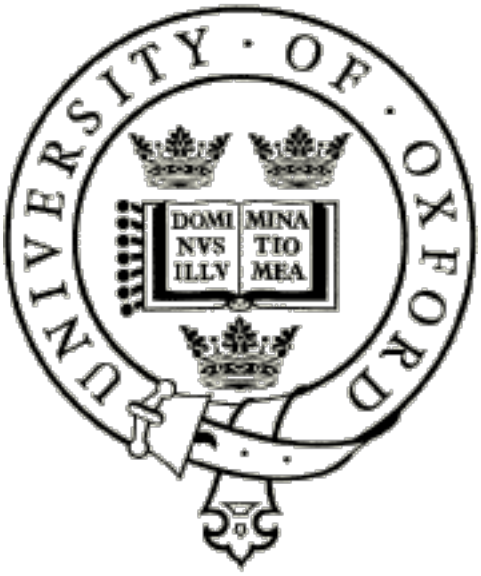
Molecular Markers of Stress



Na Cai

Clinical Medicine

St Hugh's College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Abstract

Using data from the China, Oxford and Virginia Commonwealth University Experimental Research on Genetic Epidemiology (CONVERGE) Consortium study of major depressive disorder (MDD)on 11,670 Han Chinese women, this thesis describes an investigation on the etiology of MDD, a psychiatric disease that has eluded previous genetic studies as well as investigations of its mechanistic underpinnings. It asks: what happens during stress, how may it contribute to the risk of developing MDD, and why does it increase the risk of MDD in some people but not others. It presents three main findings.

First, a GWAS on MDD conducted on 10,640 samples (5,303 cases and 5,337 controls) in the CONVERGE dataset found two genome wide significant associations with MDD, one lying at the 5' side of SIRT1, and the other in an intron of LHPP. Both signals have been replicated in a completely independent cohort of severe MDD cases and matched controls from Northern China, making them the first replicated association loci for MDD to date. Second, I found there are more copies of mtDNA in cases of MDD than controls and while the increase can be induced by stress, it is contingent on the depressed state. Further analyses of results from animal experiments showed stress increases mtDNA levels in a dose-dependent, reversible and tissue specific way that is mediated partly by stress steroids. Third, the total amount of heteroplasmy was found to increase with increasing mtDNA levels, and therefore is higher in cases of MDD than controls, consistent with a change in mitochondrial function observed in animal models of chronic stress.

All three findings suggest stress causes changes in mtDNA, and this change may be larger in cases of MDD than controls. This difference between cases and controls may be due to differences in their regulation of mtDNA levels and sequence mutation during stress, and this may be genetically determined. This study provides a new perspective to the etiology of depression, suggesting it may have origins in metabolic regulation.

Word count: Approx. 46,000 words

To my parents

Acknowledgements

Many people have played big roles in my PhD, and I am very grateful for having them.

My supervisor Jonathan tried talking me out of doing a PhD when I first came to see him – he said it's a hard life, full of insecurities and disappointments, and gave me an alternative: don't do the PhD, have a nice normal job, and have time for exercise, spouse and beautiful children. He then asked if I still wanted to do a PhD, and I said "yes". Since then, Jonathan told me uncomfortable truths when they could help me and kept them from me when they were too damaging, pushed me to take on responsibilities before helping me become ready for them, made me skeptical of every claim including his own but trusting of my own crazy ideas that no one believed in, and taught me to face problems and uncertainties with a big bright positive smile. He also told me to write in short sentences, which I am still working on. Thank you Jonathan for showing me your way of doing science and handling situations, for the incredible amount of pushiness and support for my work, for all the life tips, jokes and shared grumpiness towards collaborators' holidays, and for being my friend.

Richard saw me through CTC in my first year and GRC in my final year, the two conferences that gave me confidence in facing the community. Thank you Richard for being so dependably calm and supportive, even after seeing my horrible slides 30 minutes before the GRC talk.

Simon and GJ did all of the wonderful work stressing mice, which formed a central part of my PhD, gave me my first publication, and great stories to tell at conferences! Thank you both for your expertise and your hard work, it was a pleasure working with scientists like yourselves, and I wish we could collaborate in the future.

Colleagues in BGI (Jieqin, Jingchu, Songli, Qibin) were vital to all the sequencing, variant calling, imputation and validation efforts. Thank you all for being so supportive and responsive to my requests, and thank you all for becoming so good at understanding me after working together for so long. Colleagues in VCU (Ken, Tim, Roseann) were with me during for the reviews on the depression paper – the time difference worked well as at any point in time some of us were working on it! Thank you for your wonderful camaraderie.

Members of the Flint and Mott groups (Amelie, Jessie, Jerome, Leo, Xiangchao, Caroline, Amarjit, Polinka, Freddie, Wharton, Sanja) helped me in countless ways since the start of my PhD. Thank you for teaching me programming, involving me in projects, and helping me with endless qPCRs.

My friends (Amelie, Jessie, Carme, Robbie, Winni, Kiran, Maria, Martha, Pavitra, Roos Clare, Holly), some of whom also my colleagues on projects, have been my inspiration and comfort for the past four years. Thank you for the geeky conversations, pranks, post-its, football, zombies-full-of-screaming, board games, chats on the staircase, and balloons in my room.

Doug has read through all of my manuscripts and presentations, including this thesis, and motivated me with his exceptional work ethic. He has also made ordinary days feel special, amazing deeds look natural, and things I worry about seem little. Thank you for your quiet (sneaky) and constant support.

My parents believed in me when I didn't, and believed in their choices to leave home behind for my development. Their hard work and sacrifice are the foundations of everything I am able to achieve. Mum and Dad, I cannot thank you enough, for everything

CONTENTS

1. Introduction

This thesis describes an investigation into molecular markers of stress, identified from a genetic study on major depressive disorder (MDD), a psychiatric disease that has eluded previous genetic studies as well as investigations of its mechanistic underpinnings. It asks: what happens during stress, how may it contribute to the risk of developing MDD, and why does it increase the risk of MDD in some people but not others. In other words, what does environmental stress do at physiological and cellular levels that may be related to the etiology of MDD, and what genetic components does that interact with to confer the variability between people in susceptibility to disease under the same environmental stimuli?

Answering these questions requires one to look at both the environmental and genetics components that may be contributing to MDD. A dataset that contains both is needed, and the more details the dataset provides, the better the chances of answering the questions. The China, Oxford and Virginia Commonwealth University Experimental Research on Genetic Epidemiology (CONVERGE) Consortium study of MDD collected detailed personal, family and medical histories of 11,670 Han Chinese women from 60 hospitals across 30 cities in China. Half were cases of severe MDD, and the other half were matched controls. All subjects were whole-genome sequenced to 1.7x coverage using next generation paired-end sequencing using DNA extracted from their saliva samples.

The CONVERGE dataset represents the most coherent and comprehensive dataset of MDD to date, the scale of which is second only to the combined dataset from many different cohorts collected from Europe and the US as part of the Psychiatric Genomics Consortium (PGC), and it is the only of its kind in an Asian population. The depth of phenotypic information collected from both cases of MDD and controls as well as the selectivity in sample inclusion are unparalleled.

Having this dataset presented me with many opportunities to look into both genetic contributions and consequences of environmental factors on MDD. In this thesis I describe both genetics and molecular approaches taken to unravel the relationship between genetics, environmental stress, and MDD.

1.1 Major Depressive Disorder

Major depressive disorder (MDD) is a medical condition that includes abnormalities of affect and mood, neuro-vegetative functions (sleep and appetite disturbances), cognition (inappropriate feeling of guilt and worthlessness) and psychomotor activity (agitation or retardation). For most affected individuals, MDD is a life-long episodic disorder with multiple recurrences (averaging one in every 5 years).

MDD is one of the most common psychiatric disorders and one of the world's leading causes of morbidity, posing enormous burdens on health care systems[1]. Lifetime prevalence of MDD varies, from 3% in Japan and China to 16.9% in the U.S., with most countries falling in the range 8-12% [1,2]. Lower rates in Japan and China are at least in part because individuals who report depressive illness in East Asian cultures are more impaired than those in Western cultures[3].

The current diagnostic criteria for MDD, according to American Psychiatric Association's DSM-IV manual, represent a clinical and historical consensus on the important symptoms and signs of depressive illness. Affected individuals display a range of symptoms of varying disability, and there is ongoing debate whether MDD is best seen categorically as a single type of ailment, or a continuum in extent of dysfunction in affect regulation. Severity can be measured multiple ways, including number of episodes, length of episodes, the number of symptoms, and combination of clinical features, though there is no general consensus[4]. Current treatments are entirely symptomatic, and there are continuing disputes over the value of medication and psychological therapies[5,6]. As such, MDD is commonly referred to as a "disorder", which is an ailment that disrupts normal function, and not a "disease", which is a disorder with a known pathophysiological basis.

1.1.2 Known risk factors of MDD

The field of psychiatric epidemiology has identified a substantial list of putative risk factors for MDD, among which three stand out in consistency of their association with MDD and evidence suggesting at least some degree of causality: gender, stressful life events (including adverse childhood experiences), and personality traits[4].

Epidemiological studies of MDD have consistently shown higher prevalence of MDD in women as compared to men [7]. Twin studies investigating whether heritability of MDD differed and genetic components contributing to MDD were shared between the two sexes have also shown MDD was more heritable in women than men (40-42% in women compared to 29-30% in men), with genetic correlations of +0.55-+0.63 [8,9]. While there is substantial sharing of genetic risk factors for MDD in men and women, the equally appreciable proportion of genetic contribution of sex-specific effects points to different genetic variants being important in the etiology of MDD in men and women.

A wide range of environmental adversities or stressful life events, such as job loss, marital difficulties, major health issues and loss of close personal relationships, is associated with risk of MDD [10]. In particular, adversities in childhood such as sexual abuse, poor parent-child relationships and parental discord are thought to have chronic effects on a person and predict later onset MDD. Childhood sexual abuse (CSA), for example, has repeatedly been shown to have a persistent and large effect on the risk of MDD: depressed patients who report CSA had been shown to have more severe depressive symptoms, longer duration of episodes, earlier onset, and higher levels of comorbidity [11-13]. Two studies have shown that compared to unexposed co-twins, twins exposed to CSA have significantly elevated risk for MDD[14,15], indicating the overwhelmingly environmental contribution of CSA to MDD risk.

Lastly, certainly personality traits are associated with MDD, the best example being “Neuroticism”, which encompasses the predisposition to develop emotional upset

under stressful situations. Levels of neuroticism strongly predicted the risks for both lifetime and new-onset MD; twin modeling indicated that the association between neuroticism and MD resulted largely from shared genetic risk factors, with a genetic correlation of +0.46 to +0.47 [9] .

1.1.3 Biological theories of MDD

Theories on biological pathways leading to MDD are mostly derived from observations of symptoms of MDD that overlap with other biological states, or their responses to antidepressant drugs. One clue comes from the similarity between MDD and depressive symptoms that occur in the context of other medical conditions, such as endocrine disturbances (hyper- or hypocortisolemia, hyper- or hypothyroidism), collagen vascular diseases, Parkinson's Disease, traumatic head injury, asthma, diabetes, and certain cancers. This suggests there may be shared biological pathways between MDD and what is causing the depressive symptoms in these instances, with additional factors contributing to these symptoms being more easily reached, persistent and recurrent in individuals with MDD. Though no conclusive mechanistic understanding of MDD had been achieved, there is mounting evidence for a few theories. If supplemented with evidence from genetic studies, a new understanding of MDD could be reached.

1.1.3.1 The serotonergic and noradrenergic system hypothesis

Much of the current understanding of MDD comes from effective treatments, which were discovered serendipitously. Two classes of agents were discovered to be effective antidepressants. The first class, tricyclic antidepressants (such as imipramine), arose from antihistamine research. Their mechanisms of action involve the inhibition of serotonin or norepinephrine reuptake transporters. The second class, monoamine

oxidase inhibitors, were discovered from work on antitubercular drugs, and act by inhibition of monoamine oxidase, which is a major catabolic enzyme for monoamine neurotransmitters (including catecholamines such as adrenaline, dopamine, norepinephrine, as well as tryptamines such as serotonin). These discoveries lead to the development of serotonin-selective reuptake inhibitors (SSRIs) and norepinephrine-selective reuptake inhibitors widely used today.

Both classes of antidepressants effectively increase the length of time and concentration of monoamine neurotransmitters such as serotonin and noradrenaline at synapses and therefore increase neurotransmission. However, all antidepressants exert their mood-elevating effect only after prolonged administration, inconsistent with the fast action of neurotransmission enhancements by their acute activities. Despite efforts involving knockout and chronic stress animal models and human functional imaging and genotyping, abnormalities in the serotonergic and noradrenergic systems such as the serotonin transporter gene 5-HTT, its proximal 5' regulatory region 5-HTT gene-linked polymorphic region (5-HTTLPR), and the gene encoding monoamine oxidase A (MAOA), had not been conclusively shown to confer risk to MDD by direct genetic effect, gene-environment interactions, or gene-gene interactions [16-25]

1.1.3.2 The corticosteroid receptor hypothesis

Elevated circulating levels of stress hormones among depressives were recognized long before antidepressants were discovered: while it was first dismissed as epiphenomena reflecting the stressful experience of being in the depressed state, evidence has accumulated to show that aberrant stress hormone regulation could lead to depression and antidepressants may act through normalization of changes in the hypothalamic-pituitary axis (HPA) regulating the synthesis and release of the stress hormones.

Stress is a risk factor for a variety of mood disorders, including MDD and post-traumatic stress disorder, which can manifest years after the stressful event. A prominent mechanism by which the brain reacts to stress is activation of the HPA axis. The HPA axis starts in the paraventricular nucleus (PVN) of the hypothalamus, where neurons secrete corticotropin releasing factor (CRF). CRF stimulates the synthesis and release of adrenocorticotropic hormone (ACTH) in the anterior pituitary, which then stimulates the synthesis and release of glucocorticoids (cortisol in humans, corticosterone in rodents) into the blood stream from the adrenal cortex. Glucocorticoids are steroid hormones that bind to specific receptors present in a variety of tissue, and in particular exert profound effects on general metabolism ("gluco" came from its role in regulation of metabolism of glucose) and behavior due to its actions on many brain regions.

The corticosteroid receptor hypothesis for MDD was derived from various clinical observations, including: a) the number of ACTH and cortisol secretory pulses is increased in depressives [26], b) levels of CRF in the cerebral spinal fluid are elevated in depressives [27], and c) the number of CRF secreting neurons is increased in depression [28], and lastly d) depressives are not responsive to the suppressive modulatory effects of synthetic glucocorticoid dexamethasone on the secretion of ACTH [29], suggesting dysregulation in their negative feedback system with regards to ACTH and cortisol production.

While these effects are relatively transient, inconsistent with presentation of disease long after the stressful event, many long-term effects had been discovered. Theories how increased cortisol may cause MDD later in life fall in two categories: i) sustained elevation of cortisol (hypercortisolemia) could cause damage to hippocampal neurons causing dysregulation of its negative feedback control on the secretion of CRH by the hypothalamus [30], resulting therefore in escalating levels of cortisol production and a prolonged state of hypercortisolemia, ii) impairment of neuronal growth mediated

by the brain-derived neurotrophic factor (BDNF) mediated partly via stress-induced and glucocorticoids and partly via increased neurotransmission at serotonergic synapses [31].

1.2 Biomarkers of MDD

There are defining somatic symptoms and metabolic features in MDD, yet there are not many molecular insights into their causes and effects. Adverse life experiences, particularly those in childhood, are potent predictors of later onset depression and disease morbidity and mortality [32-34]. There is considerable interest in understanding the mechanisms through which they do so, as it remains unclear how illness becomes apparent decades after the presumed initiating event. Long-standing hypotheses include chronic activation of the hypothalamic-pituitary-adrenal axis (HPA) [35-37] and alterations of neuroimmune function [38].

Causal associations between stressful life events and early adversities such as childhood sexual abuse and MDD are well documented [14,39]. If the molecular markers of stress were permanent, or reversible only after a period of time, or cause downstream changes that could persist longer than the initial signal, then they should be visible in sufferers of MDD, especially those who are in depressive episodes, and who may have previously experienced more severe adversities than others. Molecular signatures of stressful life experiences and their relation to disease are therefore of special interest to clarify the causal relationship between signature, disease and stress. In chapter 4 I explore the relationship with two molecular markers that had gathered the interest of researchers in the past few years in their relationship with stress and psychiatric diseases – mean chromosomal telomere length, and mitochondrial DNA (mtDNA) levels.

1.2.1 Telomere length

There are many reports of association between leukocyte telomere length and psychiatric conditions. It has been reported that people who had suffered childhood trauma have shorted telomeres, increased rates of telomeric attrition, and other metabolic signs of "premature aging" as compared to controls. Telomere length declines with age as a function of cell division [40], and telomere shortening has been associated with age-related medical conditions such as heart disease and cancer [41].

However, aging *per se* is not the only determinant of telomere length. Prior studies have reported shorter telomeres under circumstances of higher chronicity of stress [42], higher degree of perceived stress [43], cumulative childhood stress exposure [44], longer exposure to depression [45,46], and depression severity [43,45]. Given that these associations remain after controlling for age, telomere length is perhaps more accurately conceptualized as a marker of biological rather than chronological aging [47].

Telomere length and telomeric sequences have been studied with molecular methods such as terminal restriction fragment (TRF) length analysis and qPCR (and more sophisticated versions of it). Though well-calibrated and accurate, and for some even so at the per-chromosome resolution, these are low-throughput methods that cannot be applied to large-scale studies with thousands of samples. With whole genome sequencing data, one can simply use the number of reads containing the telomeric sequence in a sample relative to a non-repetitive region of the same sample to estimate the mean length of telomeres among all of the sample's chromosomes. Though recent studies would argue that the signal for premature aging responses is not the mean telomere length but the shortest telomere length among all chromosomes, a measure of mean telomere length easily attainable with whole genome sequencing is still meaningful for comparison of average rate of telomeric attrition between samples and its association with metabolic and psychiatric phenotypes.

1.2.2 Mitochondrial DNA

One of the unexpected observations reported in this thesis is that the amount of mitochondrial DNA is increased in patients with MDD. Here I provide a brief introduction to the relevant mitochondrial biology. The mitochondrion is a double-membrane enclosed organelle present in variable numbers in cells of different tissue types. Mitochondria function in an integrated reticulum that is dynamically remodeled by fission and fusion events that occur alongside the biogenesis of new as well as biodegradation and autophagy of impaired or surplus mitochondria [48,49].

Each functional unit of mitochondria contains its own circular 16.5kb genome in multiple copies, and the replication of which can occur both in and out of synchrony of the cell cycle [50]. Despite the apparent autonomy, the number of copies of mitochondrial DNA (mtDNA) appears to be under tight regulation; previous studies have demonstrated that while the number of functional units of mitochondria are under tissue-specific regulation such that they vary widely between cell types but not within [51], mtDNA levels per functioning unit of mitochondrion are relatively constant both in a single cell type and between cell types even across different mammalian species [51]. Intriguingly, studies have found that the number of mitochondria units in daughter cells of mitosis is a better predictor of the time it takes for them to enter G1 phase of cell cycle than cell size. Taken together with there having been no reports of a "mitochondrial cancer" where the replication of mtDNA or the organelle has gotten out of control and taken over the cellular machinery, one can reasonably hypothesize that while mtDNA replication does not necessarily happen in synchrony with that of the nuclear genome, it is under inescapable regulation by both cellular metabolic demands and checkpoints in the cell cycle.

Changes in the amount of mtDNA has been associated with age, aneuploidy[52], deficiencies in respiratory chain subunits[53], myocardial infarction[54], cancer[55,56] and

recent studies have also shown that physical trauma can cause increased release of mtDNA from cells, which elicits immune reactions akin to that in sepsis[57,58].

There is an increasing effort to look into the mitochondrial genome for its potential to influence metabolism, stress responses and behaviour. The analysis of sequence variants in the mitochondrial genome has been traditionally dominated with haplogroup classifications of samples after sequencing specific haplogroup tagging variant sites, and only more recently had whole sequences of mitochondria been used to identify sequence variants that may be related to diseases. Decreased mitochondrial functioning has been reported in aging [59], due in part to oxidative stress [60]. Increased oxidative stress has also been reported in MDD and other psychiatric and psychological conditions [61]. As mtDNA content has been demonstrated to increase [62] in aging, it has been proposed that this physiological stressor has accelerates cellular aging, mimicking the effects of accelerated chronological aging [63]. The association between mtDNA and MDD is inconsistent, with some studies reporting an increase in mtDNA among MDD cases [43], another reporting the opposite [64], and another reporting no association[65]. Few studies are available on history of adverse experience in cases of MDD and their relationship with mtDNA, though Tyrka, et al. [43], using a far smaller sample, reported no association between mtDNA copy number and depressive symptoms or perceived stress.

1.3 Genetic studies of MDD

Family and twin studies have been used to investigate whether susceptibility to MDD has a heritable component and together yield an estimate of the additive genetic effects between 31%–42% [9,66], with no evidence to suggest shared environmental factors contribute meaningfully to familial aggregation. Thus there is good evidence that MDD is heritable and that the heritability is about the same as type 2 diabetes, a condition to which GWAS strategies have been successfully applied. Genetic studies of MDD could lead to the identification of inherited risk variants, genes and pathways associated with the disease. Findings from these studies could improve the understanding of the disease mechanisms and therefore guide the development of prevention and treatment strategies. However, genetic studies on MDD, more than those of other psychiatric diseases studied extensively to date, have not been successful. A short review of the findings from these genetic studies illustrates the frustration in genetic studies of MDD.

1.3.1 Linkage studies

To date there have been four whole genome linkage scans of major depression [67-70]. Three of these studies reported genome-wide significant linkage signals, but disagreed about their location (each study finding evidence for a single locus on respectively chromosomes 2q, 12q and 15q). The fourth study [69] was unable to detect any locus that exceeded a genome wide-significant threshold, but provided some support for loci on 12q and 15q.

1.3.2 Candidate gene studies

Many candidate gene association studies of MDD have been performed, mostly informed by pharmacological and hormonal hypotheses of the origins of depression [71]. A

relatively recent report used a reasonably large sample (1738 cases and 1802 controls) and attempted to replicate all reported associations that were supported by at least two positive findings[72]. The authors found minimal support for 4 out of 55 candidate genes: C5orf20 (P=0.038), NPY (P=0.034), TNF (P=0.0034) and SLC6A2 (P=0.039). They considered that these modest findings were likely false positives.

1.3.3 Genome wide association studies

Genome wide association studies (GWAS) involve the typing of a dense array of genetic markers, capturing a substantial proportion of variation in the genome sequence, in a set of samples informative of a trait of interest and the testing of associations between trait and genotype to map the location of susceptibility effects. Genotypes are usually obtained from genotyping arrays containing a few hundred thousand to a million markers, each of which will be tested for association with the trait of interest. Due to the large number of independent statistical tests performed simultaneously on the same data, the threshold below which a P-value of association could be considered significant is adjusted by a Bonferroni correction for the total number of tests, usually approximated to a million for convenience (such that the significance threshold for P values would be 5×10^{-8}).

Successful GWAS of common diseases, such as GWAS from the Wellcome Trust Case Control Consortium [73], has improved our understanding of the genetic basis of many common diseases: type 1 [74] and type 2 diabetes [75], coronary heart disease [76], Crohn's disease [77], prostate cancer [78] and breast cancer [79], as well as quantitative traits like height [80], body mass index [81], and blood lipid levels [82,83]. GWAS on psychiatric diseases have since 2007 been championed by the Psychiatric Genomics Consortium (PGC), a confederation involving more than 500 investigators from 80 institutions in 25 countries. Diseases of interest are autism, attention-deficit hyperactivity disorder,

bipolar disorder, major depressive disorder and schizophrenia. While there is moderate success in GWAS on bipolar disorder [84] (4 loci found with 7000 cases and 9000 controls) and a recent breakthrough in schizophrenia research, finding 108 associated genetic loci using 37,000 cases and 113,000 controls in a multi-stage GWAS [85], no genome-wide significant association had been found for ADHD and MDD.

Why are some diseases more amenable to GWAS than others? The power to detect a true association that exists at a particular false discovery rate using a set sample size depends on the effect size of the locus, which in the case of a single association test between a trait and one genetic variant, is the proportion of individual differences for a trait in the population that can be accounted for by the genetic variant tested. For a given heritability the more genetic factors involved, the smaller their individual effect sizes will be, and the lower the power to detect any one of them. Moreover, a large number of genetic factors would result in a spectrum of cumulative genetic variant effects in a gene pool (population) containing these genetic variants, the skewness of which could be controlled by relative allele frequencies of genetic variants, which in turn can be affected by natural selection or genetic drift. If the trait of interest is a disease, then this spectrum underlies the genetic liability to the disease, and if the disease is a complex disease affected by environmental factors, a similar spectrum of environmental liability needs to be added to it to confer the disease liability.

The inheritance of disease susceptibility can therefore be explained by a liability model where a combination of genes and environmental effects contribute to one's liability to the disease and those whose liability are above a particular threshold, which is determined by the population prevalence of the disease[86] are considered cases, while others are considered "normal" controls. This division can be symptomatic instead of quantitative or distinct, as it is in the case of MDD. An inappropriate definition of this division, drawing samples from the entire of the liability spectrum but artificially classifying them as cases and controls, or wrongly classifying cases or near-cases as

controls, could reduce the apparent odds ratio of any genetic variant truly associated with the disease liability, decreasing the power to detect such an association. Conversely, being selective of both disease cases (hence reducing population prevalence of the disease) and controls (hence reducing misclassification) for genetic studies could increase power. MDD suffers from both being a complex disease with many genetic factors of small effect sizes, and having symptomatic diagnostic criteria that may be inconsistent across studies and thwarted by self-reporting biases.

Six GWAS of MDD have failed to find common variants (those with allele frequencies greater than 5%)[87-90]. Combined analysis of 9,420 cases and 9,519 controls found no P-value that exceeded genome-wide significance; furthermore, there was no enrichment of significant results above that expected by chance [91]. However two GWAS with substantially smaller samples report positive findings: a discovery sample of 604 cases implicated HOMER1 [92], and an even smaller discovery sample of 353 cases was used to identify neuron-specific neutral amino acid transporter (SLC6A15) [93]. Both findings still lack replication from multiple independent sources and cannot be ruled out as false positives.

1.4 Limitations of past approaches to genetic studies of MDD

Despite the lack of a conclusive finding, genetic studies of MDD have provided a wealth of information regarding the genetic architecture of MDD that could inform new studies. Through assessing the amount of sharing of common variants between cases of MDD and controls in GWAS and computing a "SNP-based" or "narrow-sense" heritability (h^2), it was found that an appreciable proportion of the genetic variants involved are likely to be common (variants above 5% in frequency in population, with estimated h^2 of 21-30%, would account for more than 50% of heritability estimated from twin studies), acting independently and additively. These findings all point to the theoretical possibility of finding loci robustly associated with MDD given a well-considered study design.

