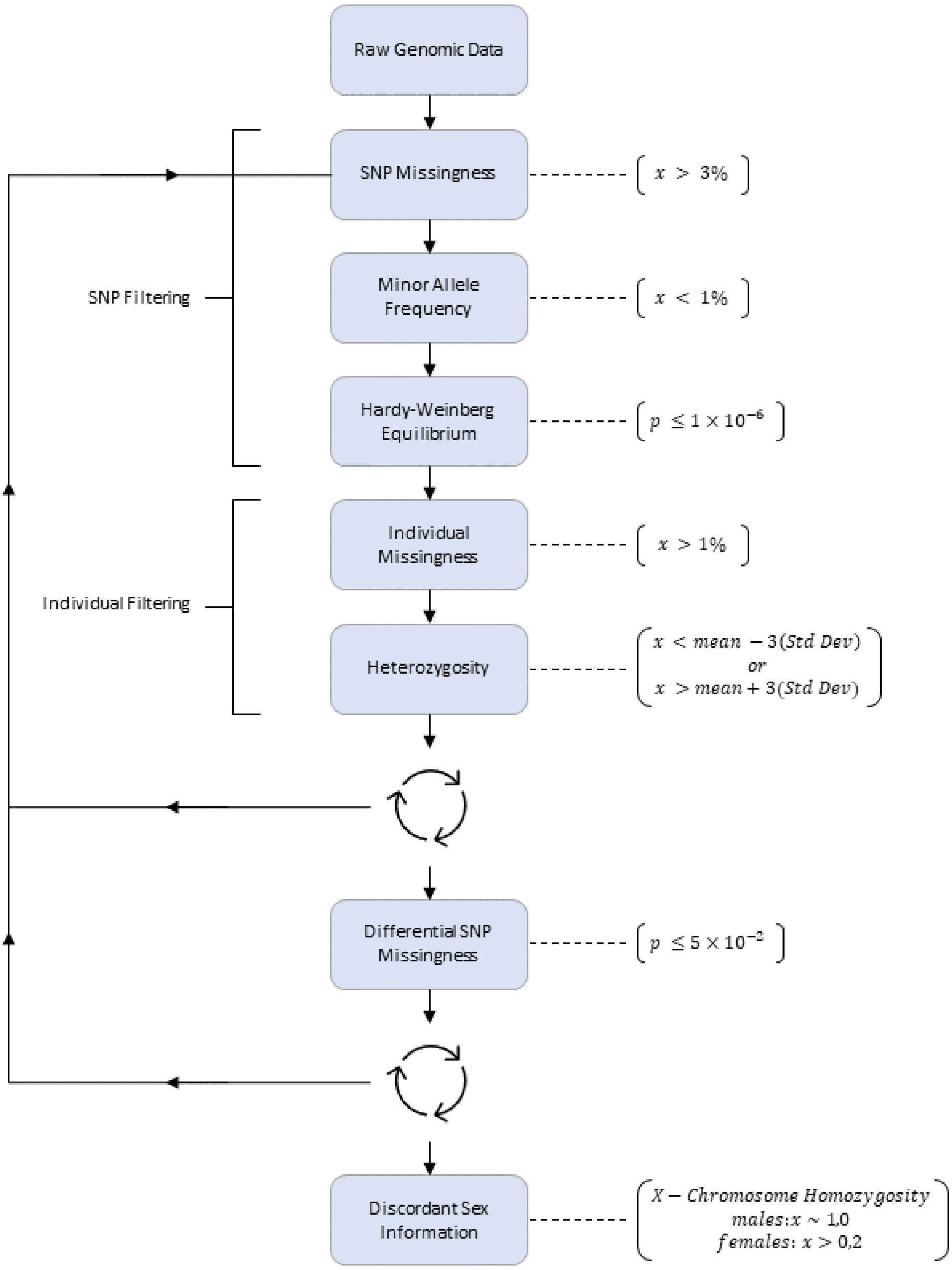

Raw Genomic Data
SNP Missingness
$x > 3\%$
Minor Allele Frequency
$x < 1\%$
Hardy-Weinberg Equilibrium
$p \leq 1 \times 10^{-6}$
SNP Filtering
Individual Missingness
$x > 1\%$
Heterozygosity
$x < mean - 3(Std\ Dev)$ or $x > mean + 3(Std\ Dev)$
Individual Filtering
Differential SNP Missingness
$p \leq 5 \times 10^{-2}$
Discordant Sex Information
$X - Chromosome\ Homozygosity$ $males: x \sim 1{,}0$ $females: x > 0{,}2$

Figure 1: Quality control of genotyping data. A visual depiction of the quality control pipeline used to prepare genomic data for analysis. The overall cleaning procedure consists of three broad stages, two of which utilise iterative elimination as progression barriers. Stages one and two, defined as the initial filtering of SNPs and individuals followed by testing for differential SNP missingness, are repeated until no further elements are removed. Once all iterative requirements are met, the cleaning process is finalised with a test to verify reported sex. The specific thresholds used to identify objects qualifying for exclusion are indicated by the dashed line to the right of each subtest.

Single nucleotide polymorphisms were removed if found to present more than 3% missingness, a minor allele frequency (MAF) below 1% or if found to be in violation of the standard Hardy-Weinberg equilibrium threshold ($p \leq 1 \text{ x } 10^{-6}$). Missingness refers to the proportional degree to which SNPs are absent across individual genotypes (87). For example, if a SNP was only observed in 95% of a cohort, that particular SNP would present a missingness of 5%. MAF measures reflect the variation found within SNPs present in a sample population. Specifically, MAF examines the prevalence at which the least common, or minor, allele for a given SNP occurs in the population under consideration (85,87). Traditionally, a MAF below 1% has been used to distinguish rarer mutations from more established SNPs (85,88). Lastly, Hardy-Weinberg equilibrium is a biological law that places theoretical limits on the extent to which the genetic architecture of a non-evolving population, as encapsulated by allele and genotype frequencies, can change across generations (84,87,88). Ensuring Hardy Weinberg conformity, which is tested for by using population allele frequencies to compare the expected and observed occurrence of individual genotypes, is essential for the identification of genotyping errors within a genomic dataset (87,88).

Quality control was conducted at the individual level by removing participants displaying more than 1% missingness or heterozygosity to an excessive degree of deviancy. In contrast to SNP missingness, individual missingness serves as a proxy for the strength of genotyping; or rather, as a measure of how accurately individual genotypes were captured (87). In like manner, heterozygosity, which refers to the natural state in which a SNP exists as two different alleles at the same locus, addresses a different facet of overall data quality (85,87). Abnormal heterozygosity, as identified by individuals falling further than three standard deviations away from the mean heterozygosity observed in the cohort, may suggest that the DNA sample used for initial genotyping was of poor quality (85,87). A graphical representation depicting the relationship between individual missingness and observed heterozygosity has been provided as a proxy for overall sample quality in Supplementary Figure 1.

Together, the steps outlined in the preceding paragraphs represent a singular phase of filtering; the completion of which was accompanied by a test for differential SNP missingness between cases and controls, wherein any offending SNPs ($p \leq 0.05$) were noted for subsequent removal. Differential SNP missingness functions as a modified extension of classic SNP missingness in that instead of assessing whether a SNP is equally prevalent across all individuals, the test examines how SNPs are distributed between cases and controls (84). Paired in tandem, these measures were conducted in an iterative manner until no further SNPs or individuals qualified for exclusion.

Lastly, any individual found to present discordant sex information was removed from all ensuing analysis (Supplementary Figure 2). Sex was confirmed by comparing reported identity labels to the occurrence of heterogeneity within X-chromosome associated SNPs. As males only possess a single X-chromosome, they should not present as heterozygous for any X-chromosome associated SNPs. By calculating homozygosity (i.e. the presence of two identical alleles at the same locus) for each individual X-chromosome marker and comparing the observed mean homozygosity rate to that which would be expected in males (~ 1.00) and females ($x < 0.2$); one could identify study participants presenting sex information that conflicts with that recorded during the initial interview process (84,85,87,88).

While not traditionally considered integral to the quality control process – it should be noted that all ensuing analysis was limited to autosomal chromosomes as sex-linked testing did not fall within the scope of this study.

5. Principal component analysis

Before proceeding with subsequent analysis, it was deemed necessary to address the potential effects that population stratification may have on the association testing conducted at a later point in this study. At a fundamental level, association tests are simply statistical methods designed to identify relationships between variants and a particular trait or characteristic of interest. However, maintaining the statistical integrity of said methods requires that the tests adhere to several predefined assumptions, the most important of which is the assumption that all variants observed are independently distributed across the sample population (89).

Violations to this assumption primarily arise when participants are derived from more than one ancestral population. This is encapsulated in the term ‘population stratification’, which refers to the occurrence wherein, due to a myriad of historical, geographical and cultural contributions, different populations may present varying allele frequencies for the same genetic loci (90). As such, failing to account for innate differences in allele frequencies across ancestry groups may increase the degree to which false positives are detected – where the ancestrally-driven prevalence of certain variants is mistakenly identified as being indicative of a causal relationship with the trait under study.

One of the more common approaches to address such concerns is to conduct a principal component analysis (PCA) spatially contextualising the sample against a range of population references. Briefly, PCA is a statistical method designed to condense high-density data into its fewest number of non-correlated components; effectively allowing one to quantify isolated sources of variation within a dataset (91). With respect to genomics studies, the non-correlated components identified theoretically reflect the genetic diversity observed across sampled individuals - where the clustering of individuals within a component indicates a similar degree of SNP variation when compared to the mean. Consequently, plotting multiple versions of said non-correlated components (termed principal components) against each other would allow one to visually assess how the genetic diversity observed in a study cohort relates to that of known population references.

With regards to selecting the appropriate reference for accurate contextualisation, the sample population utilised here poses a unique challenge – in that the SAC community traces its roots to five original source populations: African San, African non-San, European, South Asian and East Asian (92). To best assure adequate representation, we elected to combine data provided by two existing references: (i) Phase 3 of the 1000 Genomes project (which contains information pertaining to individuals of African, European, South Asian, East Asian and Admixed American descent) (93); and (ii) that of a recent study by Uren *et al.*, (2016) which provided genomic data specific to individuals of KhoeSan heritage (94).

In order to implement PCA with the intentions described above, the existing study data were merged with that of the population references to create a temporary dataset for further analysis. Importantly, sample homogeneity was maintained throughout the merging process by restricting imported SNPs to those already present within our initial cohort. Once successfully merged, each individual was tagged with a new identifier indicative of their representative population group (African, Admixed American, East Asian, European, KhoeSan, SAC or South Asian) and PCA was conducted using PLINK v1.9 and R v4.0.2. A graphical aid depicting the results of said testing, has been provided in Figure 2 below.

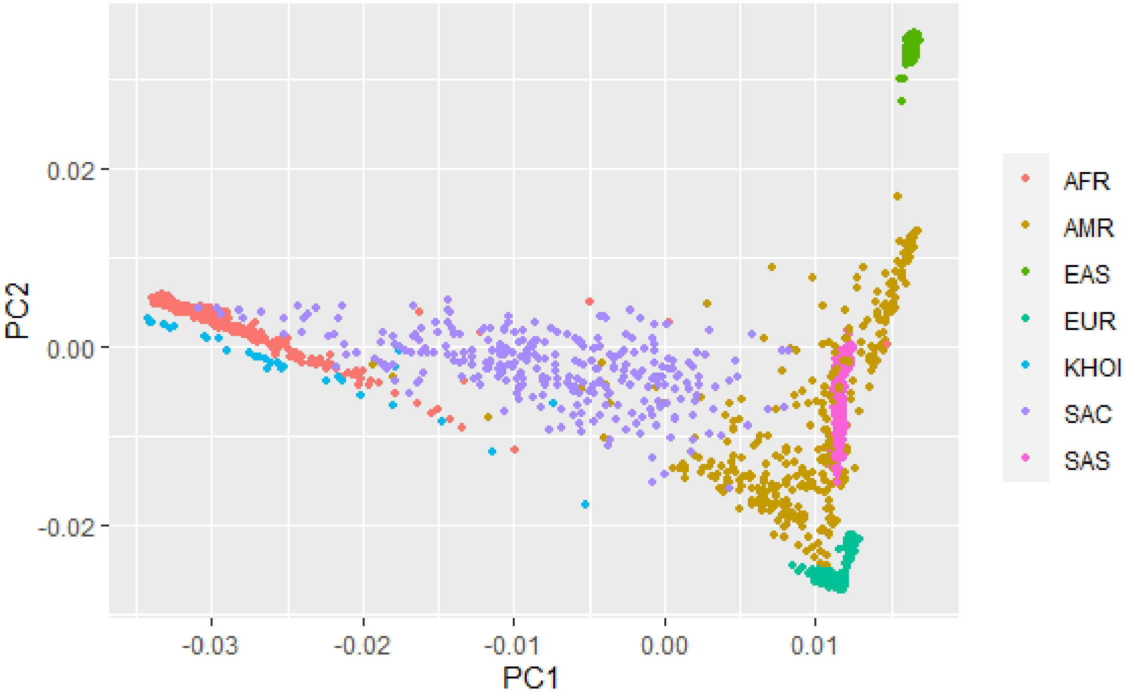


Figure 2: Principal component analysis depicting relationship between study cohort and population references. A visual aid illustrating the genetic diversity of our sample population. Obtained by plotting the first two principal components, which represent the two biggest sources of variation identified, against one another - the above graph depicts the relationship between our sample population and that of the references in the form of shared genetic variation when compared to the dataset mean. Each population group was assigned a unique visual identifier; wherein *AFR = African, AMR = Admixed American, EAS = East Asian, EUR = European, KHOI = KhoeSan, SAC = South African Coloured, SAS = South Asian*; and spatial proximity was indicative of similar genetic diversity.

Subsequent visual assessment of the spatial relationship between our sample and that of the population references confirmed that a substantial degree of genetic diversity was contributed by multiple source populations, and thus highlighted the need to account for the effects of population stratification in ensuing association tests. To ensure methodological consistency between the techniques described here and that of the studies used as guiding references, we elected to utilise methods proposed by Coleman *et al.*, (2016) - wherein the appropriate covariates were identified by conducting PCA on the initial study data (Supplementary Figure 3) and subsequently using linear regression models to determine which principal components were most closely associated with the outcome under consideration (85) (Supplementary Table 1). After regressing the first twenty principal components against PTSD status – where twenty serves as the standard output for PLINK v1.9's inbuilt PCA function – it was determined that the components corresponding to the second (p = 0.078) and eighteenth (p = 0.077) largest sources of variation were noticeably distinguishable from that of the field and consequently, should both be employed as genomic covariates in ensuing analyses.

6. Imputation of genotyping data

Once subjected to quality control, the genotyping data was submitted to the Sanger Imputation Service (SIS) for imputation (95). While recent years have seen remarkable advancements in the development of genotyping arrays; when employed alone, such methods present inherent limitations in the statistical power they lend to association studies. Genotyping arrays are not designed with blanket coverage of the human genome in mind. Rather, they scan for a select number of common and rare genetic variants to generate a profile curated to maximize downstream imputation accuracy (86). Broadly speaking, imputation refers to the process through which missing genotypes are statistically inferred by comparing genotyped data to a population-specific reference panel (96). This is primarily done through the use of linkage disequilibrium patterns: the phenomenon where particular groups of alleles are more frequently observed together than traditionally expected should the laws of mendelian inheritance be in effect (97). In light of this occurrence, one could infer the state of SNPs falling outside the detection scope of genotyping arrays by comparing successfully genotyped loci to a suitable reference panel reflective of the linkage disequilibrium patterns most prevalent in your sample population (86). Therefore, by increasing the number of genotyped loci available for subsequent analysis, imputation serves to dramatically improve one’s ability to detect statistically meaningful findings through association testing.

Imputation was performed using a combination of in-house scripts prepared by Dr. Stephanie Pitts, a previous PhD candidate of Stellenbosch University’s Division of Molecular Biology and Human Genetics, and protocols formerly detailed by Coleman *et al.*, (2016) (85) and Schurz *et al.*, (2019) (86). The process through which genotyping data is imputed occurs across three distinct phases: preliminary data preparation, submission to the SIS and post-imputation quality control. In the following paragraphs we will provide a succinct summary further characterizing the role played by each of these phases within the greater imputation pipeline. A graphical representation of this protocol has been provided in Figure 3, below.

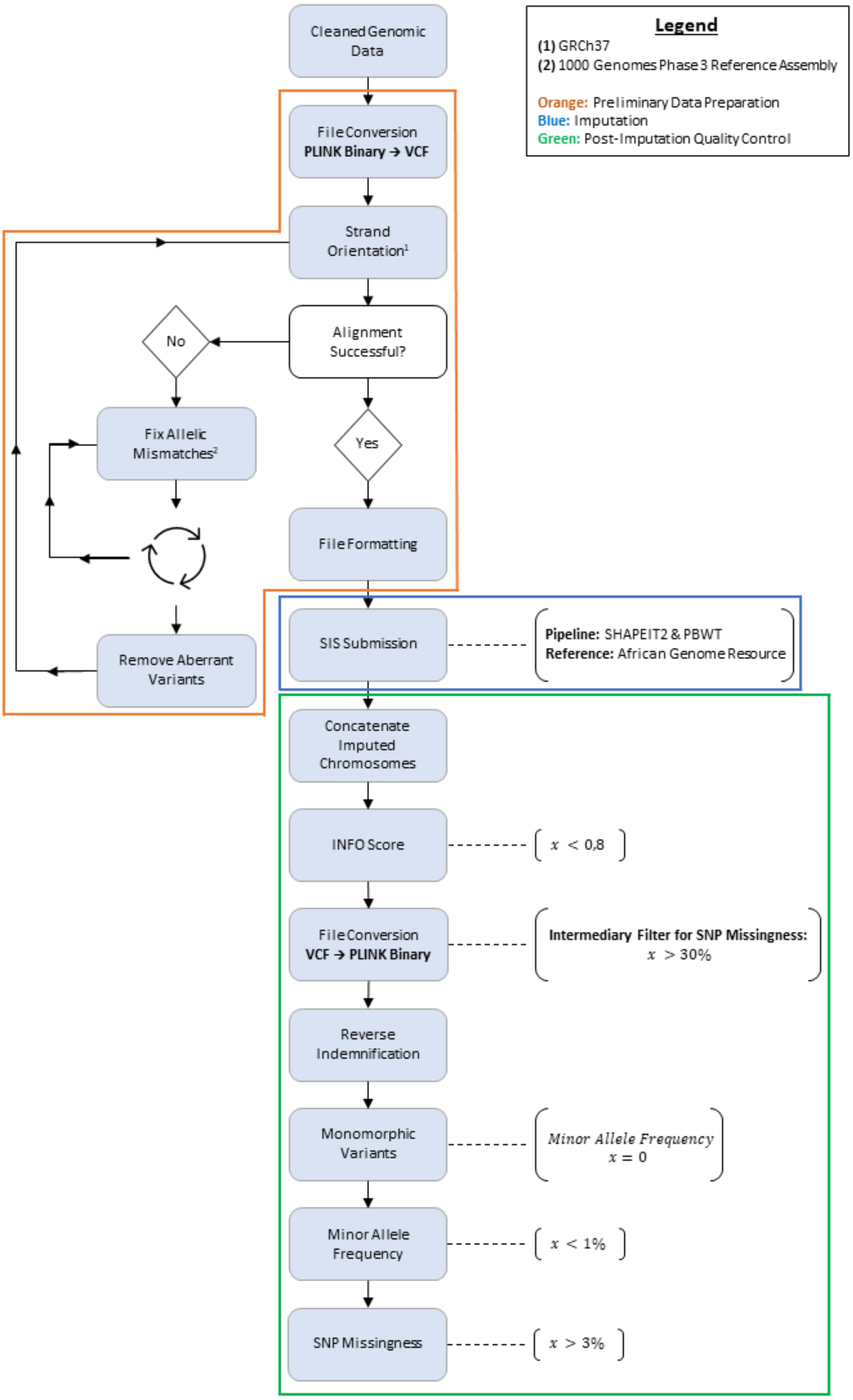


Figure 3: Imputation of genotyping data. A graphical aid illustrating the three stages through which genomic data is imputed: preliminary data preparation, imputation and post-imputation quality control. The legend on the top right provides the colour coded index distinguishing each phase and highlights which genomic reference panels were required to conduct certain subtests within the pipeline (as indicated by corresponding numerical superscripts). Dashed lines were used to incorporate additional information by further elaborating upon the parameters used to run imputation through the Sanger Imputation Service (SIS) and the variables delineating exclusionary thresholds under the imputation and post-imputation quality control phases respectively.

Prior to submitting a dataset for imputation, the SIS requests that all entries conform to a series of established formatting conditions. Namely: (i) that the submitted file be in variant calling format (VCF); (ii) that all alleles be designated in their forward strand state; (iii) that the submitted file utilise a genomic coordinate strategy aligning with that of the Genome Reference Consortium Human genome build 37 (GRCh37); (iv) that the appropriate reference alleles within the submitted file match those used in the GRCh37; (v) that the dataset be submitted as a single VCF file as opposed to one VCF per chromosome; (vi) that the submitted file be correctly sorted by genomic position; and (vii) that the submitted file utilise the chromosome naming convention described by the SIS reference index (where 1,2 and 3 correspond to chromosome 1, chromosome 2 and chromosome 3 respectively etc.). The necessary steps required to align existing genotyping data with these recommendations can be implemented using PLINK v1.9 and BCFtools v1.10.2.

Firstly, the existing dataset was converted from the PLINK binary format in which quality control is traditionally conducted to a compressed VCF through a two-step conversion process wherein the data was transformed through PLINK before undergoing BCFtools compression. Typically, a given genotyping dataset exists as three separate files when in PLINK binary format. These files, as identified by the extensions, ".bed", ".bim" and ".fam", correspond to raw genotyping data, a summary of the detected variants and anonymised sample information respectively. VCF files condense this information into a single element – wherein particular emphasis is granted to what variants were predicted and the confidence that those initial predictions were correct.