1.4.1. Heterogeneity in MDD

One often stated reason for the failure to find genetic loci associated with MDD is a lack of power with current sample sizes: the largest discovery GWAS to date used 9000 cases of MDD. As stated in the above section, it is paramount to the study of complex common diseases with a wide distribution of disease liability and symptomatic diagnostic criteria to be selective in sample collection for increasing power of genetic studies to be conducted, especially if one has a limited budget available for collecting only a particular number of samples. One way forward, as suggested below, is to reduce the heterogeneity in diseases cases collected (thereby reducing population prevalence of the "diseased") as well as screening controls to prevent misclassification.

One study addressing whether the nine DSM-IV symptomatic criteria for MDD reflect a single underlying genetic factor had found the best fitting model to explain MDD concordance in 7,500 adult twin pairs required three genetic factors, reflecting the psychomotor/cognitive, mood, and neuro-vegetative features of MDD, suggesting

heterogeneity in MDD presentation and etiology [94]. Minimizing heterogeneity in the collection of MDD and maximising genetic liability in cases could increase proportion of genetic contribution shared among cases hence increasing power in association testing.

Restricting definition of the "disease" to a subset of symptoms or clinical presentation that have low prevalence in the population but increased familial risk could increase genetic liability and chances at detecting genetic associations. Restricting sample recruitment to recurrent disease in patients identified in hospitals further enriches for severity and identifies an even rarer form of MDD. Simulations based on genetic architecture of MDD being similar to that of body mass index (similar heritability of about 40%, many loci expected to be involved [95]) showed that if only the most severe forms of disease was taken, reducing prevalence to 0.5% or lower, then fewer than 20,000 cases would be needed to provide 80% power to detect at least one GWAS locus [96].

A twin-study design revealed that familial vulnerability to illness is best characterized by intermediate levels of recurrence, long duration of episodes, high levels of impairment, and recurrent thoughts of death or suicide, suggesting these clinical features are indicative of high genetic liability [97]. One proposed subtype of depression, melancholia, was found to identify a subset of MDD cases with distinct clinical features and a particularly high heritability of liability to MDD (43%), suggesting melancholia may reflect the underlying genetic liability to MDD. Enriching for cases for those presenting melancholia, though not etiologically distinct from non-melancholic MDD, could result in increase in power to detect genetic contribution to MDD.

As described previously, two major twin studies had agreed on MDD being substantially more heritable in women (40-42%) than men (29-30%) with genetic correlations estimated at 0.55 and 0.63 respectively, indicating presence of sex specific

genetic factors [8,9]. Hence another way to reduce heterogeneity is to only look at the sex with more genetic liability to MDD.

1.4.2 Unexplored regions of the genome

The GWAS analysis of common disease, using genotypes at common single-nucleotide polymorphisms (SNPs) from genotyping arrays has yielded novel, highly significant, and replicable association findings comparable to or exceeding the number of those generated by candidate gene or positional cloning efforts[73]. There are good reasons for using genotyping arrays for GWAS, such as high accuracy of genotype calls achievable with well-established calling methods for single cohort[98] and multiple cohort studies[73], cost effectiveness in typing a small set of common variants that could capture variations in proximity by linkage disequilibrium (LD), and good resource management allowing effective pooling of data with other groups using the same array for combined analyses of larger sample sizes.

Much effort has gone into making arrays that capture variants with different allele frequencies, LD with surrounding variants, and functional categories, so as to increase the chances of capturing variants contributing to traits of interest in GWAS[99]. However, while population or disease specific arrays such as the Immunochip[100] are being developed, they are ultimately be biased towards common variants in a population or candidate regions of the genome as informed by previous successful attempts at genetic mapping. The dearth of such successes in MDD research would preclude such designs.

GWAS based on genotyping arrays relies on indirect association to represent un-typed variants. The power to detect an association between a causal locus and a trait depends not only on the size of effect the locus has on the trait and its frequency in the population of interest, but also the degree of LD between causal variant and the nearest

typed marker (usually a single nucleotide polymorphism (SNP)). Denser genotyping or imputation of genotypes at un-typed variant sites from population specific reference panels, such as that generated and carefully curated by the 1000 Genomes Project [91], will increase the ability to detect causal variants. The development of imputation methods that leverage information from both input data and reference panels could infer genotypes at both typed and un-typed SNPs for multiple single-marker association studies[101] beyond that afforded by markers of a genotyping array alone.

Limitations exist however: though a monumental and paradigm-shift effort, the numbers of haplotypes captured per major population in the 1KGP remain few (up to hundreds) and limited in representation of their respective populations. The small number of haplotypes also limits its utility in identification of variants of low allele frequencies; in fact any rare variants identified would be treated similarly to a more common variant occurring at the same frequency in this limited sample.

Moreover, variants lying below the 5% threshold for common variants elude detection in most GWAS due to lack of representation on genotyping arrays commonly used for population-based genetic studies. Similarly, regions of the genome that are not well-tagged by markers present in genotyping arrays have thus far not been studied, and other types of genetic variants such as copy number variants and structural variations cannot easily be examined. Interrogation of these variants could reveal and fill in a gap in current understanding of genetic architecture of MDD, though gaining access to them requires deployment of population-scale sequencing.

1.5 Design of the CONVERGE study

In chapter 3 of this thesis I discuss the results of GWAS on MDD in the CONVERGE study, and I attribute much of the success of the GWAS and secondary genetic analyses in this study to its design that addresses some of the problems that were outlined in the previous section.

1.5.1 Homogenous cohort of severe MDD

To reduce heterogeneity in genetic contribution to disease among cases of MDD collected and heterogeneity in all other aspects (unrelated to MDD) among all samples, the CONVERGE consortium conducted extensive screening using a two-hour long questionnaire on all 6000 cases of severe MDD and 6000 matched controls, where known sources of phenotypic and genetic heterogeneity were minimized and risk factors were documented. The criteria for recruitment are as follows:

Women only are recruited. Restricting the sample to one sex increases power over designs that analyse both sexes.

All four grandparents of both cases and controls are Han Chinese. Cases and controls are matched for location to reduce population structure effects. The choice of a Chinese hospital-acquired population of MDD cases identifies a more rare form of disease than found elsewhere in the world. MDD in China has a relatively low prevalence (e.g., life-time prevalence rate for MDD in China of 3.5% [102] versus 16.2% in the US [2]), at least in part because individuals who report depressive illness in East Asian cultures are more impaired than those in Western cultures [3]. Restricting sample recruitment to recurrent disease in patients identified in hospitals further enriches for severity and identifies an even rarer form of MDD.

All cases have multiple episodes of MDD. Cases are all clinically, rather than community ascertained, again increasing the chances that they have a genetic aetiology

[103]. The clinical features of CONVERGE cases show that the sample is indeed enriched for severe disease. In CONVERGE rates of a severe subtype of MDD, melancholia proved to be high: 85% met diagnostic criteria. Other measures of severity were also enriched: mean number of episodes is 5.3, mean number of DSM-criteria in the most severe episode is 8.3 (out of 9), and median episode length is 24 weeks.

CONVERGE cases were selected to be free of confounds that afflict other samples. For example, none of the CONVERGE cohort abuse alcohol or illicit drugs, and the rates of smoking are low (only 333 ever-smokers out of 6,120 MDD cases). Smoking, and alcohol and drug abuse are commonly comorbid with MDD and may act as causes of MDD, as well as consequences [104,105]; these important confounds have negligible impact on CONVERGE. To reduce the probability that cases with recurrent MDD would go on to develop bipolar disorder (an exclusion criteria), a minimum age is set at 30.

Controls are selected so as to maximize their value for genetic studies. They are screened by personal interview to ensure that they have not experienced a prior depressive episode and are interviewed to obtain data on environmental and other risk factors. They have a minimum age of 40 to reduce the chances that they subsequently develop MDD.

1.5.2 Next generation sequencing

Rather than relying on indirect association in GWAS using genotyping arrays, where power to detect an association depends not only on size of effect the locus has on the trait and its frequency in the cohort but also the distance and LD between it and the nearest typed marker, next generation sequencing (NGS) allows the identification of all variants in the genome and allow the direct identification of causal variants in association studies[106]. In the CONVERGE whole genome sequence was acquired to a mean coverage of 1.7X (where every base in the genome is covered, on average, 1.7 times, 95% CIs 0.7-4.3) per individual from which variants sites were identified and a

select, high quality set was imputed to arrive at allele dosages for use in association testing.

To date, whole genome NGS has been limited to samples of at most a few hundred individuals from a single population (GoNL) [107]. While this method can deliver high quality genotypes for single nucleotide polymorphisms (SNPs) and structural variants (SVs) (with intermediate ~13X coverage > 99% accuracy for a genotype call), the majority of variants occurring at a frequency of under 1% will be missed in such small sample sizes. Also, while deep sequencing has been successful in studies of Mendelian disorders[108-112], its application to studies of complex traits that require thousands of samples is forbidding due to high sequencing costs. Sequencing at low coverage (below 1x) could make a larger number of samples fit the budget of a single project[113] without incurring a loss in power for association studies if imputation was applied to augment the sequencing in inferring genotypes at variants both directly sequenced and missing[101,113,114]. With NGS, it is also possible to look for changes in the whole genome, especially "reactive" parts such as the mitochondrial genome and the telomeric DNA and their associations with MDD or its predisposing factors.

1.6 Thesis outline

This thesis begins with a description of the process through which genotype data were extracted from whole-genome NGS on 11,670 samples in the CONVERGE cohort for genetic analyses. Chapter 2 details the processing of NGS data, the calling and validation of variants using sequences from all samples, the imputation strategy for variants for which reference haplotypes are present and for those for which no such information is available, and finally the quality control on results of imputation. The chapter then goes through the precautions and considerations necessary for a successful GWAS, such as removing contaminated samples with unreliable sequenced and inferred genotypes, excluding samples that are more related to each other than rest of the samples, and detecting and controlling for population structure appropriately.

Chapter 3 shows how the whole spectrum of genetic variants, from the very common to those private to a single sample, may contribute to MDD. It first presents the results from GWAS on MDD in CONVERGE, the largest and most homogenous discovery cohort of severe MDD cases collected to date, along with analyses further reducing heterogeneity in the cases of MDD and maximizing genetic contribution. It reports the finding of the first two genome-wide significant associations for MDD that successfully replicated in a substantially sized independent cohort, obtained from the same population and also enriched for severe MDD cases, and a third genome-wide significant locus upon removing MDD cases with potentially high environmental contributions to their disease etiology. The chapter then goes on to examine the genetic architecture of MDD, and attempts to find the "missing heritability" of MDD using whole-genome variations obtained from NGS and methods accounting for LD between the markers appropriately. Finally, it investigates the involvement of private deleterious variants in coding regions of the genome in MDD, and specifically asks if particular candidate pathways are enriched for such deleterious variants.

Chapter 4 begins with the observation that there are associations between the amount of mtDNA and mean telomere length with MDD, as well as with life adversities. It describes how these findings were validated by eliminating the technical and biological artefacts that could explain the association between MDD and mtDNA levels. It then explores the mechanisms of the molecular changes in mtDNA with functional studies on laboratory mice that directly test the causal relationship between stress and changes in mtDNA levels for time dependence, reversibility, tissue specificity, and their metabolic consequences.

As the discovery of changes in mtDNA in MDD and stress begged the question what mechanisms might be controlling normal regulation of amount of mtDNA and the transition between normal and stressed states, Chapter 5 begins with my findings of genetic factors likely controlling steady state levels of mtDNA. It then asks whether changes in mtDNA level during stress were accompanied by sequence changes, and if yes, whether these sequence changes could provide the link between stress and MDD. Chapter 5 concludes on analyses done to show that sequence changes in the mtDNA as mtDNA levels increase may be important as functional switches between metabolic strategies during stress, and such a switch may be underlying the physical symptoms of MDD.

Finally, the concluding chapter attempts to summarise the insights from each chapter, and forms an outlook on the nature of MDD based on all of the findings presented in this thesis.

2. Considerations for GWAS on MDD using NGS data

This chapter describes the considerations needed for appropriately using low coverage sequencing to genotype the CONVERGE cohort. It details how I identified SNPs from NGS for imputation, checked imputation results to extract accurate ones for GWAS, picked out samples showing signs of DNA contamination and those who were more related than others from the cohort, and using clean genotyping information from uncontaminated and unrelated samples constructed a picture of population structure in the cohort, and systematically tested the different methods for controlling for population structure in GWAS.

2.1 Next generation sequencing data

Using next generation sequencing (NGS) to directly genotype all samples in a study cohort at all variant (and non-variant) sites in the genome would be an ideal design, as it allows the identification of all SNPs, as well as structural variants and variants in repeat-rich parts of the genome such as centromeric and telomeric regions. However accurate genotyping requires coverage of about 40x[115] rendering NGS too expensive for studies that need thousands of samples. Using lower coverage NGS by building barcoded libraries and pooling samples' DNA on sequencing lanes could cut costs substantially, though the lower the coverage per sample, the more regions are missed per sample, and the harder it is to call genotypes, especially heterozygous genotypes, with any confidence. However, the combination of direct inference of genotype or genotype probabilities from sequencing reads and imputation makes low-coverage NGS a cost-effective way to generate genotyping information for association studies[113].

The CONVERGE project used low-coverage sequencing for generation of genotypic data for GWAS on MDD. Figure 2.1 shows the coverage obtained for 11,670 sample to be, on average, 1.7X. This makes pre- and post-variant calling quality control

paramount to identifying variant sites and estimating genotype likelihoods at these sites for imputation, and for imputation to perform well. Also, it necessitates many post-imputation quality checks to ensure the imputed genotype probabilities and allele dosages were reflections of the samples' true genotypes, so as to be useful in association testing. This chapter describes the variant calling process, as well as the steps taken to arrive at genotype data usable for GWAS.

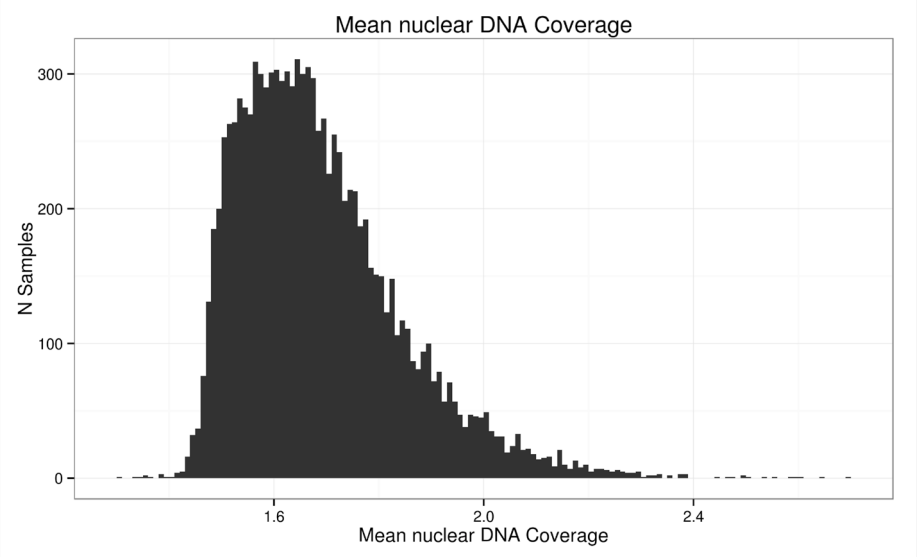


Figure 2.1 | Sequencing coverage per sample in CONVERGE This figure shows the mean sequencing coverage per site in the nuclear genome per sample for 11,670 samples in CONVERGE; the mean sequencing coverage is 1.7X.

2.2 Variant calling on low-coverage NGS data

There are challenges in extracting variant information from NGS data, and even more so from low-coverage NGS data. Studies like CONVERGE using NGS at low coverage benefit from making variant calls among all samples in the cohort together, which presents computation and logistics challenges due to the large sample size. A tradeoff between sensitivity and selectivity needs to be achieved to arrive at an appropriately accurate set of variant calls– hence a two step approach, first resulting in a high sensitivity raw call set then tuned for selectivity to minimize positives calls, was adopted.

I performed pre-processing and variant calling from low-coverage sequencing from all 11,670 samples in CONVERGE using "best-practices" approaches[116] in processing next-generation sequencing data (Methods). Variant quality score recalibration (VQSR) was then performed with GATK's[116-118] VariantRecalibrator (v2.7-4-g6f46d11) in SNP variant calls using the SNPs in 1000 Genomes Phase 1 ASN Panel [91] as the known, truth and training sets to exclude likely false positive SNP calls from further analysis as potential variant calling errors. Sets or "tranches" of SNPs with increasing VQSLOD scores (increasing stringency in exclusion of outliers and concomitant decreasing sensitivity to known SNPs) were generated at the specified sensitivities to known SNPs and transition-to-transversion (TiTv) ratios were calculated for each tranche.

Table 2.1 summarizes this information. A sensitivity to known SNPs of 90.0% was chosen for the inclusion of novel SNPs in downstream analyses as it balances optimization of TiTv ratio in novel SNPs with retaining as many potential novel SNPs as possible so as not to lose variant information. This gave a total of 21,356,798 (9,053,391 known in 1000 Genomes Phase 1 ASN Panel and 11,486,024 novel), biallelic SNPs identified from all chromosomes and unassembled contigs.

Sensitivity to Known SNPs (%)	Number of SNPs		TiTv Ratio	
	Known	Novel	Known	Novel
79	7946865	9361731	2.1783	2.1998
80	8047459	9502607	2.1786	2.2011
85	8550425	10357018	2.1817	2.204
87	8751611	10853838	2.1838	2.2042
89	8952798	11310561	2.1862	2.1961
90	9053391	11486024	2.187	2.1868
95	9556357	13089943	2.1849	1.9637
98	9858137	15716679	2.1759	1.6553
99	9958730	17362038	2.1721	1.5296
99.9	10049264	20940831	2.1683	1.353
100	10059324	22722016	2.168	1.2839

Table 2.1 | Transition-Transversion (TiTv) ratio for known and novel SNPs discovered in CONVERGE in different tranches of sensitivities to known SNPs in 1000 Genomes Phase 1 ASN Panel. This table shows the results of VQSR on all SNPs called in CONVERGE using default annotations in GATK and biallelic SNPs in 1000 Genomes Phase 1 ASN Panel as "known", "true", and "training" sets. The first column shows the tranche defined by sensitivity to known set, the second and third column shows the number of known and novel SNPs in CONVERGE falling into each tranche, and the last two columns show the TiTv ratio of known and novel SNPs in each tranche respectively.

20,539,441 SNPs from the autosomes and chromosome X were considered potential "true-positive" biallelelic variants, and were put forth for imputation of genotype probabilities and downstream analyses. In addition, all known sites in 1000 Genomes Phase 1 ASN panel[91] were included in all further analyses regardless of their VQSLOD scores and if they were identified as polymorphic in CONVERGE. Numbers of SNPs imputed in each chromosome and the percentage of SNPs in each category are shown in Table 2.2.

Chr	All Imputed	Known in 1000 Genomes Phase 1 ASN Panel						Novel in CONVERGE		Used in GWAS	
		Imputed	CONVERGE		Not in CONVERGE		% of Imputed	Count	% of Imputed	Count	% of Imputed
			Count	(%)	Count	(%)					
chr1	1906566	1049943	709178	67.54	340765	32.456	55.07	856623	44.93	473126	24.82
chr2	2080626	1141277	768281	67.32	372996	32.682	54.85	939349	45.15	504269	24.24
chr3	1730976	971680	660330	67.96	311350	32.042	56.14	759296	43.87	444601	25.69
chr4	1683635	963695	655100	67.98	308595	32.022	57.24	719940	42.76	452165	26.86
chr5	1552865	866733	586899	67.71	279834	32.286	55.82	686132	44.19	395356	25.46
chr6	1542585	895135	602872	67.35	292263	32.65	58.03	647450	41.97	416919	27.03
chr7	1386630	780333	519815	66.62	260518	33.385	56.28	606297	43.72	357113	25.75
chr8	1352488	743135	500602	67.36	242533	32.636	54.95	609353	45.05	330591	24.44
chr9	1049711	587643	394219	67.09	193424	32.915	55.98	462068	44.02	264762	25.22
chr10	1204931	669919	456560	68.15	213359	31.848	55.6	535012	44.4	310998	25.81
chr11	1203686	662829	443108	66.85	219721	33.149	55.07	540857	44.93	297679	24.73
chr12	1141658	644071	435882	67.68	208189	32.324	56.42	497587	43.59	294448	25.79
chr13	858815	486316	334096	68.7	152220	31.301	56.63	372499	43.37	225668	26.28
chr14	793487	445316	303662	68.19	141654	31.81	56.12	348171	43.88	201734	25.42
chr15	722748	397917	266748	67.04	131169	32.964	55.06	324831	44.94	175810	24.33
chr16	793223	418892	276272	65.95	142620	34.047	52.81	374331	47.19	178608	22.52
chr17	674848	363982	238795	65.61	125187	34.394	53.94	310866	46.07	150405	22.29
chr18	683315	386244	265213	68.67	121031	31.335	56.53	297071	43.48	173158	25.34
chr19	542027	299517	192281	64.2	107236	35.803	55.26	242510	44.74	128700	23.74
chr20	552234	294881	200864	68.12	94017	31.883	53.4	257353	46.6	127726	23.13
chr21	324063	182937	123793	67.67	59144	32.33	56.45	141126	43.55	83062	25.63
chr22	334001	180274	119161	66.1	61113	33.9	53.98	153727	46.03	79845	23.91
chrX	943020	404837	265078	65.48	139759	34.522	42.93	538183	57.07	175876	18.65
Total	25058138	13837506	9318809	67.35	4518697	32.655	55.22	11220632	44.78	6242619	24.91

Table 2.2 | Imputed SNPs in CONVERGE. This table shows the number of imputed SNPs in CONVERGE in each chromosome. The first two columns show the chromosome and total number of SNPs imputed in that chromosome. The next six columns show the composition of these SNPs in relation to the 1000 Genomes Phase 1 ASN Panel: the number of SNPs imputed that were found to be polymorphic in 1000 Genomes Phase 1 ASN Panel, the number and percentage of SNPs polymorphic in CONVERGE, the number and percentage of SNPs not polymorphic in CONVERGE, and the percentage of SNPs in 1000 Genomes Phase 1 ASN Panel among all imputed SNPs. The following two columns show the number and percentage of imputed SNPs that were novel in CONVERGE. The final two columns show the number and percentage of imputed SNPs that were include in the GWAS for MDD.

2.3 Imputation at novel SNPs

As the variants in the final call set generated from 12,000 CONVERGE samples way outnumber the 13,837,179 biallelic SNPs polymorphic in the Asian (ASN) reference panel in 1KGP Phase 1[91] (including Chinese-South (CHS), Chinese-Beijing (CHB), Chinese-Dai (CHD) and Japanese (JPN)), inferring of genotypes at the loci not present in 1KGP Phase 1 required the generation of a pseudo-reference assuming the ASN panel presented the reference allele at all these loci, or a reference-free approach. I first evaluated the merits of the two approaches, and when decided on the latter, quantified the accuracy of imputation through comparing the imputed allele dosages at sets of loci at which some or all of the CONVERGE samples were genotyped with genotyping arrays, custom chips, or higher coverage sequencing.

The SNPs for imputation can be divided into those present in 1000 Genomes Phase 1 ASN panel, which can be imputed using the 1000 Genomes Phase 1 ASN haplotypes as reference haplotypes, and those that were not. There were two conceivable ways to impute and infer genotypes at loci in the latter set, the "novel" SNPs from the joint variant calling performed on the low-coverage sequencing of all CONVERGE samples. The first is a reference-based method, where the genotypes at the novel loci discovered in CONVERGE were assumed to be homozygous for the reference alleles (in the human reference genome GRCh37.p5) in the 1KGP Phase 1 ASN panel. A new, "filled in" 1KGP Phase 1 ASN panel would be created with homozygous reference alleles for all novel sites, and used as reference panel for imputation into CONVERGE samples. This approach would be convenient, as imputation would be run only once on all 6 million SNPs identified in CONVERGE.

The second is a two step approach that involves imputing the "known" SNPs identified in CONVERGE and also present in 1KGP using the 1KGP Phase 1 ASN panel, then imputing the "novel" SNPs without a reference panel leveraging only the input data,

then combining the results of the two rounds of imputation by lifting the results of the latter into the former.

Using a 5MB region on chromosome 10 (65-70MB), I assessed the relative merits of the two methods together with a colleague working on CONVERGE, Warren Kretzschmar. Warren wrote the imputation pipeline for both approaches, and implemented the analysis: genotype likelihoods at variant sites passing the above criteria were estimated using SNPTools[119]; for the first approach, BEAGLE[120] (version 3.3.2) was run for six iterations in parallel on chunks containing 3000 SNPs each, with neighboring chunks sharing 600 SNPs that were used for ligation post-imputation, with the "filled in" 1000 Genomes Phase 1 ASN haplotypes as reference panel; for the second approach, BEAGLE was the same way as above except using the original 1KGP Phase 1 ASN panel on only the "known" SNPs, while it was run for ten iterations on chunks of 3000 SNPs each with 600 SNPs of overlap without a reference panel on all "novel" SNPs in CONVERGE, before the results were combined.

I assessed the relative merits of the two methods by examining their respective aggregate r2 between imputed dosages at both known and novel SNPs in the 5MB region with genotypes from a Illumina ZhonghuaOmni BeadChip performed on 16 samples (Methods, Figure 2.3a) and genotypes called from 10x coverage sequencing on 8 samples (Figure 2.3b) per minor allele frequency bin.

As the two-imputation approach gave higher aggregate Pearson r2 between imputed allele frequency and genotypes from both the Zhonghua array and 10x coverage sequencing at all alternative allele frequency bins, we used the two-imputation strategy for imputation over the whole genome for all samples in CONVERGE.

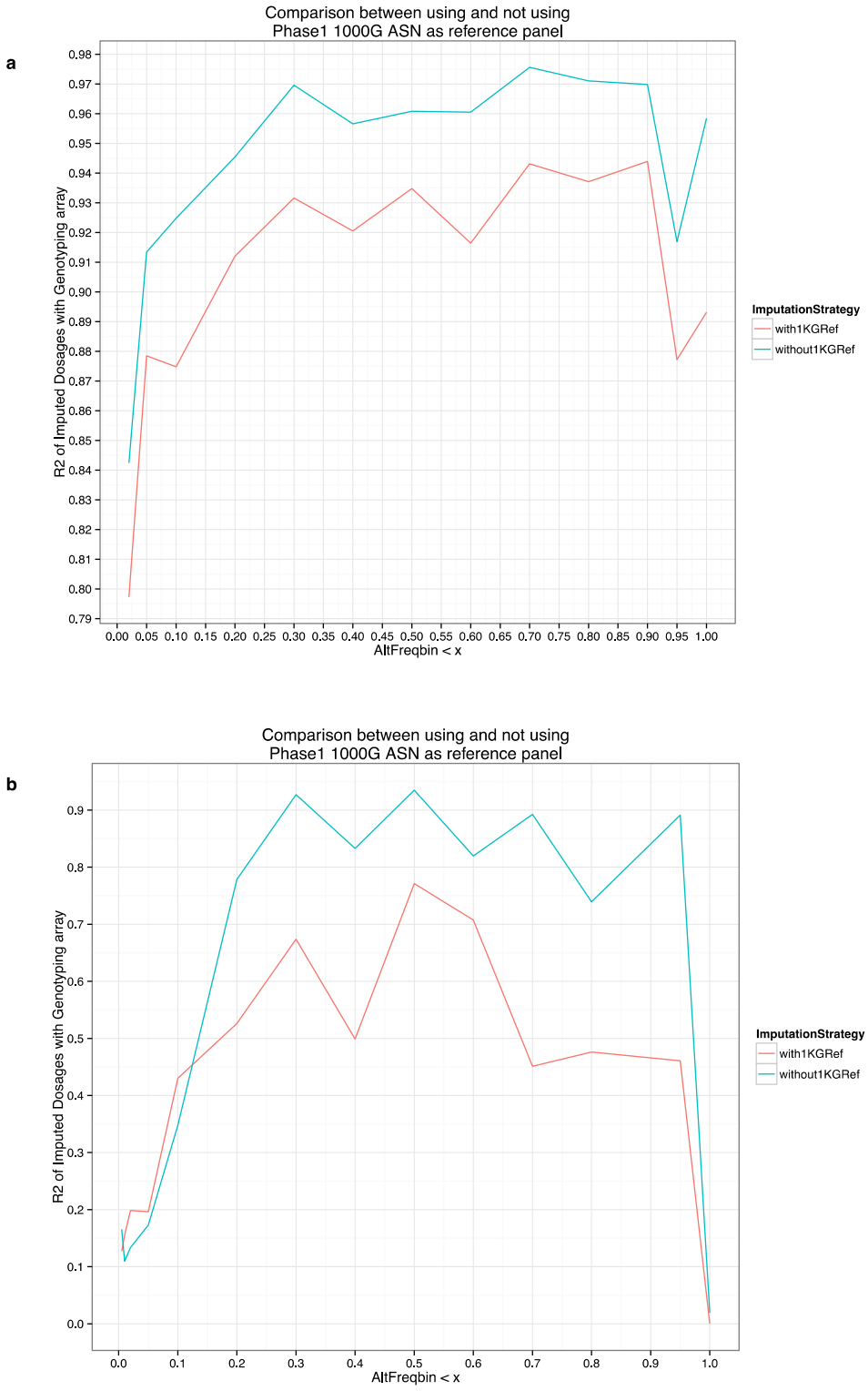


Figure 2.3 | Comparing "filled in" reference and reference-free methdos for imputation at novel SNPs in CONVERGE. This figure shows the aggregate Pearson r^2 between imputed allele dosages and genotypes from **a.** Illumina ZhonghuaOmni BeadChip on 16 samples and **b**. 10x coverage sequencing on 8 samples per alternative allele frequency bin, at both known and novel SNPs, using the "filled in" 1KGP Phase 1 ASN haplotype panel as reference for imputation (line in red) and without using a reference (line in blue).

2.4 Assessing imputation quality

2.4.1High coverage (10x) whole-genome sequencing on 9 samples

Nine samples in CONVERGE were sequenced to an average of 10x coverage using the same Illumina Hiseq platform used for the low-coverage sequencing samples. Mapping and processing of the 10x sequencing reads were performed in an identical fashion as the low-coverage sequences. Variant calling (for both SNPs and INDELs) and VQSR for SNPs were performed on the high coverage dataset using sequencing reads from all nine samples with UnifiedGenotyper in GATK[116-118] (v2.7-2-g6bda569, Methods); this gave a total of 5,914,716 (5,272,185 known in 1000 Genomes Phase 1 ASN Panel and 642,525 novel) biallelic SNPs identified in high coverage sequencing data with which we are able to check for squared Pearson correlation coefficient (r^2) and concordance with imputed allele dosages and hard-called genotypes of the same samples respectively.

I calculated the percentage concordance between hard-called genotypes from imputed genotype probabilities (where the genotype with the maximum imputed genotype probability > 0.9 was called) and genotypes called from the 10x coverage sequencing data at all imputed variant sites for each of the nine samples. The mean concordance per sample across all sites was 98.1% (standard error = 0.12%), and the percentage concordance calculated for each alternative allele frequencies was shown in Figure 2.4. I also calculated r^2 between imputed allele dosages and genotypes from the 10x coverage sequencing data at all imputed variant sites for each of the nine samples. The mean r^2 per sample across all sites was 0.938 (standard error = 0.003). The r^2 calculated for each alternative allele frequencies was shown in Figure 2.5. Table 2.3 combines all these information and shows the mean percentage concordance and Pearson r^2 among 9 samples for each genotype class (homozygous reference, heterozygous and homozygous alternative) for each alternative allele frequency bin.

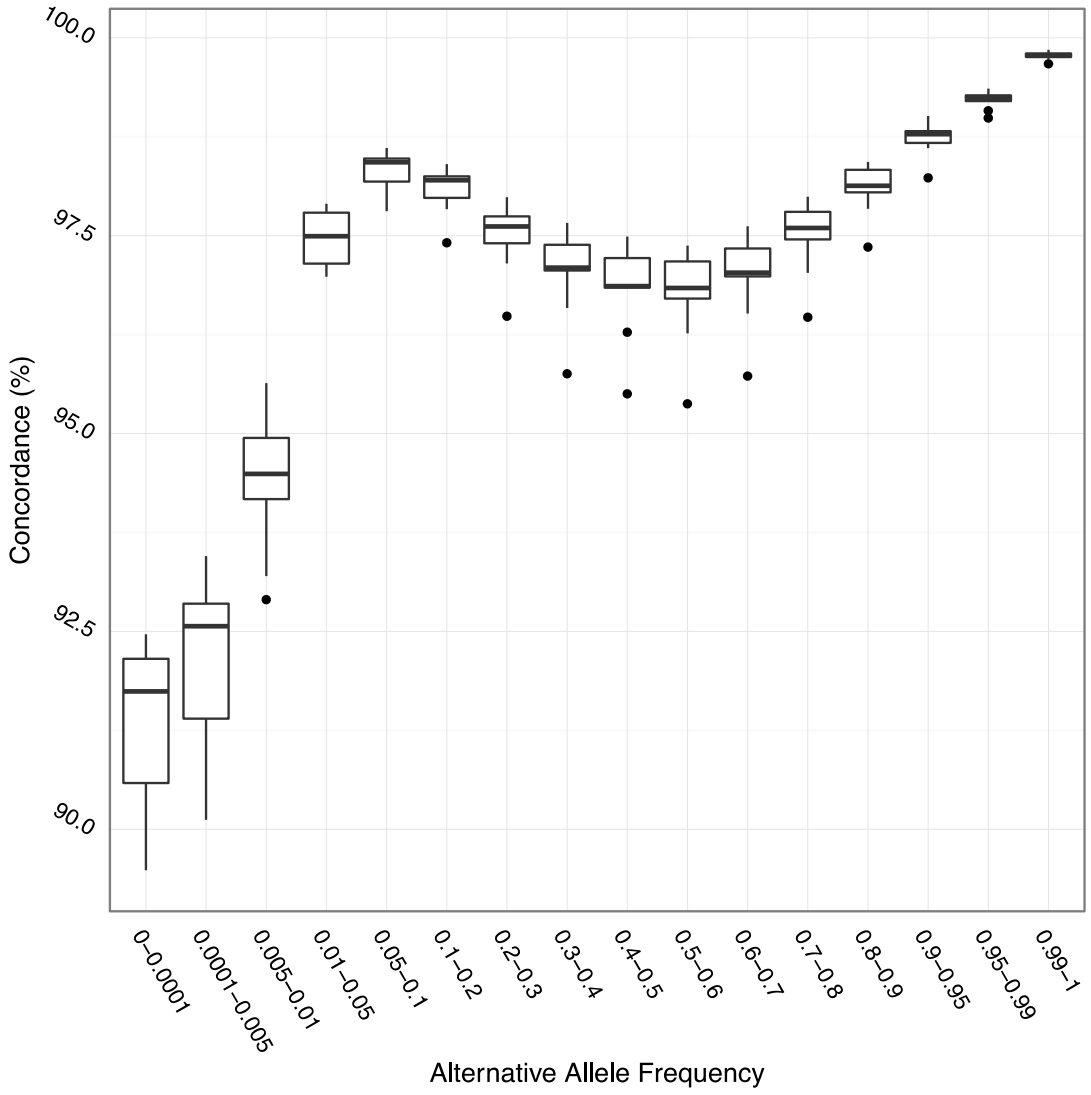


Figure 2.4 | Percentage concordance between hard-called genotypes from imputed genotype probabilities and genotypes called from 10x coverage sequencing in nine samples This figure shows the percentage concordance between genotypes from imputation (hard-called as the genotype with the maximum imputed genotype probability, barring those where the maximum genotypes probability was smaller than 0.9 in which case the genotypes was considered missing) and genotypes directly called from 10x coverage sequencing data in nine samples. Each box contains data points for nine samples; the vertical axis shows the percentage of SNPs per sample in each alternative allele frequency bracket (horizontal axis) with concordant genotypes between the two datasets.

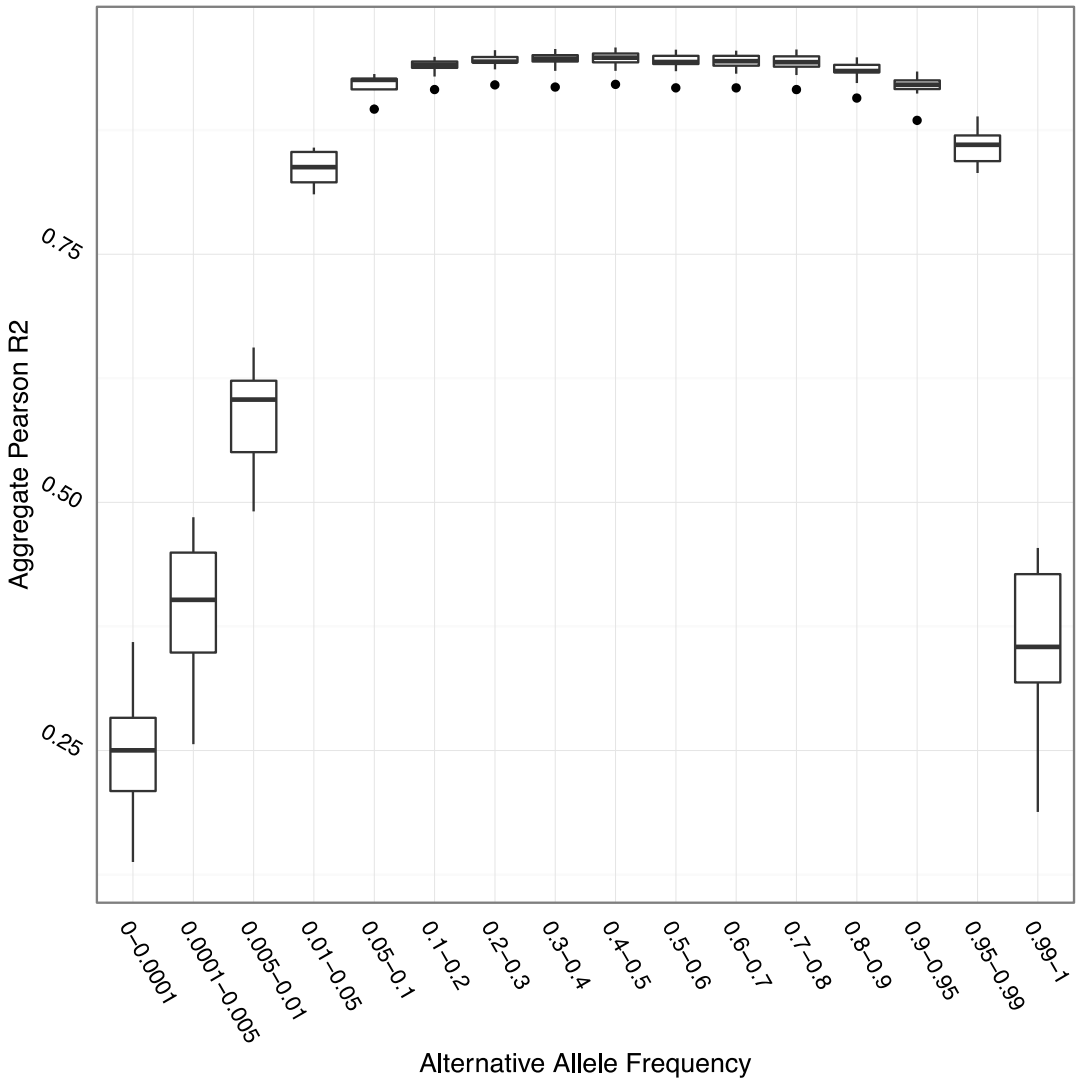


Figure 2.5 | Aggregate Pearson r^2 between imputed allele dosages and genotypes called from 10x coverage sequencing in nine samples This figure shows the aggregate Pearson r^2 between imputed alternative allele dosages with genotypes called directly from 10x coverage sequencing data in nine samples. Each box contains data points for nine samples; the vertical axis shows aggregate Pearson r^2 of all SNPs in each alternative allele frequency bracket (horizontal axis) with concordant genotypes between that imputed alternative allele dosages and directly called from 10x coverage sequencing data. The lower concordance this figure shows as compared to that in Figure 2.7 is due to higher missingness among the nine samples as compared to 72 samples genotyped on the Zhonghua BeadChip.

Alternative Allele Frequency Bin	Homozygous Reference		Heterozygous		Homozygous Alternative		All SNPs				
	Mean SNPs	Mean Con (%)	Mean SNPs	Mean Con (%)	Mean SNPs	Mean Con (%)	N SNPs	Mean Con (%)	SE Con (%)	Mean r^2	SE r^2
0 to 1e-04	7971	100.00	995	22.86	16	0.00	8983	91.3	0.35	0.248	0.027
1e-04 to 0.005	111799	99.98	14222	31.76	293	0.79	126315	92.07	0.38	0.385	0.027
0.005 to 0.01	74457	99.93	9933	54.98	240	4.03	84630	94.38	0.29	0.58	0.020
0.01 to 0.05	399671	99.65	64629	86.32	2610	40.85	466910	97.46	0.12	0.834	0.006
0.05 to 0.1	400397	99.44	92770	95.69	6292	67.11	499459	98.33	0.08	0.921	0.004
0.1 to 0.2	608665	99.08	226295	97.40	27668	81.86	862627	97.51	0.10	0.944	0.003
0.2 to 0.3	384386	98.41	249721	97.91	52223	89.03	686329	97.05	0.15	0.944	0.004
0.3 to 0.4	238076	97.54	244163	98.04	77195	92.44	559434	96.83	0.19	0.945	0.004
0.4 to 0.5	149370	96.46	227250	98.07	105556	94.72	482177	96.76	0.20	0.943	0.004
0.5 to 0.6	83882	94.86	194108	98.03	128809	96.11	406799	96.99	0.21	0.942	0.004
0.6 to 0.7	45250	92.66	154084	98.00	152864	97.28	352198	97.5	0.19	0.942	0.004
0.7 to 0.8	21825	89.35	113565	97.83	182797	98.27	318188	98.1	0.16	0.934	0.004
0.8 to 0.9	7450	82.02	68168	97.27	203198	98.97	278817	98.74	0.11	0.918	0.005
0.9 to 0.95	1083	68.11	17749	96.14	115843	99.43	134675	99.21	0.08	0.86	0.006
0.95 to 0.99	197	42.75	6364	91.25	124004	99.71	130566	99.78	0.04	0.356	0.027
0.99 to 1	13	8.39	1040	37.85	325044	99.98	326096	97.53	0.02	0.967	0.002
0 to 1	2534492	98.49	1685057	96.37	1504654	97.21	5724203	98.08	0.12	0.938	0.003

Table 2.4 | Validation of imputation results with 10x-coverage whole-genome sequencing data on 9 samples. The table shows concordance between SNP genotypes imputed from low coverage sequence (median 1.7X) and from 10X coverage. Percentage concordances (Mean Con (%)) are presented for each genotype class (Homozygous Reference, Heterozygous and Homozygous Alternative), as well as for all genotype classes (All SNPs) along with the mean number of SNPs (Mean SNPs) per sample in each genotype class for all Alternative Allele Frequency Bins. Mean aggregate Pearson r^2 (Mean r2) between all imputed allele dosages and genotypes at all sites (All SNPs) were also computed for each sample at alternative allele frequency bins. The table also shows the standard error (SE) for percentage concordance and Pearson r^2 computed for SNPs in all genotype classes (All SNPs) among the 9 samples for each alternative allele frequency bin.

2.4.2 HumanOmniZhongHua-8 (v1.0) BeadChip on 72 samples

72 samples from CONVERGE were genotyped on the Illumina HumanOmniZhongHua-8 (v1.0B) BeadChip which contains 898,122 SNPs, optimized for the Chinese population using tag SNPs from all three HapMap phases and the 1000 Genomes Project. Sample genotypes with GenCall score of 0.15 or lower were considered no-calls.

857,766 out of the 895,623 SNPs on the array were included in our imputation (95.77%). Upon examining why 37,857 SNPs on the array were not imputed, I found that SNPs that were on the array, but not imputed, were not polymorphic in either CONVERGE samples or 1000 Genomes Phase 1 ASN panel. None of the 72 samples genotyped were polymorphic at these 37,857 sites. Table 2.5 shows the number of SNPs genotyped on the Zhonghua array on each chromosome, and the call rate among the 72 samples on each chromosome.

855,132 SNPs that showed biallelic polymorphism in CONVERGE or the 1000 Genomes Phase1 ASN Panel and passed genotyping quality filters were used for comparison with imputation results. I calculated the percentage concordance and Pearson r^2 between hard-called genotypes and allele dosages from imputation with genotype calls from the ZhongHua8 genotyping array for all 855,132 SNPs on 72 samples. These are shown in Figures 2.6 and 2.7. The overall percentage concordance for all 855,132 imputed SNPs genotyped on the array was 98.0% (standard error = 0.04%) and the overall r^2 was 0.974 (standard error = 0.0005).

Chromosome	All Array SNPs	No Calls	Not Polymorphic	Used For Comparison
chr1	69485	5	3199	66126
chr2	70071	3	2958	66928
chr3	60192	7	2608	57516
chr4	53419	1	2478	50807
chr5	52477	1	2142	50307
chr6	62652	15	2555	59930
chr7	47158	4	1909	45157
chr8	45906	0	1824	44062
chr9	40597	0	1521	38996
chr10	46134	1	1971	44131
chr11	42894	1	1971	40888
chr12	42849	3	1822	40950
chr13	32399	1	1451	30931
chr14	28777	1	1167	27594
chr15	27769	3	1099	26647
chr16	29448	3	1276	28146
chr17	25411	1	1145	24252
chr18	27117	1	1104	26003
chr19	18919	4	956	17942
chr20	21918	2	843	21057
chr21	12613	1	500	12103
chr22	14056	1	480	13538
chrX	23362	1	869	21121
chrY	2041	1208	833	0
chrM	151	0	151	0
chrXY	307	307	0	0
Total	898122	1575	38832	855132

Table 2.5 | SNPs genotyped per chromosome on Illumina HumanOmniZhongHua Beadchip. This table shows in the first two columns the chromosome and the number of SNPs on each chromosome on the Illumina HumanOmniZhongHua Beadchip. In the next three columns are the number of SNPs with no calls in any of the 72 samples, that are not polymorphic in the 72 samples, and that were polymorphic and genotypes at which were used for comparison with genotypes hard-called from imputed genotype probabilities and imputed alternative allele dosages.

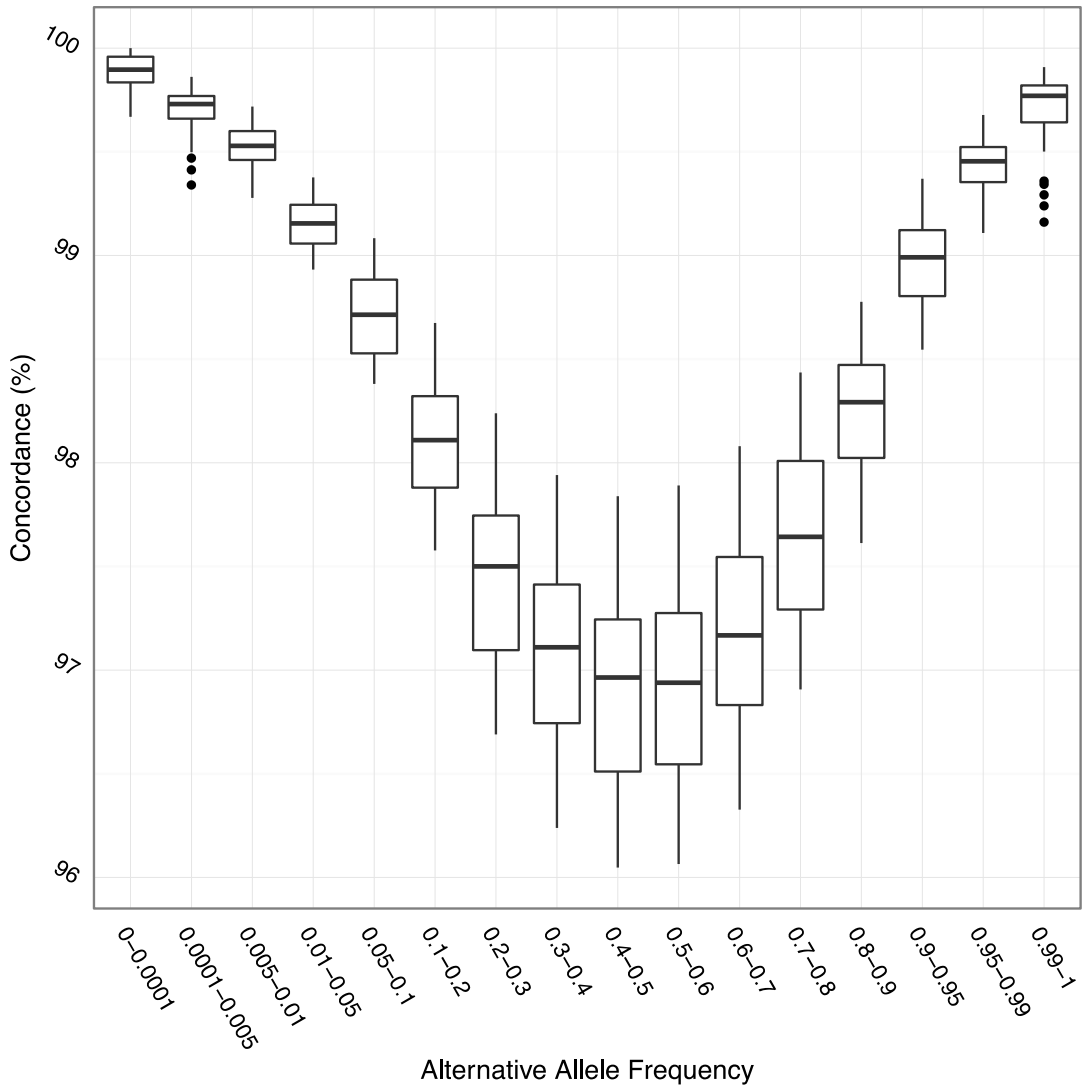


Figure 2.6 | Percentage concordance between hard-called genotypes from imputed genotype probabilities and genotypes from the Illumina HumanOmniZhongHua Beadchip in 72 samples. This figure shows the percentage concordance between genotypes from imputation (hard-called as the genotype with the maximum imputed genotype probability, barring those where the maximum genotypes probability was smaller than 0.9 in which case the genotypes was considered missing) and genotypes directly called from Illumina HumanOmniZhongHua Beadchip in 72 samples . Each box contains data points for nine samples; the vertical axis shows the percentage of SNPs per sample in each alternative allele frequency bracket (horizontal axis) with concordant genotypes between the two datasets.

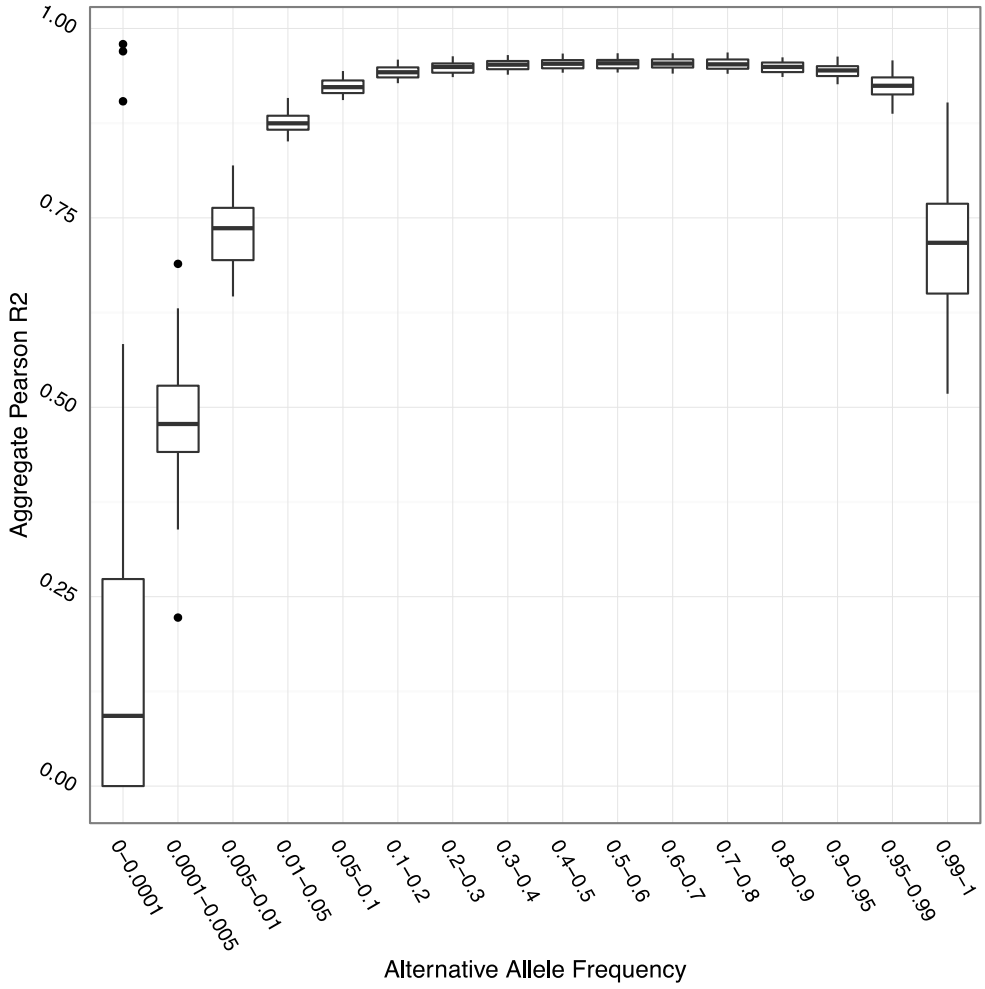


Figure 2.7 | Aggregate Pearson r² between imputed allele dosages and genotypes called from Illumina HumanOmniZhongHua Beadchip in 72 samples. This figure shows the aggregate Pearson r² between imputed alternative allele dosages with genotypes called directly from 10x coverage sequencing data in nine samples. Each box contains data points for nine samples; the vertical axis shows aggregate Pearson r² of all SNPs in each alternative allele frequency bracket (horizontal axis) with concordant genotypes between that imputed alternative allele dosages and directly called from Illumina HumanOmniZhongHua Beadchip in 72 samples.