Once converted, strand orientation was assessed to ensure that all variants called were correctly reported in their forward strand state. Briefly, strand orientation serves as a directional reference frame through which to discern the sense and antisense states in which variants occur (98). The forward, or sense, strand is the polynucleotide chain which is not transcribed and thus most closely approximates any mRNA produced. Strand orientation was determined by attempting to align the dataset against the GRCh37 reference; wherein alignment efficacy was evaluated through the reported number of allelic mismatches and non-biallelic sites identified. Within the context described above, the term allelic mismatches does not possess a strict biological definition – but rather serves as the appropriate nomenclature for when the variant detected at a given loci differs from that registered at its corresponding genomic position in the reference. Alternatively, the frequency at which non-biallelic sites are observed is indicative of the degree to which multiallelic loci are present in a dataset. Multiallelic loci are genetic sites where more than one variable allele exists alongside the predominant wild type (99). While recognised as a naturally occurring genomic state, multiallelic loci tend to be excluded from analysis pipelines – as historical uncertainty regarding their prevalence and distribution throughout the genome has rendered it difficult to accurately account for their presence (100).

Allelic mismatches identified while attempting to verify strand orientation were corrected by fixing the dataset against the concatenated 1000 Genomes Phase 3 reference assembly – an internationally generated dataset serving as the premier catalogue of human genetic variation (93). Corrective alignment was conducted in an iterative manner, where strand orientation assessment and assembly adjustment were cyclically repeated until no further mismatches could be rectified. Once automated mismatch reduction had plateaued, the remaining allelic mismatches and non-biallelic sites were manually removed before ultimately authenticating that strand orientation aligned with that established by the GRCh37 reference. Final preparatory modifications prior to submitting the dataset to the SIS primarily consisted of formatting alterations: (i) ensuring that all chromosomes had been named in accordance with Ensembl conventions (as defined by the European Bioinformatics Institute) (101); and (ii) numerically sorting the integrated dataset by genomic position.

Upon submission to the SIS, an imputation job was created requesting that the dataset be processed through the SHAPEIT2 & PBWT pre-phasing and imputation pipeline using the African Genome Resource reference panel (86).

When employed alone, the size of the reference panel against which one's dataset is compared renders imputation highly computationally intensive. Therefore, pre-phasing tools, such as SHAPEIT2, are commonly utilised for initial haplotype estimation – in which statistical models are used to generate preliminary genotype probabilities so as to alleviate later computational burden (102,103). Once the dataset was adequately prepared, imputation was performed using a modified version of the Positional Burrows-Wheeler Transform, or PBWT, algorithm. While a detailed explanation as to the functioning of this algorithm falls outside the scope of this thesis, Durbin *et al.*, (2014) has provided a comprehensive analysis describing how the method can be applied to haplotype matching in their paper *Efficient haplotype matching and storage using the positional Burrows-Wheeler transform (PBWT)* (104). In order to ensure that the reference panel implemented was ancestrally representative of our sample population, we elected to utilise the African Genome Resource – which serves as a conglomerate template consisting of the 1000 Genomes Phase 3 reference release supplemented with samples sourced from Uganda, Ethiopia, Egypt, Namibia and South Africa (86,95). However, it should be noted that the South African contributions were strictly limited to individuals of Zulu (African Non-San) ancestry.

Once the imputation job was successfully completed, the SIS returned the dataset in the form of a series of compressed VCF files indicative of each imputed chromosome. Prior to conducting post-imputation quality control, the imputed chromosomes were concatenated into a singular file and assessed to verify whether all submitted chromosomes were present and had been adequately imputed. Quality control was initiated by first examining the condition to which imputation was implemented across all presented variants. Upon processing an imputation request, the pipeline employed provides a software-specific variable representative of the confidence that each variant was correctly imputed. Within the context of our analysis, the SHAPEIT2 & PBWT pipeline supplied an INFO score as an emblematic measure of performance. INFO scores are not traditionally restricted to a recommended threshold but are rather limited depending on the frequency at which they occur across individual datasets. To further elaborate, existing standards advocate that one generates a frequency distribution of INFO scores across all chromosomes and then subsequently select an exclusionary threshold at the point of inflection (85). A graphical example illustrating how this was done has been provided in Supplementary Figure 4. For the purpose of this analysis, we elected to utilise an INFO score threshold of 0.8 to delineate poorly imputed variants.

After filtering according to INFO score, the dataset was restored to PLINK binary format through a two-step conversion process – during which, variants were further restricted to those presenting less than 30% missingness while in an intermediary PLINK format (".ped" & ".map") before being returned to binary format for the application of additional quality control measures (86). Firstly, the indemnification applied by the SIS during imputation was rectified by reincorporating the appropriate identity and phenotype variables at the individual level. At which point SNPs were removed if found to present a MAF below 1%, an excessive degree of SNP missingness or if found to exist in a monomorphic state. In contrast to previous steps attempting to correct for SNP missingness, the iteration implemented here does not rely on a predetermined threshold, but rather uses the same frequency distribution-based method applied for the selection of an INFO score threshold outlined in the paragraph above. When assessing the degree to which SNP missingness occurred across all remaining variants, it was determined that the inflection point, and thus the appropriate exclusionary threshold, fell at SNPs presenting a missingness greater than 3% (Supplementary Figure 5). Contrarily, the steps followed to address MAF and monomorphism are theoretically indistinguishable – wherein both could be done simultaneously through the application of the aforementioned 1% MAF threshold. Monomorphic variants serve as the antithesis of multiallelic loci in that they are predominately found to occur in a single phase across all individuals in a population (105). As such, they are commonly omitted from analysis considering genetic variation due to the lack of informativeness derived from their homogenous state (85,88). Distinguishing between variants removed due to monomorphism and notable MAF, as determined by applying MAF thresholds of 0% and 1% respectively, remains solely for academic purposes – in that the findings do not meaningfully contribute to subsequent analysis but rather serve to further characterise the dataset under study.

7. Genome-wide association analysis

Once the genotyping data had successfully undergone standard quality control and subsequent imputation, the expanded dataset could be prepared for genome-wide association testing. First, the imputed dataset was revaluated, and any individual found to present a PI_HAT value greater than 0.1875 was removed from all subsequent analysis. PI_HAT, which represents the proportion of identity by descent, is the statistic produced by PLINK's inbuilt test for duplicated samples and relatedness (106). If so prompted, PLINK is capable of detecting hidden familial relationships by creating pairwise identity by descent matrices across all individuals in one's sample. Simply, the matrices serve as a method of identifying every manner in which a unique pair of individuals can be grouped and subsequently examining how similar the two are as compared to that which would be expected by random chance. PI_HAT effectively encapsulates this observed similarity or dissimilarity in the form of a numeric ratio indicative of suspected degrees of relatedness; where a PI_HAT value of one typically represents duplicated samples, or monozygotic twins, and genetic distinctiveness increases as the statistic approaches zero (84). A PI_HAT value of 0.1875 corresponds to approximately half-way that which one would expect to observe between second and third degree relatives (0.25 and 0.125 respectively) and is commonly regarded as being sufficiently stringent so as to address the overrepresentation of familial variants when assessing vulnerabilities to the trait under study (70,85). It should be noted, that relatedness is traditionally accounted for during the final stages of quality control (often as an accompaniment to testing for discordant sex information); however, we elected to initially refrain from implementing said measures so as to maximise the amount of genetic data available for imputation (86).

The existing dataset was further refined by restricting our analysis to SNPs that had been assigned unique identifiers, or rsIDs. rsIDs are a product of the National Centre for Biotechnology Information's Single Nucleotide Polymorphism Database (dbSNP) – wherein their allocation highlights SNPs that have been thoroughly mapped and annotated to the degree that they are deemed to be reasonably stable references for investigative and reporting purposes (107).

Potential covariates were identified by using a series of statistical tests to investigate whether information obtained through the neuropsychological assessments and demographic questionnaire differed according to PTSD case-control status. Specifically, recorded variables pertaining to participant age and gender were deemed most likely to serve as genetic confounders. Statistical tests were carried out in a manner appropriate to the nature of each variable, with assumptions appertaining to normality and variance guiding which analytical methods were implemented. Neither participant age (t = -0.651; *p* = 0.515) nor participant sex (X^2 = 0.013; *p* = 0.908) were found to significantly diverge between PTSD cases and TEC. As such, genomic covariates were limited to the two principal components (PC2 & PC18) previously included to account for the effects of population stratification.

A GWAS was performed by running the cleaned imputed dataset through PLINK v1.9's inbuilt association function while using principal components 2 and 18 as covariates. Subsequent results were visually assessed through a R v4.0.2. generated Manhattan plot and a p-value threshold of 5e-8 was used to delineate genome-wide significance. To ensure adequate elaboration upon any variants which may trend towards suggestive associations, corresponding summary statistics were submitted to the online tool FUMA (Functional Mapping and Annotation for Genome-Wide Association

Studies) (108) for subsequent annotation of the resulting findings. All FUMA analysis were conducted with the standard recommended settings.

8. Polygenic risk score analysis

PRS analysis was primarily performed using *PRSice* (69), a specialized command line tool tailored to automate the necessary adjustments and calculations required to generate PRS for each individual in one's sample. The software package is designed in such a manner that all relevant parameters are concatenated into a single command with minimal user requirements. Despite the automated nature of the aforementioned process, there was one particular facet unto which additional considerations were granted: the selection of an appropriate reference dataset.

Inarguably, the most critical component relating to the accurate calculation of PRS is the selection of a reference dataset indicative of variants previously associated with one's phenotype of interest. The reference, or base, dataset provides the summary statistics from which the estimated effect sizes for individual variants are garnered; effectively rendering it the sole arbiter of the weighting system used to calculate risk. Current norms dictate that the ideal reference is simply the largest publicly available GWAS assessing the trait under study (68,70). For the purpose of our analysis, we elected to utilise the PGC PTSD Work Group's most recent data release (PTSD Freeze 2) – which contains data pertaining to 206,655 multi-ethnic individuals sourced from sixty independent GWAS (6).

The greater PGC PTSD Freeze 2 data release can be broadly differentiated into three distinct subgroupings: individuals of European ancestry (89.25%), individuals of Latino/Native American ancestry (2.91%) and individuals of African American ancestry (7.84%). Moreover, it should be noted that there was no participant overlap between the sample population utilised here and that of the PGC PTSD Freeze 2 data release. To determine which combination of population references would maximise predictive performance, we elected to conduct two preliminary runs comparing the African American data and the collective overall dataset as our respective templates. Predictive performance was assessed across a predetermined range of p-value thresholds, wherein *PRSice* applied p-value cut-offs of 0.001, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5 and 1 to the references so as to identify which combination of PTSD-associated variants could best distinguish between cases and controls in our sample. Goodness-of-fit was then subsequently evaluated through a two-metric system, in which the overall performance of each run was judged by the proportion of variation explained (as represented by Nagelkerke's pseudo R^2 value) and the corresponding significance level of the most predictive model. Multiple testing adjustments were enforced through Bonferroni correction, wherein the threshold for significance was set as the standard significance threshold divided by the number of p-value cut-offs tested (p = 0.05 / 8 cut-offs tested = 6.25e-3).

PRS constructed using a combination of all three population subgroupings were found to consistently outperform those employing solely the African American data – with the most predictive combined PRS explaining approximately 4.41% of the variation observed in PTSD outcome (R^2 = 4.41e-2; p = 0.064). The overall validity of this model was further confirmed by recreating *PRSice*'s Nagelkerke's pseudo R^2 calculation in R v4.0.2. Where an identical value was produced when using the PRS scores assigned to each individual to simulate the logistic regression model through which the metric was inferred.

Ultimately, *PRSice* determined that PTSD status could best be predicted by assessing the degree to which each individual presented combinations of a select 1,444 variants identified by applying a p-value cut-off of 0.001 to the combined PGC PTSD Freeze 2 reference dataset.

9. Quality control & functional normalisation of DNA methylation data

The DNA methylation data made available for 120 of the genotyped individuals (61 PTSD cases & 59 TEC) was prepared for subsequent analysis through a host of packages run through R v4.0.2. Quality control and functional normalization were performed using *meffil* (109) equipped with parameters previously described by Dr. Sylvanus Toikumo, a previous PhD candidate of Stellenbosch University's Division of Molecular Biology and Human Genetics (110). Developed by Min *et al.*, (2018) (109), *meffil* is tailored to maximise computational efficiency and ensuing repeatability for the processing of DNA methylation data derived from Illumina Methylation BeadChip microarrays. The overarching package consists of several self-contained modules, in which each component can independently generate a report documenting how the algorithm addressed one particular facet of the analysis. In the following paragraphs, we will provide a brief explanation as to the underlying mechanisms behind two such components, those considering quality control & functional normalization, and the manner in which they were applied to the dataset in question.

Before moving forward, a quick introduction may be needed as to the technical aspects through which BeadChip microarrays detect DNA methylation. Illumina products utilizing BeadChip-based detection methods rely on preassembled libraries of silica microbeads – where individual beads have been laced with unique oligonucleotides corresponding to predetermined loci throughout the genome (111). The beads are purposefully deployed with redundancy in mind, in that multiple beads bear the same oligonucleotide sequence should adequate binding, and therefore detection, not occur at any one instance (112). However, this surplus of beads provides additional benefits alongside its primary function as a technical failsafe – wherein the performance of beads within and across samples plays an integral role in the quality control process.

Briefly, to initiate the quality control process, *meffil* requires that one modify the base algorithm by adjusting the included parameters to best suit one's analytical needs. For the purpose of this project, we elected to employ the following conditional values: (i) [*detection.threshold = 0.01*] – a p-value threshold of 0.01 should be used to distinguish between detected and undetected probes; (ii) [*detectionp.samples.threshold = 0.1*] – any sample for which more than 10% of presented probes fail to be detected should be excluded from all subsequent analysis; (iii) [*beadnum.samples.threshold = 0.1*] – any sample for which more than 10% of presented probes fail to be observed across at least three beads (henceforth referred to as the minimum bead threshold) should be excluded from all subsequent analysis; (iv) [*detectionp.cpgs.threshold = 0.1*] – any probe for which the given probe fails to be detected across more than 10% of samples should be excluded from all subsequent analysis; (v) [*beadnum.cpg.threshold = 0.1*] – any probe for which the given probe fails to meet the minimum bead threshold across more than 10% of samples should be excluded from all subsequent analysis; (vi) [*sex.outlier.sd = 5*] – any sample presenting a median X or Y chromosome intensity not within five standard deviations of the dataset median should be removed from all subsequent analysis; (vii) [*snp.concordance.threshold = 0.95*] – a concordance threshold of 0.95 should be used to select control SNP probes for concordance analysis; and (viii) [*sample.genotype.concordance.threshold = 0.95*] – any sample which fails to achieve 95% concordance when comparing genotyped loci to control SNP probes should be removed from all subsequent analysis.

To further elaborate upon two of the hitherto undiscussed commands, the manner in which *meffil* addresses discordant sex information does not differ drastically from the approach first introduced under the processing of genomic data in the chapters above. Simply, median probe intensities are calculated across all X & Y chromosome loci so as to generate a representative distribution indicative of the intensity patterns associated with each reported sex. Thus, should a particular sample present median intensity values falling further than five standard deviations away from those historically observed for their reported sex; said sample would be earmarked for removal due to sexual incongruency (Supplementary Figure 6). Furthermore, it is common practice for microarrays of this sort to incorporate a sample verification system in the form of SNP control probes; a collection of high-frequency loci for which all genotypes have been correlated to specific methylation intensities (113). When viewed in their entirety, the control probes produce an identity pattern unique to the individual from which the relevant genetic material was derived. Wherein the technical parameters described in points (vii) and (viii) establish the qualifying framework determining which loci are shared between sample & microarray and subsequently assessing concordance between the two.

In addition to the conditions outlined above, *meffil* allows one to specify a cellular composition reference to function as a theoretical representation of expected intensity measures, where considerable deviation in observed intensities may be indicative of an erroneous detection process or poor-quality sample. As all biological materials utilised throughout this thesis were originally sourced from blood, we elected to employ *meffil*'s "blood gse35069 complete" reference, which assumes the most prevalent cell populations are B-lymphocytes, CD4 T cells, CD8 T cells, eosinophils, monocytes, neutrophils and natural killer cells.

Finally, supplementary to the parameters implemented by the user, *meffil* contains its own quality control mechanisms addressing the degree to which methylation was successfully measured across samples. After assessing the distribution of intra-sample methylated and unmethylated intensities, aberrant samples were identified as those presenting median methylated intensities that fell further than three standard deviations away from that of the overall dataset (Supplementary Figure 7). Furthermore, *meffil* evaluated the general performance of the detection process by examining whether a range of quality control probes included to address specific aspects of the experimental method behaved consistently across all samples (113). Notably, exclusionary thresholds related to these quality control probes are not subjected to manufacturing recommendations but are rather determined by the distribution of outliers deviant from the mean intensity observed for each probe (109).

Once applied, the quality control process described above was conducted in a semi-iterative manner: where the erroneous probes and samples accompanying each run were either removed or, if possible, resolved until no further aberrations remained.

Finally, functional normalization was performed so as to concatenate the intensity measures generated over separate microarrays and to correct for any abnormal variation resulting from batch specific variables. A thorough overview of the statistical methods underlying normalization falls, unfortunately, outside the scope of this thesis. However, the process can broadly be described as the act of adjusting values measured across different scales by considering the degree to which technical features contribute to variation (114). Briefly, variation was first quantified through PCA; wherein it was determined that the first four principal components were most representative of the variation present in the dataset. This was deduced by assessing the associated scree plot generated by *meffil*'s built-in PCA function and assigning our threshold to the estimated point of inflection (Supplementary Figure 8). Batch variables, identifying prominent sources of technical artifacts, were created – in which array, array column and array row were deemed pertinent for inclusion. Normalization was then conducted by examining how the selected principal components, specific to the performance of known control probes, were affected by the included batch variables and adjusting the individual intensity measurements accordingly. Relationship evaluations took the form of linear regression models, where *meffil* ran each principal component against every available variable before returning a large matrix of normalized intensity values.

It should be noted that all statements in the above paragraphs that directly refer to the functioning of a *meffil* command or associated feature were sourced from the R documentation titled *Meffil: Efficient Algorithms for DNA Methylation*.

Should the reader wish to garner further information as to the functioning of the algorithms described here, the relevant package details can be found at (https://rdrr.io/github/perishky/meffil/).

10. Epigenome-wide association analysis

Once the DNA methylation data had successfully undergone quality control and functional normalization, an EWAS was performed using *meffil*'s inbuilt association function. Potential covariates were identified in a manner similar to that previously described in the preceding subheadings – where preliminary hypotheses were tested using a series of statistical measures appropriate to the conditions of normality and variance. With regards to the DNA methylation data, variables pertaining to whether the participant had a history of smoking, history of alcohol use, met the necessary requirements for suspected MetS, and had previously experienced childhood trauma, were deemed most likely to be potential confounders. The PCA covariate identification previously conducted to address population stratification was carried out solely in the context of the relevant genotyping data - as such, PC2 and PC18 were not included in any subsequent methylation analysis.

All assumptions were supported by existing evidence that each variable is both independently associated with differential DNA methylation (115–118) and closely linked with the presentation of PTSD. MetS has been well documented as presenting a high degree of PTSD comorbidity at both physiological and epidemiological levels (119). Previous studies have indicated that both PTSD and MetS often present similar aberrations in neuroendocrine and inflammatory pathways (120). Additionally, each such mechanism is thought to directly contribute to the differential DNA methylation routinely observed under both PTSD and MetS (115). Thus, to ensure that any detected epigenetic effects are attributable to PTSD alone, it is advisable to account for the potentially confounding influences of MetS. Furthermore, exposure to childhood trauma has been related to an increased risk for the development and subsequent severity of PTSD (121). Moreover, PTSD has been shown to be associated with greater rates of cigarette smoking and alcohol use than that which would be expected in the general population (122,123).

Upon subsequent analysis, neither smoking status ($X^2 = 1.091$; $p = 0.296$), alcohol use ($X^2 = 0.855$; $p = 0.355$) nor the suspected presence of MetS ($X^2 = 0.000$; $p = 1.000$) were found to differ significantly between PTSD cases and TEC. As such, covariates were limited to previous experiences of childhood trauma ($X^2 = 7.561$; $p = 5.965\text{e-}3$), supplemented by the cellular composition estimates previously generated by *meffil*. The inclusion of the latter is primarily due to the ingrained variability which commonly plagues methylomic profiles derived from whole blood samples. Whole blood is widely thought to be a highly heterogenous tissue – in that any one sample may present varying proportions of local cell lineages (124). Such a heterogenous composition becomes problematic when one considers that many cell types possess unique methylation patterns characteristic of their lineage (125). It is therefore commonly recommended that one utilise cellular composition estimates as a method of accounting for any potential confounding associated with varying cell type proportions across samples (109).

With regards to accounting for data heterogeneity, *meffil* offers four distinct approaches to address the effects of potential confounders: (i) running regression models without accounting for covariates; (ii) running regression models while accounting for the provided covariates; (iii) running regression models with the provided covariates and surrogate variable analysis (SVA); and (iv) running regression models with the provided covariates and independent surrogate variable analysis (ISVA) (109). The inclusion of SVA and ISVA serve to identify and subsequently adjust for hidden, or undefined, sources of heterogeneity within the data – wherein the sole distinction between the two lies in whether hidden confounders are searched for as linearly uncorrelated or non-linearly uncorrelated variables (126,127). In keeping with the protocols previously developed by Dr. Sylvanus Toikumo, we elected to conduct EWAS using both the provided covariates and SVA to account for potential sources of confounding.

As such, an EWAS was performed by running the normalized intensity measures through *meffil*'s inbuilt association function while using childhood trauma exposure, cellular composition estimates and SVA-determined surrogate variables as covariates. Subsequent results were visually assessed through *meffil*'s report-generated Manhattan plot and a p-value threshold of 5.95e-8 (as determined by p = 0.05 / 840,920 probes tested) was used to delineate epigenome-wide significance. To further elaborate upon any probes that may present suggestive associations, CpG sites were annotated through a combination of the R packages *missMethyl* (128) and *limma* (129). Furthermore, differentially methylated regions (DMR), genomic regions where multiple proximally associated probes present similar DNA methylation changes under the studied phenotype, were identified through *dmrff* (130). Annotations were generated using existing records sourced directly from the Illumina 450k array annotation package in R.