Table 2.6 shows the mean percentage concordance and Pearson R^2 among 72 samples for each genotype class (homozygous reference, heterozygous and homozygous alternative) for each alternative allele frequency bin. At all alternative allele frequency bins between 1 to 99%, the mean percentage concordances were above 96% and the mean R^2s were above 0.87.

Alternative Allele Frequency Bin	Homozygous Reference		Heterozygous		Homozygous Alternative		All SNPs				
	Mean SNPs	Mean Con (%)	Mean SNPs	Mean Con (%)	Mean SNPs	Mean Con (%)	Mean SNPs	Mean Con (%)	SE Con (%)	Mean r^2	SE r^2
0 to 1e-04	2410	100.0	3	13.1	0	0.0	2419	99.9	0.01	0.167	0.0269
1e-04 to 0.005	20794	100.0	111	49.6	3	4.5	20962	99.7	0.01	0.484	0.0092
0.005 to 0.01	12943	99.9	196	75.2	3	20.3	13179	99.5	0.01	0.730	0.0054
0.01 to 0.05	78562	99.7	4855	90.5	115	65.1	83806	99.2	0.01	0.876	0.0017
0.05 to 0.1	76110	99.5	12319	94.3	566	84.2	89313	98.7	0.02	0.923	0.0012
0.1 to 0.2	100316	99.2	34844	95.6	3211	91.7	138909	98.1	0.04	0.943	0.0010
0.2 to 0.3	59593	98.7	39424	96.2	6787	94.1	106138	97.5	0.05	0.949	0.0009
0.3 to 0.4	36176	98.2	38685	96.5	10582	95.5	85774	97.1	0.05	0.952	0.0009
0.4 to 0.5	21760	97.7	35205	96.7	14610	96.5	71809	96.9	0.06	0.954	0.0008
0.5 to 0.6	12263	97.0	29676	96.7	18278	97.3	60405	96.9	0.06	0.954	0.0008
0.6 to 0.7	6557	96.3	23847	96.7	22202	98.0	52845	97.2	0.05	0.954	0.0008
0.7 to 0.8	2987	95.4	17367	96.6	26037	98.6	46626	97.7	0.05	0.953	0.0009
0.8 to 0.9	1008	93.8	10545	96.1	29809	99.2	41566	98.3	0.04	0.949	0.0009
0.9 to 0.95	125	91.9	2696	95.6	16537	99.6	19443	99.0	0.02	0.944	0.0010
0.95 to 0.99	16	81.8	859	94.0	13362	99.8	14292	99.4	0.02	0.923	0.0017
0.99 to 1	0	27.4	62	70.8	7546	100.0	7646	99.7	0.02	0.706	0.0103
0 to 1	431622	99.0	250693	96.1	169644	98.0	855132	98.0	0.04	0.974	0.0005

Table 2.6 | Comparison of genotypes from a commercial genotyping array with genotypes imputed from low coverage sequence of 72 CONVERGE samples. The table shows concordance between SNP genotypes from low coverage sequence data and from a commercial genotyping array (Illumina ZhongHua-8 BeadCHIP). Percentage concordance (Mean Con (%)) was calculated between hard-called genotypes from imputed genotype probabilities (where the genotype with the maximum imputed genotype probability > 0.9 was called) and genotypes called from the same samples at the same loci from the array data. The table shows the mean percentage concordance and the mean aggregate r^2 (Mean r2) among 72 samples at each Alternative Allele Frequency Bin, along with the mean number of SNPs (Mean SNPs) per sample in each genotype class for all alternative allele frequency bins. The table also shows the standard error (SE) for percentage concordance and Pearson r^2 computed for SNPs in all genotype classes (All SNPs) among the 72 samples for each alternative allele frequency bin.

2.4.3 Sequenom at 21 sites

11,669 samples out of all CONVERGE samples were genotyped at 21 random sites in the genome using a custom Sequenom SpectroCHIP (1 sample was not genotyped as there was not enough DNA of adequate quality for genotyping on the Sequenom platform). Mass spectra were collected using the MassARRAY™ system mass spectrometer and TYPER4.0 was used to assess the reliability of genotype calls generated by SpectroREAD from the mass spectra. Default genotype call inclusion criteria were used.

In the same way I compared imputed results with genotypes called directly from 10x coverage sequencing and the Illumina HumanOmniZhongHua Beadchip, I calculated the percentage concordance and r^2 between hard-called genotypes and allele dosages from imputation with genotype calls from the Sequenom for each of the 21 sites that were genotyped. An r^2 of 0.988 and concordance of 98.19% were obtained for the 21 sites. Per site r^2 (mean = 0.985) and percentage concordance (mean = 98.16%) are shown in Table 2.7 along with the alleles and alternative allele frequency of each site.

SNP	RSID	REF	ALT	FREQ	N Samples	Con (%)	Pearson r^2
chr1:11205058	rs1057079	C	T	0.8	11645	99.85	0.998
chr1:47398743	rs3890011	G	C	0.525	11625	99.95	0.999
chr2:141751592	rs13007735	G	A	0.595	11652	99.8	0.998
chr2:204824283	rs10172036	T	G	0.434	9561	94.69	0.957
chr2:99779131	rs2516835	T	C	0.651	11533	99.58	0.997
chr3:186443018	rs1656922	T	C	0.356	11634	99.64	0.996
chr7:47968927	rs2686817	C	A	0.522	11621	99.09	0.992
chr8:143310815	rs11167136	G	A	0.498	11618	98.61	0.987
chr9:125424507	rs70156	A	C	0.5	11632	94.8	0.962
chr10:120917445	rs2275111	G	A	0.707	11651	99.54	0.995
chr10:95279506	rs2293277	A	T	0.565	11506	95.62	0.966
chr12:120995332	rs2292681	G	A	0.767	11653	99.88	0.998
chr13:31233063	rs3742302	G	A	0.153	11651	99.91	0.998
chr14:20665840	rs4981088	G	A	0.507	11508	93.92	0.954
chr15:77344793	rs11737	T	A	0.415	11628	99.69	0.997
chr15:90226947	rs7169981	C	A	0.317	11615	98.92	0.99
chr16:20986506	rs3115438	C	T	0.481	11423	99.79	0.998
chr17:5991344	rs2302836	C	T	0.829	11652	96.52	0.955
chr18:61170721	rs1455556	T	C	0.546	11631	97.73	0.983
chr20:50238545	rs2235862	A	G	0.237	11649	99.8	0.997
chr20:52786219	rs2296241	G	A	0.417	11638	93.99	0.96

Table 2.7 | Validation of imputation results with genotypes at 21 sites on custom Sequenom SpectroCHIP on all samples. The table shows concordance between SNP genotypes from low coverage sequence data and from 21 sites genotyped on a Sequenom SpectroCHIP on all samples. The first five columns show the chromosome and position (SNP), reference allele (REF) on Human Genome Reference GRCh37.p5 and alternative allele (ALT) called in CONVERGE, and the alternative allele frequency (FREQ) in CONVERGE. The next column shows the number of samples (N Samples) with genotypes from Sequenom at each of the 21 sites. The next two columns show the comparison between imputed allele dosages and genotypes from Sequenom at the 21 sites: percentage concordance (Con (%)) was calculated per site between hard-called genotypes from imputed genotype probabilities (where the genotype with the maximum imputed genotype probability > 0.9 was called) and genotypes called from the same samples at the same loci from Sequenom. Pearson r^2 was also computed per site between imputed allele dosages.

2.5 DNA contamination

In addition to ensuring accurate genotypes a series of quality control checks on the subjects are required before genetic association between variants and phenotypes can be undertaken. Standard quality control steps to be taken before association testing included: assessments of DNA sample contamination, sample relatedness, and sample stratification due to different ancestry

DNA for CONVERGE samples was extracted from saliva from all cases and controls. DNA extracted from saliva is contaminated with DNA from other sources, for example microbial DNA contamination. Assuming the number of private mutations in an individual lies within a range specific for the population, a large excess of private mutations would be suggestive of higher levels of DNA contamination. Along the same vein, mitochondrial DNA (mtDNA) is passed through the maternal lineage and, due to the replication bottleneck in the female germline, usually inherited as clonal copies. Variations in the mtDNA identified in a single individual (heteroplasmic variations) are not uncommon, and usually due to the highly mutagenic environment as well as the lack of extensive DNA repair mechanisms in the mitochondrion. A large excess of heteroplasmic mutations, however, would be suggestive of DNA contamination, especially since the mtDNA may be closer in sequence to microbial genomes due to its hypothesized bacterial origin. As NGS allows the detection of private mutations in every individual in the nuclear genome as well as heteroplasmies in the mtDNA, I used this information for identification of potentially contaminated samples.

2.5.1 Excess singletons

I first looked into both the nuclear genome and mitochondrial genome for an excess of variants called, since this would indicate cross-sample contamination due to technical issues during sequencing. I quantified the number of singleton variants called in the

genic regions of the nuclear genome (Methods) and found a mean of 71.55 singletons variants in exon regions across the whole genome supported by more than 2 sequencing reads passing sequencing quality controls per sample. Figure 2.8 shows the distribution of number of singletons detected in exonic regions across the genome from low-coverage sequencing from each sample, and while most samples fall in a normal distribution with mean of 71.55 singletons, others had excess singletons, which could be indicative of poor DNA quality that compromised sequencing and mapping qualities. I excluded 117 samples with a number of singletons greater than the 99th percentile (237.62 singletons).

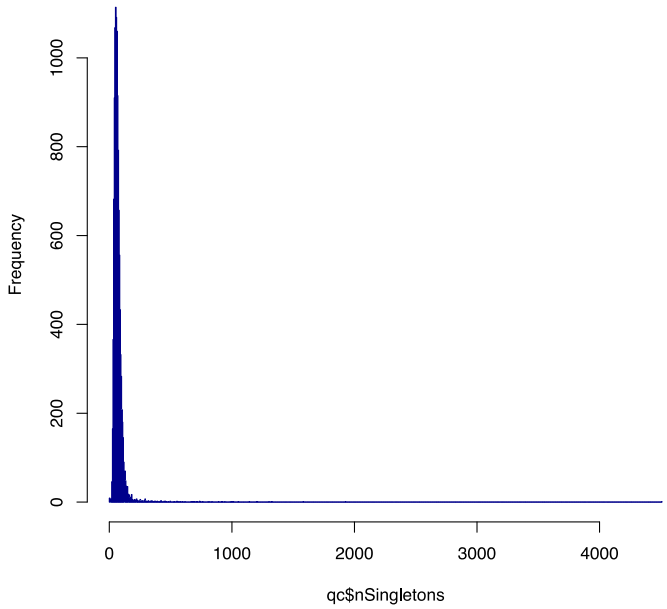


Figure 2.8 | Distribution of number of singletons in 11,670 CONVERGE samples. This figure shows the distribution of number of singletons called in exonic regions across the whole chromosomes supported by two or more high quality reads from low-coverage sequencing (mapping quality > 59, base quality > 20) in each of the 11,670 CONVERGE samples; the mean number of singletons called per sample was 71.55; 117 samples with number of singletons greater than the 99th percentile (237.62 singletons) were excluded from further analyses.

2.5.2 Excess mitochondrial heteroplasmy

I then asked if there were any samples with an excess of heteroplasmic mutations in the mitochondrial DNA per sample. As the DNA for all individuals recruited in CONVERGE were extracted from saliva samples, which would also contain the oral microbiome, external DNA contamination levels could be told from number of mismatches in mapped reads at regions of the genomes where the human genome was similar to bacterial genomes. The mitochondrial DNA in human genome, a 16kb circular DNA, is one of such instances.

Mitochondrial DNA (mtDNA) is inherited only from the mother and there is a strict bottleneck in the female germline for mtDNA, usually only one clonal copy of mitochondrial DNA is inherited at a time. Variants in the mitochondrial genome, therefore, are likely to occur between samples (homoplasmic variation) but not within samples, unless by chance, albeit a much slimmer chance, mtDNA containing different alleles at the same variant site was inherited from the mother in a single individual (heteroplasmic variation). Mismatches that occur in some sequencing reads mapping to the mtDNA region of the human genome reference in one sample would therefore indicate either this rare event of inherited heteroplasmy, the more likely accumulation of mutation in the mtDNA during the individual's lifetime, or a contamination from similar DNA sequences of bacterial origins. Samples with excessive mtDNA heteroplasmies can therefore be suspected as having a higher than normal level of bacterial contamination, and the conservative approach would be to filter them out to ensure the quality of sequencing, and all downstream analyses. The following example demonstrates the value of using mtDNA sequence to identify potential contaminants and sample mix-ups.

Figure 2.9 shows a screenshot of three samples' mtDNA reads as visualized in the integrated genome viewer[121] (IGV). Grey bars show the 100bp long sequencing reads, and each of the four types of coloured bands on the grey bars (red, green, yellow and blue) correspond to an allele at a site which is identified to be variant from the human reference genome (build hg19). Naturally occurring, bona fide SNPs usually occur once in every few hundred bases, so it not surprising to see one or two SNPs in a sequencing read. However, numerous variants on a single sequencing read and complete alignment of multiple reads with exactly the same variants in a single sample and across multiple samples suggest these are not variants from the samples, but homologous regions of the genome of a contaminant that differed from the human reference (as well as all of the samples) at the same places.

To identify the contaminant I took sequence from the reads harbouring the multiple variants, and searched for matching sequences in the National Centre for Biotechnology Information (NCBI) database, using the BLAST programme[122]. Figure 2.10 shows the BLAST results for one such read taken from one of the samples in Figure 2.9; the read came from the mtDNA of *Sus scrofa* of the Berkshire breed (domestic pig). Since the DNA samples collected from the recruited subjects were from saliva, and multiple samples showed the same contaminant DNA, it seemed I was able to forensically determine what the subjects had for meals prior to collection of their DNA.

This finding alerted me to the utility in mtDNA alignments as a good place for detecting contaminants. Due to the fact that 93% of human mtDNA is coding sequence, together with the presence of relatively high sequence homology between species, then most contaminants are likely to align to the human mtDNA reference even if they are from non-human sources. Such alignments, albeit containing many mismatches to the human mtDNA reference, would passing stringent filtering criteria based on mapping qualities, which penalizes heavily on repeat structures, poor base qualities and

unpairing of paired-end reads in alignment, but not on mismatches, especially not those that do not cause gaps in the alignments.

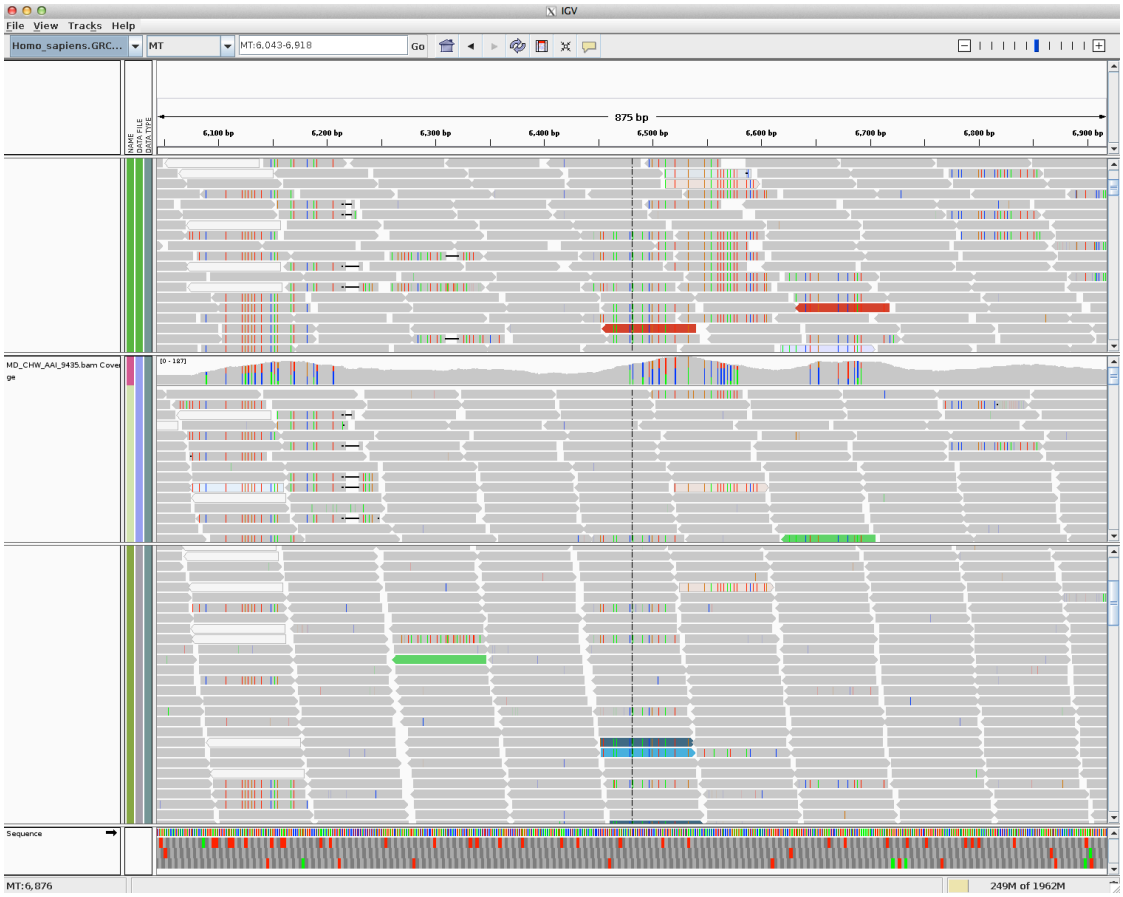


Figure 2.9 | IGV visualization of mtDNA reads from NGS on three samples in CONVERGE. This figure shows a screenshot of three samples' mtDNA reads as visualized in the IGV. Grey bars show the 100bp long sequencing reads, and each of the four types of coloured bands on the grey bars (red, green, yellow and blue) correspond to an allele at a site which is identified to be variant from the human reference genome (build hg19).

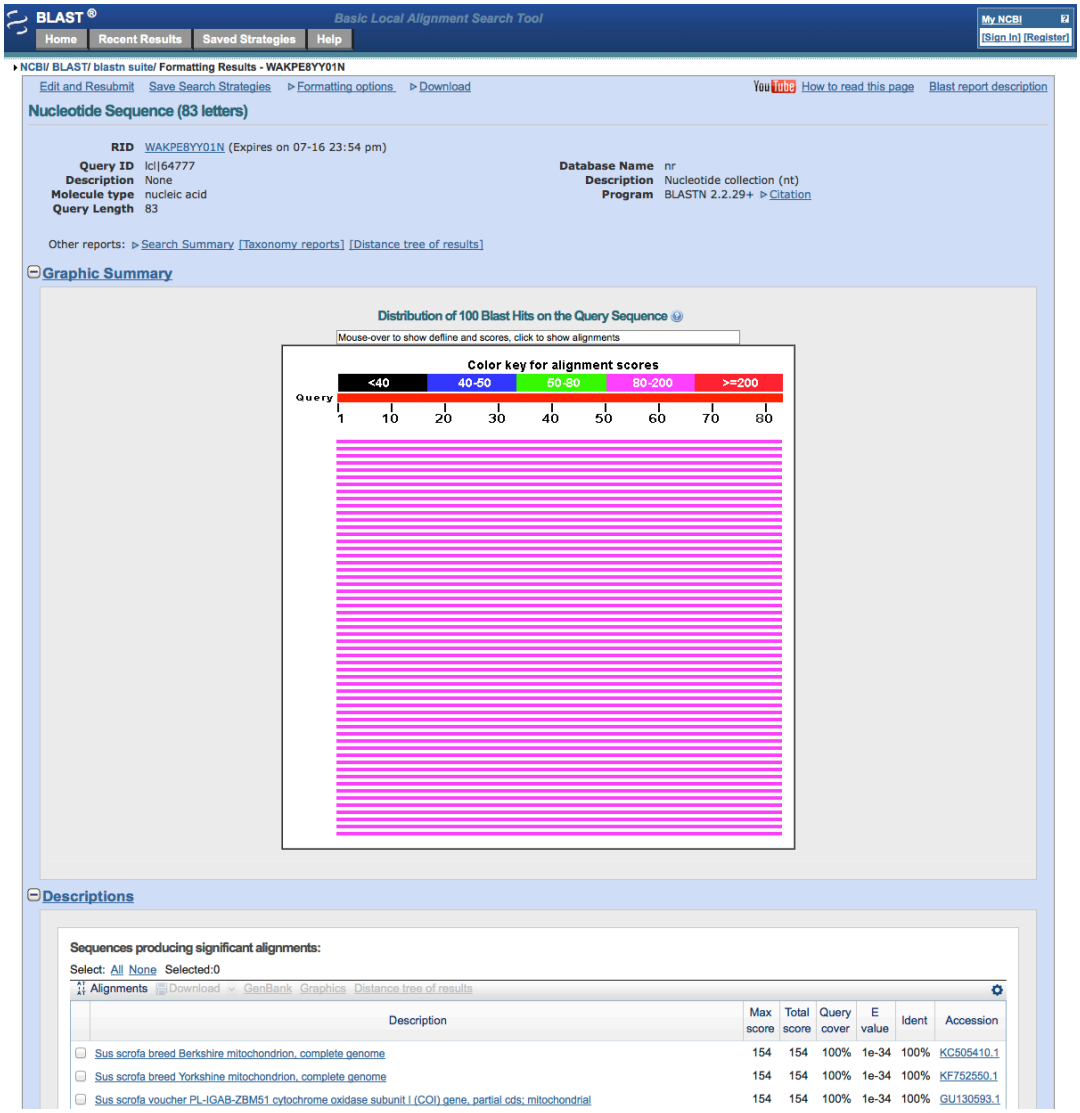


Figure 2.10 | BLAST results of sequence match for aberrant read mapping to mtDNA reference in one sample in CONVERGE. This figure shows a screenshot of the result of a BLAST search for sequence matches between one aberrant read mapping to human mtDNA reference in one sample in CONVERGE that contained numerous sequence variants. A 100% sequence similarity was found between the input sequencing read and a sequence from the mitochondrial genome of *Sus scrofa* breed Berkshire.

Mismatches that are informative of such contaminations are disguised as heteroplasmic variations (non-clonal variations in mtDNA within a single sample). Heteroplasmic variations are discussed in detail in Chapter 5 of this thesis for its significance in stress and MDD, but briefly, it is common and naturally occurring at a low level in almost every individual as a result of *de novo* mutation in the mtDNA. As

coverage of the mtDNA is on average 102X in CONVERGE, I was able to identify heteroplasmic mutations from every sample, and pick out those that deviate significantly from the rest.

I called a mean of 15.70 heteroplasmic sites per sample, and 116 samples were found to have greater than the 99th percentile of number of heteroplasmic sites (Figure 2.11). Of these 116 samples, 26 were already discarded for having excess nuclear singletons; I excluded the additional 90 from further analysis. This points to the added sensitivity at detecting contaminated samples with the examining of reads mapping to the mtDNA reference sequence in every sample.

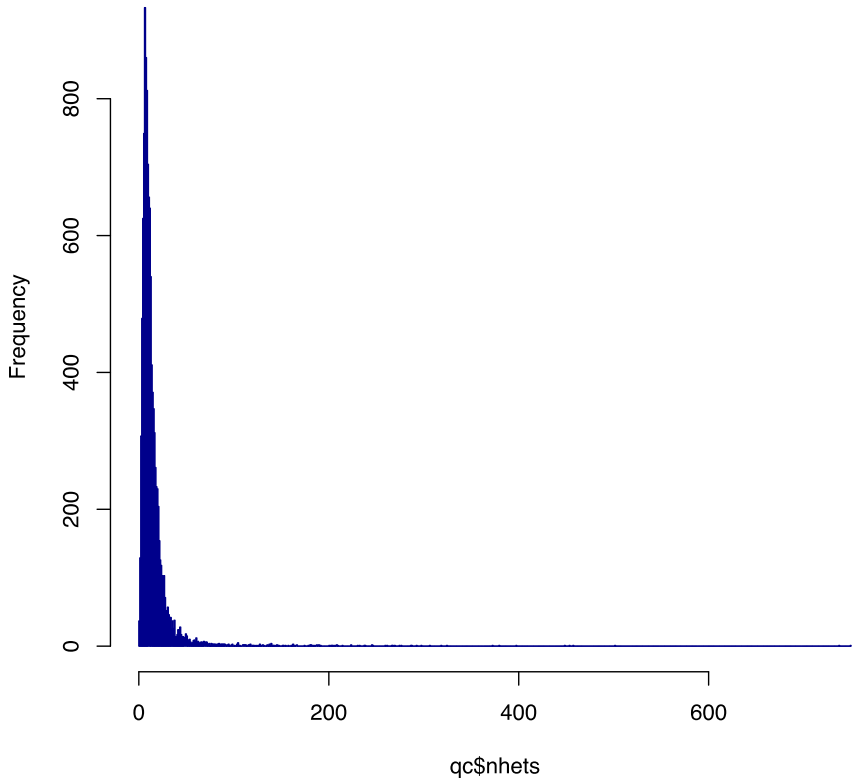


Figure 2.11 | Distribution of number of singletons in 11,670 CONVERGE samples. This figure shows the distribution of number of singletons called in exonic regions across the whole chromosomes supported by two or more high quality reads from low-coverage sequencing (mapping quality > 59, base quality > 20) in each of the 11,670 CONVERGE samples; the mean number of singletons called per sample was 71.55; 117 samples with number of singletons greater than the 99th percentile (237.62 singletons) were excluded from further analyses.

2.6 Sample relatedness

The 11,670 samples in CONVERGE were also assessed for relatedness. Varying levels of relatedness could cause inflated P-values and even false positive results in GWAS, as alleles inherited together due to high relatedness but not related to etiology of disease could appear to be so in association testing. Using standard means of estimating identity by descent (IBD) by calculating identity by state (IBS) using tagging markers in LD from the whole genome, I estimated the level of sharing between each pair of samples, and excluded all samples who had extensive sharing with any other sample, to prevent such inflations in GWAS.

Although being unrelated to other individuals recruited for the CONVERGE study was a clear criterion in our data collection process, there were instances where the same patient or a relative of the patient visited multiple hospitals and was thus recruited more than once. To exclude duplicates and first degree relatives from our sample for GWAS, I estimated pairwise genome-wide Identity by Descent (IBD) by calculating Identity by State (IBS) from hard-called genotypes from imputed genotype probabilities at 399,211 common, tagging SNPs across all autosomes (MAF >1%, LD <0.5, all known in 1000 Genomes Phase 1). I implemented this in PLINK (v1.07)[123] with the option –genome.

There were samples who had excessive IBD sharing with one or more other samples in the cohort - I calculated the proportion IBD (PIHAT, estimated as P(IBD=2)+0.5*P(IBD=1)) for every pair of samples in CONVERGE, a distribution of the maximum PIHAT between every sample and all other samples in CONVERGE is shown in Figure 2.12. While close relatives like siblings or individuals from a highly inbred population would have high PIHAT due to substantial homozygous sharing P(IBD=2), other close relatives such as parent-child pairs in an outbred population would not necessarily have a high degree of homozygous sharing but a large degree of

heterozygous sharing P(IBD=1). As such, I excluded all pairs of samples whose PIHAT exceeded 0.5 and P(IBD=1) exceeded 0.75 (removing therefore duplicates and first degree-relatives), totaling 392 samples, from the final set of samples for GWAS.

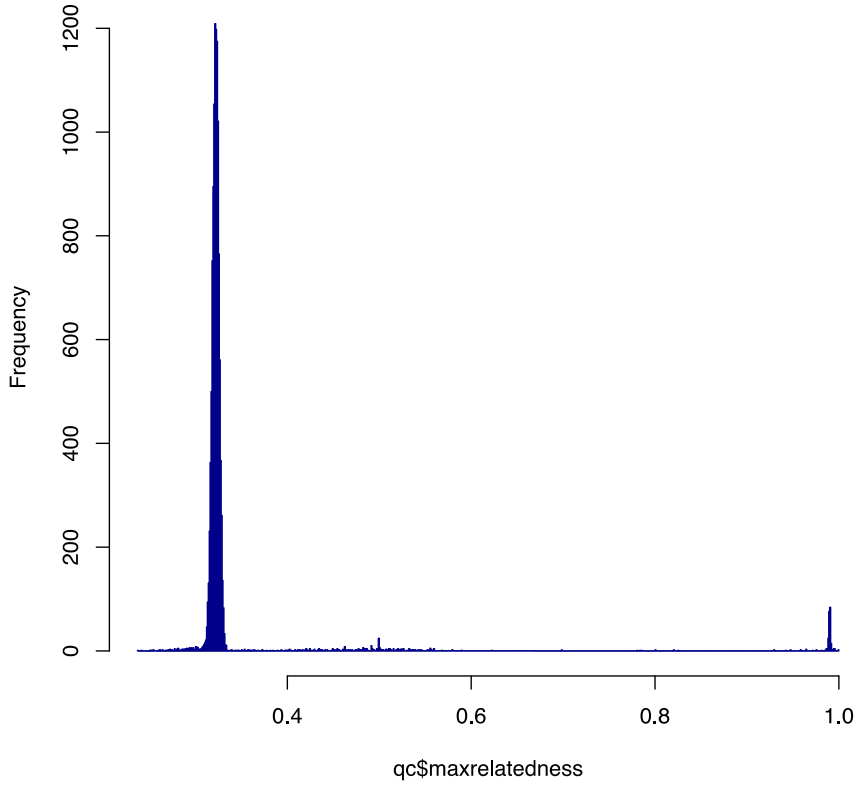


Figure 2.12 | Distribution of maximum relatedness between each sample and all other samples among the 11,670 CONVERGE samples. This figure shows the distribution of maximum PIHAT (proportion IBD, estimated as P(IBD=2)+0.5*P(IBD=1)) between each sample with all other samples in CONVERGE.

Apart from related samples, I also excluded 29 samples who had fewer than 90% of their sites with maximum genotype probabilities > 0.9, and 402 samples with incomplete phenotype information, giving a final set of 10,640 samples (5,303 cases of MDD, 5,337 controls) for GWAS on MDD. This information is summarized in Figure 2.13, which plots P(IBD=1), labeled Z1, against P(IBD=0), labeled Z0, where each point represents a pair of samples, and the colour of each point represents the reason this pair of samples were excluded from further analysis.

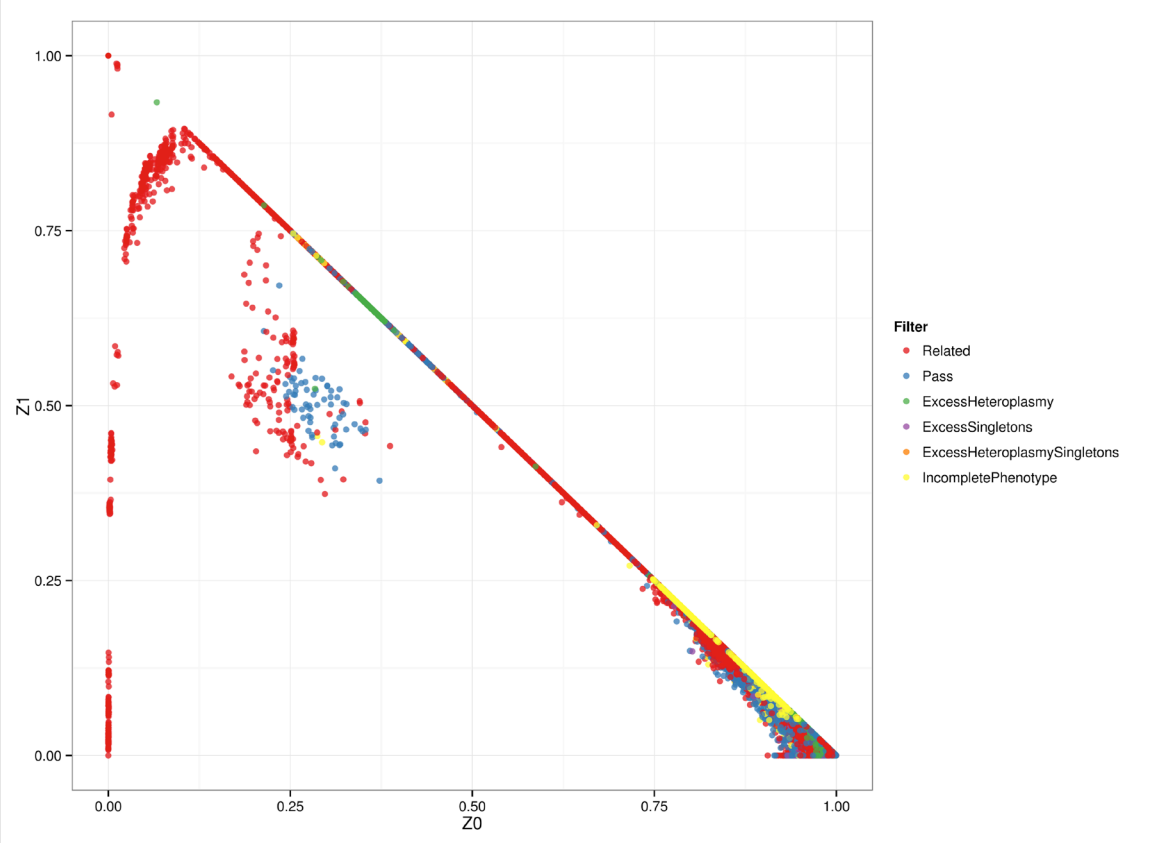


Figure 2.13 | Pairwise IBD of all samples in CONVERGE and reasons for sample exclusion. This figure plots P(IBD=1), labeled Z1, against P(IBD=0), labeled Z0, for all pairs of samples in CONVERGE. Each point represents a pair of samples, and the colour of points represent the reason this pair of samples were excluded from further analysis. In particular, pairwise IBD identified related individuals (shown in red), which were excluded from GWAS. The samples passing all quality control criteria (in blue, behind points of all other colours hence not obvious) total 10,640 samples, of which 5,303 were cases of MDD and 5,337 were controls.

2.7 Population Stratification

Population structure within GWAS samples may be confounded with the case control statuses of the samples, and hence cause false positive GWAS findings that in fact represent subtle or pronounced differences (depending on the level of confounding) between cases and controls in their ancestry. Identifying and adequately dealing with population structure remains a major technical challenge in GWAS studies and has been a reason why some have dismissed GWAS findings[124]. Here I describe the ways I used to deal with this problem.

Samples in CONVERGE were collected from 58 locations all across China, from the Northernmost province, Heilongjiang, to the Southernmost, Guangdong (Figure 2.14). Previous studies had shown that the North-South stratification in the Chinese population is striking and readily detected from genetic data, whereas East-West differences are much less discernable[125,126]. CONVERGE had collected approximately matching numbers of cases and controls from each province, so a high level of population stratification that would confound GWAS and cause false positive GWAS hits was not to be expected.

Figure 2.14 | Location of hospitals in China where CONVERGE samples were collected. Samples collected for the CONVERGE project were recruited from 58 hospitals across 30 cities in China; their locations are marked on the China map with red pins.

I studied the potential effects of population stratification on GWAS, and estimated the amount of stratification present in CONVERGE. I then assessed the different ways of controlling for population stratification in GWAS that are based on principal component analysis (PCA) [127] or linear-mixed model (LMM) approaches using a genetic relatedness matrix (GRM) constructed from tagging markers in LD from the whole genome as random effect.

2.7.1 Genomic Control

The prevailing method to evaluate the extent of population stratification is to compute the genomic control lambda (λ_{GC}), which is defined as the median χ^2 (1 degree of freedom) association statistic across SNPs divided by its theoretical median under the null distribution[128,129], where a λ_{GC} close to 1 (those less than 1.05 are usually considered benign) would indicate no stratification while one much larger than 1 would indicate so, or a variety of other confounds like family structure or cryptic relatedness. Inflation in

λ_{GC} is proportional to sample size, hence a variation of this value λ_{1000}, calculated as 1+(λ_{GC} -1) x (2000/N) where N is the sample size of the cohort, is used to evaluate the degree of inflation correcting for sample sizes. While genomic control is still widely used as a method to identify and evaluate the degree of population stratification in a GWAS cohort, is it no longer deemed adequate for correcting for it[130] due to the many causes for population stratification only some of which can be suitably addressed by dividing association statistics by λ_{GC} [131].

2.7.2 Principal component analysis

Methods that directly assess population structure through inferring genetic ancestry from whole-genome variations are emerging as effective ways to correct for population structure, among which obtaining top principal components (PCs) from principal component analysis (PCA) to be used as covariates in association testing[132] is a computationally tractable option for large, genome-wide datasets. When genome-wide data are not available, PCA can infer genetic ancestry using a small set of ancestry informative markers (AIMs) [133], though it does not provide as good an estimate of genetic ancestry as using a large number of random markers from the whole genome[130].

I therefore performed a PCA using the same 399,211 common, tagging SNPs across all autosomes (MAF >1%, LD <0.5, all known in 1000 Genomes Phase 1, computed using GCTA [134], Methods) as I had used for calculating pairwise relatedness between samples. Figure 2.15a shows all 10,640 unrelated samples in CONVERGE passing quality control criteria plotted by their top two PCs, coloured by the province in China they were from. PC1 very clearly captured the North-South cline in China, as the Northern-most provinces (Liaoning, Heilongjiang, Jilin) were on the high end of the range of PC1 while the Southern-most provinces (Guangdong, Fujian, Hainan) were on the low end.

The North-South structure is shown clearly in Figure 2.15b, where I have coloured the samples from the extreme provinces in all four directions differently from the others. I have also projected PC1 onto a map of China, as shown in Figure 2.15c, to visualize this. PC2, and in fact all other of the top 10 PCs, did not apparently capture geographical origins of the samples, which would be expected to mirror their genetic ancestry. This is in agreement with previous analyses of genetic ancestry among the Han Chinese people – analyses of whole-genome variation had resulted in clear distinctions between Northern and Southern China, but not between East and West [125,126]. While this could be due to the distribution of Han Chinese across China close to the East and particularly along the eastern coast of China, as well as the greater ethnic diversity in the West that would have caused greater extent of admixture that a careful study design would conceivably try to eliminate.

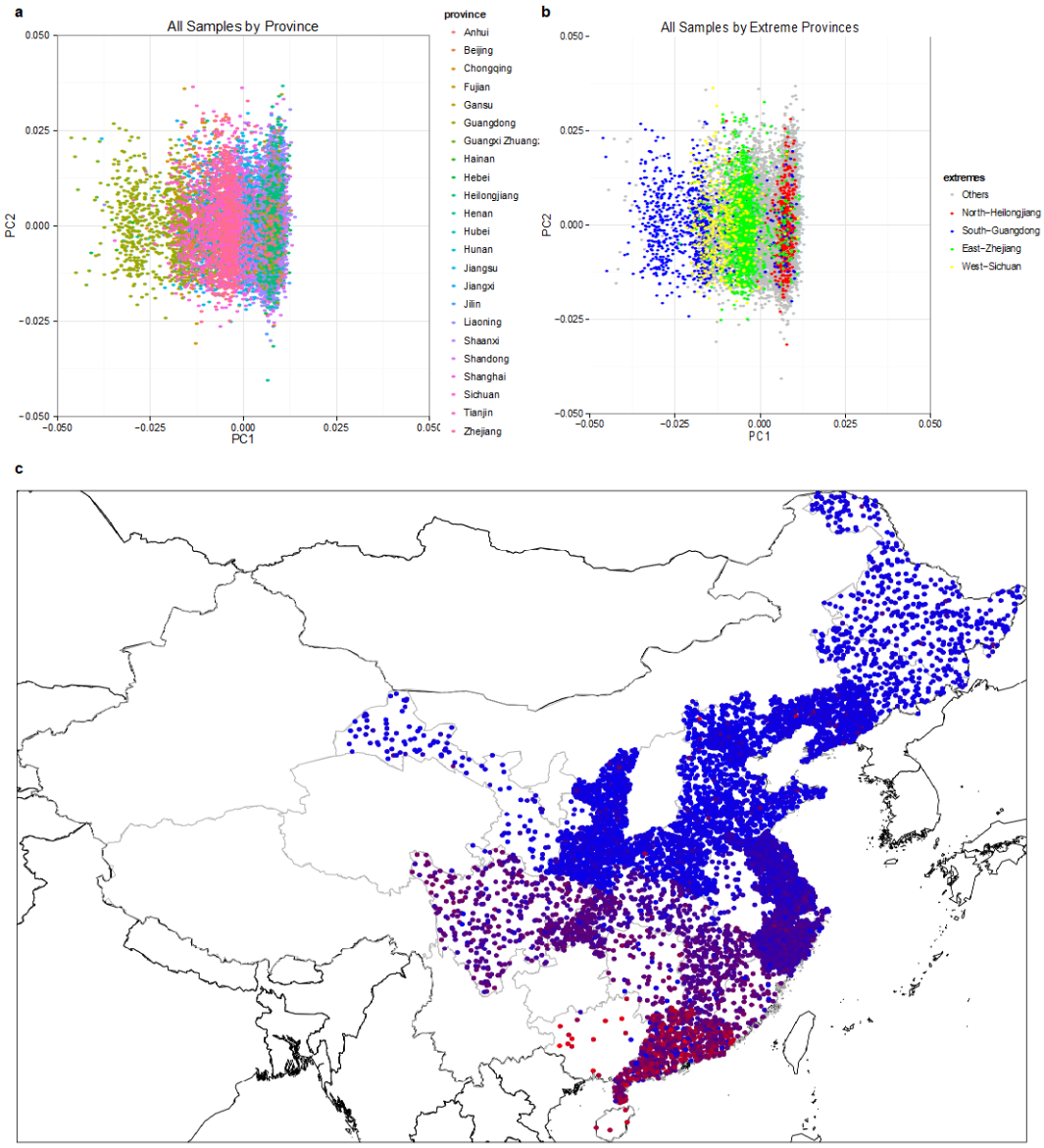


Figure 2.15 | Principal component analysis using whole-genome tagging markers on 10,640 CONVERGE samples. a. This figure shows all 10,640 unrelated samples in CONVERGE passing quality control criteria plotted by their top two PCs, coloured by the province in China they were from; b. the same plot with samples from the extreme provinces coloured different from all other samples; c. samples put on a map of China according to their provinces, and coloured by PC1. The map of China is obtained from gadm.org and plotting is done with R packages “maptools” and “sptools”.

2.8 Controlling for population structure in GWAS

2.8.1 Linear mixed model

The most efficient and conservative way to prevent false positive associations due to sample structure is to use a linear mixed model (LMM) with a covariance matrix summing the heritable and non-heritable random variation including PCs as fixed effects [130,135-139]. A number of authors have clearly demonstrated that MLMA is superior to the use of PCs [138,139]. EMMAX, for example, outperforms both principal component analysis and genomic control in correcting for sample structure. The correction for confounding is nearly perfect for common variants[140-143]. The only case where this does not hold is in the context of studies that include samples with strong familial relatedness [144]. This does not apply in this study since I had excluded any such relationships before running the association analysis.

It is known that the use of too few markers to generate the genetic relationship matrix (GRM) can compromise correction for stratification. Yang et al [145] demonstrate inflation in λ_{GC} that occurs when FaST-LMM[146] uses 4-8,000 markers to generate the GRM. However they also show that when using 100,000 markers to generate the GRM, λ_{GC} approaches 1.0, concluding that correction for confounding is nearly perfect. I had used 322,911 common SNPs (>1%) that are out of LD with each other (LD <0.5) from all autosomes, so the GRM generated should efficiently control even subtle population structure, as many reports have shown [130,135-142,144-150].

To be very conservative, I also included PCs in the association analysis, as fixed effects. This approach is recommended when using a GRM that has a relatively small set of markers and will efficiently remove population structure [151]. In this study, the inclusion is a conservative measure since, as stated above, the GRM was created with 322,911 markers.

2.8.2 Comparison of methods for controlling of population structure

To explicitly test whether I had adequately corrected for population structure between samples in CONVERGE when running the GWAS on MDD, I conducted an experiment with subsets of cases and controls from North and South of China in varying proportions to demonstrate the effects of different levels of population stratification on false positive findings from GWAS. I asked, in situations where population structure explains differences between cases and controls, did our approach (using a linear-mixed model with GRM as random effects and PCs as fixed effect covariates) eliminate its effects as I had claimed, and as shown by many to be the most robust at doing so (among other methods including genomic control, PCA and AIMs)?

To do this I created a situation where the MDD disease statuses of samples were completely confounded with population structure. I first divided our sample into equal North-South partitions using PC1 derived from 322,911 SNPs on 10,640 samples I had reported –half of our samples (n=5320, ncases = 2542, ncontrols =2778) with higher PC1 values than the median were designated "North", and those with lower PC1 values than the median (n=5320, ncases = 2761, ncontrols = 2559) were designated "South". I then took a random 2500 cases from North and 2500 controls from South to obtain a set of 5000 samples and tested for association with MDD genome-wide. In this case case-control status is completely confounded with geographical locations and population structure – the worst-case scenario for a GWAS. I expect therefore to see high inflation in test statistics that can be quantified in terms of λ_{GC}. I ran association testing with five different models:

(a) **No Correction**: no correction for population structure

(b) **AIM PCs**: correcting for population structure with 10 PCs computed from with CONVERGE sample genotypes at the 150 ancestry indicative markers (AIMs) identified by a previous study [125].

(c) **CONVERGE PCs**: correcting for population structure with 10 PCs computed from with CONVERGE sample genotypes at 322,911 SNPs across all autosomes (MAF > 1%, LD < 0.5, computed with GCTA)

(d) **LOCO GRMs + CONVERGE PCs**: correcting for population structure with a linear mixed models using leave-one-chromosome-out (LOCO) GRMs per chromosome (using the 322,911 SNPs as the base set of SNPs from which SNPs from each chromosome was removed to compute its LOCO GRM) as random effect and 10 PCs as fixed effect covariates (computed with FastLMM).

(e) **GRM + CONVERGE PCs**: correcting for population structure with a linear mixed models using a GRM computed from all 322,911 SNPs as random effect and 10 PCs as fixed effect covariates (computed with FastLMM).

For each of the five models I obtained the λ_{GC} measure of inflation and manhattan and quantile-quantile plots of association. I then created another four sets of samples with different percentages of cases and controls from North and South as listed in the table below. Cases and controls from North and South were chosen at random in each instance. Again, for each of the five models we obtained the λ_{GC} measure of inflation and manhattan and quantile-quantile plots of association. All results are summarized in Table 2.8.

Sample set	Stratification (%)	Inflation factor (λ_{GC})				
		No Correction	AIM PCs	CONVERGE PCs	LOCO GRMs + CONVERGE PCs	GRM + CONVERGE PCs
2500N Cases 2500S Controls	100	4.972	2.118	0.008	1.136	1.063
1875N 625S Cases 2500S Controls	75	3.203	1.435	0.999	1.102	1.048
1250N 1250S Cases 2500S Controls	50	2.017	1.211	1.047	1.082	1.034
1250N 1250S Cases 1875N 625S Controls	25	1.259	1.069	1.032	1.047	1.019
1250N 1250S Cases 1250N 1250S Controls	0	1.027	1.027	1.029	1.023	1.003

Table 2.8 | Inflation factor from GWAS on MDD using different means of correction for population stratification confounding disease case-control status, at varying levels of population stratification confounding disease case-control status in the GWAS sample. This table shows in the first column the number of Han Chinese cases of MDD and controls from the North (N) and South (S) and in the second column the degree of stratification confounding disease case-control status. The five columns show the inflation factor (λ_{GC}) for GWAS performed with no correction for population stratification (No Correction), correction with PCs computed from genotypes of GWAS samples at 150 Han Chinese specific AIMs (AIM PCs), correction with PCs computed from genotypes of GWAS samples at 322,911 tagging markers from all autosomes (CONVERGE PCs), correction with a linear mixed-model (LMM) using the leave-one-chromosome-out (LOCO) approach, where one genetic relatedness matrix (GRM) was constructed for per chromosome using tagging markers from the master set of 322,911 markers except those from itself to prevent proximal contamination that could reduce power of detecting true association, as well as CONVERGE PCs (LOCO GRMs + CONVERGE PCs), and finally the LMM approach using a genome-wide GRM computed using all 322,911 tagging markers as well as CONVERGE PCs, without attempting to eliminate proximal contamination (GRM + CONVERGE PCs).

From these results I observed that:

a) Population structure confounding is a problem that can be adequately corrected for by all methods below 25% confounding.

b) PCs from AIMs start to be inadequate in controlling for population structure when the confounding is greater than 50%.

c) At confounds of greater than 50%, CONVERGE PCs start to cause deflation instead of inflation. This is because of the way North and South were defined for the data. As we had assigned samples who had PC1 values above median to North and the rest to South, PC1 would correct more and more of genetic variations explaining North-South differences as the level of confounding increases.

d) Leave-one-chromosome-out (LOCO) GRMs perform poorly at levels of stratification at and above 50%, causing false positive peaks at chr6 (MHC region) and chr22 (intergenic region). These are the genomic regions shown by scans without correction for population structure to contribute to the North-South stratification. LOCO gives rise to false positive peaks because in heavily confounded samples omitting the chromosomes that contain concentrations of loci that contribute to population structure reduces correction for markers that most explain the confounding. Not correcting for them explicitly due to the LOCO approach would cause false positive association signals at those very markers. Another way to think about it is for LOCO GRMs to work in heavily confounded samples, every chromosome needs to be as good as all other chromosomes in explaining population structure differences, not too much better. This will be violated where sites on chr6 and chr22 explain more of the confounding stratification than the other chromosomes, therefore correction at chr6 and chr22 will be lacking as compared to the rest of the genome. The greater the confounding, the more subtle differences between chromosomes would show up

as false positive association signals.

e) The above, however, can be corrected well with a whole-genome GRM and PCs approach, which does not reduce proximal contamination during correction, but is able to explicitly correct for all markers involved in population structure at even high levels of confounding.

I then turned to look at the full dataset. What level of confounding was the data likely to have? To estimate that I first looked at the number of cases and controls from North and South respectively. I divided the sample into equal North-South partitions using PC1 from PCA on 322,911 SNPs on all 10,640 samples we had reported – the half of our samples (n=5320, ncases = 2542, ncontrols =2778) with higher PC1 values than the median were designated "North", and those with lower PC1 values than the median (n=5320, ncases = 2761, ncontrols = 2559) were designated "South". Taking the difference in number of cases and controls from both North and South and dividing by the total number of samples, I get (2778-2542+2761-2559)/10640 = 0.0412. Allowing room for misclassification of cases and controls to North and South, there will be around 5% confounding of case-control status with North-South stratification.

To test if this was really the case, I performed a GWAS using logistic regression for MDD in our whole dataset (10,640 samples) without any form of correction for population structure to estimate the level of confounding by comparing the λ_{GC} with the λ_{GC} from those from defined levels of confounding in the above tests. Figure 2.16 shows the manhattan and qqplots for this GWAS. The λ_{GC} without correction scan is 1.098, which falls between that of 0% confounding and 25% confounding in our test analyses. At 25%, all methods of correction, including using AIMs PCs, CONVERGE PCs, LOCO GRMs and CONVERGE PCs, and whole-genome GRMs and CONVERGE PCs, were able to adequately correct for confound from population stratification.

Nonetheless, to make sure population stratification was not a problem in this cohort, I performed the scan using AIMs PCs (λ_{GC} = 1.085), CONVERGE PCs (λ_{GC} = 1.068), and LOCO GRMs and CONVERGE PCs (λ_{GC} = 1.070). Figures 2.17 and 2.18 show the Manhattan and quantile-quantile plots for the former two models, while Figure 3.1 (in the coming chapter on GWAS on MDD) shows the latter.