11. Methylation quantitative trait loci

Having successfully determined the appropriate genomic and DNA methylation variables, mQTLs were identified through an amalgamation of packages run through R v4.0.2 - wherein the critical analysis upon which a series of accessory tests depend was primarily derived from *MatrixEQTL* (131). Briefly, *MatrixEQTL* is a statistical package tailored for the computationally efficient discovery of expression quantitative trait loci (eQTLs) (131). The package was originally designed to conduct exhaustive pairwise testing through a dose-effect matrix-based approach, where linear additive or ANOVA models are used to quantify interactions between individual variants and gene expression transcripts. Theoretically, the basic principles underlying the detection of eQTLs hold true for mQTLs as well – in that exchanging the requested gene expression matrix for one which presents normalized DNA methylation intensity measures, should allow one to identify mQTLs by independently testing for associations between every possible combination of SNP and CpG site.

Methylation quantitative trait loci were identified using a compilation of scripts first collated by Hannon *et al.*, (2018) in accompaniment to their paper: *Leveraging DNA-Methylation Quantitative-Trait Loci to Characterize the Relationship between Methylomic Variation, Gene Expression and Complex Traits* (132). Upon publishing, the aforementioned scripts were made available in a publicly accessible GitHub repository (https://github.com/ejh243/UKHLS_mQTL) – which, when employed collectively, allow one to assemble an extensive analytical framework for the detection and subsequent characterisation of mQTLs. Notably, one particularly attractive facet of this approach is that the pipeline provided is not simply limited to the identification of mQTLs; but rather, includes several accessory measures tailored specifically towards helping one interpret and further contextualise any potential associations that may arise from the raw data. The process through which all ensuing analysis was conducted can be broadly separated into two distinct components: mQTL identification, and supplementary analysis. In the following paragraphs, we will provide a concise description as to the particulars underlying each of these steps as well as briefly allude to the role that they play in the greater analytical pipeline.

Firstly, *MatrixEQTL* requires that all input adhere to a predefined template for individualised matrices – wherein the relevant genomic and DNA Methylation datasets should be formatted such that each appears as a matrix comparing sample identifiers to variant and CpG sites, respectively. In such manner, each point of convergence conveys an individual's status in relation to a specific SNP or CpG site; where the relationship in question is represented by either the genotype observed at a particular locus or the normalized intensity measure detected at a corresponding methylation probe.

Data appropriate to the generation of the genomic matrix was obtained by using the most predictive PRS model to extract informative SNPs from the imputed dataset. Briefly, previous analysis had determined that the most predictive PRS was one that had been constructed to utilise summary statistics derived from all three population subgroupings available under the PGC PTSD Freeze 2 data release. Furthermore, *PRSice* determined that PTSD status could best be predicted through a combination of 1,444 variants identified by applying a p-value threshold of 0.001 to the PGC reference (R^2 = 4.41e-2; p = 0.064).

Using PLINK v1.9, the imputed dataset was filtered such that all remaining individuals were restricted to the 1,444 SNPs highlighted in the paragraph above. Critically, in order to enable the dosage-based calculations upon which the functioning of *MatrixEQTL* depends, the newly filtered dataset should be recoded so that individual genotypes are represented as a limited series of numerical values. Such a conversion was implemented through PLINK v1.9's inbuilt recode function – which transforms one's dataset by assigning values of 0, 1 or 2 to indicate whether a particular SNP is absent, present in a heterozygotic state or present in a homozygotic state under the individual in question. As *meffil* presents normalized intensity measures in a format similar to that requested by *MatrixEQTL*, subsequent preparation was limited to the identification of individuals for whom both genomic and DNA methylation data remained available (97 Individuals – 51 PTSD cases & 46 TEC).

Initial mQTL estimates were obtained by using *MatrixEQTL* to test for associations between all possible pairings of SNP and CpG sites, while employing childhood trauma exposure and cellular composition estimates as covariates. Testing was conducted using a linear additive model approach wherein each calculation attempted to determine whether the genotype observed at a particular locus predicted fluctuations in a corresponding normalized intensity value. Due to the sheer number of iterations required to implement exhaustive pairwise testing, all analysis was accompanied by multiple testing correction in the form of the Benjamini–Hochberg false discovery rate (FDR) procedure. Furthermore, a Bonferroni-corrected p-value threshold was used for post-hoc interpretation – wherein the threshold for significance was set as the standard threshold for GWAS divided by the number of CpG sites tested (p = 5.00e-8 / 840,920 CpG probes = 5.95e-14) (132).

In order to reduce computational intensity, the genomic matrix was first subset by physical position before testing for associations on a chromosome-by-chromosome basis. The subsequent results, which were produced in the form of individual text files, were then manually concatenated, and thereafter merged with data pertaining to the latest version of the Illumina Infinium MethylationEPIC Manifest File (v1.0 B5). The MethylationEPIC manifest effectively serves as a comprehensive reference through which to contextualise any notable CpG sites; the most important components of which are the physical and chromosomal positions at which each occurs. Upon combining the newly obtained CpG positional data with that previously generated for the included SNPs, potential mQTLs were classified according to the following criteria: (i) any instance where a SNP falls within 500kb (500,000 base pairs) of a CpG site was termed to be a cis-acting, or local, mQTL; and (ii) any SNP-CpG pairing in which either component falls greater than 500kb away or originates from a different chromosome was defined as a trans-acting, or distant, mQTL. (133).

At this point, Hannon *et al.*, (2018) (132) recommends that one utilise PLINK's inbuilt clumping function to verify whether any associations detected are truly independent in nature. If employed with the provided parameters, such a measure would utilise linkage disequilibrium patterns to group correlated SNPs (250kb window; r^2 = 0.1) in a manner ranked by the degree to which they are associated with the phenotype under study. As such, instances where multiple SNPs are found to be associated with the same CpG site would be further refined by eliminating candidates simply identified due to their close proximity to a more significant SNP. While we elected to adhere to the recommended instructions for the sake of thoroughness – it should be noted that this step is rendered largely redundant due to *PRSice* performing the same LD-based clumping method during its initial calculations.

Lastly, to further elaborate upon any detected findings, we elected to conduct Bayesian Colocalisation analysis to test whether multiple mQTLs could be refined down to a single genomic signal. Doing so parses through the raw findings and highlights notable mQTL interactions for subsequent exploratory analysis. Independently, the individual SNP-CpG interactions reported by MatrixEQTL provide relatively minimal molecular context. However, identifying variants that affect multiple different CpG sites would suggest that a much broader, and potentially more impactful, effect may be in play. Briefly, the dataset was restricted to SNP-CpG pairs that occurred on the same chromosome and surpassed a relaxed p-value threshold of 1e-10 – thereafter, mQTL analysis was repeated in a series of windows were every possible paring of CpG sites within 250kb of one another were tested against all SNPs within 500kb (132,134). Dataset formatting was done through Hannon *et al.*, (2018)'s provided scripts and colocalisation was performed through the *coloc* package (134).

12. Statistical analysis

All data analysis was conducted in R v4.0.2 at a significance level of α = 0.05. Differences in the demographic and clinical data were assessed against the dataset which they were suspected to have the greatest confounding effect. Participant age and sex were investigated as a function of PTSD status within all individuals for whom genotyping data were available and MetS, history of smoking, history of alcohol use and CTQ scores were evaluated as a function of PTSD status within all individuals for whom DNA methylation data were available. Current MDD was assessed as a function of PTSD status across all recruited individuals. Overall differences were assessed through the application of Wilcoxon rank sum tests, independent sample t-tests or Pearson's Chi-squared tests where deemed appropriate by the Shapiro-Wilk test for normality.

Results

1. Statistical analysis of demographic and clinical data

Among the 295 individuals for whom genotyping data were available, the average age of PTSD cases and TEC did not differ significantly at 41 years (σ = 10.938) and 42 years (σ = 13.317), respectively (t = -0.651; *p* = 0.515) (Table 4). Furthermore, the reported proportion of participant sex was not significantly different across both outcomes – with females outnumbering males in a 3:1 ratio across both considered subgroupings (X^2 = 0.013; *p* = 0.908). Moreover, PTSD status was shown to be associated with higher median scores on the CAPS-5 (*p* = 2.200e-16), with PTSD cases and TEC presenting median severity scores of 37 (interquartile range (IQR) = 30 – 44) and 4 (IQR = 0 – 11) respectively. Upon examination, it was determined that neither age nor reported sex warranted inclusion as a genomic covariate due to neither presenting any significant divergence between the phenotypes under study.

Table 4: Demographic and clinical variables pertaining to genotyping and DNA methylation data

	Total	**PTSD**	**TEC**				
Genotyping Data:	**(n = 295)**	**(n = 153)**	**(n = 142)**	**W**	**t**	**X^2**	**p**
Mean Age	41.701	41.254	42.182		-0.651		0.515
(σ)	(12.129)	(10.938)	(13.317)				
Reported Sex (%)							
Male	77 (26.102)	39 (25.490)	38 (26.760)			0.013	0.908
Female	218 (73.898)	114 (74.510)	104 (73.240)				
Median CAPS-5 Severity Score	27	37	4	17176			***2.200e-16***
(IQR)	(6 - 38)	(30 - 44)	(0 - 11)				
Metabolic Syndrome (%)							
Present	84 (28.475)	46 (30.065)	38 (26.761)			0.249	0.618
Absent	211 (71.525)	107 (69.935)	104 (73.239)				
Smoked Previously (%)							
Yes	207 (70.169)	108 (70.588)	99 (69.718)			1.28e-3	0.971
No	88 (29.831)	45 (29.412)	43 (30.282)				
Previous Alcohol Use (%)							
Yes	262 (88.814)	135 (88.235)	127 (89.437)			0.020	0.887
No	33 (11.186)	18 (11.765)	15 (10.563)				
Median CTQ Score	47	54	40	15544			***1.582e-10***
(IQR)	(35 – 66.5)	(40 – 78)	(32 – 53.5)				
Moderate to Extreme (CTQ ≥ 41) Childhood Trauma Present (%)						20.975	***4.652e-6***
Yes	182 (61.695)	114 (74.510)	68 (47.887)				
No	113 (38.305)	39 (25.490)	74 (52.113)				
Major Depressive Disorder Present (%)						38.511	***5.445e-10***
Yes	63 (21.356)	55 (35.948)	8 (5.634)				
No	232 (78.644)	98 (64.052)	134 (94.366)				

** W, t and X^2 represent the test statistics for the Wilcoxon rank sum test with continuity correction, Welch's two sample t-test and Pearson's Chi-squared test with Yates continuity correction, respectively.*

Table 4: Demographic and clinical variables pertaining to genotyping and DNA methylation data

DNA Methylation Data:	**Total** **(n = 118)**	**PTSD** **(n = 61)**	**TEC** **(n = 57)**	**W**	**t**	**X^2**	**p**
Mean Age	43.394	42.700	44.136		-0.731		0.467
(σ)	(10.653)	(10.706)	(10.640)				
Reported Sex (%)							
Male	34 (28.814)	16 (26.230)	18 (31.579)			0.192	0.662
Female	84 (71.186)	45 (73.770)	39 (68.421)				
Median CAPS-5 Severity Score	26	36	4	3474.5			***2.200e-16***
(IQR)	(6 – 36)	(30 – 42)	(1 – 13)				
Metabolic Syndrome (%)							
Present	57 (48.305)	29 (47.541)	28 (49.123)			0.000	1.000
Absent	61 (51.695)	32 (52.459)	29 (50.877)				
Smoked Previously (%)							
Yes	75 (63.559)	42 (68.852)	33 (57.895)			1.091	0.296
No	43 (36.441)	19 (31.148)	24 (42.105)				
Previous Alcohol Use (%)							
Yes	100 (84.746)	54 (88.525)	46 (80.702)			0.855	0.355
No	18 (15.254)	7 (11.475)	11 (19.298)				
Median CTQ Score	46.50	50.00	40.00	2423.5			***2.256e-4***
(IQR)	(35 – 61.50)	(40 - 71)	(31 - 51)				
Moderate to Extreme (CTQ ≥ 41) Childhood Trauma Present (%)						7.561	***5.965e-3***
Yes	72 (61.017)	45 (73.770)	27 (47.368)				
No	46 (38.983)	16 (26.230)	30 (52.632)				
Major Depressive Disorder Present (%)						6.422	***0.011***
Yes	20 (16.949)	16 (26.230)	4 (7.017)				
No	98 (83.051)	45 (73.770)	53 (92.983)				

** W, t and X^2 represent the test statistics for the Wilcoxon rank sum test with continuity correction, Welch's two sample t-test and Pearson's Chi-squared test with Yates continuity correction, respectively.*

The 295 individuals for which genotyping data were available presented no significant divergence for the suspected presence of MetS, smoking status and previous alcohol use. However, PTSD status was observed as being associated with both higher scores on the CTQ and current MDD. Analysis determined that there was no significant difference in the frequency at which MetS was observed between PTSD cases and TEC ($X^2 = 0.249$; $p = 0.618$). Moreover, PTSD cases did not differ from TEC in both smoking history ($X^2 = 1.28e\text{-}3$; $p = 0.971$) and alcohol use ($X^2 = 0.020$; $p = 0.887$). PTSD cases reported having previously smoked and used alcohol at a prevalence of 70.588% (n = 108) and 88.235% (n = 135), respectively. Similarly, 69.718% (n = 99) of TEC were identified as having a history of smoking and 89.437% (n = 127) had previously engaged in alcohol intake.

Individuals falling under the PTSD subgrouping were deemed more likely to have previously experienced moderate to extreme childhood trauma ($X^2 = 20.975$; $p = 4.652e\text{-}6$), with PTSD cases presenting an elevated median CTQ score of 54 (IQR = 40 – 78) as compared to the median score of 40 (IQR = 32 – 53.5) documented amongst TEC ($p = 1.582e\text{-}10$). Furthermore, PTSD status was associated with an increased likelihood of having current MDD ($X^2 = 38.511$; $p = 5.445e\text{-}10$). According to initial M.I.N.I v6.0 assessment, 35.948% (n = 55) of PTSD cases and 5.634% (n = 8) of TEC were highlighted as exhibiting evidence of current MDD upon participant recruitment.

Within the subset of individuals for whom DNA methylation data had passed quality assurance (n = 118; 61 PTSD cases & 57 TEC), neither the suspected presence of MetS, smoking status, nor history of alcohol use were found to significantly differ between cases and controls. However, a PTSD diagnosis was consistently shown to be associated with higher scores on the CTQ both when considering median CTQ score and when distinguishing between participants who had experienced at least a moderate degree of childhood trauma (CTQ ≥ 41).

Analysis indicated that there was no significant difference in the frequency at which MetS occurred between the PTSD and TEC groups ($X^2 = 0.000$; $p = 1.000$). Furthermore, 68.852% (n = 42) of PTSD cases had engaged in smoking at some point in their lifetime – an observation which holds true for only 57.895% (n = 33) of TEC ($X^2 = 1.091$; $p = 0.296$). Moreover, similar analysis as to lifetime alcohol intake indicated that 88.525% (n = 54) of PTSD cases presented a history of alcohol use, where 80.702% (n = 46) of TEC fell under the same criteria ($X^2 = 0.855$; $p = 0.355$).

According to the clinical assessments, a PTSD diagnosis was associated with higher median scores on the CTQ ($p = 2.256e\text{-}4$); wherein 73.770% (n = 45) of PTSD cases were determined to have experienced at least a moderate degree of childhood trauma (CTQ ≥ 41). Alternatively, only 47.368% (n = 27) of TEC were noted as having met the same screening threshold ($X^2 = 7.561$; $p = 5.965e\text{-}3$). Considering the analysis conducted above, it was determined that methylomic covariates should be limited to previous experiences of moderate to extreme childhood trauma, supplemented by the cellular composition estimates previously generated by *meffil*.

In addition to the potential confounding elements considered, the grouping of individuals for which DNA methylation data were available did not differ significantly in age nor sex. However, the PTSD subgrouping was found to be associated with elevated median severity scores on the CAPS-5 and the presentation of current MDD. The average age of PTSD cases and TEC was 42 ($\sigma = 10.706$) and 44 ($\sigma = 10.640$), respectively ($t = -0.731$; $p = 0.467$). Moreover, the reported proportion of participant sex was similar to that of the genotyping dataset, with females outnumbering males in the same 3:1 ratio ($X^2 = 0.192$; $p = 0.662$).

PTSD status was observed as being associated with higher median severity scores on the CAPS-5, with PTSD cases and TEC presenting median scores of 36 (IQR = 30 – 42) and 4 (IQR = 1 – 13), respectively ($p = 2.200e\text{-}16$). Furthermore, individuals suspected of experiencing current MDD were revealed as being more likely to fall under the PTSD subgrouping, with 26.230% (n = 16) of PTSD cases and 7.017% (n = 4) of TEC identified as potentially experiencing current MDD under M.I.N.I v6.0 criteria ($X^2 = 6.422$; $p = 0.011$).

2. Genotyping data

GWAS data were initially available for 327 individuals (164 PTSD cases & 163 TEC). *A priori* analysis indicated that the raw dataset contained data on 1,713,748 SNPs, with a genotyping rate of 99.92%. After implementing quality control as per the methods described in the quality control of genotyping data section, the dataset was reduced to 319 individuals and 975,841 SNPs at a genotyping rate of 99.95% (Table 5). SNPs were initially removed if found to violate conditions for SNP missingness (11,797 SNPs omitted), minor allele frequency (697,101 SNPs omitted), Hardy Weinberg Equilibrium (4 SNPs omitted) or differential SNP missingness (309 SNPs omitted). Furthermore, an additional 28,696 SNPs were removed when filtering out all sex-chromosome associated data. Individuals were subsequently removed if found to violate conditions for individual missingness (n = 0), present as excessively heterozygous (n = 6) or present discordant sex information (n = 2).

Table 5: Components tested during quality control of genotyping data

Component Assessed	Exclusion Threshold	Resulting Loss
SNP-Level Filtering:		
SNP Missingness	x > 3%	11,797 SNPs
Minor Allele Frequency	x < 1%	697,101 SNPs
Hardy Weinberg Equilibrium	$p \leq 1 \text{ x } 10^{-6}$	4 SNPs
Differential SNP Missingness	$p \leq 5 \text{ x } 10^{-2}$	309 SNPs
Individual-Level Filtering:		
Individual Missingness	x > 1%	0 Individuals
Heterozygosity	$x < (\bar{x} - 3(\sigma))$ or $x > (\bar{x} + 3(\sigma))$	6 Individuals
Discordant Sex Information *(X – Chromosome Homozygosity)*	*Males:* x ~ 1.00 *Females:* x > 0.20	2 Individuals

** Post hoc removal of sex-chromosome associated data further excluded 28,696 SNPs.*

Performing imputation through the SHAPEIT2 & PBWT pre-phasing and imputation pipeline resulted in a preliminary net gain of 88,916,709 imputed variants. Applying the appropriate exclusionary thresholds reduced this to 11,331,335 SNPs at a genotyping rate of 98.88% (Table 6). Prior to conducting imputation, SNPs were removed if found to present allelic mismatches when aligned against the 1000 Genomes Phase 3 reference assembly (54,462 SNPs omitted). Once imputed, subsequent filtering removed SNPs found to violate conditions for INFO score (28,663,434 SNPs omitted), monomorphic variants (40,731,074 SNPs omitted), minor allele frequency (7,840,533 SNPs omitted) or SNP missingness (1,271,712 SNPs omitted). After accounting for hidden familial relationships (PI_HAT > 0.1875; 24 individuals lost) and further restricting analysis to SNPs that possessed unique rsIDs (778,004 SNPs lost), the prepared dataset consisted of 295 individuals (153 PTSD cases & 142 TEC) and 10,553,331 SNPs at a genotyping rate of 99.35%.

Table 6: Components tested during preparation & subsequent cleaning of imputed data

Component Assessed	Exclusion Threshold	Resulting Loss
Preliminary Data Preparation:		
Allelic Mismatches	-*	54,462 SNPs
Post-Imputation Quality Control:		
INFO Score	x < 0.8	28,663,434 SNPs
Monomorphic Variants (*Minor Allele Frequency*)	x = 0	40,731,074 SNPs
Minor Allele Frequency	x < 1%	7,840,533 SNPs
SNP Missingness	x > 3%	1,271,712 SNPs[1]

* *Allelic mismatches were identified through iterative corrective alignment against the 1000 Genomes Phase 3 reference assembly* (93).

[1] *It should be noted that no variants were removed upon applying an intermediary filter for SNP Missingness (x > 30%) during the VCF to PLINK binary file conversion.*

3. Genome-wide association testing

GWAS was performed on the imputed dataset (n = 295; 153 PTSD cases & 142 TEC) using principal components 2 and 18 as genomic covariates (Figure 4, Supplementary Figure 9). Upon initial assessment, it was determined that no single variant surpassed the necessary threshold for achieving genome-wide significance ($p \leq 5.00e\text{-}8$). A summary documenting 8 of the most notable associations identified has been provided in Supplementary Table 2.

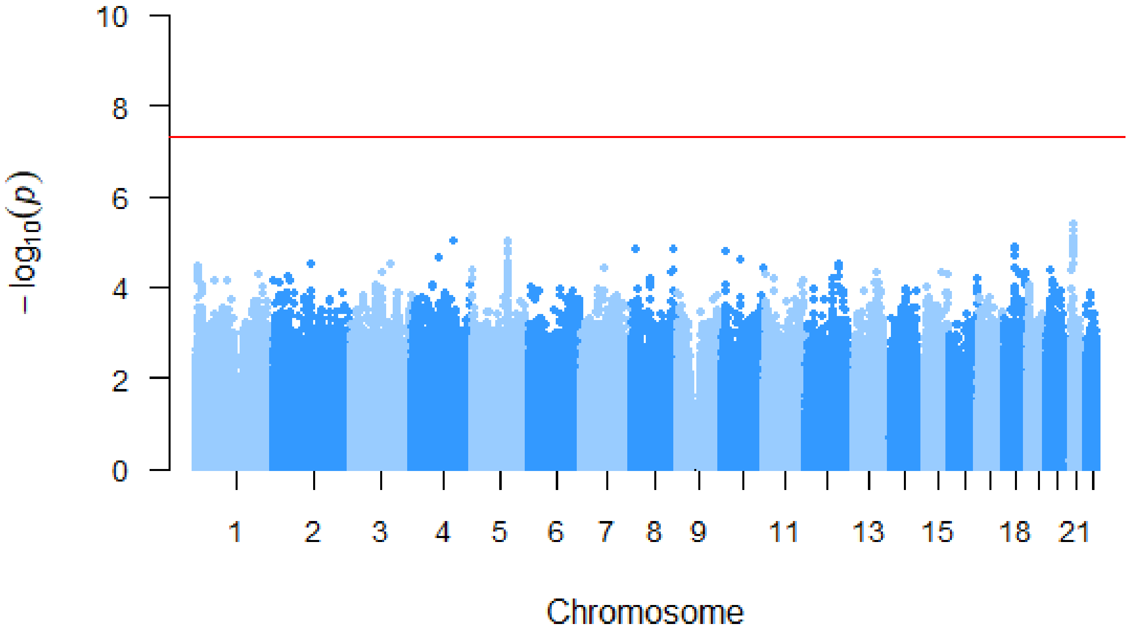


Figure 4: Manhattan plot depicting the results of a genome-wide association test conducted on PTSD. The graph represents the degree to which individual SNPs are associated with PTSD in the cohort under study. The x-axis serves as a positional scale for the physical location at which each SNP occurs and the y-axis reflects the corresponding p-values in -$\log_{10}$ form. Critically, the unbroken red line is the commonly implemented threshold for delimiting genome-wide significance ($p \leq 5.00\text{e-}8$).

4. Polygenic risk scores

PRS were initially calculated using two combinations of the PGC PTSD Freeze 2 population subgroupings as references: one using solely African American data and the other employing European, Latino/Native American and African American data (Figure 5). PRS constructed using a combination of European, Latino/Native American and African American data consistently outperformed those using solely African American data across all thresholds tested. *PRSice* determined that the most predictive PRS utilized 1,444 variants (*p*-value threshold = 0.001) derived from the combined reference to calculate weighted scores for each individual. At this threshold, the proportion of variance explained by the isolated PRS model versus that explained by the genomic covariates (principal components 2 & 18) was 1.54% (Nagelkerke's pseudo R^2 = 0.0154) and 2.87% (Nagelkerke's pseudo R^2 = 0.0287), respectively. As such, the combined total proportion of variance explained was 4.41% (Nagelkerke's pseudo R^2 = 0.0441). However, the model was not able to significantly distinguish between PTSD cases and TEC both before ($p \leq 0.05$) and after ($p \leq 6.25\text{e-}3$) adjusting for Bonferroni multiple testing correction ($p = 0.064$).

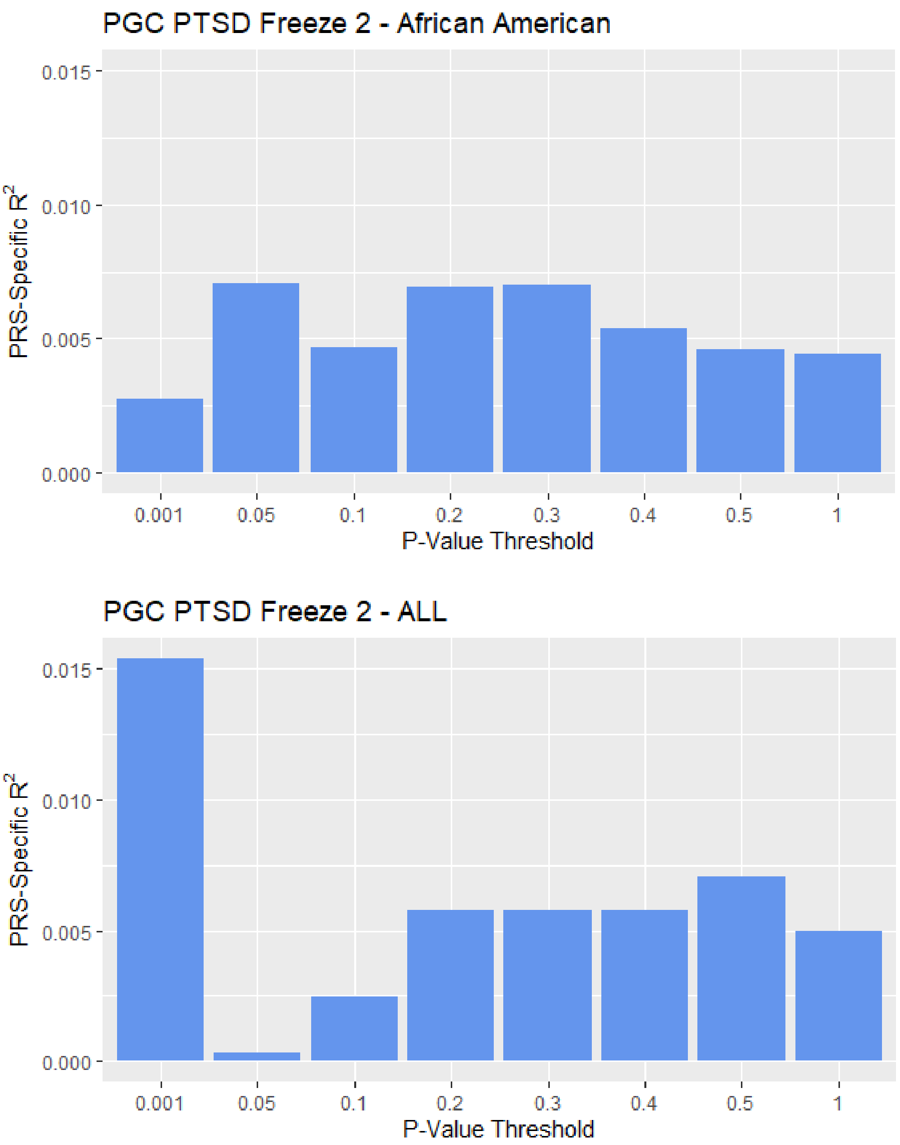


Figure 5: Comparison of the proportion of variance explained by two PGC PTSD Freeze 2 references. Model predictiveness was assessed by applying a predetermined range of p-value thresholds (0.001, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5 and 1.00) to the African American and combined European, Latino/Native American & African American summary statistics to evaluate their ability to distinguish between PTSD cases and TEC within the imputed dataset. Reference performance was measured through the proportion of variance (in the form of Nagelkerke's pseudo R^2) explained across all tested thresholds. *PRSice* determined that PTSD status could best be predicted through a combination of 1,444 variants identified by applying a p-value threshold of 0.001 to the PGC-All summary statistics (Nagelkerke's pseudo $R^2 = 0.0441$; $p = 0.064$).

5. DNA methylation data

Prior to performing quality control, it was determined that the unfiltered dataset presented data pertaining to 865,859 CpG probes. Upon applying the appropriate parameters in *meffil* (as described in the quality control and functional normalisation of DNA methylation data subheading), the dataset was reduced to 118 individuals (61 PTSD cases and 57 TEC) and a corresponding 840,920 CpG probes (Table 7). CpG probes were removed due to failing conditions for probe p-values (1,078 probes omitted) and probe bead numbers (4,299 probes omitted), as well as due to being associated with sex-chromosome linked data (19,562 probes omitted). No individual samples were found to violate conditions for sample p-values or sample bead numbers, nor present discordant sex information – however, two individuals were removed due to aberrant median methylated intensities. Moreover, both procedure- and identity-based control probes were found to perform in line with manufacturing recommendations and all methylation profiles fell within the theoretical boundaries established by the gse35069 complete whole blood reference.

Table 7: Components tested during quality control of DNA methylation data using *meffil*

Component Assessed	Exclusion Threshold	Resulting Loss
Individual-Level Filtering:		
Sample P-Value	More than 10% of sample probes fail to be detected	0 Individuals
Sample Bead Number	More than 10% of sample probes fail to be detected across at least 3 beads	0 Individuals
Discordant Sex Information (*Median X- and Y-Chromosome Intensities*)	x < (Dataset Median – 5(σ)) or x > (Dataset Median + 5(σ))	0 Individuals
Median Methylated Intensities	x < (Dataset Median – 3(σ)) or x > (Dataset Median + 3(σ))	2 Individuals
Probe-Level Filtering:		
Probe P-Value	Probes fail to be detected across more than 10% of samples	1,078 Probes
Probe Bead Number	Probes fail to be detected across at least 3 beads for more than 10% of samples	4,299 Probes

** A p-value threshold of 0.01 was used to distinguish between detected & undetected probes.*

*** Post hoc removal of sex-chromosome associated data further excluded 19,562 probes.*

6. Epigenome-wide association testing

EWAS was performed on the normalized intensity measures (n = 118; 61 PTSD cases & 57 TEC) using childhood trauma, cellular composition estimates and SVA-determined surrogate variables as covariates (Figure 6, Supplementary Figure 10).

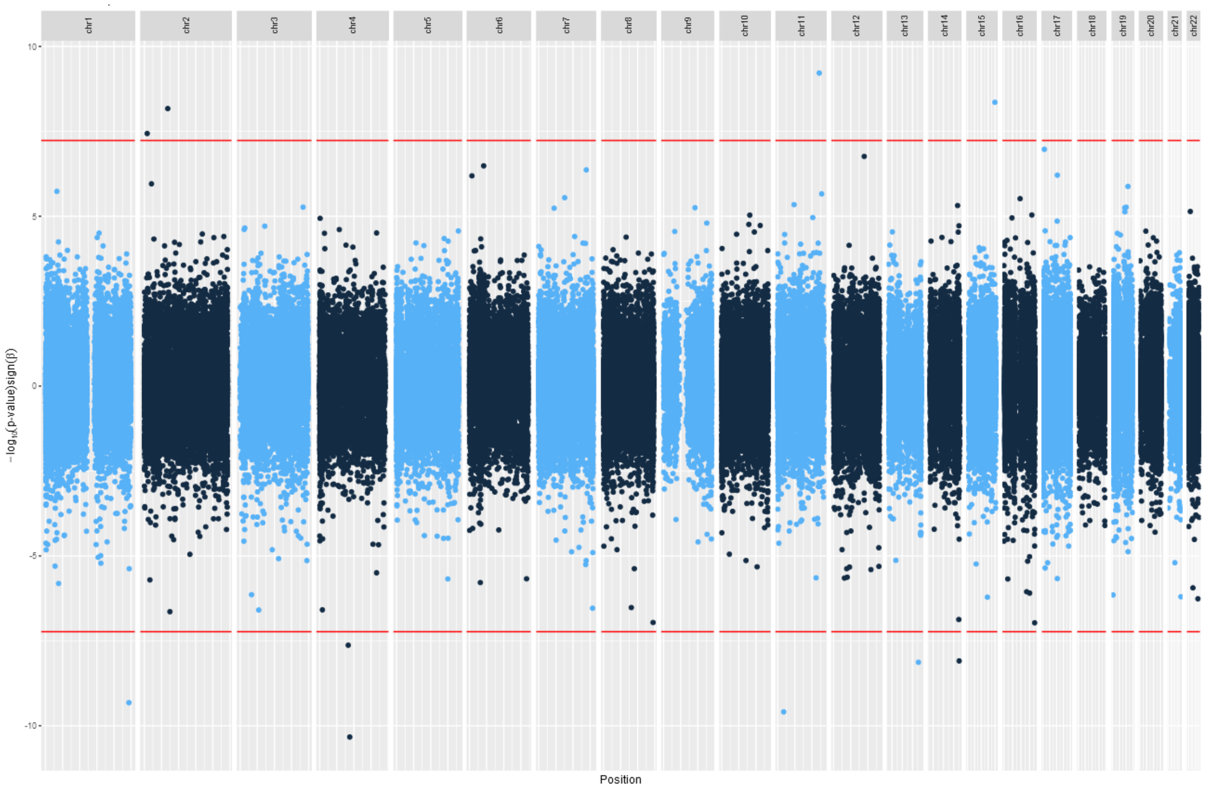


Figure 6: Manhattan plot depicting the results of an epigenome-wide association test conducted on PTSD. The *meffil* – generated graph represents the degree to which individual CpG probes are associated with PTSD in the cohort under study. The x-axis serves as a positional scale for the physical location at which each CpG probe occurs and the y-axis reflects the corresponding p-values in -$\log_{10}$ form. Upwards and downwards trajectories are reflective of hyper- (increased) and hypo- (decreased) methylation relative to TEC, respectively. The unbroken red lines represent the upper and lower bounds for achieving epigenome-wide significance ($p \leq 5.95e\text{-}8$). 10 CpG probes surpassed the necessary threshold for achieving epigenome-wide significance.

Upon initial assessment, it was determined that 10 CpG probes surpassed the necessary threshold for achieving epigenome-wide significance ($p \leq 5.95e\text{-}8$) (Table 8). Furthermore, 3 of the significant probes could be annotated to proximally located genes: (i) cg13981804 was annotated to *chromosome 4 open reading frame 36* (*C4orf36*) (*p* = 4.654e-11); (ii) cg10245330 was annotated to *SHH signalling and ciliogenesis regulator SDCCAG8* (*SDCCAG8*) and AKT serine/threonine kinase 3 (*AKT3*) (*p* = 4.781e-10); and (iii) cg11724557 was annotated to *tublin epsilon and delta complex 1* (*TEDC1*) (*p* = 8.105e-09). One DMR was detected at chr4: 53588360–53588374 (Supplementary Table 3) (*p* = 4.038e-9) but shared no overlap with the 10 significant probes identified.

Table 8: Annotation of notable CpG probes identified through epigenome-wide association testing

CpG Site	Chromosome	Position	Coefficient	Nearest Gene	p
cg13981804	4	87797509	-0.034	*C4orf36*	4.654e-11
cg22155376	11	17591837	-0.070	-	2.559e-10
cg10245330	1	243652681	-0.006	*SDCCAG8* *&* *AKT3*	4.781e-10
cg07172637	11	121427672	0.020	-	6.057e-10
cg26847571	15	99664049	0.061	-	4.391e-9
cg20919705	2	68446054	0.032	-	6.729e-9
cg03696393	13	107027491	-0.005	-	7.387e-9
cg11724557	14	105965005	-0.009	*TEDC1*	8.105e-9
cg18579761	4	83323585	-0.021	-	2.347e-8
cg00511884	2	8374459	0.022	-	3.644e-8

** - or + coefficient values are indicative of hypo- and hyper-methylation relative to TEC, respectively.*

*** Nearest gene determined through missMethyl & limma mediated annotation (Illumina 450k array annotation manifest).*

7. Methylation quantitative trait loci

Methylation quantitative trait loci were initially identified for 97 individuals (51 PTSD cases & 46 TEC) by assessing potential relationships between 1,444 variants, comprising the PRS, and 840,920 CpG probes while using childhood trauma and cellular composition estimates as covariates. In total, we detected 44,614 mQTLs as a product of interactions between 250 SNPs and 26,349 CpG probes ($p \leq 5.95e\text{-}14$) (Table 9). Each CpG probe was found to be associated with a median of 2 different variants (IQR = 1 – 2), whereas each SNP was found to be associated with a median of 3 CpG probes (IQR = 1 – 10.5). However, while the most interconnected CpG probe only interacted with 4 different variants, its SNP counterpart was associated with 7,519 mQTLs (rs144798302). Furthermore, the CpG probes associated with mQTLs spanned a wide variety of genomic regions, wherein: 0.144% occurred in genes, 0.108% occurred in non-gene regions, 8.213% occurred in promoter regions, 0.238% occurred in cell-type specific genes, 0.004% occurred in non-gene regions attributed to cell-type, 0.543% occurred in cell-type specific promotor regions, and 90.750% were unclassified.

Table 9: mQTLs detected

	mQTLs Detected	Total SNPs Involved	Total CpG Probes Involved	Mean Effect (σ)
Total	44,614	250	26,349	0.230 (0.136)
Cis-Acting	95	62	91	0.194 (0.119)
Trans-Acting	44,519	197	26,273	0.230 (0.136)

** SNP-CpG parings that occurred within 500kb of each other were defined as cis-acting and pairings that occurred over a distance greater than 500kb or on separate chromosomes were defined as trans-acting.*

Of the 44,614 mQTLs detected, 95 were defined as cis-acting mQTLs (encompassing interactions between 62 SNPs and 91 CpG probes) and 44,519 were termed to be trans-acting mQTLs (interactions between 197 SNPs and 26,273 CpG probes). Mean effect sizes were calculated across the absolute value of the change in methylation intensity induced by the presence of the tested SNP so as to generate a proxy representative of the net effect associated with the various mQTL subtypes. SNPs were associated with a mean 0.194 (σ = 0.119) and 0.230 (σ = 0.136) change in methylation intensity at cis-acting and trans-acting mQTLs, respectively. The mean effect change detected across all registered mQTLs was 0.230 (σ = 0.136). Furthermore, restricting trans-acting mQTLs to those with SNP-CpG parings which occurred on the same chromosome, indicated that mQTLs with interactive distances between 500kb - 1Mb (1,000,000 base pairs) and those occurring across greater than 1Mb presented mean effect sizes of 0.190 (σ = 0.114) and 0.219 (σ = 0.127), respectively (Figure 7). Upon attempting to implement Bayesian Colocalisation, no colocalising effects were detected.

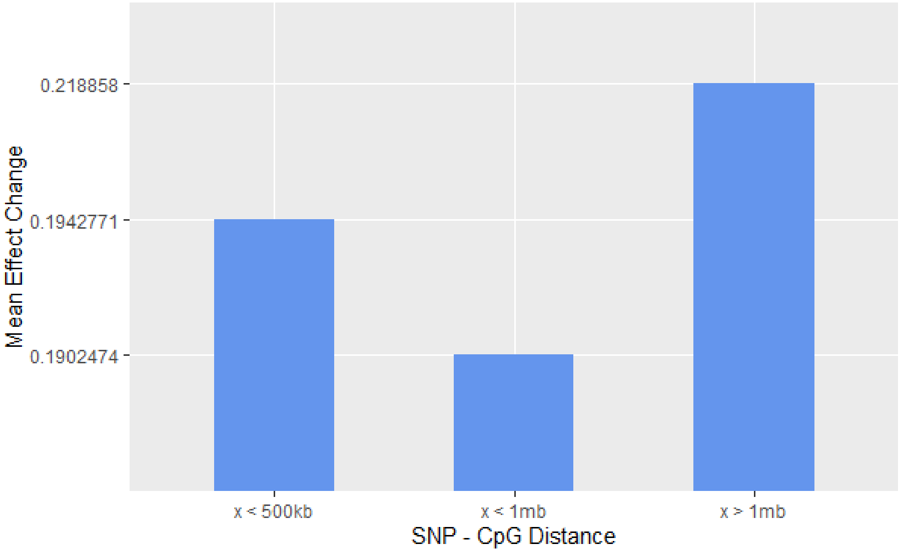


Figure 7: Comparing the mean effect sizes of cis- and trans-acting mQTLS. Mean effect sizes were calculated as the absolute value of the change in methylation intensity induced by the presence of the tested SNP.

Discussion

In this thesis, we proposed that a multi-omics, or systems biology, based approach presented a promising avenue through which to assess the molecular mechanisms underpinning the development of PTSD. We briefly outlined evidence suggesting that the presentation of PTSD is predicated upon a network of molecular interactions – and drew particular attention to the potential concordance of aberrant genetic and epigenetic effects (13,49,65). Furthermore, we suggested that the identification of mQTLs, genomic loci where specific genetic variants are capable of directly influencing DNA methylation patterns, served as a method to detect such genetic-epigenetic interactions and advocated that PRS could be used to circumvent traditional issues posed by limited sample sizes.

As such, the aim of this study was twofold: (i) to use PRS modelling for the construction of a variable capable of predicting PTSD case-control status within a South African population; and (ii) to integrate PRS and DNA methylation data for the identification of mQTLs associated with PTSD.

Prior to implementing mQTL analysis, we elected to conduct independent association tests so as to further characterise how both the genomic and DNA methylation datasets were related to the phenotype under study. Applying a GWAS approach to the genotyping data revealed that no single variant surpassed the necessary threshold for achieving genome-wide significance ($p \leq 5.00\text{e-}8$) (Figure 4). Furthermore, assessing the 8 most notable SNPs highlighted by FUMA offered no link to any genes that had previously been found to be associated with PTSD (Supplementary Table 2). However, FUMA's inbuilt MAGMA gene-set analysis indicted that the dataset was enriched for a negative regulation of neurotransmitter transport gene set derived from the Molecular Signatures Database v7.2 (Bonferroni-corrected p = 0.011).

This finding is important for two reasons: firstly, previous studies have shown that both PTSD, as well as a dysfunctional behavioural and physiological response to extreme stress, are associated with abnormal neurotransmitter profiles and reactivity (135). Specifically, fluctuations in the homeostatic state of neurotransmitters involved in the stress- and fear-response are believed to greatly affect both one's initial risk of developing PTSD as well as the severity of subsequent symptoms experienced (136). Moreover, of particular importance is the observation that PTSD is not consistently attributed to deviations in a single neurotransmitter class – but rather, is thought to be the product of widespread dysregulation through an interconnected network of neurobiological interactions (137). As such, while one would not necessarily expect to observe highly significant loci in such a small sample, indirect evidence of a mechanism previously associated with PTSD is encouraging as to the possibility of uncovering additional molecular underpinnings with further analysis.

Upon initially constructing PRS, we identified the PGC's PTSD Freeze 2 data release as the large-scale GWAS best suited for the prediction of PTSD status within our sample population. Preliminary runs were conducted using the African American and a combination of all European, Latino/Native American, and African American population data as the reference datasets through which the *PRSice* algorithm selected PTSD-associated variants for score generation (Figure 5). Models constructed using a combination of all population subgroupings were consistently found to explain a greater proportion of phenotypic variance than those only utilising African American references. The most predictive PRS could use weighted compositions of 1,444 variants derived from the combined European, Latino/Native American, and African American references to explain 4.41% ($R^2 = 0.0441$) of the phenotypic variance observed. However, this model (p = 0.064) did not survive Bonferroni multiple testing correction.

This raises a particularly pertinent issue with regards to the effectiveness of PRS constructed using reference datasets that were derived from population groups distinct from one's own cohort. PRS are fundamentally dependent on the accurate estimation of variant effect sizes – a property which largely relies on the relationship between the population from which the reference draws and that in which the scores are calculated (138). Due to innate differences in allele frequencies and linkage disequilibrium patterns between populations, variant effect estimates tend to translate poorly across ancestry groups; often to the degree that one frequently observes a considerable loss in predictive performance when failing to ensure base-target portability (139).