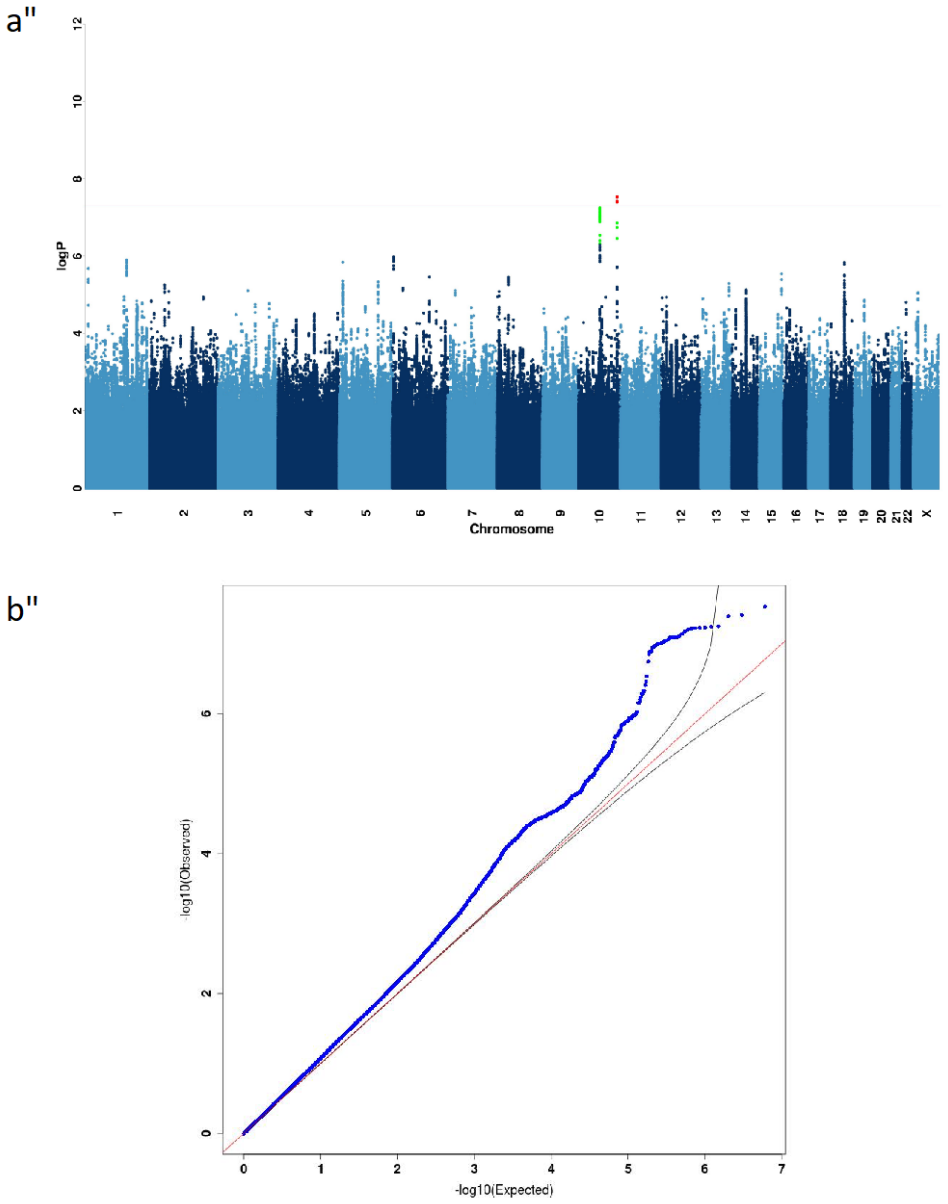


Figure 2.16 | GWAS on MDD in the CONVERGE with no correction of population structure. (A) Manhattan plot of genome wide association for MDD; (B) Quantile-quantile plot of GWAS for MDD using logistic regression with no PCs as covariates on 10,640 samples (5,303 cases, 5,337 controls), genomic inflation factor lambda (λ) = 1.098.

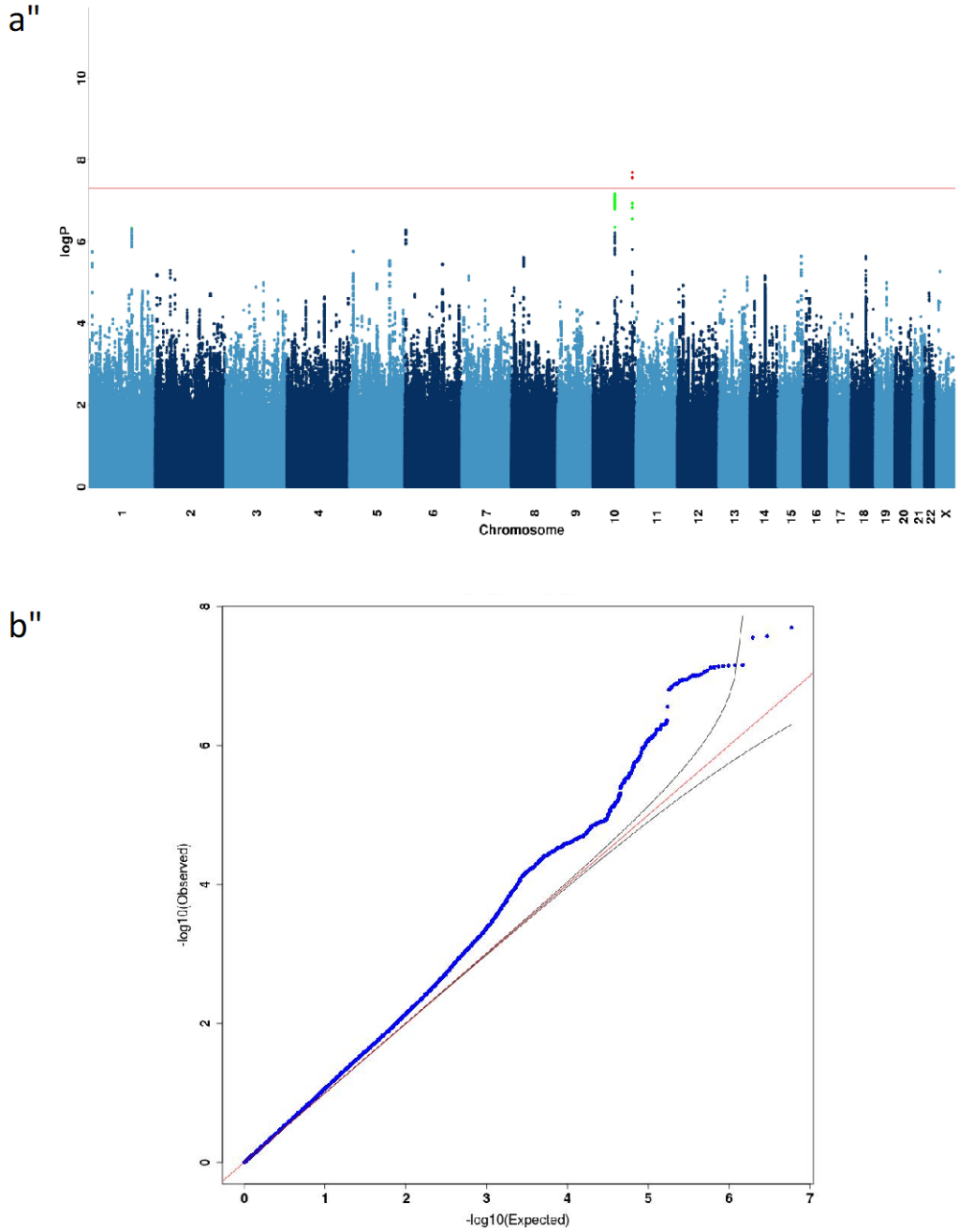


Figure 2.17 | GWAS on MDD in the CONVERGE with no correction of population structure. (A) Manhattan plot of genome wide association for MDD; (B) Quantile-quantile plot of GWAS for MDD using logistic regression with no PCs as covariates on 10,640 samples (5,303 cases, 5,337 controls), genomic inflation factor lambda (λ) = 1.085.

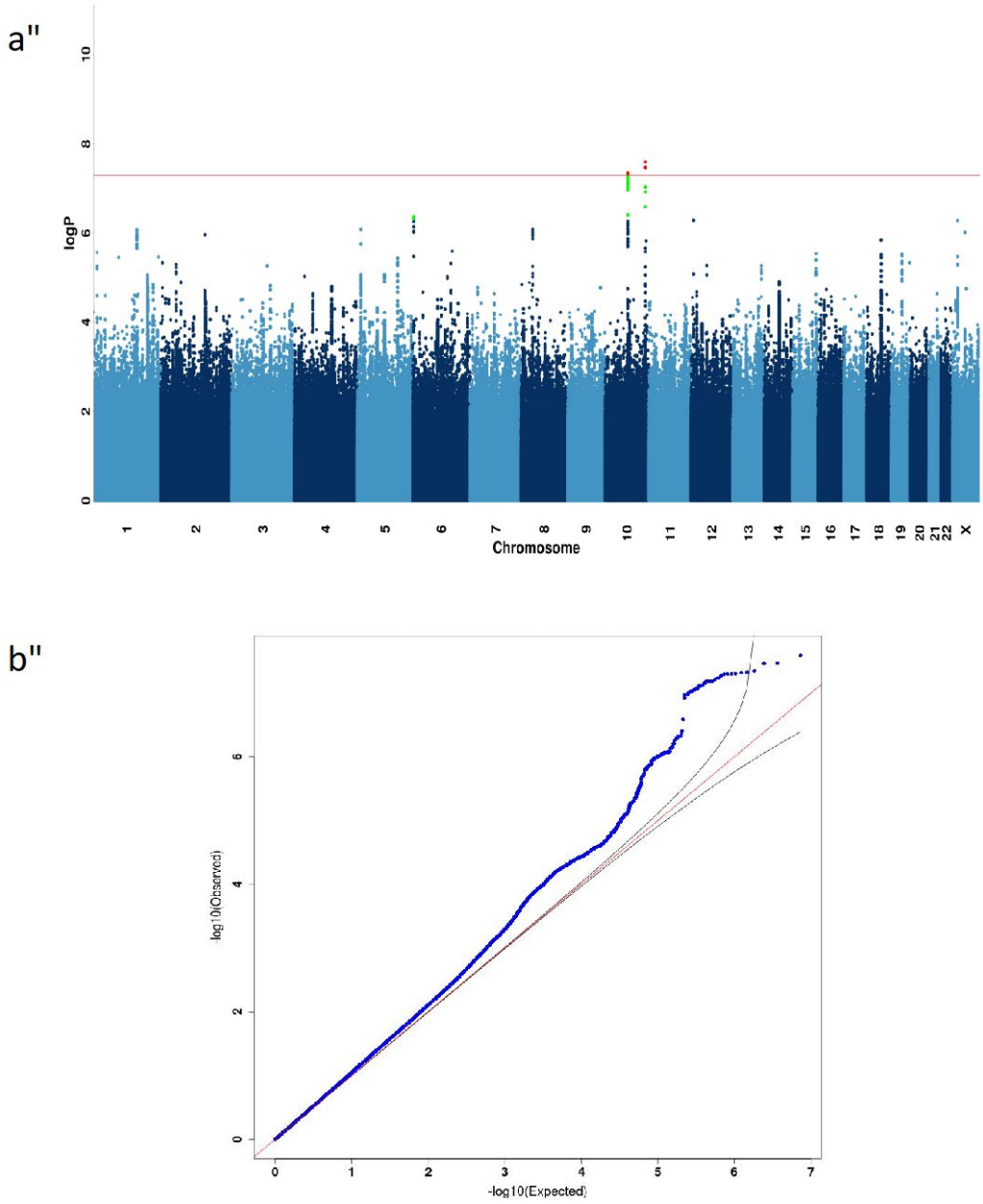


Figure 2.18 | GWAS on MDD in the CONVERGE with no correction of population structure. (A) Manhattan plot of genome wide association for MDD; (B) Quantile-quantile plot of GWAS for MDD using logistic regression with no PCs as covariates on 10,640 samples (5,303 cases, 5,337 controls), genomic inflation factor lambda (λ) = 1.068.

2.9 Discussion

This chapter detailed the steps I had taken to maximize the reliability of results from a GWAS on MDD: I performed both pre-variant calling processing of the raw NGS data and post-variant calling filtering of the raw variant calls for false positives; I checked for excess of private variants in the nuclear genome and heteroplasmic variants in the mtDNA as indicators of contamination of source DNA and excluded samples with such indications; I computed the pairwise relatedness between all samples in CONVERGE, and removed samples that were more related than others; I then constructed a picture of the population structure in CONVERGE, and having ensured population stratification was slight in CONVERGE and checked the various ways I could control for population structure in GWAS, picked a conservative approach that balances eliminating population effects and preserving power to detect truly associated loci.

How important were these steps in the success of a GWAS? I had performed these steps as they were theoretically the correct things to do, and while each step was backed up strongly with empirical evidence that there were false positives, contamination, uneven relatedness between samples and slight population stratifications in the data, I had not at every stage examined the effect of my quality control steps. For comparison therefore I performed the same GWAS on MDD, but without taking a number of the above quality control steps. Having not excluded samples that had potentially contaminated DNA and who were related to one or a few other more than all others in the cohort, and having not controlled for population structure in any of the ways I had explored in this chapter, I ran the GWAS on MDD with all 11,670 imputed CONVERGE samples using a logistic regression model, same as used in the GWAS I show in Figure 2.16. If removing samples with contaminated DNA and related samples had no effect at all, the results of this new GWAS I had performed should look almost identical to that shown in Figure 2.16.

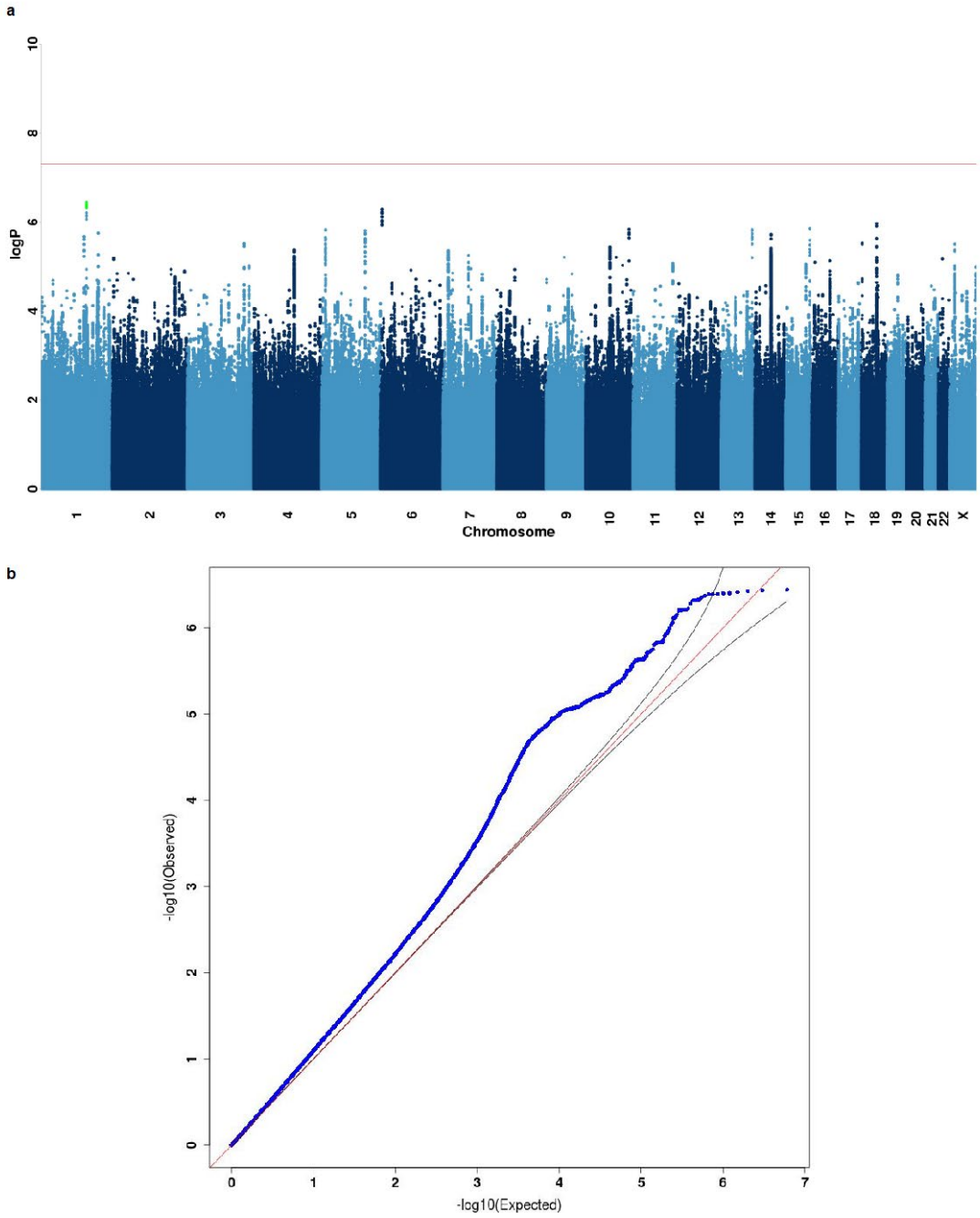


Figure 2.19 | GWAS on MDD in the CONVERGE with no exclusion of samples and no correction for population structure. (A) Manhattan plot of genome wide association for MDD using the logistic regression model; (B) Quantile-quantile plot of GWAS for MDD using logistic regression with no PCs as covariates on all CONVERGE samples, genomic inflation factor lambda (λ) = 1.126.

However, that is not the case. I show in Figure 2.19 the manhattan plot and quantile-quantile plot for this final GWAS done without a few of the quality control steps. It shows a higher inflation factor than that in Figure 2.16, and absolutely no sign of genome wide significant associations.

While all other considerations (variant calling and imputation qualities, sample relatedness, population stratification) had been long known and discussed in many GWAS studies, methods for identification of potentially contaminated samples using count of private variants detected in the nuclear genome and heteroplasmic variants in the mtDNA would not have been applicable to any study that had not used NGS as the genotyping strategy. These methods can potentially be applied to future datasets to identify DNA contamination and help genetic associations that would otherwise had been hidden.

The implementation of such rigorous quality controls is likely important when the power of detecting a signal is low. The fact that good quality control steps could uncover GWAS loci showed that the CONVERGE cohort was on the borderline in term of sample size and phenotypic specificity for detecting genome wide significant associations. Other studies may be in the same situation, where sample sizes were determined based on power calculations and lessons from previous studies. Utilizing as many ways as possible for quality control, and developing new methods to identify possible confounds, and discarding data of poor quality would in fact be prudent as it would allow full utilization of the data collected.

3. Genetic contribution to MDD

In this chapter I report the results of GWAS of MDD CONVERGE, the first of its kind performed in the Chinese population. As detailed in the Introduction chapter of this thesis, the CONVERGE study was designed to answer the question: to what extent would assaying all variants with frequencies greater than 0.1%, in a sample chosen to reduce known causes of heterogeneity, increase power to discover genetic effects?

Heritability of MDD had been estimated to be around 40% in twin studies[9], though attempts to find contributing genetic loci in the genome wide association study (GWAS) framework using genotyping arrays and more recently imputation using the 1000 Genomes to capture more variation had not yielded replicable genome-wide significant loci or capture the 40% heritability estimated by twin studies. Chasing down the "missing" heritability for MDD and understanding the polygenic architecture of MDD hence became important for informing designs for future genetic studies on MDD, as knowing the number, effect sizes, penetrance and frequency of occurrence of causal variants for MDD would tell us what sample sizes and prevalence of disease phenotypes we would need to have a chance of finding genetic loci involved. As such, in this thesis I also describe analyses leveraging dense set of markers in CONVERGE and recent advances in statistical and bioinformatics methodologies to delineate the genetic architecture of MDD.

Prior analyses indicated that the genetic basis of MDD arises from the joint effect of many loci each contributing a small effect. While all existing data from both CONVERGE and previous studies are consistent with this assumption, they are in fact also consistent with there being genetic loci contributing larger effects, under two (not mutually exclusive) hypotheses: first, genetic loci of larger effects would be observed if more homogenous phenotypic classification of MDD could be achieved, and each subgroup analysed independently for their shared genetic factors; second, some of the

heritability of MDD is explained by genetic loci of larger effects that are very rare in the population.

While the selectivity of sample inclusion in the CONVERGE study attempts to address the first hypothesis, whole-genome NGS on all CONVERGE samples has given me the chance to identify private variants from every sample in CONVERGE. I was able to ask if particular gene sets were enriched for rare, deleterious variants in cases of MDD as compared to controls, adding to the understanding of the genetic architecture of MDD beyond variants segregating at appreciable frequencies in the population.

3.1GWAS of MDD

3.1.1 GWAS of MDD finds two genome-wide significant loci

I performed the GWAS of MDD on 10,640 samples (5,303 cases of MDD, 5,337 controls) using imputed allele dosages at 6,242,619 SNPs with a linear mixed model with a genetic relatedness matrix (GRM) as a random effect and principal components from eigen-decomposition of the GRM as fixed effects (Methods). Figure 3.1 shows the Manhattan and quantile-quantile plots for this GWAS. The genomic-control inflation factor (λ, the ratio of the observed median χ^2 to that expected by chance) for association with MDD was 1.070 (for common SNPs > 2 %, $\lambda = 1.074$), and the adjusted measure for sample size to that of 1,000 cases and 1,000 controls (λ_{1000}) was 1.013, indicating minimal inflation due to population stratification.

Two loci exceeded genome-wide significance in association with MDD: one 5' to the sirtuin1 (*SIRT1*) gene on chromosome 10 (SNP = rs12415800, chr10:69624180, minor allele frequency (MAF) = 45.2%, P = 1.92×10^{-8}, Figure 3.2a) and the other in an intron of the phosphorlysine phosphohistidine inorganic pyrophosphate phosphatase (*LHPP*) gene (SNP = rs35936514, chr10:126244970, MAF = 26.0%, P = 1.27×10^{-8}, Figure 3.2b). *LHPP* had been previously associated with MDD, though its potential function as a phosphoamidase is only speculative[152]

SIRT1 is an NAD+-dependent deacetylase that has a central role in mitochondrial biogenesis, stress tolerance and fat metabolism[153]. It influences processes that feature among the vegetative symptoms of MDD: alterations in food intake, wakefulness and circadian rhythms. *SIRT1*, through interaction with *PGC-1α*, induces a gluconeogenic transcriptional program and associated mitochondrial biogenesis[154].

To ensure the associations were true and not driven by erroneous imputation, I checked the accuracy of the imputed genotypes at 12 SNPs each representing a peak in the GWAS (though only two were genome-wide significant) with P < 1×10^{-5} by re-

genotyping all of the CONVERGE samples using a custom Sequenom MassARRAY™ system mass spectrometer. Table 3.1 shows that the correlations between the imputed allele dosages used for association testing and the genotypes from the Sequenom assay at all 12 sites were high (mean $R^2 = 0.984$), and direct testing of association between the genotypes at the 12 sites from Sequenom with MDD showed odds ratios for the two genome-wide significant SNPs assessed by the two methods were almost identical with highly overlapping confidence intervals (rs12415800 odds ratios: 1.167 vs 1.167; rs35936514 odds ratios: 0.845 vs 0.842).

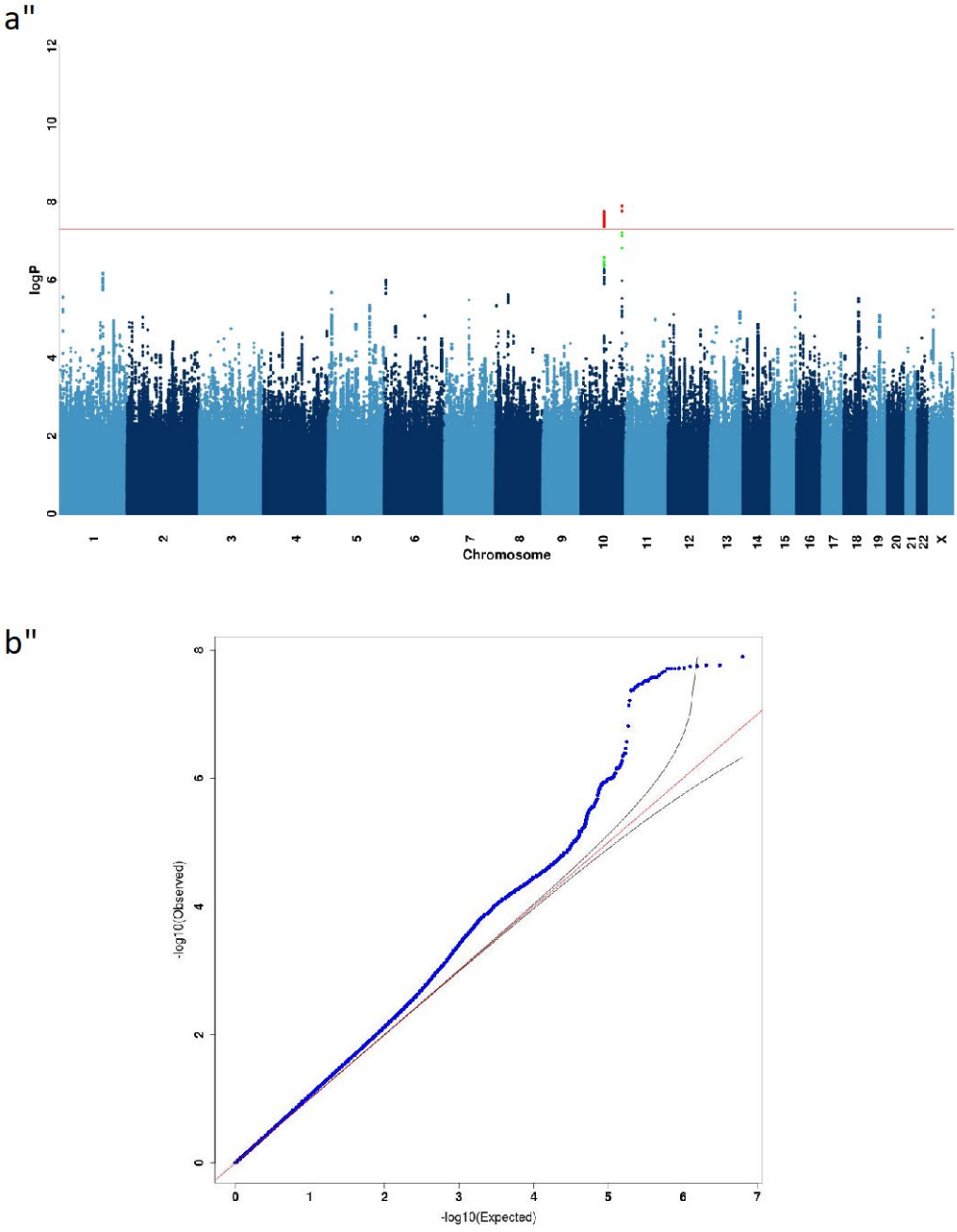


Figure 3.1 | GWAS on MDD in the CONVERGE sample using the LOCO-MLMA method. (A) Manhattan plot of genome wide association for MDD; (B) Quantile-quantile plot of GWAS for MDD using the Mixed-Linear-Model with exclusion of chromosome marker is on (MLMe) method implemented in FastLMM on 10,640 samples (5,303 cases, 5,337 controls), genomic inflation factor lambda (λ) = 1.070, rescaled for an equivalent study of 1,000 cases and 1,000 controls (λ_{1000}) = 1.014.