Superficially, the most apparent solution would simply be to select a reference dataset originating from the same ancestry group as one's sample population. However, historical bias towards Euro-centric studies have generated an artificial dearth of large-scale non-European GWAS – wherein dramatically reduced sample sizes lend towards an inevitable loss of predictive accuracy (140). Given the severity of such complications, considerable efforts have been devoted to the development of statistical techniques tailored specifically to maximising the performance of PRS in underrepresented populations. In fact, two methods in particular, one advocating for the utilisation of multiple multi-ethnic reference datasets and the other for the incorporation of functional-annotation based weighting; have become increasingly well established in recent years (140–143). While we elected to limit all optimisation efforts to the boundaries established by the largest available reference dataset, it is encouraging to note the development of new methods which may help further refine our analyses moving forward.

Conducting an EWAS on the DNA methylation data indicated that 10 CpG probes passed the necessary threshold for achieving epigenome-wide significance ($p \leq 5.95\text{e-}8$) (Figure 6, Table 8). Of the significant methylation sites, *missMethyl* and *limma* annotated three to the genes *chromosome 4 open reading frame 36* (*C4orf36*), *SHH signalling and ciliogenesis regulator SDCCAG8 (SDCCAG8*), *AKT serine/threonine kinase 3* (*AKT3*) and *tubulin epsilon and delta complex 1* (*TEDC1*). Methylation site cg13981804 was found to be hypomethylated in association with *C4orf36* (*p* = 4.654e-11); a protein coding gene whose function has yet to be determined (144). Furthermore, methylation site cg10245330, which occurs in a region overlapped by both of its affected genes, was found to be hypomethylated in association with both *SDCCAG8* and *AKT3* (*p* = 4.781e-10). *SDCCAG8* encodes a centrosome-associated protein involved in the arrangement of cellular machinery during mitosis and directly contributes to the assembly of cellular cilia (145). Alternatively, *AKT3* encodes a protein kinase falling under the greater serine/threonine class and existing literature indicates that the resulting gene product plays a role in cellular signalling as well as being closely linked to regulation of the cell cycle, glycogen synthesis and glucose uptake (146). Lastly, methylation site cg11724557 was found to be hypomethylated in association with *TEDC1* (*p* = 8.105e-9); which encodes a gene product involved in ciliary signalling as well as centriole and actin filament stability (147). However, the significant DMR identified at chr4:53588360–53588374 (*p* = 4.038e-9), did not correspond with any known gene or regulatory element.

Of the significant probes and their respective annotations, all but the *AKT3* annotation were found to have no previous association with PTSD. Furthermore, the mechanistic pathways underlying the regulation of centriole-based organelles, cilia functioning, and mitosis seem superficially unlikely to contribute directly to mediating risk and resilience to PTSD. However, *AKT3* was previously identified by Xie *et al.,* (2013) as a suggestive association (*p* = 5.11e-6) in a GWAS conducted on 1,578 European Americans (16). Due to their associated role in specific cellular signalling pathways, the AKT protein kinases have long been considered potential exploratory targets for PTSD (148). The extended signalling network associated with AKT glycogen synthesis currently serves as one of the major molecular targets for lithium- and antidepressant-based psychiatric pharmacological interventions (149). Moreover, previous studies have suggested that both normal and aberrant neurotransmitter profiles regulate behavioural changes through this same pathway (150). Additionally, a shared genetic foundation may already have been established through recent observations associating altered *AKT1* activity with an increased prevalence of schizophrenia and mood disorders (151).

It should be noted, that previous EWAS have attributed *AKT3* hypomethylation to the effects of smoking-related behaviours (152). As such, while our preliminary analysis indicated that smoking status did not differ significantly between cases and controls (X^2 = 1.091; *p* = 0.296), any association detected herein may potentially be due to the unaccounted for effects of smoking. However, the nature of this finding indirectly mirrors that of a similar observation recently made by Smith *et al.*, (2020) – wherein, in a large EWAS meta-analysis the authors noted several epigenome-wide significant associations reflecting hypomethylation in *aryl-hydrocarbon receptor repressor* (*AHRR*), a gene which had previously been associated with smoking-related hypomethylation (12). While the findings initially lost significance upon introducing new controls for smoking status, subsequent assessment comparing the methylated state of known smoking loci and *AHHR* revealed that the detected hypomethylation associations were predominantly concentrated in non-smoking cases (12). As such, the authors highlighted a new approach through which to possibly untangle independent observations from mistaken associations to known confounding effects.

mQTLs were investigated in 97 individuals (51 PTSD cased & 46 TEC) by assessing potential relationships between 1,444 variants and 840,920 CpG probes. We identified 44,614 mQTLs, 95 of which were cis-acting and 44,519 of which were trans-acting, at a *p*-value of 5.95e-14 (Table 9). Preliminary analysis seemed to indicate that the genomic and epigenetic datasets presented a high degree of interconnectivity; wherein both the SNP and CpG portions of each SNP-CpG paring were often found to interact with multiple other mQTLs in both a cis- and trans-acting manner. Each CpG probe was found to interact with a median of 2 different variants (IQR = 1 – 2) whereas the inverse indicated that each SNP was found to interact with a median of 3 different CpG probes (IQR = 1 – 10.5). However, the genomic dataset was skewed towards a much higher upper bound, with one variant (rs144798302) interacting with 7,519 unique mQTLs. The variant in question exists in an largely isolated state on chromosome 1 (chr1:35789171) and has yet to be assigned any functional or regulatory annotation (153).

Nevertheless, the high degree of reciprocity between genetic and epigenetic sites appears to support our initial hypothesis that PTSD could best be represented as the product of an extensively integrated network of molecular interactions (13). This notion is further bolstered by the finding that the distribution of mQTL-associated CpGs is widely spread across a variety of genomic regions. While the vast majority of CpGs are, admittedly, unclassified – the second (8.213%) and third (0.543%) largest categories pertain to general promotor regions and cell-type specific promoter regions, respectively. This is particularly critical as they represent the two methods through which DNA methylation can arguably exert its greatest effects.

The occurrence of differential DNA methylation in promoter regions has become one of the most widely documented methods through which epigenetic influences can affect gene expression (154). Specifically, methylation-induced alterations to either the general chromatin structure or transcription factor binding mechanisms is closely associated with reduced expression in a corresponding gene (155). Moreover, one of the other primary functions of DNA methylation pertains to the regulation of tissue- and cell-type specific differentiation. Maintaining cell-type specific differentiation is simply a prolonged and highly localised method of influencing gene expression – wherein the collective status of a series of cell-type specific promotors is critical to the development and overall functioning of a particular cell (156,157). The benefits of such an effect lies in its potential to further extend the molecular reach of an mQTL to affect additional biological domains. To this extent, it would theoretically be possible for a single mQTL-associated SNP to induce differential DNA methylation and differential gene expression both locally and across considerable distances of the genome.

Interestingly, none of the ten significant CpG sites, or, for that matter, any of the eight suggestive SNPs observed in this study, were found to take part in any significant mQTL associations. Furthermore, reducing the p-value threshold from 5.95e-14 to the 1e-10 utilised for Bayesian Colocalisation analysis, still failed to result in the incorporation of any of the previously identified CpG sites or suggestive SNPs. Such a discrepancy may potentially be due to the data loss encountered upon trimming the genotyping and DNA methylation datasets down to an overlapping sample. However, given concerns as to whether the *AKT3* finding should be attributed to genomic or environmental influences – this raises an interesting question as to whether mQTL analysis could be utilised to help distinguish between molecularly- and environmentally-induced differential methylation changes.

One unexpected result was the frequency with which mQTLs were distributed, as well as the differences in accompanying mean effect sizes, across cis- and trans-acting mQTLs. Hannon *et al.*, (2018) briefly summarised the existing literature in saying that mQTLs are: (i) expected to occur more frequently in cis-acting regions than in their trans-acting counterparts; and (ii) expected to grow in both effect size and significance as the distance between the SNP-CpG pairing decreases (132). However, we found the opposite to be true for both conditions. First, our integrated dataset was found to present a substantially larger proportion of trans-acting mQTLs which in turn were also associated with a larger mean effect size than their cis-acting counterparts. Furthermore, restricting trans-acting mQTLs to those with SNP-CpG parings which occurred on the same chromosome (Figure 7), indicated that while mean effect sizes increased as the genomic distance fell from above 500kb – a greater increase was observed once the interactive distance was greater than 1Mb.

This considerable deviation from the established norm serves as a reminder that these results may not be truly representative of the molecular mechanisms underlying PTSD in our sample population. Importantly, the crux of the mQTL analysis depended on the PTSD-associated variants identified by our best-fit PRS model. As such, the study effectively made two allowances in exchange for a more statistically powerful association proxy: (i) Hannon *et al.*, (2018)'s adjusted *MatrixEQTL* scripts only implemented clumping procedures after initially calculating mQTLs and their associated effect sizes – therefore, utilising a PRS-mediated approach forces the dataset to undergo premature clumping; and (ii) concerns amid the loss of predictive accuracy associated with using both an unoptimised reference dataset and non-significant PRS model.

The premature clumping may explain some of the discrepancies observed in the expected cis-to-trans ratio simply due to methodological differences in the manner which clumping is conducted. *PRSice* implements clumping by using local linkage disequilibrium patterns to group every SNP which surpasses a predetermined correlation threshold in a set window of base pairs (69,70). However, while the mQTL clumping procedure is based on the same underlying principals, it differs in that clumping is only conducted on an assembled list of SNPs that were shown to present at least one mQTL association (132). Once the genomic dataset has been trimmed down to the 1,444 variants best suited for the prediction of PTSD case-control status, it is dramatically less likely that any of the remaining variants will occur in close proximity with one another. As such, the aberrant cis-to-trans ratio is most likely the product of chance amplified by a strict clumping procedure.

By extension, this may also explain why we failed to detect any overarching causal variants when attempting to implement Bayesian Colocalisation analysis. *Coloc*'s functioning requires that it conduct pairwise mQTL tests for all possible CpG pairings within 250kb from one another against all SNPs within 500kb of the pairing (132,134). Therefore, the relative depletion of closely grouped variants due to clumping may have drastically reduced the number of instances for which such a calculation is even possible.

However, the concerns surrounding the predictive accuracy of the PRS model implemented are markedly more difficult to account for. When considering that the variants utilized for mQTL analysis were unable to significantly distinguish between cases and controls, one should be hesitant to regard these findings as much more than a proof-of-concept paving the way for future research.

Yet, while this thesis did not achieve results with a high degree of certainty, it did highlight certain limitations that should be addressed by other studies moving forward. Firstly, regardless of the exact molecular mechanisms through which PTSD develops, aberrant effects are widely thought to originate within the central nervous system (13). Moreover, epigenetic modifications have shown a marked ability to present both tissue- and cell- specific differentiation patterns (12). As such, there is growing concern that existing sampling methods fail to accurately capture a representative image as to the inner workings of the brain (158). While the non-invasive nature of peripheral blood measurements render them ideal for biomarker-based studies, they are at most an indirect approximation of neuropathological features (49). While several online databases, such as the University of Essex Blood Brain DNA Methylation Comparison Tool (159) and iMethyl (160,161), have been developed to help assess how well peripheral findings translate to the brain, such methods are only a temporary solution in the face of greater a problem. Future efforts need to be dedicated towards developing an appropriate proxy for physical brain tissue that can be readily compared against a more accessible medium.

Secondly, PTSD’s inherent dependency on an initial traumatic event can greatly complicate efforts to assess the epigenetic mechanisms underlying the disorder. Very few individuals live in such conditions that researchers have accesses to extensive clinical and biological data prior to the triggering event occurring (12). As such, it may prove difficult to differentiate novel epigenetic changes from the effects of unknown confounders as well as from the initial traumatic or stressful event.

Furthermore, experimental design poses several issues when current diagnostic criteria account for the duration at which symptoms are experienced and a substantial proportion of the developmental risk is dependent on individual perception of the traumatic event (11). The primary concern is maintaining strict phenotypic boundaries between PTSD cases and TEC; however, both the alleviation of past symptoms and delayed development of future symptoms can readily introduce confounding elements in form of past or subthreshold cases recruited to the control group. Moreover, the basic principles of experimental design dictate that one would ideally like to match recruited participants across demographic data and the outcome under study. However, quantifying the different types and severity of trauma as well as the individual experience thereof, may ultimately prove too challenging to maintain a homogenous sample.

One potentially notable limitation lies in how we elected to utilise childhood trauma exposure and MDD as analytical covariates. Both PTSD and MDD are highly colocalising and childhood trauma has previously been found to be comorbid with a wide variety of psychopathological outcomes (162,163). Within the subset of individuals for which genotyping data were available (n =295; 153 PTSD cases and 142 TEC), PTSD diagnosis was significantly associated with previous experiences of moderate to extreme childhood trauma ($X^2 = 20.975$; $p = 4.652e\text{-}6$) and current MDD ($X^2 = 38.511$; $p = 5.445e\text{-}10$). Although PTSD, MDD and childhood trauma exist as distinct entities, each shares a similar cluster of symptoms or resulting effects such that it is impractical for either MDD or childhood trauma to be used as exclusion criteria in the current study (164). However, these shared molecular underpinnings have raised concerns as to how best to account for the effects of MDD and childhood trauma without introducing additional confounding variables to PTSD-associated analysis.

We elected to utilise an approach described by Contractor *et al.,* (2018) as the causality explanation – where childhood trauma is incorporated as a covariate due to serving as a risk factor for PTSD development, whereas MDD is not controlled for as PTSD is subjectively deemed a more direct risk factor for the development of MDD than vice versa (165). Moreover, previous analysis done by Dr Patricia Swart within Stellenbosch University's Neuropsychiatric Genetics Research Group used PCA to assess MDD aggregation within the genotyping cohort, wherein it was determined that current MDD presented no spatial relationship that was overtly indicative of hidden confounding. Nevertheless, how best to account for MDD and childhood trauma in PTSD studies remains poorly understood, and future studies may benefit from further refining of how each occurrence is defined at a molecular level.

Moreover, post-hoc evaluation of the statistical power underlying the mQTL analysis highlights additional limiting factors. Current consensus indicates that one's statistical power to detect both cis- and trans-acting mQTLs extends from a sufficiently powered variant discovery phase – wherein the detection of mQTLs, and indeed any additional characterisation efforts, essentially serve as exploratory rather than identificatory analyses (132). While constructing PRS effectively circumvented initial sample size concerns, such predictive approaches are also subject to similar statistical uncertainties. Using sufficiently large training datasets typically increases confidence in the predictive accuracy of PRS, with the caveat that the selected training dataset is appropriate for both the trait and population under study (166). As such, considering the aforementioned concerns regarding the suitability of the PTSD Freeze 2 data release as a training dataset for our sample population, as well as the statistical insignificance of the most predictive model, the mQTL analysis may by extension lack sufficient statistical power to accurately detect cis- and trans-acting effects.

The potential of using a PRS-mediated proxy for the detection of mQTLs associated with PTSD remains a promising concept. To better characterise the molecular mechanisms underlying PTSD; future studies could expand upon our suggested multi-omics proposal by further incorporating additional biological domains to maximise the amount of available information. Specifically, supplementing existing research with neuroimaging-based approaches would help to begin circumventing the inaccessibility of brain tissue by generating proxy measures that could be correlated to peripheral findings (49). Moreover, there exists a dire need to further encourage the establishment and subsequent curation of brain-based biobanks dedicated towards the recruitment of individuals who have experienced psychiatric disorders throughout their lifetime. Such efforts should ideally be accompanied by longitudinal assessments wherein some degree of basal clinical data is provided from the point of recruitment. In such manner, one would increase the available resources for verifying molecular findings post-mortem while having the necessary supplementary information to formulate observations appropriately.

Furthermore, the vast majority of studies have thus far approached PTSD from a purely binary perspective, where PTSD cases are collectively pooled into a single representative phenotype. The primary disadvantage therein is that individual symptom clusters as well as sub- and intermediate-phenotypes remain relatively understudied (13). Such isolated components may allow for more refined analysis of the molecular mechanisms underlying PTSD by not masking weaker sources of phenotypic heterogeneity (6). Future studies could thus further explore a quantitative approach towards analysing PTSD, wherein participants are classified in a continuous manner according to symptom severity. Adopting such a probabilistic method may grant greater sensitivity and help better define PTSD's biological underpinnings (167).

Additionally, while PCA-based covariates are commonly used to address potential population stratification, incorporating increasingly complex techniques may allow future studies to more accurately adjust for any confounding effects related to genetic ancestry (168). Principal component analysis utilises an uninformed approach to identifying variation in a given dataset – where intricate genetic contributions are often missed in a broader search for large swaths of hidden variation

(169). Programs such as the ancestry estimation tool ADMIXTURE avoid such pitfalls by using population-specific references to estimate the degree to which historical source populations contribute to individual ancestry (170). Adopting such an approach would allow one to quantitatively assess population stratification at the individual-level, thus further increasing the accuracy of analysis conducted in ancestrally diverse populations (171).

Lastly, on numerous occasions throughout this thesis, we noted the disproportionate degree to which genomic studies had been conducted in European populations - often highlighting the need for future contributions using multi-ethnic samples. Much of this conversation used the term "ancestry" to refer to broad geographical and historical differences between different population groups. While not incorrect, using such a descriptor greatly oversimplifies the role of diversity in molecular studies. Recent work by Peterson *et al.,* (2019) describes how using ancestral divisions often overlooks the contributory effects of ethnic factors (172). The authors further mention that ethnicity may serve as a surrogate variable for several social, cultural and environmental considerations of interest to the study of disease aetiology (172). While there is a broader need for more analysis conducted in individuals of non-European ancestry, future studies may greatly benefit from maximising diversity on a smaller scale. Where ensuring adequate representation both across, as well as within, ancestral divisions may drastically improve our understanding of genetic variation under psychiatric disorders.

Conclusion

Thus far, there are no valid biomarkers capable of granting us insight into the functional manner in which PTSD develops in the brain (173). However, prior twin and epidemiological studies have indicated that the disorder presents a heritability of 40-50%, which would suggest that there is at least some partial genetic component associated with the risk of developing PTSD (60). Additionally, past studies investigating the epigenetic patterns underlying the disorder have shown that PTSD presents methylation alterations associated with disparate functioning in immune-, stress-, and neurotransmitter-pathways that mediate risk and resilience to PTSD (174). Attempts to quantify these findings on a molecular scale have thus far been met with limited success – wherein the highly polygenic nature of PTSD has rendered it difficult to obtain sample sizes large enough for findings to withstand multiple-testing correction and subsequent replication in independent cohorts (6,12). Furthermore, extant knowledge gaps have been further amplified by the disproportional degree to which previous research has focused on European and military-derived cohorts (49). As such, there exists a dire need for an increase in research attempting to unravel the molecular underpinnings of PTSD across ancestrally diverse populations and within general society.

This study utilised a PRS-based approach to integrate genomic and epigenetic data for the identification of mQTLs associated with PTSD. PRS were constructed using combined European, Latino/Native American and African American summary statistics, derived from the PGC’s largest multi-ethnic GWAS, to predict PTSD case-control status in a local South African population. Although the PRS model was unable to significantly distinguish between PTSD cases and TEC, we isolated 1,444 variants from the most predictive p-value cut-off to serve as an analytical proxy for PTSD risk. Upon integrating the isolated variants and DNA methylation data, the study was able to identify 44,614 mQTLs acting across 250 SNPs and 26,344 CpG probes. Moreover, the study identified evidence of substantial interconnectivity between the discovered mQTLs, wherein CpG sites were found to interact with a median of 2 different variants (IQR = 1 – 2) and each variant was found to interact with a median of 3 CpG probes (IQR = 1 – 10.5). Our results further support the hypothesis that the development of PTSD is dependent on an interconnected network of molecular interactions and highlight the need for future studies dedicated towards optimising PRS construction in multi-ethnic populations.

Moreover, the methods implemented here serve as an important proof-of-concept for genomic studies conducted in resource-limited settings. The increased utility of techniques that allow one to quantify genetic risk is of pivotal importance to populations historically limited by small sample sizes; not only due to the potential elaboration as to how genetic underpinnings contribute to suspected risk - but also due to the translational potential of a representative predictor variable. By addressing statistical power concerns, genetically supported predictor variables may greatly expand the analytical capabilities of similar genomic studies. Such predictor variables are a rapidly growing field, and while predictive accuracy may still need to improve to qualify for routine use in clinical settings, the associated data produced remains highly biologically relevant in its ability to shed light on the underlying nature of complex diseases. Further testing predictive risk estimates both within local and similarly understudied populations would allow one to greatly expand our current understanding of the genetic and molecular mechanisms underlying PTSD. Additionally, refining the implementation of said estimates may potentially lead to the discovery of novel genetic variation relating to PTSD susceptibility in South African populations.

List of Abbreviations

SASHS – South African Stress and Health Survey

PTSD – Posttraumatic Stress Disorder

GWAS – Genome-Wide Association Study

PGC – Psychiatric Genomics Consortium

EWAS – Epigenome-Wide Association Study

SNP – Single Nucleotide Polymorphism

mQTL – Methylation Quantitative Trait Loci

PRS – Polygenic Risk Score

TEC – Trauma-Exposed Controls

SAC – South African Coloured

CAPS-5 – Clinician Administered Posttraumatic Stress Disorder Scale for DSM-5

CTQ – Childhood Trauma Questionnaire

MDD – Major Depressive Disorder

M.I.N.I v6.0 – Mini-International Neuropsychiatric Interview

MetS – Metabolic Syndrome

HDL-C – HDL Cholesterol

STEPS – STEPwise Approach to Surveillance

MAF – Minor Allele Frequency

PCA – Principal Component Analysis

SIS – Sanger Imputation Service

VCF – Variant Call Format

GRCh37 – Genome Reference Consortium Human genome build 37

dbSNP – National Centre for Biotechnology Information's Single Nucleotide Polymorphism Database

FUMA – Functional Mapping and Annotation for Genome-Wide Association Studies

SVA – Surrogate Variable Analysis

ISVA – Independent Surrogate Variable Analysis

DMR – Differentially Methylated Region

eQTL – Expression Quantitative Trait Loci

IQR – Inter Quartile Range

References

1. Atwoli L, Stein DJ, Koenen KC, McLaughlin KA. Epidemiology of posttraumatic stress disorder: Prevalence, correlates and consequences. Curr Opin Psychiatry. 2015;28(4):307–11.