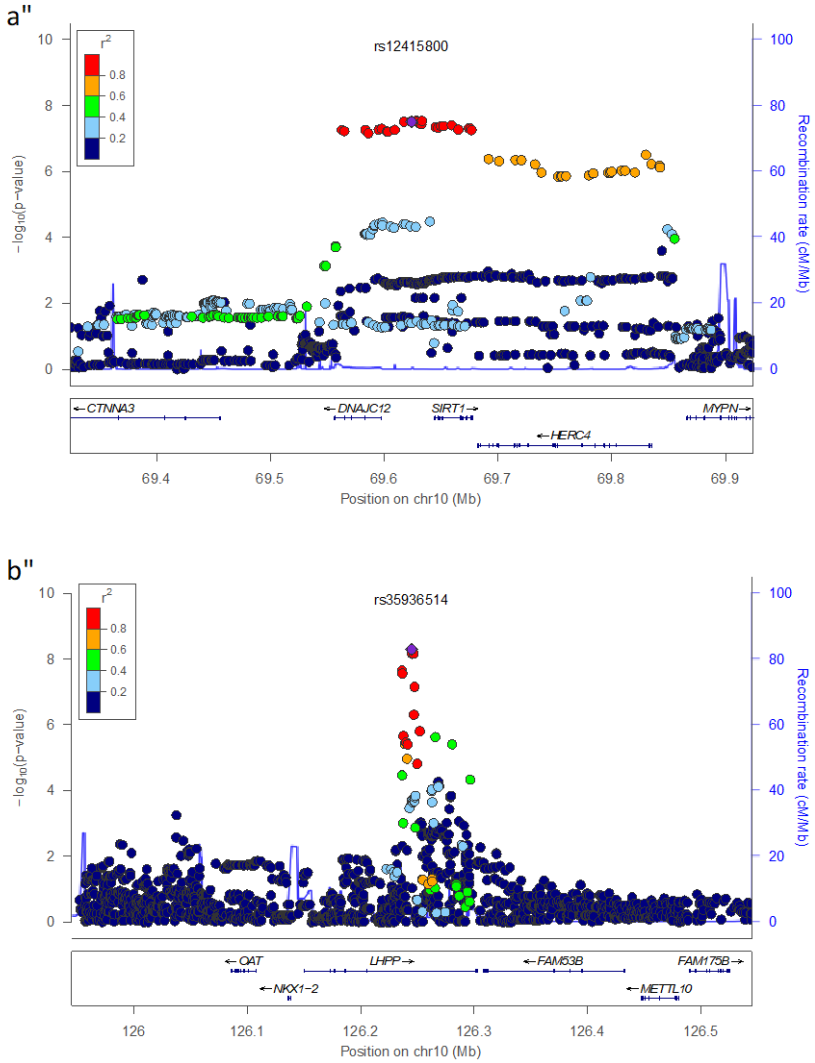


Figure 3.2 | Two loci associated with major depressive disorder in the CONVERGE sample. Detailed views of two loci associated with MDD: (a) the SIRT1 locus on chromosome 10 at 69.6 Mb; (b) the LHPP locus on chromosome 10 at position 126.2 Mb. The −log10 P-values of imputed SNPs associated with MD by logistic regression are shown on the left y axes together the recombination rates (NCBI Build GRCh37), represented by light-blue lines, with scales on the right y axes. Genes within the regions are shown in the bottom panels. The horizontal axis gives the chromosomal position in megabases (Mb). The index SNPs are shown as larger purple diamonds labeled by their marker names; linkage disequilibrium with the remaining SNPs is indicated by different colors. The plots were drawn using LocusZoom[155].

SNP					Imputed Dosages (N=9,921)				Sequenom genotypes			
CHR	**POS**	**RSID**	**REF**	**ALT**	**OR**	**SE**	**P**	**N**	**R^2**	**OR**	**SE**	**P**
1	11493832	rs2922240	C	T	1.141	0.029	5.82E-06	9,864	0.991	1.141	0.029	5.72E-06
1	175151950	rs3766688	C	T	0.870	0.029	1.93E-06	9,901	0.995	0.871	0.029	2.32E-06
1	228052027	rs57047840	G	A	1.141	0.032	4.13E-05	9,724	0.974	1.141	0.032	3.91E-05
5	9161674	rs55713588	G	A	1.302	0.052	3.34E-07	9,636	0.925	1.263	0.050	2.87E-06
6	4386107	rs55800092	T	C	0.819	0.040	6.35E-07	9,881	0.992	0.817	0.040	5.52E-07
10	**69624180**	**rs12415800**	**A**	**G**	**1.167**	**0.029**	**8.44E-08**	**9,689**	**0.993**	**1.167**	**0.029**	**9.30E-08**
10	**126244970**	**rs35936514**	**T**	**C**	**0.845**	**0.033**	**2.91E-07**	**9,915**	**0.993**	**0.842**	**0.033**	**1.40E-07**
13	107659212	rs61967003	T	C	1.765	0.116	9.64E-07	9,914	0.997	1.748	0.116	1.53E-06
14	66833851	rs17827252	G	C	0.896	0.029	1.12E-04	8,562	0.999	0.897	0.031	3.94E-04
19	34493757	rs11880240	G	C	1.281	0.056	1.08E-05	9,912	0.996	1.281	0.056	1.14E-05
X	24656658	rs1921918	A	G	0.877	0.032	3.59E-05	9,899	0.994	0.880	0.032	6.39E-05
X	25011374	rs11573525	T	C	1.158	0.033	9.60E-06	9,912	0.969	1.144	0.032	3.21E-05

Table 3.1 | Comparison between association results using imputed dosages and directly genotyped markers. The table reports results for association between MDD and 12 SNPs. The first five columns give the chromosome (CHR), genomic position (POS), SNP identifier (RSID), reference allele (REF) on Human Genome Reference GRCh37.p5 and alternative allele (ALT) called in CONVERGE. The next three columns show results for imputed allele dosages at 12 SNPs (odds ratio (OR) of association with MDD with respect to the alternative allele and standard error (SE); P-values of association (P)). The next two columns present the number of samples successfully genotyped using the Sequenom platform (a high sensitivity and specificity assay), and the Pearson correlation (R^2) between the imputed allele dosages and the genotypes from Sequenom. The final three columns present results of association with MDD using genotypes from the Sequenom genotyping platform. Bold type indicates the genome-wide significant markers: Extended Data Table 1b gives further information on the results for these markers.

3.1.2 Replication in an independent cohort of MDD in Han Chinese

To replicate the associations, I obtained genotypes at the same 12 SNPs in a separate Han Chinese cohort of 3,231 cases with recurrent MDD, and 3,186 controls (both sexes) using the Sequenom MassARRAY™ system mass spectrometer to replicate the results in the CONVERGE discovery GWAS for MDD. The replication sample was obtained from five hospitals in the north of China. Patients were diagnosed as having MDD by at least two consultant psychiatrists by DSM-IV criteria. Samples were of both sexes, and all four grandparents were Han Chinese. Cases were aged between 30 and 60, and had two or more episodes of MDD meeting DSM-IV criteria. Exclusion criteria were pregnancy, severe medical conditions, abnormal laboratory baseline values, unstable psychiatric features (e.g., suicidal), a history of alcoholism or drug abuse, epilepsy, brain trauma with loss of consciousness, neurological illness, or a concomitant Axis I psychiatric disorder. Control subjects were recruited from local communities and provided information about medical and family histories. Exclusion criteria were a history of major psychiatric or neurological disorders, psychiatric treatment or drug abuse, or family history of severe forms of psychiatric disorders.

Two SNPs at the peaks of association for *SIRT1* and *LHPP* loci (rs12415800 and rs35936514 respectively) for MDD in CONVERGE were significantly associated with MDD (Table 3.2). Analysis of the combined samples gave P-values for association with MDD at these two SNPs of 2.53×10^{-10} and 6.45×10^{-12}, respectively. Table 3.3 shows the genotype distribution and p-values for violation of Hardy-Weinberg Equilibrium (HWE) in both CONVERGE and the replication cohort at both SNPs.

SNP					CONVERGE (N=10,640)					Replication (N=6,417)			Joint (N=17,057)		
CHR	**POS**	**RSID**	**REF**	**ALT**	**FREQ**	**INFO**	**OR**	**SE**	**P**	**OR**	**SE**	**P**	**OR**	**SE**	**P**
1	11493832	rs2922240	T	C	0.385	1.018	1.141	0.028	2.80E-06	0.949	0.037	1.54E-01	1.070	0.022	2.46E-03
1	175151950	rs3766688	T	C	0.394	1.003	0.875	0.028	1.83E-06	0.991	0.037	8.15E-01	0.918	0.022	1.34E-04
1	228052027	rs57047840	A	G	0.284	0.970	1.138	0.031	4.64E-05	1.001	0.041	9.90E-01	1.088	0.025	5.57E-04
5	9161674	rs55713588	A	G	0.096	0.893	1.278	0.050	6.04E-07	1.054	0.062	3.93E-01	1.042	0.035	2.08E-01
6	4386107	rs55800092	C	T	0.151	1.001	0.824	0.039	1.35E-06	0.962	0.052	4.49E-01	0.876	0.031	1.82E-05
10	**69624180**	**rs12415800**	**G**	**A**	**0.452**	**0.992**	**1.164**	**0.028**	**1.92E-08**	**1.130**	**0.036**	**7.71E-04**	**1.150**	**0.022**	**2.37E-10**
10	**126244970**	**rs35936514**	**C**	**T**	**0.260**	**0.993**	**0.839**	**0.032**	**1.27E-08**	**0.838**	**0.041**	**1.68E-05**	**0.842**	**0.025**	**6.43E-12**
13	107659212	rs61967003	C	T	0.017	0.999	1.645	0.109	6.70E-06	0.788	0.150	1.11E-01	1.277	0.087	4.81E-03
14	66833851	rs17827252	C	G	0.463	1.011	0.887	0.028	1.44E-05	0.962	0.041	3.41E-01	0.907	0.023	2.20E-05
19	34493757	rs11880240	C	G	0.068	1.019	1.291	0.055	8.02E-06	1.048	0.072	5.12E-01	1.184	0.043	9.15E-05
X	24656658	rs1921918	A	G	0.721	0.995	0.883	0.031	3.22E-05	0.994	0.047	9.01E-01	0.917	0.026	1.09E-03
X	25011374	rs11573525	C	T	0.260	0.971	1.160	0.032	5.86E-06	1.011	0.047	8.19E-01	1.100	0.027	2.18E-04

Table 3.2 | Genetic association between MDD and 12 variants in CONVERGE and a replication sample The table reports results for 12 SNPs in the CONVERGE and replication samples. The first five columns give the chromosome (CHR), genomic position (POS), SNP identifier (RSID), reference allele (REF) on Human Genome Reference GRCh37.p5 and alternative allele (ALT) called in CONVERGE. The next five columns show the alternative allele frequency (FREQ) and results of association testing with MDD using imputed allele dosages in 10,640 CONVERGE samples (5,303 cases, 5,337 controls); information scores (INFO), odds ratio (OR) of association with MDD with respect to the alternative allele and standard error (SE) in the odds ratio were obtained from a logistic regression model; P values of association (P) obtained from a linear-mixed model with a genetic relatedness matrix containing all samples. The next three columns present results of association with MDD in the replication cohort of 6,417 samples (3,231 cases, 3,186 controls) from a logistic regression model. The final three columns present results of association with MDD in a joint analysis with both CONVERGE and replication cohort from a logistic regression model. Bold type indicates the genome wide significant markers.

MDD Disease State	SNP	CONVERGE		Replication Cohort	
		HomRef/Het/HomAlt	HWE P-value	HomRef/Het/HomAlt	HWE P-value
All	rs12415800	2151/5301/3169	0.445	1212/3037/1920	0.857
	rs35936514	705/4054/5794	0.919	422/2400/3398	0.974
Cases	rs12415800	1178/2626/1490	0.741	654/1538/918	0.829
	rs35936514	318/1919/3027	0.549	190/1136/1783	0.627
Controls	rs12415800	973/2675/1679	0.106	558/1499/1002	0.971
	rs35936514	387/2135/2767	0.389	232/1264/1615	0.503

Table 3.3 | Genotype distribution and p-values for violation of Hardy-Weinberg Equilibrium in CONVERGE and replication cohorts. This table shows the number of samples with the homozygous reference genotype (HomRef), heterozygous genotypes (Het), and homozygous alternative genotype (HomAlt), as well as p-values for violation of Hardy-Weinberg equilibrium (HWE) for both CONVERGE study samples and the replication cohort from North China at the top SNPs rs12415800 in the *SIRT1* locus and rs35936514 in the *LHPP* locus from the GWAS on MDD. The top two rows show these measures for all samples in both the CONVERGE and replication study, the next two rows show that for just cases in CONVERGE and the replication cohort, and the last two shows show that for just the controls. The genotype distributions for CONVERGE are obtained from hard-called genotypes from maximum imputed genotype probabilities at for each sample at each of the two sites. As a genotype will not be called if the maximum genotype probability at a site is lower than 0.9 for any single sample, the total number of CONVERGE samples showing called HomRef/Het/HomAlt genotypes does not equal 10,640 for either SNP. For rs12415800, 19 (9 cases, 10 controls) samples have no genotype calls due to maximum genotype probability being smaller than 0.9, giving a total of 10621 CONVERGE (5,294 cases, 5,327 controls) samples with genotype calls. For rs35936514, 87 (39 cases, 48 controls) samples have no genotype calls due to maximum genotype probability being smaller than 0.9, giving a total of 10,553 (5,264 cases, 5,289 controls) CONVERGE samples with genotype calls.

3.1.3 Power to detect association increases with the severity of disease

In the CONVERGE cohort, 85% of cases met DSM-IV melancholia criteria [156]. Prior research has suggested that MDD patients with melancholia have more impairing, recurrent episodes and that risk for MDD is increased in the co-twins of probands with the melancholic subtype [157]. This increase is greater in monozygotic than dizygotic pairs, as would be expected if the subtype were associated with greater genetic risk [157]. If this classification represents a genetically more homogenous subtype of depression, then restricting the GWAS to only those cases diagnosed with melancholia should increase power to detect genetic association with disease status. To test this hypothesis, I performed the GWAS with only those cases diagnosed with melancholia and all controls in the cohort (4,509 cases, 5,337 controls).

The GWAS identified the same two loci that exceeded genome-wide significance on chromosome 10 (Figure 3.3). Even though the sample for melancholia was smaller than for MDD, at the *SIRT1* locus the significance of association was two orders of magnitude greater than for (top SNP = rs80309727, chromosome 10:69617347, MAF = 45.2%, P = 2.95×10^{-10}). To determine whether the increased association might have arisen by chance, I generated an empirical distribution of odds ratios by randomly selecting 4,509 cases from the total set and re-analysing the association with each of the genome-wide significant variants, and found that the observed value lay on the 98.8 percentile at the *SIRT1* locus, but at the 61.6 percentile at the *LHPP* locus (Figure 3.4). This supports the interpretation that power to detect association increases with the selectivity for disease phenotype, in this case in terms of severity of disease.

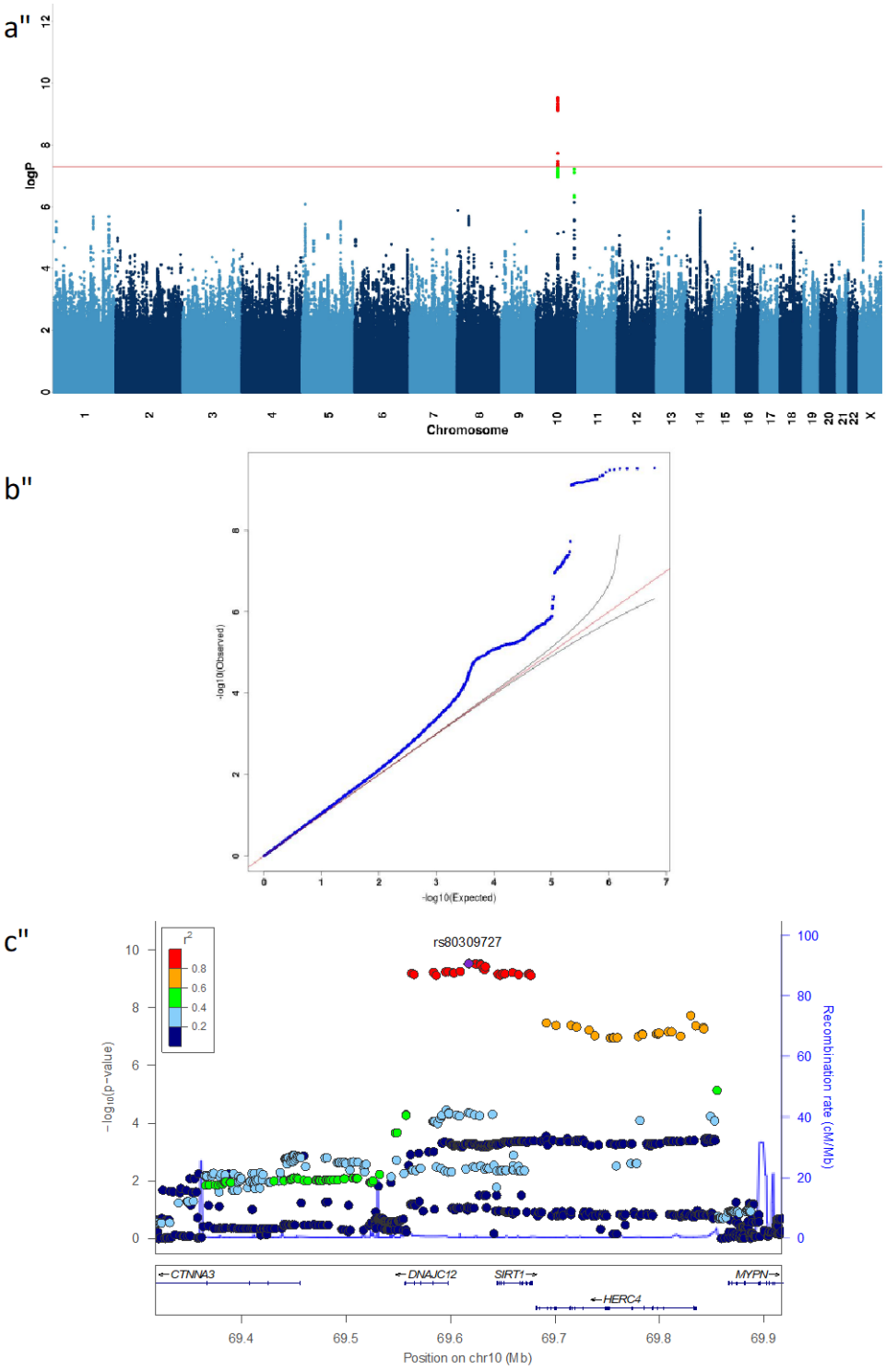


Figure 3.3 | Manhattan and quantile quantile plots for melancholia. a, Manhattan plot of GWAS for melancholia using the Mixed-Linear-Model with exclusion of chromosome marker is on (MLMe) method implemented in FastLMM on 9,846 samples (4,509 cases, 5,337 controls). **b**, Quantile-quantile plot of GWAS for melancholia, genomic inflation factor lambda (λ) = 1.069, rescaled for an equivalent study of 1,000 cases and 1,000 controls (λ_{1000}) = 1.014. **c**, Regional association plot on GWAS hit on chromosome 10 focusing on top SNP rs80309727 at 5' of SIRT1 gene, generated with LocusZoom.

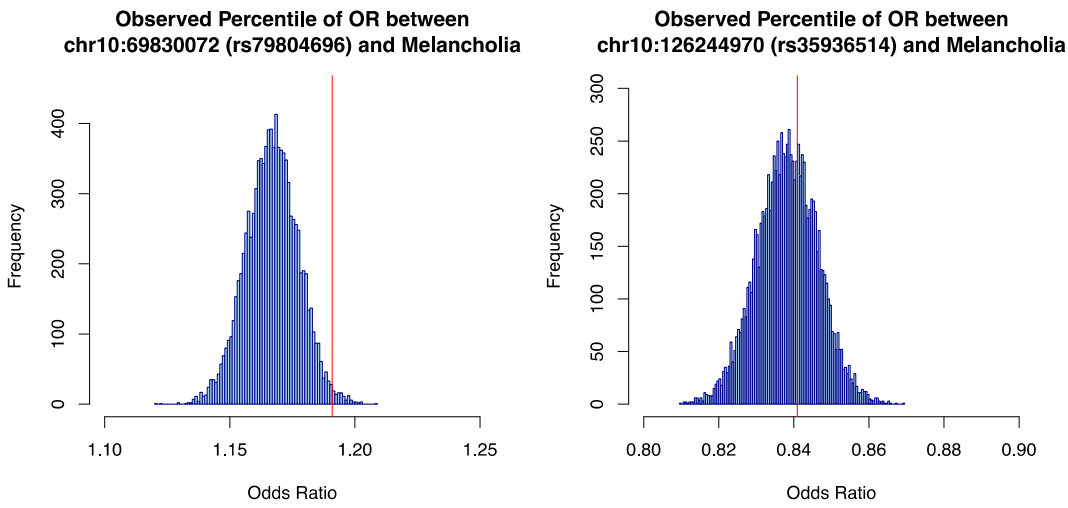


Figure 3.4 | Empirical estimation of the odds ratio increases due to the removal of cases not falling under the diagnostic class of melancholia from an association analysis with major depression. The figures show the empirical distributions of the odds ratios for association with each of two SNPs (rs79804696, rs35936514), after removing a random set of 796 samples, equal to the number of cases of MDD not diagnosed as being melancholic. The horizontal axis is the odds ratio for each analysis, and the vertical axis the frequency of occurrence of the odds ratio in 10,000 analyses. The vertical red line is the observed odds ratio after removing cases of MDD not diagnosed as melancholic.

3.1.4 Power to detect associations increase with removing environmental factors

I explored the hypothesis that MDD might be etiologically heterogeneous by considering whether our genetic signals were strengthened by the exclusion of cases where liability to illness may have arisen largely from adverse environmental experiences. One of the strongest and best-validated risk factors for MDD is childhood sexual abuse (CSA) [158,159]. I therefore removed 856 subjects reporting CSA and repeated the GWAS for MDD (Figure 3.5) and melancholia (Figure 3.6) on the remaining samples.

Similar to the phenomenon seen in the GWAS for melancholia (restricting the definition of disease to only the severe cases of the melancholia subtype), the odds ratio for association with MDD at the *SIRT1* locus increased from 1.167 to 1.195 (top SNP = rs79804696, chromosome 10:69830072, P = 5.72×10^{-9}) and with melancholia from 1.191 to 1.225 (P = 3.10×10^{-10}), while the *LHPP* locus showed no change in odds ratio.

In addition, this GWAS detected an additional genome-wide significant locus on chromosome 8, 3' of *SLC25A37*, for both MDD (top SNP = rs59341197, chromosome 8:23461233, MAF = 27.5%, P = 4.26×10^{-8}, Figure 3.5c) and melancholia (top SNP = rs2942219, chromosome 8:23439308, MAF = 27.6%, P = 2.37×10^{-8}, Figure 3.6c). *SLC25A37*, also called mitoferrin, is an iron transporter protein in embedded in the inner mitochondrial membrane that specifically mediates iron uptake in developing erythroid cells, having hence an essential role in heme biogenesis and iron homeostasis.

To determine whether the increased association after exclusion of cases with CSA might have arisen by chance, I generated an empirical distribution of odds ratios by randomly deleting 856 individuals from the total set and re-analysing the association with each of the genome-wide significant variants. For MDD, the increase was significant at the *SIRT1* (P = 0.0013) and *SLC25A37* loci (P = 0.0004), but not at *LHPP* (P = 0.74); for melancholia, the odds ratios were also significantly increased at *SIRT1* (P = 0.0006) and *SLC25A37* loci (P = 0.0003), but not at *LHPP* (P = 0.48) (Figure 3.7). Thus the

enrichment in signal observed, after removal of individuals with CSA, likely results from a reduction in the etiologic heterogeneity of cases, and is not due to fluctuations in association statistics upon removal of a few hundred samples.

These results show that the inclusion of cases of MDD with high levels of environmental risk masks genetic signal, as removing cases with a severe environmental risk, CSA, increased genetic signal at one locus and unmasked a previously undiscovered locus. As genetic signal at the LHPP locus did not change significantly upon restricting disease severity nor removing environmental signal, it suggests there are genetic factors contributing to MDD whose effects may be mimicked by environmental stress, and other factors whose effects may not. Notably, the fact that two replicated risk loci lie close to genes involved in mitochondrial function (*SIRT1* and *SLC25A37*), together with an enrichment of mutations in genes involved in mitochondrial function, suggest an unexpected origin for at least some of the phenotypic manifestations of MDD.

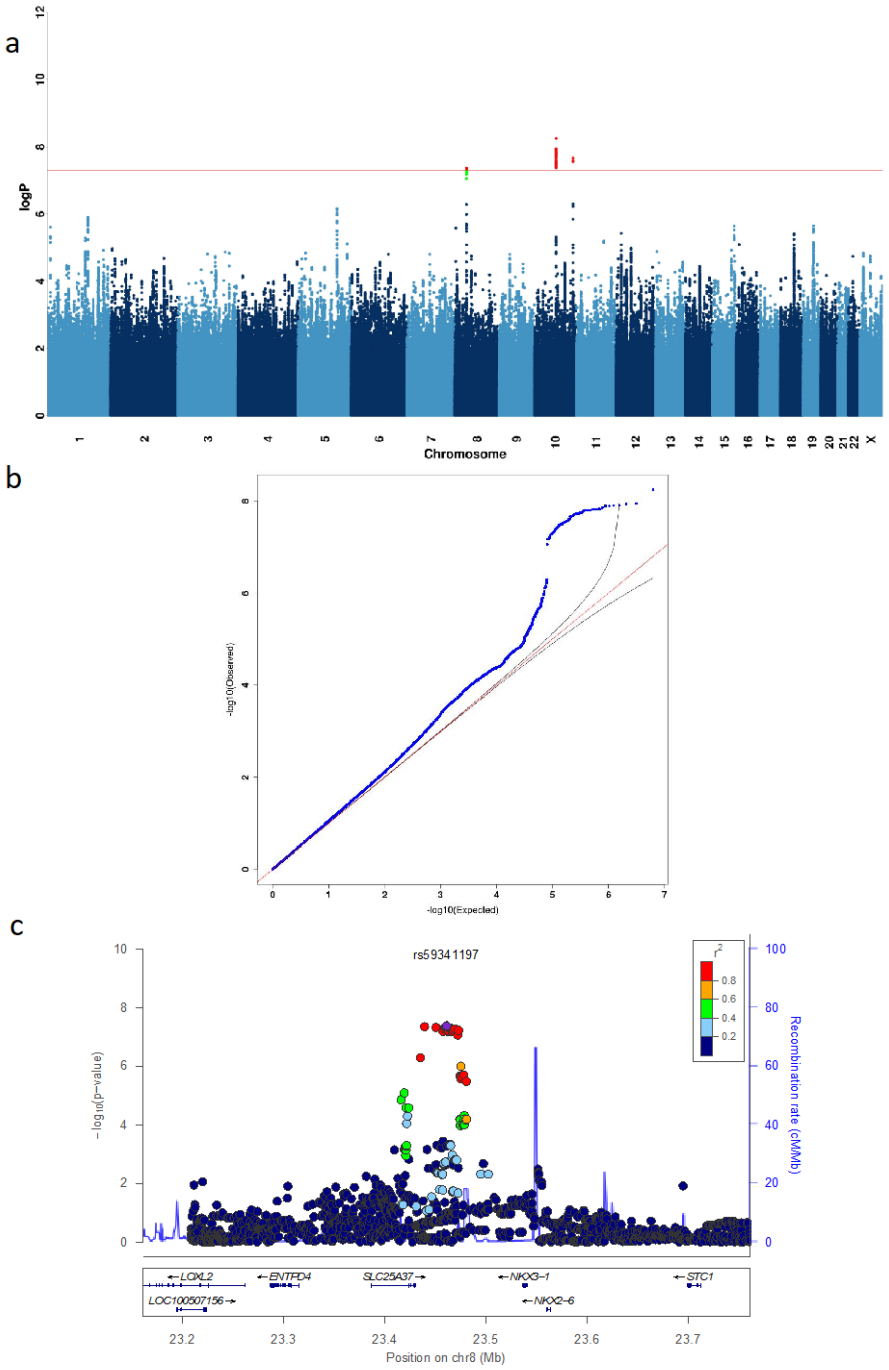


Figure 3.5 | Manhattan and quantile quantile plots for major depressive disorder excluding cases with childhood sexual abuse. a, Manhattan plot of GWAS for MDD excluding 856 samples reporting childhood sexual abuse (CSA) and 188 samples with no information on state of CSA, using the Mixed-Linear-Model with exclusion of chromosome marker is on (MLMe) method implemented in FastLMM. This restricted set consists of 9,784 samples (4,613 cases, 5,171 controls). **b**, Quantile-quantile plot of GWAS for MDD with the restricted data set, genomic inflation factor lambda (λ) = 1.069, rescaled for an equivalent study of 1,000 cases and 1,000 controls (λ_{1000}) = 1.014. **c**, Regional association plot on GWAS hit on chromosome 8 focusing on top SNP rs59341197 at 3' of SLC25A37 gene, generated with LocusZoom.