2. Hinsberger M, Holtzhausen L, Sommer J, Kaminer D, Elbert T, Seedat S, et al. Long-term effects of psychotherapy in a context of continuous community and gang violence: Changes in aggressive attitude in high-risk South African adolescents. Behav Cogn Psychother. 2020;

3. Atwoli L, Stein DJ, Williams DR, Mclaughlin KA, Petukhova M, Kessler RC, et al. Trauma and posttraumatic stress disorder in South Africa: Analysis from the South African Stress and Health Study. BMC Psychiatry. 2013;

4. Kirkpatrick HA, Heller GM. Post-Traumatic Stress Disorder: Theory and Treatment Update. Int J Psychiatry Med. 2014;47(4):337–46.

5. Ratanatharathorn A, Boks MP, Maihofer AX, Aiello AE, Amstadter AB, Ashley-Koch AE, et al. Epigenome-wide association of PTSD from heterogeneous cohorts with a common multi-site analysis pipeline. Am J Med Genet Part B Neuropsychiatr Genet. 2017;174(6):619–30.

6. Nievergelt CM, Maihofer AX, Klengel T, Atkinson EG, Chen CY, Choi KW, et al. International meta-analysis of PTSD genome-wide association studies identifies sex- and ancestry-specific genetic risk loci. Nat Commun. 2019;

7. Qi W, Gevonden M, Shalev A. Prevention of Post-Traumatic Stress Disorder After Trauma: Current Evidence and Future Directions. Current Psychiatry Reports. 2016.

8. Koirala R, Søegaard EGI, Ojha SP, Hauff E, Thapa SB. Trauma related psychiatric disorders and their correlates in a clinical sample: A cross-sectional study in trauma affected patients visiting a psychiatric clinic in Nepal. PLoS One. 2020;

9. Fenster RJ, Lebois LAM, Ressler KJ, Suh J. Brain circuit dysfunction in post-traumatic stress disorder: from mouse to man. Nature Reviews Neuroscience. 2018.

10. Karam EG, Friedman MJ, Hill ED, Kessler RC, McLaughlin KA, Petukhova M, et al. Cumulative traumas and risk thresholds: 12-month ptsd in the world mental health (WMH) surveys. Depress Anxiety. 2014;

11. Wilker S, Schneider A, Conrad D, Pfeiffer A, Boeck C, Lingenfelder B, et al. Genetic variation is associated with PTSD risk and aversive memory: Evidence from two trauma-Exposed African samples and one healthy European sample. Transl Psychiatry. 2018;

12. Smith AK, Ratanatharathorn A, Maihofer AX, Naviaux RK, Aiello AE, Amstadter AB, et al. Epigenome-wide meta-analysis of PTSD across 10 military and civilian cohorts identifies methylation changes in AHRR. Nat Commun. 2020;

13. Nievergelt CM, Ashley-Koch AE, Dalvie S, Hauser MA, Morey RA, Smith AK, et al. Genomic Approaches to Posttraumatic Stress Disorder: The Psychiatric Genomic Consortium Initiative. Biol Psychiatry [Internet]. 2018;83(10):831–9. Available from: https://doi.org/10.1016/j.biopsych.2018.01.020

14. Misganaw B, Guffanti G, Lori A, Abu-Amara D, Flory JD, Hammamieh R, et al. Polygenic risk associated with post-traumatic stress disorder onset and severity. Transl Psychiatry. 2019;

15. Logue MW, Baldwin C, Guffanti G, Melista E, Wolf EJ, Reardon AF, et al. A genome-wide association study of post-traumatic stress disorder identifies the retinoid-related orphan receptor alpha (RORA) gene as a significant risk locus. Mol Psychiatry. 2013;

16. Xie P, Kranzler HR, Yang C, Zhao H, Farrer LA, Gelernter J. Genome-wide association study identifies new susceptibility loci for posttraumatic stress disorder. Biol Psychiatry. 2013;

17. Guffanti G, Galea S, Yan L, Roberts AL, Solovieff N, Aiello AE, et al. Genome-wide association study implicates a novel RNA gene, the lincRNA AC068718.1, as a risk factor for post-traumatic stress disorder in women. Psychoneuroendocrinology. 2013;

18. Wolf EJ, Rasmusson AM, Mitchell KS, Logue MW, Baldwin CT, Miller MW. A genome-wide association study of clinical symptoms of dissociation in a trauma-exposed sample. Depress Anxiety. 2014;

19. Nievergelt CM, Maihofer AX, Mustapic M, Yurgil KA, Schork NJ, Miller MW, et al. Genomic predictors of combat stress vulnerability and resilience in U.S. Marines: A genome-wide association study across multiple ancestries implicates PRTFDC1 as a potential PTSD gene. Psychoneuroendocrinology. 2015;

20. Almli LM, Stevens JS, Smith AK, Kilaru V, Meng Q, Flory J, et al. A genome-wide identified risk variant for PTSD is a methylation quantitative trait locus and confers decreased cortical activation to fearful faces. Am J Med Genet Part B Neuropsychiatr Genet. 2015;168(5):327–36.

21. Ashley-Koch AE, Garrett ME, Gibson J, Liu Y, Dennis MF, Kimbrel NA, et al. Genome-wide association study of posttraumatic stress disorder in a cohort of Iraq-Afghanistan era veterans. J Affect Disord. 2015;

22. Stein MB, Chen CY, Ursano RJ, Cai T, Gelernter J, Heeringa SG, et al. Genome-wide association studies of posttraumatic stress disorder in 2 cohorts of US army soldiers. JAMA Psychiatry. 2016;

23. Duncan LE, Ratanatharathorn A, Aiello AE, Almli LM, Amstadter AB, Ashley-Koch AE, et al. Largest GWAS of PTSD (N=20 070) yields genetic overlap with schizophrenia and sex differences in heritability. Mol Psychiatry. 2018;

24. van der Merwe C, Jahanshad N, Cheung JW, Mufford M, Groenewold NA, Koen N, et al. Concordance of genetic variation that increases risk for anxiety disorders and posttraumatic stress disorders and that influences their underlying neurocircuitry. J Affect Disord. 2019;

25. Shen H, Gelaye B, Huang H, Rondon MB, Sanchez S, Duncan LE. Polygenic prediction and GWAS of depression, PTSD, and suicidal ideation/self-harm in a Peruvian cohort. Neuropsychopharmacology. 2020;

26. Oosthuizen F, Wegener G, Harvey BH. Nitric oxide as inflammatory mediator in post-traumatic stress disorder (PTSD): evidence from an animal model. Neuropsychiatr Dis Treat. 2005;

27. Tamura G, Olson D, Miron J, Clark TG. Tolloid-like 1 is negatively regulated by stress and glucocorticoids. Mol Brain Res. 2005;

28. Vitureira N, Andrés R, Pérez-Martínez E, Martínez A, Bribián A, Blasi J, et al. Podocalyxin is a novel

polysialylated neural adhesion protein with multiple roles in neural development and synapse formation. PLoS One. 2010;

29. Gemechu JM, Sharma A, Yu D, Xie Y, Merkel OM, Moszczynska A. Characterization of Dopaminergic System in the Striatum of Young Adult Park2 -/- Knockout Rats. Sci Rep. 2018;

30. Kalisch R, Gerlicher AMV, Duvarci S. A Dopaminergic Basis for Fear Extinction. Trends in Cognitive Sciences. 2019.

31. Liberzon I, King AP, Ressler KJ, Almli LM, Zhang P, Ma ST, et al. Interaction of the ADRB2 gene polymorphism with childhood trauma in predicting adult symptoms of posttraumatic stress disorder. JAMA Psychiatry. 2014;

32. Berardis D, Marini S, Serroni N, Iasevoli F, Tomasetti C, Bartolomeis A, et al. Targeting the Noradrenergic System in Posttraumatic Stress Disorder: A Systematic Review and Meta-Analysis of Prazosin Trials. Curr Drug Targets. 2015;

33. Mellon SH, Gautam A, Hammamieh R, Jett M, Wolkowitz OM. Metabolism, Metabolomics, and Inflammation in Posttraumatic Stress Disorder. Biological Psychiatry. 2018.

34. Wang Z, Caughron B, Young MRI. Posttraumatic stress disorder: An immunological disorder? Frontiers in Psychiatry. 2017.

35. Dey BK, Mueller AC, Dutta A. Long non-coding rnas as emerging regulators of differentiation, development, and disease. Transcription. 2014.

36. Gunaratne PH, Creighton CJ, Watson M, Tennakoon JB. Large-scale integration of MicroRNA and gene expression data for identification of enriched microRNA-mRNA associations in biological systems. Methods Mol Biol. 2010;

37. Youssef NA, Lockwood L, Su S, Hao G, Rutten BPF. The effects of trauma, with or without PTSD, on the transgenerational DNA methylation alterations in human offsprings. Brain Sciences. 2018.

38. Rusconi F, Battaglioli E. Acute stress-induced epigenetic modulations and their potential protective role toward depression. Front Mol Neurosci. 2018;

39. Chatterjee N, Gim J, Choi J. Epigenetic profiling to environmental stressors in model and non-model organisms: Ecotoxicology perspective. Environ Health Toxicol. 2018;

40. Martin C, Cho YE, Kim H, Yun S, Kanefsky R, Lee H, et al. Altered DNA Methylation Patterns Associated With Clinically Relevant Increases in PTSD Symptoms and PTSD Symptom Profiles in Military Personnel. Biol Res Nurs. 2018;

41. Roberts AL, Gladish N, Gatev E, Jones MJ, Chen Y, MacIsaac JL, et al. Exposure to childhood abuse is associated with human sperm DNA methylation. Translational Psychiatry. 2018.

42. Smith AK, Conneely KN, Kilaru V, Mercer KB, Weiss TE, Bradley B, et al. Differential immune system DNA methylation and cytokine regulation in post-traumatic stress disorder. Am J Med Genet Part B Neuropsychiatr Genet. 2011;

43. Rutten BPF, Vermetten E, Vinkers CH, Ursini G, Daskalakis NP, Pishva E, et al. Longitudinal analyses of the DNA methylome in deployed military servicemen identify susceptibility loci for post-traumatic stress disorder. Mol Psychiatry. 2018;

44. Kuan PF, Waszczuk MA, Kotov R, Marsit CJ, Guffanti G, Gonzalez A, et al. An epigenome-wide DNA methylation study of PTSD and depression in World Trade Center responders. Transl Psychiatry. 2017;

45. Mehta D, Bruenig D, Carrillo-Roa T, Lawford B, Harvey W, Morris CP, et al. Genomewide DNA methylation analysis in combat veterans reveals a novel locus for PTSD. Acta Psychiatr Scand. 2017;

46. Uddin M, Ratanatharathorn A, Armstrong D, Kuan PF, Aiello AE, Bromet EJ, et al. Epigenetic meta-analysis across three civilian cohorts identifies NRG1 and HGS as blood-based biomarkers for post-traumatic stress disorder. Epigenomics. 2018;

47. Mehta D, Pelzer ES, Bruenig D, Lawford B, McLeay S, Morris CP, et al. DNA methylation from germline cells in veterans with PTSD. J Psychiatr Res. 2019;

48. Snijders C, Maihofer AX, Ratanatharathorn A, Baker DG, Boks MP, Geuze E, et al. Longitudinal epigenome-wide association studies of three male military cohorts reveal multiple CpG sites associated with post-traumatic stress disorder. Clin Epigenetics. 2020;

49. Logue MW, Miller MW, Wolf EJ, Huber BR, Morrison FG, Zhou Z, et al. An epigenome-wide association study of posttraumatic stress disorder in US veterans implicates several new DNA methylation loci. Clin Epigenetics. 2020;

50. Katrinli S, Zheng Y, Gautam A, Hammamieh R, Yang R, Venkateswaran S, et al. PTSD is associated with increased DNA methylation across regions of HLA-DPB1 and SPATC1L. Brain Behav Immun. 2021;

51. An N, Bassil K, Al Jowf GI, Steinbusch HWM, Rothermel M, de Nijs L, et al. Dual-specificity phosphatases in mental and neurological disorders. Progress in Neurobiology. 2020.

52. Sayad A, Ghafouri-Fard S, Omrani MD, Taheri M. Associations Between Two Single-Nucleotide Polymorphisms in NINJ2 Gene and Risk of Psychiatric Disorders. J Mol Neurosci. 2020;

53. Krzyzewska IM, Ensink JBM, Nawijn L, Mul AN, Koch SB, Venema A, et al. Genetic variant in CACNA1C is associated with PTSD in traumatized police officers. Eur J Hum Genet. 2018;

54. Feng DY, Guo BL, Liu GH, Xu K, Yang J, Tao K, et al. Nerve growth factor against PTSD symptoms: Preventing the impaired hippocampal cytoarchitectures. Prog Neurobiol. 2020;

55. Lane JM, Liang J, Vlasac I, Anderson SG, Bechtold DA, Bowden J, et al. Genome-wide association analyses of sleep disturbance traits identify new loci and highlight shared genetics with neuropsychiatric and metabolic traits. Nat Genet. 2017;

56. Dallacasagrande V, Hajjar KA. Annexin A2 in Inflammation and Host Defense. Cells. 2020.

57. Cueto FJ, del Fresno C, Sancho D. DNGR-1, a Dendritic Cell-Specific Sensor of Tissue Damage That Dually Modulates Immunity and Inflammation. Frontiers in Immunology. 2020.

58. Moen SH, Ehrnström B, Kojen JF, Yurchenko M, Beckwith KS, Afset JE, et al. Human Toll-like receptor 8

(TLR8) is an important sensor of pyogenic bacteria, and is attenuated by cell surface TLR signaling. Front Immunol. 2019;

59. Yamagata M, Sanes JR. Expression and roles of the immunoglobulin superfamily recognition molecule sidekick1 in mouse retina. Front Mol Neurosci. 2019;

60. Afifi TO, Asmundson GJG, Taylor S, Jang KL. The role of genes and environment on trauma exposure and posttraumatic stress disorder symptoms: A review of twin studies. Clin Psychol Rev. 2010;30(1):101–12.

61. Dean KR, Hammamieh R, Mellon SH, Abu-Amara D, Flory JD, Guffanti G, et al. Multi-omic biomarker identification and validation for diagnosing warzone-related post-traumatic stress disorder. Mol Psychiatry. 2019;

62. Heckmann BL, Zhang X, Xie X, Liu J. The G0/G1 switch gene 2 (G0S2): Regulating metabolism and beyond. Biochimica et Biophysica Acta - Molecular and Cell Biology of Lipids. 2013.

63. Daskalakis NP, Cohen H, Cai G, Buxbaum JD, Yehuda R. Expression profiling associates blood and brain glucocorticoid receptor signaling with trauma-related individual differences in both sexes. Proc Natl Acad Sci U S A. 2014;

64. Bam M, Yang X, Zumbrun EE, Zhong Y, Zhou J, Ginsberg JP, et al. Dysregulated immune system networks in war veterans with PTSD is an outcome of altered miRNA expression and DNA methylation. Sci Rep. 2016;

65. Maddox SA, Kilaru V, Shin J, Jovanovic T, Almli LM, Dias BG, et al. Estrogen-dependent association of HDAC4 with fear in female mice and women with PTSD. Mol Psychiatry. 2018;

66. Wang Z, Qin G, Zhao TC. HDAC4: Mechanism of regulation and biological functions. Epigenomics. 2014.

67. Smith AK, Kilaru V, Kocak M, Almli LM, Mercer KB, Ressler KJ, et al. Methylation quantitative trait loci (meQTLs) are consistently detected across ancestry, developmental stage, and tissue type. BMC Genomics. 2014;15(1).

68. Lambert SA, Abraham G, Inouye M. Towards clinical utility of polygenic risk scores. Human Molecular Genetics. 2019.

69. Euesden J, Lewis CM, O’Reilly PF. PRSice: Polygenic Risk Score software. Bioinformatics. 2015;

70. Choi SW, Mak TSH, O’Reilly PF. A guide to performing Polygenic Risk Score analyses. bioRxiv. 2018.

71. Schür RR, Schijven D, Boks MP, Rutten BPF, Stein MB, Veldink JH, et al. The effect of genetic vulnerability and military deployment on the development of post-traumatic stress disorder and depressive symptoms. Eur Neuropsychopharmacol. 2019;

72. Waszczuk MA, Docherty AR, Shabalin AA, Miao J, Yang X, Kuan PF, et al. Polygenic prediction of PTSD trajectories in 9/11 responders. Psychol Med. 2020;

73. Weathers FW, Bovin MJ, Lee DJ, Sloan DM, Schnurr PP, Kaloupek DG, et al. The clinician-administered ptsd scale for DSM-5 (CAPS-5): Development and initial psychometric evaluation in military veterans. Psychol Assess. 2018;30(3):383–95.

74. Bernstein DP, Ahluvalia T, Pogge D, Handelsman L. Validity of the childhood trauma questionnaire in an adolescent psychiatric population. J Am Acad Child Adolesc Psychiatry. 1997;

75. Spies G, Ahmed-Leitao F, Fennema-Notestine C, Cherner M, Seedat S. Effects of HIV and childhood trauma on brain morphometry and neurocognitive function. J Neurovirol. 2016;

76. Liebschutz JM, Buchanan-Howland K, Chen CA, Frank DA, Richardson MA, Heeren TC, et al. Childhood Trauma Questionnaire (CTQ) Correlations with prospective violence assessment in a longitudinal cohort. Psychol Assess. 2018;

77. Sheehan D V., Lecrubier Y, Sheehan KH, Amorim P, Janavs J, Weiller E, et al. The Mini-International Neuropsychiatric Interview (M.I.N.I.): The development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. In: Journal of Clinical Psychiatry. 1998.

78. Pettersson E, Larsson H, Lichtenstein P. Common psychiatric disorders share the same genetic origin: A multivariate sibling study of the Swedish population. Mol Psychiatry. 2016;

79. Alberti KGMM, Eckel RH, Grundy SM, Zimmet PZ, Cleeman JI, Donato KA, et al. Harmonizing the Metabolic Syndrome. Circulation. 2009;

80. Matsha TE, Hassan MS, Hon GM, Soita DJ, Kengne AP, Erasmus RT. Derivation and validation of a waist circumference optimal cutoff for diagnosing metabolic syndrome in a South African mixed ancestry population. Int J Cardiol. 2013;

81. Riley L, Guthold R, Cowan M, Savin S, Bhatti L, Armstrong T, et al. The world health organization STEPwise approach to noncommunicable disease risk-factor surveillance: Methods, challenges, and opportunities. Am J Public Health. 2016;

82. Bien SA, Wojcik GL, Zubair N, Gignoux CR, Martin AR, Kocarnik JM, et al. Strategies for enriching variant coverage in candidate disease loci on a multiethnic genotyping array. PLoS One. 2016;

83. Moran S, Arribas C, Esteller M. Validation of a DNA methylation microarray for 850,000 CpG sites of the human genome enriched in enhancer sequences. Epigenomics. 2016;

84. Anderson CA, Pettersson FH, Clarke GM, Cardon LR, Morris AP, Zondervan KT. Data quality control in genetic case-control association studies. Nat Protoc. 2010;

85. Coleman JRI, Euesden J, Patel H, Folarin AA, Newhouse S, Breen G. Quality control, imputation and analysis of genome-wide genotyping data from the Illumina HumanCoreExome microarray. Brief Funct Genomics. 2016;

86. Schurz H, Müller SJ, Van Helden PD, Tromp G, Hoal EG, Kinnear CJ, et al. Evaluating the accuracy of imputation methods in a five-way admixed population. Front Genet. 2019;

87. Marees AT, de Kluiver H, Stringer S, Vorspan F, Curis E, Marie-Claire C, et al. A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. Int J Methods Psychiatr Res. 2018;

88. Turner S, Armstrong LL, Bradford Y, Carlsony CS, Crawford DC, Crenshaw AT, et al. Quality control procedures for genome-wide association studies. Curr Protoc Hum Genet. 2011;

89. Sul JH, Martin LS, Eskin E. Population structure in genetic studies: Confounding factors and mixed models. PLoS Genetics. 2018.

90. Ma S, Shi G. On rare variants in principal component analysis of population stratification. BMC Genet. 2020;

91. Byun J, Han Y, Gorlov IP, Busam JA, Seldin MF, Amos CI. Ancestry inference using principal component analysis and spatial analysis: A distance-based analysis to account for population substructure. BMC Genomics. 2017;

92. Daya M, Der Merwe L, Galal U, Möller M, Salie M, Chimusa ER, et al. A panel of ancestry informative markers for the complex five-way admixed South African Coloured population. PLoS One. 2013;

93. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Bentley DR, Chakravarti A, et al. A global reference for human genetic variation. Nature. 2015.

94. Uren C, Kim M, Martin AR, Bobo D, Gignoux CR, Van Helden PD, et al. Fine-scale human population structure in Southern Africa reflects ecogeographic boundaries. Genetics. 2016;

95. McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, Teumer A, et al. A reference panel of 64,976 haplotypes for genotype imputation. Nat Genet. 2016;

96. Shi S, Yuan N, Yang M, Du Z, Wang J, Sheng X, et al. Comprehensive Assessment of Genotype Imputation Performance. Hum Hered. 2019;

97. Slatkin M. Linkage disequilibrium - Understanding the evolutionary past and mapping the medical future. Nature Reviews Genetics. 2008.

98. Subramanian H, Gatenby RA. Evolutionary advantage of anti-parallel strand orientation of duplex DNA. Sci Rep. 2020;

99. Nietlisbach P, Keller LF, Postma E. Genetic variance components and heritability of multiallelic heterozygosity under inbreeding. Heredity (Edinb). 2016;

100. Campbell IM, Gambin T, Jhangiani SN, Grove ML, Veeraraghavan N, Muzny DM, et al. Multiallelic Positions in the Human Genome: Challenges for Genetic Analyses. Hum Mutat. 2016;

101. Yates AD, Achuthan P, Akanni W, Allen J, Allen J, Alvarez-Jarreta J, et al. Ensembl 2020. Nucleic Acids Res. 2020;

102. Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat Genet. 2012;

103. Delaneau O, Marchini J, Zagury JF. A linear complexity phasing method for thousands of genomes. Nat Methods. 2012;

104. Durbin R. Efficient haplotype matching and storage using the positional Burrows-Wheeler transform (PBWT). Bioinformatics. 2014;

105. Grabowska AD, Lacerda EM, Nacul L, Sepúlveda N. Review of the Quality Control Checks Performed by Current Genome-Wide and Targeted-Genome Association Studies on Myalgic Encephalomyelitis/Chronic

Fatigue Syndrome. Frontiers in Pediatrics. 2020.

106. Ellingson SR, Fardo DW. Automated quality control for genome wide association studies. F1000Research. 2016;

107. Alag S. Unique insights from ClinicalTrials.gov by mining protein mutations and RSids in addition to applying the Human Phenotype Ontology. PLoS One. 2020;

108. Watanabe K, Taskesen E, Van Bochoven A, Posthuma D. Functional mapping and annotation of genetic associations with FUMA. Nat Commun. 2017;

109. Min JL, Hemani G, Davey Smith G, Relton C, Suderman M. Meffil: efficient normalization and analysis of very large DNA methylation datasets. Bioinformatics. 2018;

110. Toikumo S, Malan-Müller S, Swart PC, Womersley JS, van den Heuvel L, Tromp G, et al. Epigenome-wide association study on PTSD diagnosis and Metabolic Syndrome in a South African population. Stellenbosch University; 2019.

111. Pidsley R, Zotenko E, Peters TJ, Lawrence MG, Risbridger GP, Molloy P, et al. Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. Genome Biol. 2016;

112. Bacolod MD, Das SK, Sokhi UK, Bradley S, Fenstermacher DA, Pellecchia M, et al. Examination of Epigenetic and other Molecular Factors Associated with mda-9/Syntenin Dysregulation in Cancer Through Integrated Analyses of Public Genomic Datasets. In: Advances in Cancer Research. 2015.

113. Heiss JA, Just AC. Identifying mislabeled and contaminated DNA methylation microarray data: An extended quality control toolset with examples from GEO. Clin Epigenetics. 2018;

114. Liu J, Siegmund KD. An evaluation of processing methods for HumanMethylation450 BeadChip data. BMC Genomics. 2016;

115. Samblas M, Milagro FI, Martínez A. DNA methylation markers in obesity, metabolic syndrome, and weight loss. Epigenetics. 2019.

116. Suderman M, Borghol N, Pappas JJ, Pinto Pereira SM, Pembrey M, Hertzman C, et al. Childhood abuse is associated with methylation of multiple loci in adult DNA. BMC Med Genomics. 2014;

117. Brückmann C, Islam SA, MacIsaac JL, Morin AM, Karle KN, DI Santo A, et al. DNA methylation signatures of chronic alcohol dependence in purified CD3+ T-cells of patients undergoing alcohol treatment. Sci Rep. 2017;

118. Tsai PC, Glastonbury CA, Eliot MN, Bollepalli S, Yet I, Castillo-Fernandez JE, et al. Smoking induces coordinated DNA methylation and gene expression changes in adipose tissue with consequences for metabolic health 06 Biological Sciences 0604 Genetics. Clin Epigenetics. 2018;

119. Lihua M, Tao Z, Hongbin M, Hui W, Caihong J, Xiaolian J. Metabolic syndrome risk in relation to posttraumatic stress disorder among trauma-exposed civilians in Gansu Province, China. Med (United States). 2020;

120. Michopoulos V, Vester A, Neigh G. Posttraumatic stress disorder: A metabolic disorder in disguise? Experimental Neurology. 2016.

121. McLaughlin KA, Koenen KC, Bromet EJ, Karam EG, Liu H, Petukhova M, et al. Childhood adversities and post-traumatic stress disorder: Evidence for stress sensitisation in the World Mental Health Surveys. British Journal of Psychiatry. 2017.

122. Subbie-Saenz de Viteri S, Pandey A, Pandey G, Kamarajan C, Smith R, Anokhin A, et al. Pathways to post-traumatic stress disorder and alcohol dependence: Trauma, executive functioning, and family history of alcoholism in adolescents and young adults. Brain Behav. 2020;

123. Kearns NT, Carl E, Stein AT, Vujanovic AA, Zvolensky MJ, Smits JAJ, et al. Posttraumatic stress disorder and cigarette smoking: A systematic review. Depression and Anxiety. 2018.

124. Rahmani E, Schweiger R, Shenhav L, Wingert T, Hofer I, Gabel E, et al. BayesCCE: a Bayesian framework for estimating cell-type composition from DNA methylation without the need for methylation reference. Genome Biol. 2018;

125. Hicks SC, Irizarry RA. MethylCC: Technology-independent estimation of cell type composition using differentially methylated regions. Genome Biol. 2019;

126. Leek JT, Storey JD. Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genet. 2007;

127. Teschendorff AE, Zhuang J, Widschwendter M. Independent surrogate variable analysis to deconvolve confounding factors in large-scale microarray profiling studies. Bioinformatics. 2011;

128. Phipson B, Maksimovic J, Oshlack A. MissMethyl: An R package for analyzing data from Illumina's HumanMethylation450 platform. Bioinformatics. 2016;

129. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;

130. Suderman M, Staley JR, French R, Arathimos R, Simpkin A, Tilling K. Dmrff: Identifying differentially methylated regions efficiently with power and control. bioRxiv. 2018.

131. Shabalin AA. Matrix eQTL: Ultra fast eQTL analysis via large matrix operations. Bioinformatics. 2012;

132. Hannon E, Gorrie-Stone TJ, Smart MC, Burrage J, Hughes A, Bao Y, et al. Leveraging DNA-Methylation Quantitative-Trait Loci to Characterize the Relationship between Methylomic Variation, Gene Expression, and Complex Traits. Am J Hum Genet. 2018;

133. Volkov P, Olsson AH, Gillberg L, Jørgensen SW, Brøns C, Eriksson KF, et al. A genome-wide mQTL analysis in human adipose tissue identifies genetic variants associated with DNA methylation, gene expression and metabolic traits. PLoS One. 2016;

134. Giambartolomei C, Vukcevic D, Schadt EE, Franke L, Hingorani AD, Wallace C, et al. Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. PLoS Genet. 2014;

135. Rasmusson AM, Pineles SL. Neurotransmitter, Peptide, and Steroid Hormone Abnormalities in PTSD:

Biological Endophenotypes Relevant to Treatment. Current Psychiatry Reports. 2018.

136. Sherin JE, Nemeroff CB. Post-traumatic stress disorder: The neurobiological impact of psychological trauma. Dialogues Clin Neurosci. 2011;

137. Pitman RK, Rasmusson AM, Koenen KC, Shin LM, Orr SP, Gilbertson MW, et al. Biological studies of post-traumatic stress disorder. Nature Reviews Neuroscience. 2012.

138. Amariuta T, Ishigaki K, Sugishita H, Ohta T, Koido M, Dey KK, et al. Improving the trans-ancestry portability of polygenic risk scores by prioritizing variants in predicted cell-type-specific regulatory elements. Nat Genet. 2020;

139. Martin AR, Gignoux CR, Walters RK, Wojcik GL, Neale BM, Gravel S, et al. Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. Am J Hum Genet [Internet]. 2017;100(4):635–49. Available from: http://dx.doi.org/10.1016/j.ajhg.2017.03.004

140. Márquez-Luna C, Loh PR, Price AL. Multiethnic polygenic risk scores improve risk prediction in diverse populations. Genet Epidemiol. 2017;41(8):811–23.

141. Grinde KE, Qi Q, Thornton TA, Liu S, Shadyab AH, Chan KHK, et al. Generalizing polygenic risk scores from Europeans to Hispanics/Latinos. Genet Epidemiol. 2019;43(1):50–62.

142. Hu Y, Lu Q, Liu W, Zhang Y, Li M, Zhao H. Joint modeling of genetically correlated diseases and functional annotations increases accuracy of polygenic risk prediction. PLoS Genet. 2017;13(6):1–22.

143. Hu Y, Lu Q, Powles R, Yao X, Yang C, Fang F, et al. Leveraging functional annotations in genetic risk prediction for human complex diseases. PLoS Comput Biol. 2017;13(6):1–16.

144. Bethesda (MD): National Library of Medicine (US) NC for BI. C4orf36 chromosome 4 open reading frame 36 [Homo sapiens (human)] [Internet]. [cited 2021 Feb 6]. Available from: https://www.ncbi.nlm.nih.gov/gene/132989

145. Bethesda (MD): National Library of Medicine (US) NC for BI. SDCCAG8 SHH signaling and ciliogenesis regulator SDCCAG8 [Homo sapiens (human)] [Internet]. [cited 2021 Feb 6]. Available from: https://www.ncbi.nlm.nih.gov/gene/10806

146. Bethesda (MD): National Library of Medicine (US) NC for BI. AKT3 AKT serine/threonine kinase 3 [Homo sapiens (human)] [Internet]. [cited 2021 Feb 6]. Available from: https://www.ncbi.nlm.nih.gov/gene/10000

147. Bethesda (MD): National Library of Medicine (US) NC for BI. TEDC1 tubulin epsilon and delta complex 1 [Homo sapiens (human)] [Internet]. [cited 2021 Feb 6]. Available from: https://www.ncbi.nlm.nih.gov/gene/283643

148. Miao X-R, Chen Q-B, Wei K, Tao K-M, Lu Z-J. Posttraumatic stress disorder: from diagnosis to prevention. Mil Med Res. 2018;

149. Losenkov IS, Vyalova NM, Simutkin GG, Bokhan NA, Ivanova SA. An association of AKT1 gene polymorphism with antidepressant treatment response. World J Biol Psychiatry. 2016;

150. Beaulieu JM. A role for Akt and glycogen synthase kinase-3 as integrators of dopamine and serotonin

neurotransmission in mental health. Journal of Psychiatry and Neuroscience. 2012.

151. Zhang K, Qu S, Chang S, Li G, Cao C, Fang K, et al. An overview of posttraumatic stress disorder genetic studies by analyzing and integrating genetic data into genetic database PTSDgene. Neuroscience and Biobehavioral Reviews. 2017.

152. Joehanes R, Just AC, Marioni RE, Pilling LC, Reynolds LM, Mandaviya PR, et al. Epigenetic Signatures of Cigarette Smoking. Circ Cardiovasc Genet. 2016;

153. Bethesda (MD): National Library of Medicine (US) NC for BI. rs144798302 [Internet]. [cited 2021 Feb 7]. Available from: https://www.ncbi.nlm.nih.gov/snp/rs144798302

154. Anastasiadi D, Esteve-Codina A, Piferrer F. Consistent inverse correlation between DNA methylation of the first intron and gene expression across tissues and species. Epigenetics and Chromatin. 2018;

155. Deng GX, Xu N, Huang Q, Tan JY, Zhang Z, Li XF, et al. Association between promoter DNA methylation and gene expression in the pathogenesis of ischemic stroke. Aging (Albany NY). 2019;

156. Brown AJ, James DC. Constructing strong cell type-specific promoters through informed design. In: Methods in Molecular Biology. 2017.

157. Scott CA, Duryea JD, MacKay H, Baker MS, Laritsky E, Gunasekara CJ, et al. Identification of cell type-specific methylation signals in bulk whole genome bisulfite sequencing data. Genome Biol. 2020;

158. Walton E, Hass J, Liu J, Roffman JL, Bernardoni F, Roessner V, et al. Correspondence of DNA methylation between blood and brain tissue and its application to schizophrenia research. Schizophr Bull. 2016;

159. Hannon E, Lunnon K, Schalkwyk L, Mill J. Interindividual methylomic variation across blood, cortex, and cerebellum: Implications for epigenetic studies of neurological and neuropsychiatric phenotypes. Epigenetics. 2015;

160. Komaki S, Shiwa Y, Furukawa R, Hachiya T, Ohmomo H, Otomo R, et al. IMETHYL: An integrative database of human DNA methylation, gene expression, and genomic variation. Hum Genome Var. 2018;

161. Hachiya T, Furukawa R, Shiwa Y, Ohmomo H, Ono K, Katsuoka F, et al. Genome-wide identification of inter-individually variable DNA methylation sites improves the efficacy of epigenetic association studies. npj Genomic Med. 2017;

162. McLaughlin KA, Lambert HK. Child trauma exposure and psychopathology: mechanisms of risk and resilience. Current Opinion in Psychology. 2017.

163. Barbano AC, van der Mei WF, deRoon-Cassini TA, Grauer E, Lowe SR, Matsuoka YJ, et al. Differentiating PTSD from anxiety and depression: Lessons from the ICD-11 PTSD diagnostic criteria. Depress Anxiety. 2019;

164. Chu DA, Bryant RA, Gatt JM, Harris AWF. Cumulative childhood interpersonal trauma is associated with reduced cortical differentiation between threat and non-threat faces in posttraumatic stress disorder adults. Aust N Z J Psychiatry. 2019;

165. Contractor AA, Greene T, Dolan M, Elhai JD. Relations between PTSD and depression symptom clusters in samples differentiated by PTSD diagnostic status. J Anxiety Disord. 2018;

166. Dudbridge F. Power and Predictive Accuracy of Polygenic Risk Scores. PLoS Genet. 2013;

167. Radhakrishnan K, Aslan M, Harrington KM, Pietrzak RH, Huang G, Muralidhar S, et al. Genomics of posttraumatic stress disorder in veterans: Methods and rationale for Veterans Affairs Cooperative Study #575B. Int J Methods Psychiatr Res. 2019;

168. Wu C, Dewan A, Hoh J, Wang Z. A Comparison of Association Methods Correcting for Population Stratification in Case-Control Studies. Ann Hum Genet. 2011;

169. Liu CC, Shringarpure S, Lange K, Novembre J. Exploring population structure with admixture models and principal component analysis. In: Methods in Molecular Biology. 2020.

170. Alexander DH, Lange K. Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. BMC Bioinformatics. 2011;

171. Gaspar HA, Breen G. Probabilistic ancestry maps: A method to assess and visualize population substructures in genetics. BMC Bioinformatics. 2019;

172. Peterson RE, Kuchenbaecker K, Walters RK, Chen CY, Popejoy AB, Periyasamy S, et al. Genome-wide Association Studies in Ancestrally Diverse Populations: Opportunities, Methods, Pitfalls, and Recommendations. Cell. 2019.

173. Bharadwaj RA, Jaffe AE, Chen Q, Deep-Soboslay A, Goldman AL, Mighdoll MI, et al. Genetic risk mechanisms of posttraumatic stress disorder in the human brain. J Neurosci Res. 2018;96(1):21–30.

174. Daskalakis NP, Rijal CM, King C, Huckins LM, Ressler KJ. Recent Genetics and Epigenetics Approaches to PTSD. Curr Psychiatry Rep. 2018;20(5).

Supplementary Tables & Figures

Supplementary Table 1: Regressing the first twenty principal components against PTSD status

Variable	Estimate	Std Error	t	p
PC 1	0.094	0.502	0.186	0.852
PC 2	0.888	0.502	1.768	0.078*
PC 3	-0.536	0.502	-1.067	0.287
PC 4	0.099	0.502	0.197	0.844
PC 5	-0.826	0.502	-1.645	0.101
PC 6	-0.375	0.502	-0.747	0.456
PC 7	0.290	0.502	0.578	0.564
PC 8	0.375	0.502	0.746	0.456
PC 9	0.286	0.502	0.569	0.570
PC 10	-0.402	0.502	-0.800	0.425
PC 11	0.399	0.502	0.794	0.428
PC 12	-0.417	0.502	-0.830	0.407
PC 13	0.169	0.502	0.336	0.737
PC 14	0.009	0.502	0.018	0.985
PC 15	-0.544	0.502	-1.083	0.280
PC 16	-0.165	0.502	-0.328	0.743
PC 17	0.267	0.502	0.532	0.595
PC 18	0.891	0.502	1.775	0.077*
PC 19	0.738	0.502	1.470	0.143
PC 20	0.194	0.502	0.387	0.699

** Due to their noticeable separation against the field, the second and eighteenth principal components were employed as genomic covariates to account for the effects of population stratification.*

Supplementary Table 2: Annotation of notable SNPs identified through genome-wide association testing

rsID	Chromosome	Position	Nearest Gene	p
rs6534683	4	129382308	PRPF31	9.859e-6
rs382260	5	112210344	SRP19	9.743e-6
rs28493191	21	22094730	LINC00320	8.993e-6
rs34997358	21	22088371	LINC00320	7.904e-6
rs2826490	21	22098294	LINC00320	5.865e-6
rs1871923	21	22098717	LINC00320	5.865e-6
rs9980899	21	22099354	LINC00320	5.865e-6
rs9984307	21	22095495	LINC00320	3.985e-6

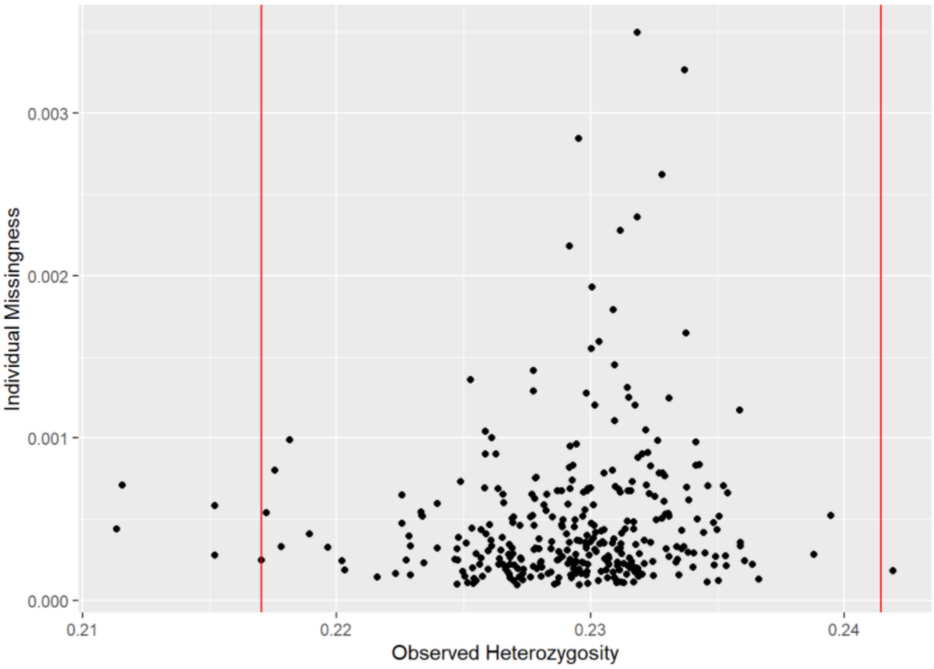
** Nearest gene determined through FUMA (Functional Mapping and Annotation for Genome-Wide Association Studies).*

Supplementary Table 3: Detection of differentially methylated regions through *dmrff*

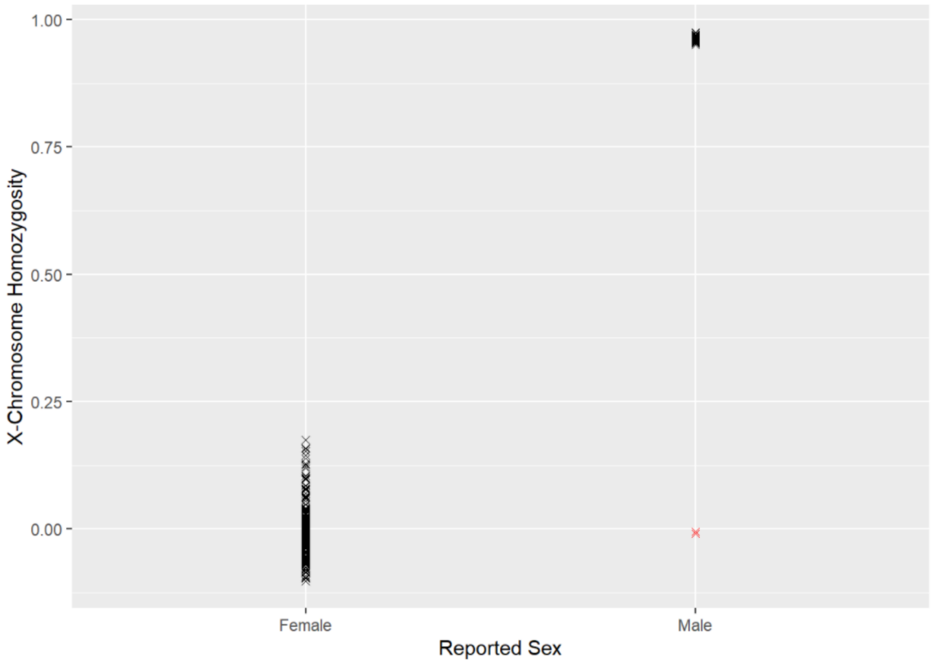
Chromosome	Start	End	Coefficient	p
4	53588360	53588374	-0.038	4.038e-9

** p-value threshold for epigenome-wide significance = p ≤ 5.95e-8*

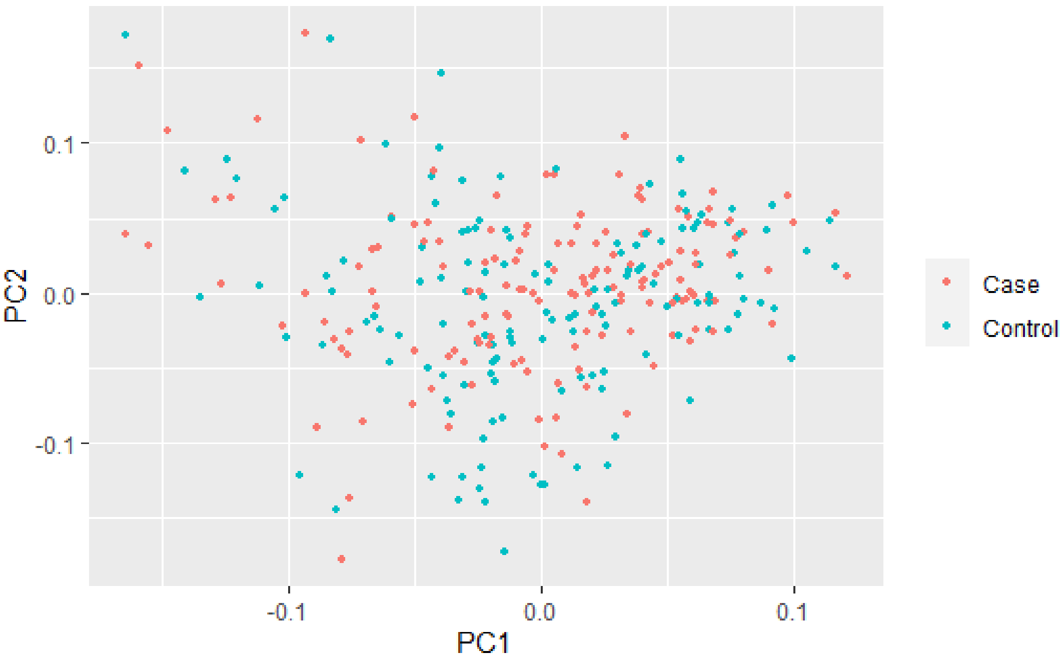
*** - or + coefficient values are indicative of relative hypo- and hyper-methylation, respectively.*



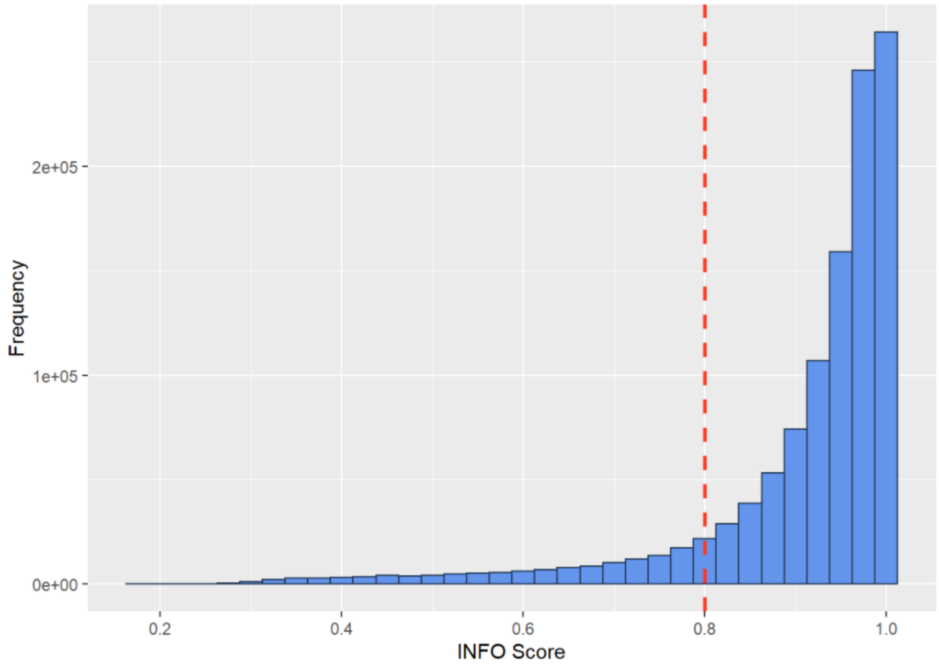
Supplementary Figure 1: Distribution of individual missingness and observed heterozygosity across sample cohort. Observed heterozygosity was calculated for each individual according to the formula: O(Het) = [N(NM) – O(Hom)]/N(NM); where O(Het), O(Hom) & N(NM) correspond to observed heterozygosity, observed homozygosity and the number of non-missing genotypes, respectively. The mean observed heterozygosity was 0.229 (σ = 0.004) and the exclusionary boundaries (x̄ +/- 3σ) were defined as falling at 0.217 and 0.241 (as depicted by the red lines). Six individuals were removed due to presenting excessive heterozygosity while all other participants were found to fall within an acceptable range of individual missingness (x < 1%).



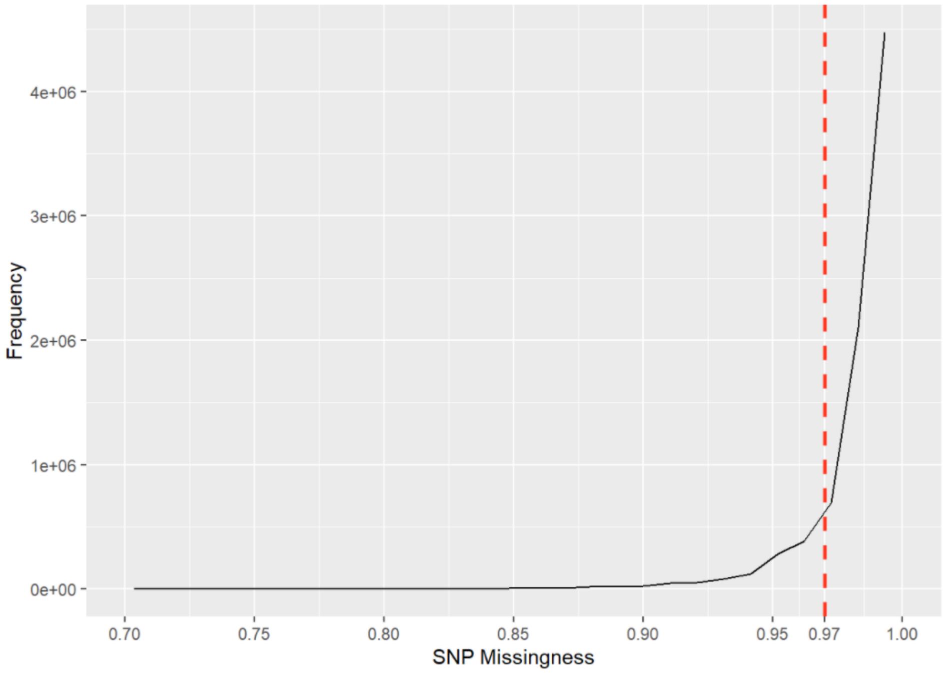
Supplementary Figure 2: Comparing reported sex against SNP-inferred sex. Participants presenting discordant sex information were identified by comparing reported identity labels to the mean homozygosity observed across X-chromosome associated SNPs. Analytical consensus suggests that one would typically expect males to display X-chromosome homozygosity rates at approximately 1.00 and females at less than 0.20, respectively. The two individuals for whom reported sex did not agree with SNP-based estimates have been denoted in red, above.



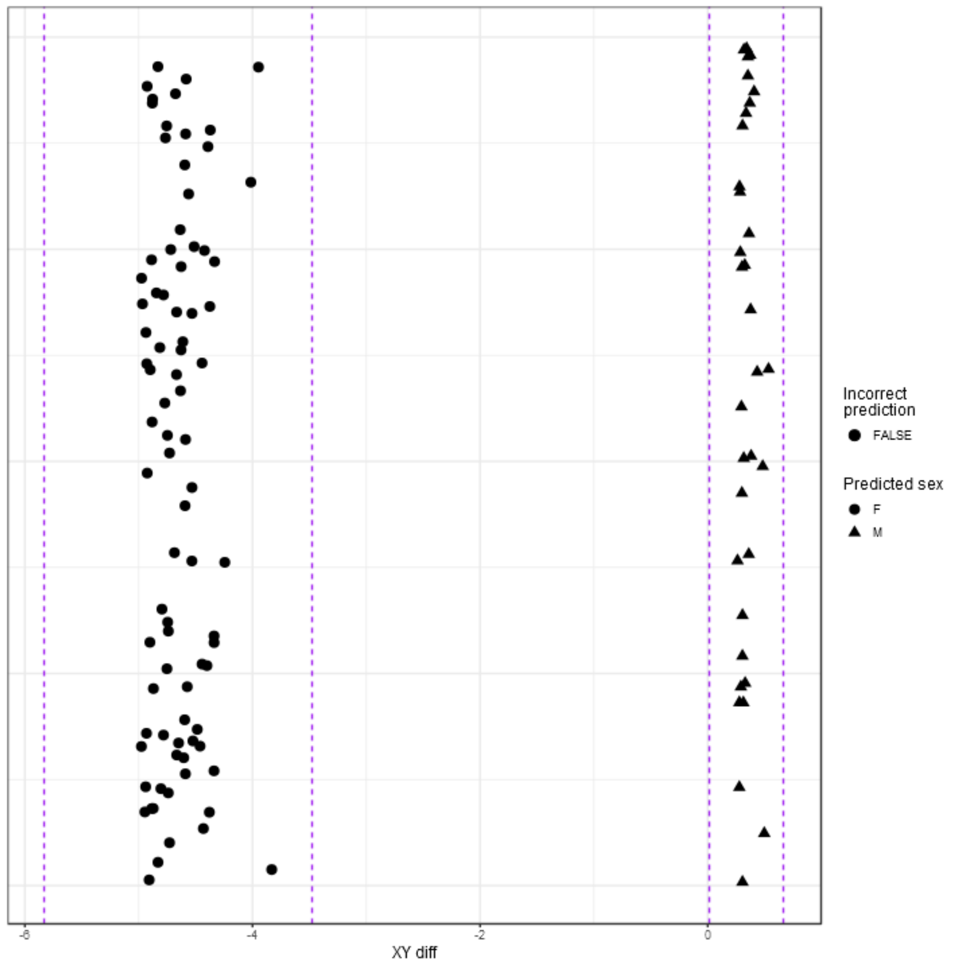
Supplementary Figure 3: Principal component analysis depicting relationship between PTSD cases and trauma-exposed controls. A visual aid illustrating the distribution of genetic variation within the sample cohort. Obtained by plotting the first two principal components against one another – the above graph depicts shared genetic variation compared to the dataset mean. Optically confirming that neither cases nor controls aggregate in an independent manner serves as a method of assessing whether underlying confounders may potentially be present.



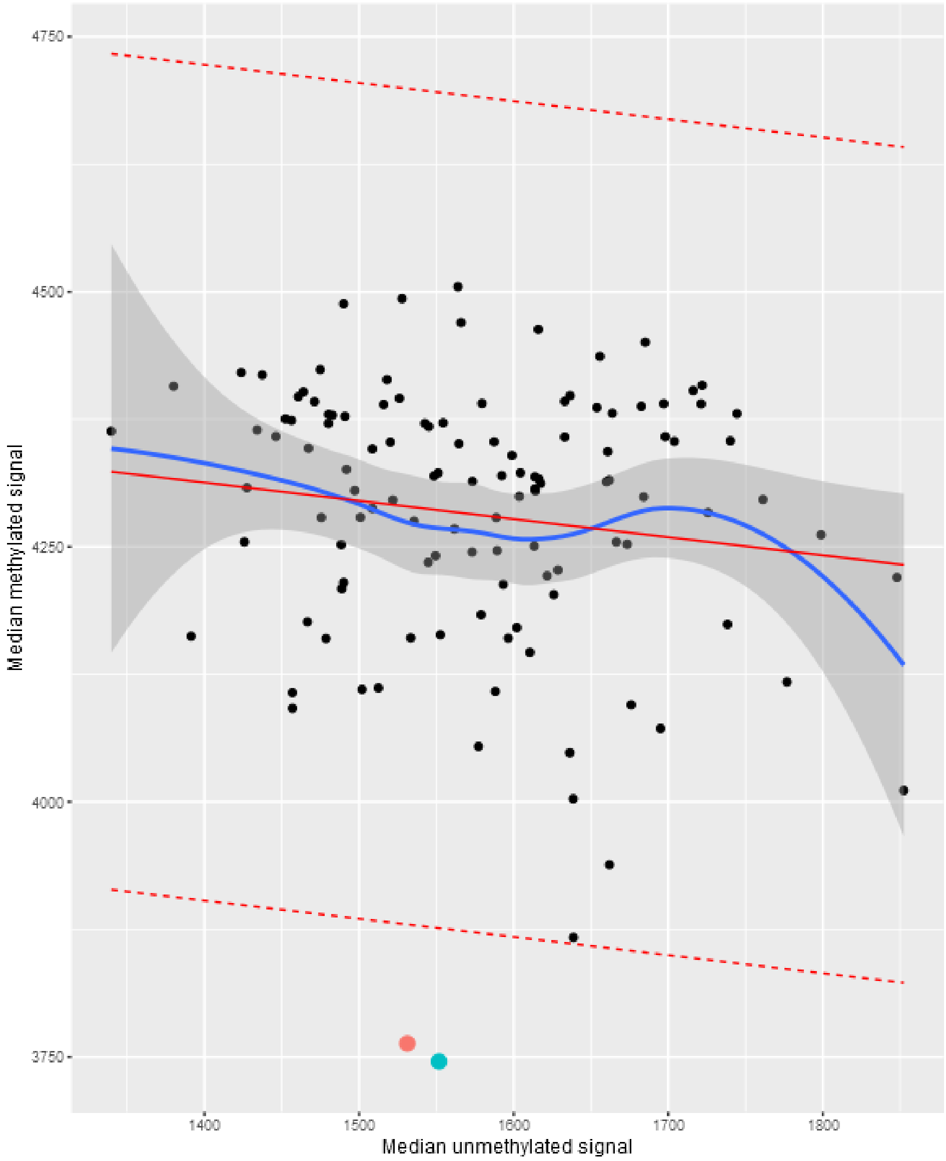
Supplementary Figure 4: Frequency distribution of INFO scores for chromosome-1 associated SNPs. Upon completing imputation, the SHAPEIT2 & PBWT pipeline provided INFO scores (ranging from 0.00 – 1.00) as a representative measure for assessing the confidence that each variant was imputed correctly. Traditionally, one can select the most appropriate cut-off value by generating frequency distributions depicting the spread of INFO scores across each chromosome (such as that provided for chromosome 1, above) and visually identifying the point of inflection. In order to limit analysis to SNPs that were more likely to survive subsequent quality control, the aforementioned figures were restricted to variants with a minor allele frequency greater than 1%. For the purpose of this study, we elected to employ an exclusionary threshold that would remove all SNPs presenting an INFO score of less than 0.8 (as depicted by the red line).



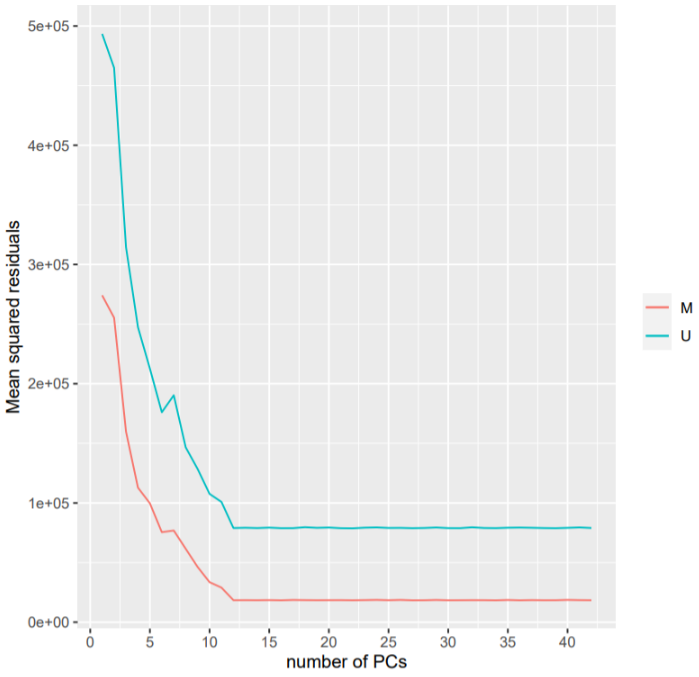
Supplementary Figure 5: Frequency distribution depicting SNP missingness across all imputed variants. SNPs presenting an excessive degree of missingness were identified by generating a frequency distribution for SNP call rate across all imputed chromosomes and visually determining an estimated point of inflection. Establishing a call rate threshold of 0.97, as depicted by the red line, translates to the removal of all variants presenting a missingness greater than 3%.



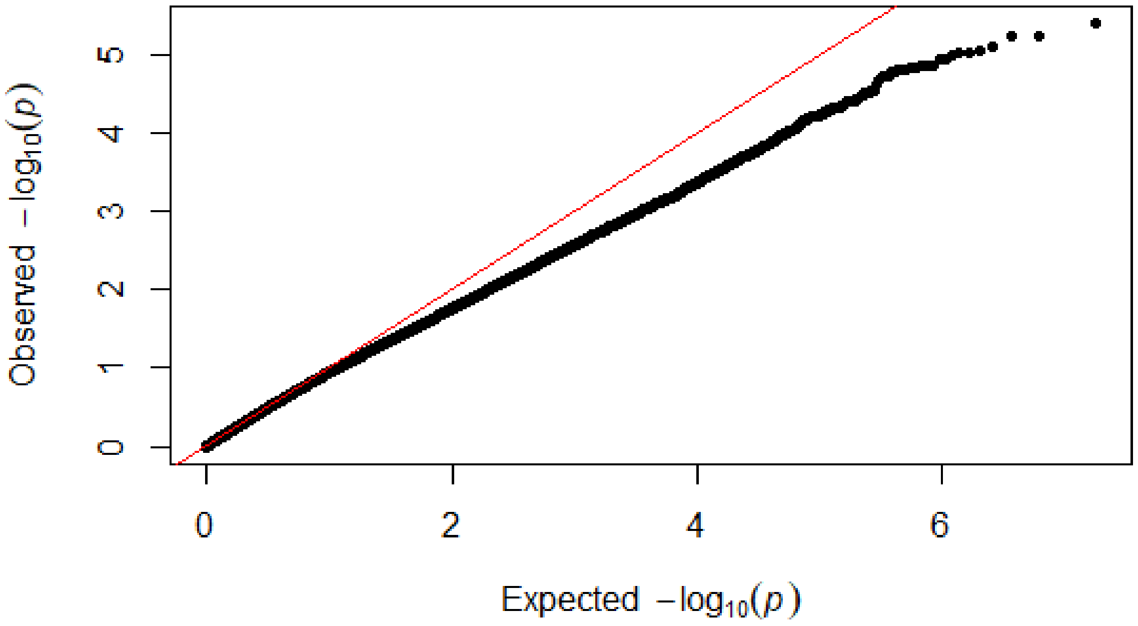
Supplementary Figure 6: Identifying discordant sex information amongst the DNA methylation data. The above is a *meffil* generated graph demonstrating the software package's inbuilt mechanism for identifying discordant sex information. While the Y-axis is simply for sample differentiation, spacing along the X-axis is indicative of the observed difference between median X- and Y-chromosome intensities across all samples. The occurrence of clustering on the left and right represents the spatial arrangement of female and male measurements within our dataset – wherein an exclusionary threshold of five standard deviations away from each cluster's median (as depicted by the dashed lines) was used to identify participants for subsequent removal.



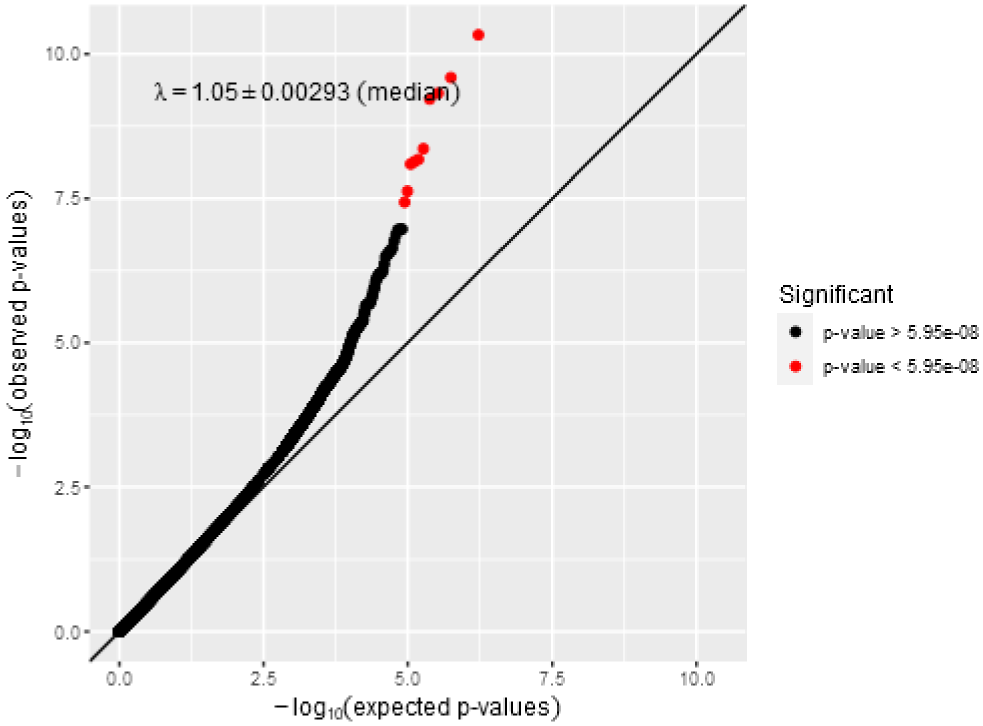
Supplementary Figure 7: Assessing sample quality through median methylated intensities. In addition to the user-defined variables, *meffil* presents a series of ingrained methods tailored to independently verifying sample quality. One such method is to plot median methylated and unmethylated intensities against one another, before subsequently identifying outliers relative to the adjusted median methylated intensity expected by regression. The linear regression models describing the relationship between median methylated and unmethylated intensities have been depicted by the solid red line above – wherein the dashed counterparts are reflective of the default outlier threshold of three standard deviations away from the expected methylated signal. As is depicted in the figure above, *meffil* identified two individuals for whom the median methylated intensities differed from that expected by the dataset norm.



Supplementary Figure 8: Corresponding scree plot for DNA methylation principal component analysis. In attempting to quantify sources of variation against which to regress the selected technical variables, *meffil* generated the above scree plot denoting the mean squared residuals associated with each principal component. Briefly, the mean squared residuals effectively represent the degree to which each principal component contributes to the variation observed in the dataset. Upon assessing the distribution of mean squared residuals across both the methylated and unmethylated control probes (as depicted by the red and blue lines, respectively), it was determined that the appropriate point of inflection lay at approximately four principal components.



Supplementary Figure 9: Q-Q plot pertaining to the genome-wide association test conducted on PTSD. Q-Q plots are graphical aids which allow one to formulate a superficial estimate as to the confidence that a model's results fall within the realms of reasonable plausibility. Briefly, each plot compares the distribution of observed p-values to that which would be expected should the greater dataset be derived from a normal distribution. Any deviation from the perfect norm (as depicted by the red line), can be addressed upon individual discretion. As can be seen in the figure above, the genome-wide association test conducted on PTSD generated p-values marginally lower than that dictated by the expected normal distribution.



Supplementary Figure 10: Q-Q plot pertaining to the epigenome-wide association test conducted on PTSD. Q-Q plots are graphical aids which allow one to formulate a superficial estimate as to the confidence that a model's results fall within the realms of reasonable plausibility. Briefly, each plot compares the distribution of observed p-values to that which would be expected should the greater dataset be derived from a normal distribution. Any deviation from the perfect norm (as depicted by the black line), can be addressed upon individual discretion. As can be seen in the *meffil* – generated figure above, the epigenome-wide association test conducted on PTSD generated p-values higher than that dictated by the expected normal distribution.