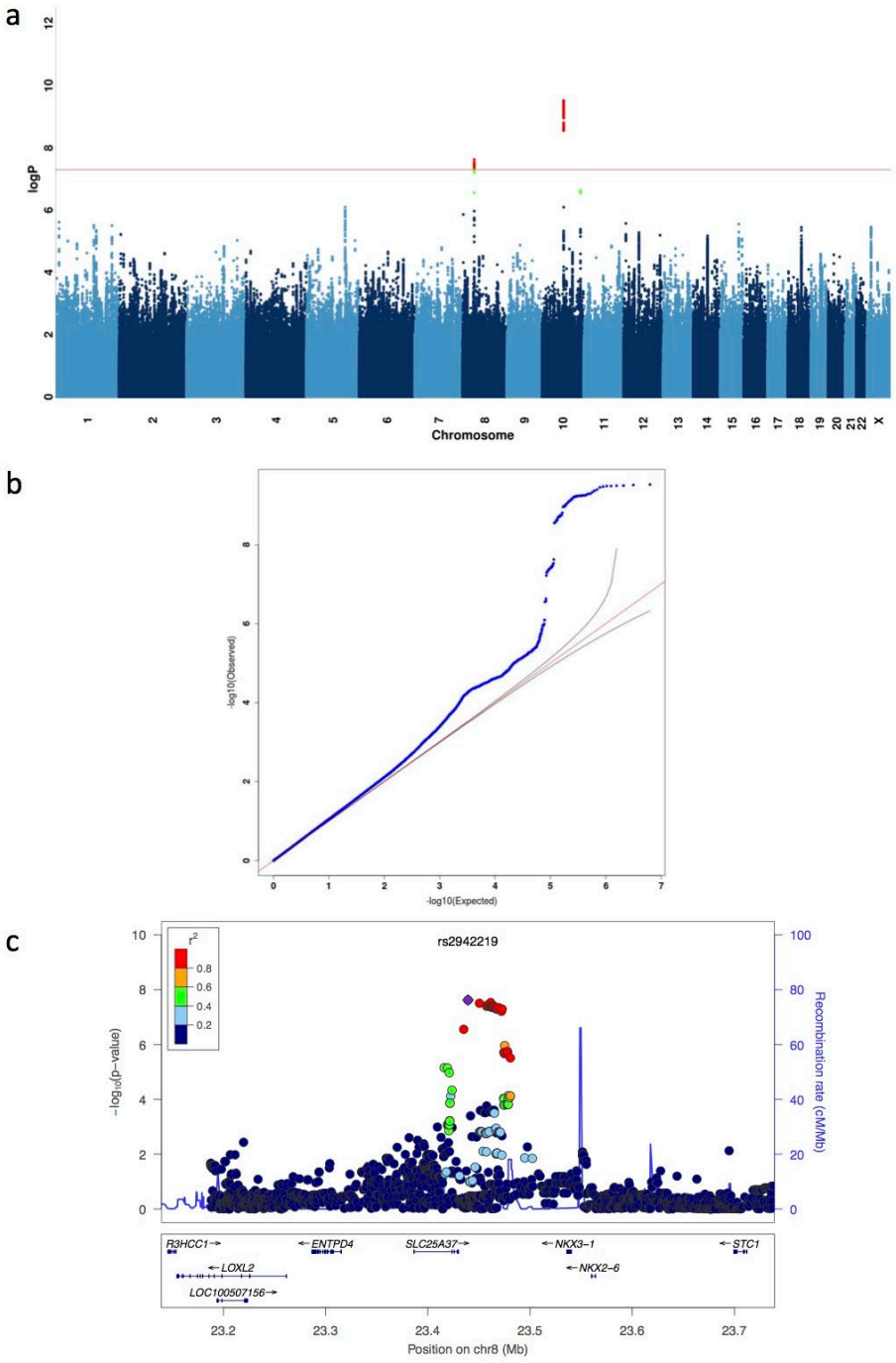


Figure 3.6 | Manhattan and quantile quantile plots for melancholia excluding cases with childhood sexual abuse. a, Manhattan plot of GWAS for Melancholia excluding samples reporting childhood sexual abuse (CSA) and samples with no information on state of CSA, using the Mixed-Linear-Model with exclusion of chromosome marker is on (MLMe) method implemented in FastLMM. This restricted set consists of 9,126 samples (3,955 cases, 5,171 controls). **b**, Quantile-quantile plot of GWAS for MD with the restricted data set, genomic inflation factor lambda (λ) = 1.062, rescaled for an equivalent study of 1,000 cases and 1,000 controls (λ_{1000}) = 1.014. **c**, Regional association plot on GWAS hit on chromosome 8 focusing on top SNP rs2942219 at 3' of SLC25A37 gene, generated with LocusZoom.

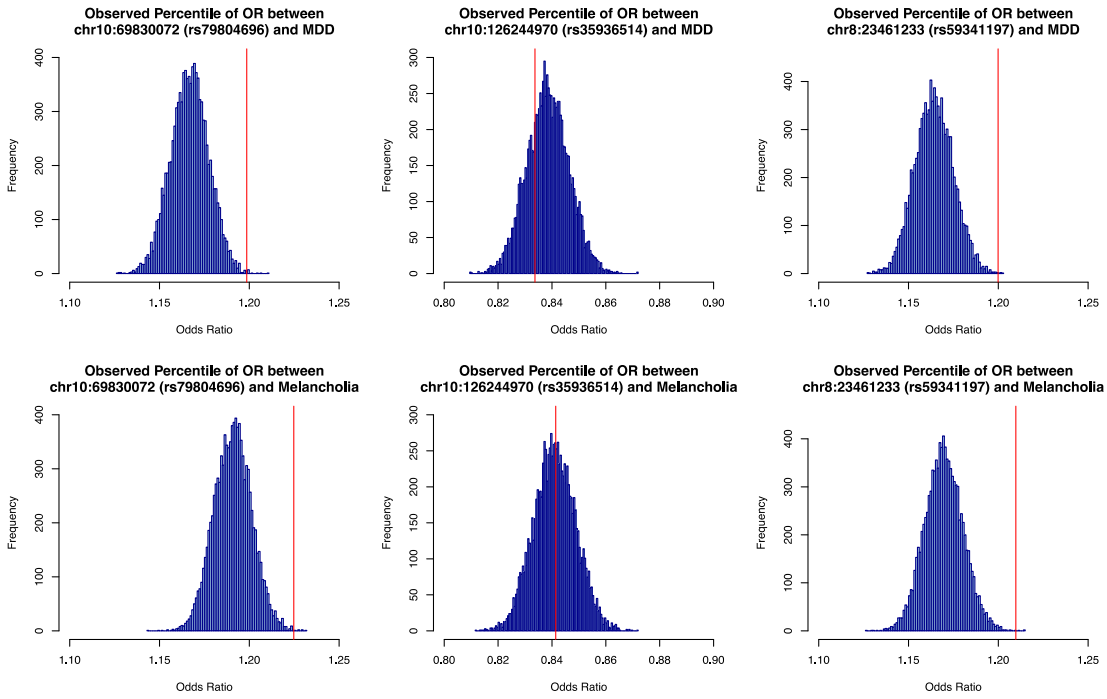


Figure 3.7 | Empirical estimation of the odds ratio increases due to the removal of cases with childhood sexual abuse (CSA) from an association analysis with major depression and melancholia The figures show the empirical distributions of the odds ratios for association with each of three SNPs (rs79804696, rs35936514, and rs59341197), after randomly removing a number of individuals equal to that affected with CSA. The horizontal axis is the odds ratio for each analysis, and the vertical axis the frequency of occurrence of the odds ratio in 10,000 analyses. The vertical red line is the observed odds ratio after removing CSA cases.

3.2 Genetic architecture of MDD

Family and twin studies had concurred in identifying a significant heritable component to MDD; high-quality twin studies estimate the additive genetic effects between 31%–42% [9,66], with no evidence to suggest shared environmental factors contributed meaningfully to the familial aggregation.

Methods have been established in the past five years to estimate "narrow sense heritability" (h2) from SNP data[134,160]. These methods involve constructing a genetic relatedness matrix (GRM) from whole-genome variant data and restricted maximum likelihood (REML) estimation of additive effects of all variants. Assessing the amount of sharing of common variants between cases of MDD and controls in GWAS though computing h2 found an appreciable proportion of the genetic variants involved are likely to be common (variants above 5% in frequency in population, with estimated h^2 of 21-30%, would account for more than 50% of heritability estimated from twin studies), acting independently and additively. Each of these loci was likely to exert a small risk or protective effect on susceptibility to MDD (odds ratio < 1.1), as none of the GWAS to date, using as many as 9000 cases of MDD and matched number of controls, were adequately powered to detect them [66].

One of the possible reasons for this discrepancy is that causal variants for MDD are not well captured or are poorly tagged by genotyping arrays used for genetic studies of MDD, hence their effects were partially or completely missed in the h2 estimations. With whole genome NGS data on all samples in CONVERGE, I was able to address this concern directly, and for the first time, estimate h2 of MDD using not just tagging but all variants in the genome.

3.2.1 Estimation of h2 of MDD using whole-genome variations

I calculated h2 for MDD using genotypes at all 6,242,619 SNPs across all autosomes in 10,442 samples using GCTA [134](Methods), converting the disease status to the liability scale with population prevalence set at 8% to account for the ascertainment of cases in our cohort (where there were about 50% cases). This is a dense set of markers that would be able to capture both more common and rare variation than present on a genotyping array, and would in theory allow the closing of gap on the "missing heritability" of MDD if it was previously underestimated due to the consideration of only common variants that were either directly genotyped on genotyping arrays, or preferentially chosen post-imputation due to the likely higher quality of imputation from a limited reference panel at common variants.

I obtained a h2 estimate of 28.1% (SE = 3.21%) for MDD, higher than previous estimates from Cross-Disorder Group of the Psychiatric Genomics Consortium (PGC) with SNPs from genotyping arrays 21% [161].

To ensure there was no inflation in my estimates because of relatedness between our samples, I constructed GRMs using GCTA (version 1.24.7) per chromosome for all autosomes and performed the estimation of h2 using each GRM, and with all GRMs jointly, getting one random effect estimate for each chromosome in each case. If there was inflation due to relatedness, which would show up as long range linkage disequilibrium (LD) between chromosomes, we would see higher per chromosome h2 estimates from the per GRM estimates of h2 than the joint estimation of per chromosome random effects using all GRMs.

Figure 3.8 plots the h2 estimated per chromosome against the chromosome length for all autosomes, for both methods described. The significant correlation between chromosome length and h2 (Pearson r2 = 0.506, P- 0.000143 for the joint REML analysis, Pearson r2 = 0.495, P-value = 0.000176 for the separate REML analysis)

shows that there are causal loci distributed across the genome, roughly evenly across all chromosomes, and the similar slope between the two methods (r = 0.711 for the joint REML analysis, r = 0.704 for the separate REML analysis) showed that there wasn't cryptic relatedness in the CONVERGE cohort, and the h2 estimates from both per-chromosomal estimates and all-autosomal estimates were unlikely to be inflation due to relatedness between samples.

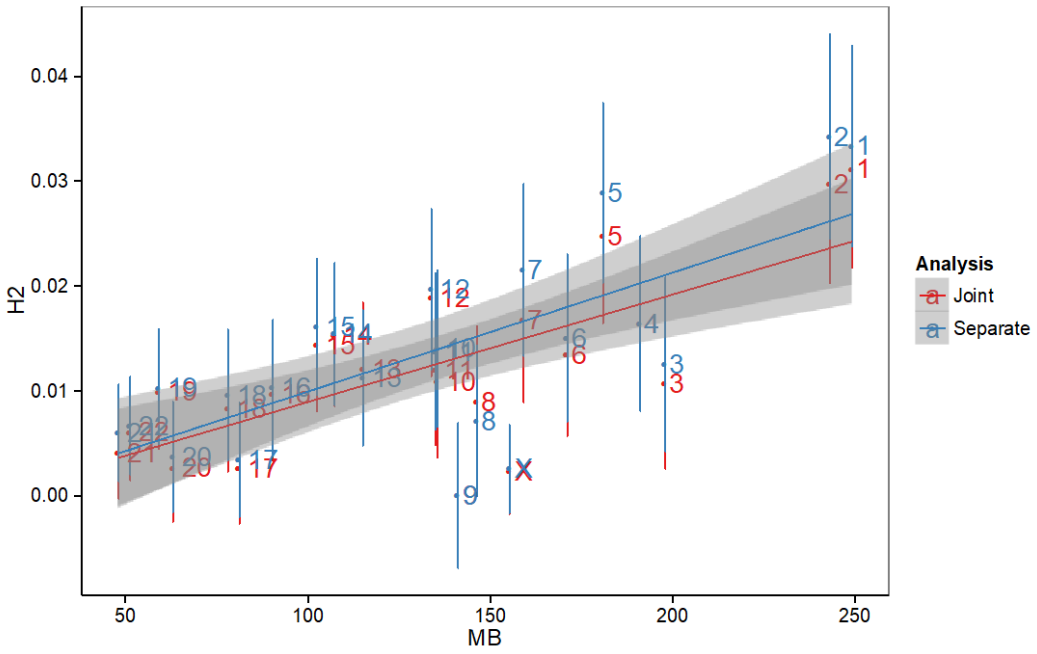


Figure 3.8 | Estimates of h2 of MDD contributed by SNPs per chromosome against the chromosome This figure shows the the h2 estimated per chromosome against the chromosome length for all autosomes and chromosome X, estimated separately using per chromosome genetic relatedness matrices (GRMs) generated from all SNPs in each chromosome, as well as with all per-chromosome GRMs jointly, getting one random effect estimate for each chromosome in each case. There is significant correlation between chromosome length and h2 for both analyses (Pearson r2 = 0.506, P- 0.000143 for the joint REML analysis, Pearson r2 = 0.495, P-value = 0.000176 for the separate REML analysis), suggesting genetic factors contributing to MDD are distributed across the genome such that sampling from a larger portion of the genome would give higher h2, and the similar slope between the two methods (r = 0.711 for the joint REML analysis, r = 0.704 for the separate REML analysis) showed that there wasn't cryptic relatedness in the CONVERGE cohort.

3.2.1 Controlling for local LD

As the genetic data in CONVERGE was imputed at all sites that were likely true SNPs called from whole-genome low coverage sequencing, the number of markers with which the GRM could be constructed with and the degree of tagging of the whole genome were higher than if they came from a genotyping array. As I was using an unprecedentedly dense set of markers for estimating h2, it was important that LD between markers was taken into account.

I first asked, were there regions of high LD among SNPs used in analysis in CONVERGE? A very simple way to visualize if this was a problem was to construct a GRM using all SNPs, where each SNP was given equal weight, and perform principal component analysis (PCA) on the GRM. If there were regions of high LD, such as relatively recent structural variations accumulating in particular sub-populations in the cohort, and many SNPs were in those regions, then they would be driving the PCA and the top PCs were likely to separate samples based on their genotypes at the high LD regions. If there were no such regions, and the genomes in the CONVERGE cohort were homogeneous in SNP content and LD landscape, then the top PCs should be driven by variants accumulating in different sub-populations in the CONVERGE cohort and reflect ancestral differences that could mirror their geographical origins, but not the genotypes at any set of markers. I therefore performed this analysis, constructing an unweighted GRM using all SNPs across all autosomes using GCTA (version 1.24.7), and performing a PCA on the GRM.

Figure 3.9 shows the PCs 1 to 3 from the unweighted plotted against each other – while PC1 reflects the North-South cline, which was the dominant signal in the PCs calculated for correction of population, PCs 2 and 3 clearly reflected variations that were able to separate samples into three distinct genotypes. While this analyses could not

quantify the extent to which there was uneven LD across the genome, it showed how a few high LD regions could dominate the GRM.

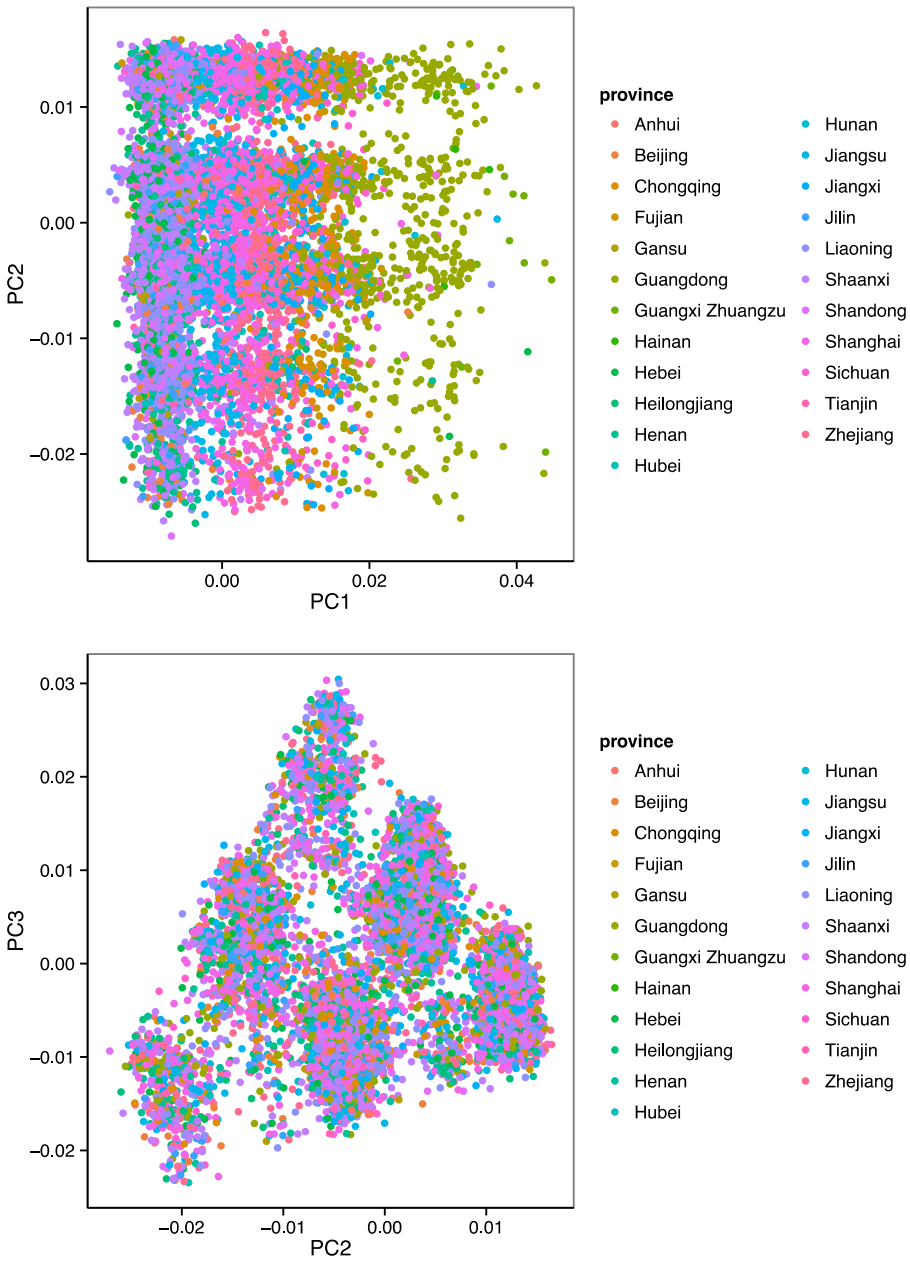


Figure 3.9 | PCA on CONVERGE samples based on GRM constructed with 6M whole-genome imputed SNPs without weighting for local LD This figure plots PC1 against PC2 of a PCA based on a GRM constructed using 6M whole-genome imputed SNPs using GCTA (version 1.24.7) without weighting for local LD, coloured by provinces where the samples were from, in the upper panel; the lower panel shows the same plot for PCs 2 and 3. If the 6M SNPs represent the genome equally, with no uneven LD between them, PCA should extract structure in the population that reflect genetic ancestry, which is usually mirrored by geographical origin of the samples. In this case, as the PCs do not cluster the samples based on their geographical origin, but instead cluster

them into groups recognizable as genotype categories comprising two segregating alleles, it demonstrates the presence of high LD regions in the genome where a disproportionate number of SNPs lie and dominate the PCA.

Why should uneven LD be a concern? Speed et al 2012 had shown that if all variants were given equal weight in the generation of GRMs, such that there is a much higher representation of high LD regions of the genome as compared to the low LD regions (assuming relatively even spacing of variants across the genome), then having the majority of causal loci in regions of high LD are likely to upwardly bias h2 estimates from the resulting GRM, and conversely causal loci in regions of low LD are likely to cause a downward bias[160]. In accordance with this, pruning by LD (decreasing the representation of SNPs in high LD regions but not in low LD regions) would cause decreases in h2 estimates in the former case, and increase h2 estimates in the latter.

I investigated this first by pruning the set of autosomal SNPs by pairwise LD for maximum pairwise r2 between SNPs of 0.999, 0.998, 0.995, 0.99, 0.98, 0.95, 0.9, 0.8, 0.5, 0.2, and 0.1 and constructing GRMs using the resultant set of SNPs for estimating h2 as shown in Table 3.4. This analysis shows that a) about half the SNPs are in perfect LD with another SNP in the dataset – merely pruning the set of SNPs with a LD r2 threshold of 0.999 shaves off more than half the SNPs, and b) pruning to decreasing LD r2 thresholds caused firstly an increase in h2 estimates, reaching a maximum at LD r2 threshold at 0.8, showing that causal SNPs lie mostly in lower LD regions than the thresholds imposed, and further pruning caused a decrease in h2 estimates showing either the causal SNPs lie mostly in LD regions above the threshold imposed, or simply the loss of information on variation in the cohort.

LD r2 threshold	h2	SE	N
0.1	0.219752	0.033848	78118
0.2	0.291947	0.039941	130516
0.5	0.393083	0.044148	312725
0.8	0.409411	0.042003	616445
0.9	0.391321	0.040567	829691
0.95	0.394609	0.039599	1044577
0.98	0.379291	0.038586	1348669
0.99	0.368249	0.037921	1608914
0.995	0.361813	0.037288	1906353
0.998	0.355303	0.0363	2344128
0.999	0.34845	0.035509	2686906
1	0.280616	0.031621	6242619

Table 3.4 | Estimates of h2 of MDD with whole-genome SNP sets pruned by local pairwise LD This table shows the effect of pruning the set of autosomal SNPs by pairwise LD for maximum pairwise r2 between SNPs (thresholds of 0.999, 0.998, 0.995, 0.99, 0.98, 0.95, 0.9, 0.8, 0.5, 0.2, and 0.1) on h2 estimates from the unweighted GRMs constructed with the resulting set of genome-wide SNPs. The first column shows the LD threshold, the second column shows the h2 estimate for MDD, the third column shows the standard error (SE) of the estimate, and the final column shows the number of SNPs in the SNP set after pruning that went into constructing the GRMs (N).

I asked what LD bracket do most causal SNPs for MDD lie in. To prevent SNPs in complete LD with other SNPs from dominating the top quartiles of SNPs when ranked on LD, I calculated the "LD scores", the sum of an unbiased estimator of r2 between each SNP and all other SNPs in a 100kb window[162] (Methods), for all SNPs in the dataset pruned at LD r2 threshold of 0.999 (hence removing only those SNPs that were in

complete LD with each other), then grouping SNPs based on which quartile their respective LD scores lie among that of all SNPs. Then, I computed a GRM for each LD quartile, and estimated the h2 for MDD from each of the GRMs. Table 3.5 shows the results of this analysis – h2 for MDD was similarly in the top 3 quartiles of LD, and drops though not sharply for the lowest quartile, attesting that among all SNPs used in this analysis, most of those contributing to MDD were in mid to high-LD regions of the genome.

In a response to Speed et al 2012, Lee et al 2013 suggested that stratification by MAF should adequately capture the differences between LD landscape different SNPs lie in, and adjust for LD appropriately when estimating h2[163]. To prevent confusing common SNPs with simply SNPs in high LD, which though could be highly overlapping represent different concepts, I stratified all SNPs used in this analysis by both MAF quintiles and LD quartiles, asking therefore what the contribution of SNPs in 20 different stratifications were to MDD. This result is summarized in Table 3.6. The h2 estimate, from a joint REML of GRMs constructed the 20 SNP sets, was 33.3% (SE = 4.24%) should therefore effectively be equivalent to an estimate that is "weighted" for MAF and uneven LD.

LD quartile	h2	SE
1	0.123257	0.040386
2	0.090228	0.032425
3	0.109301	0.026829
4	0.069512	0.016598
Total	0.392298	0.044606

Table 3.5 | Estimates of h2 of MDD with whole-genome SNP sets pruned by local pairwise LD This table shows the effect of pruning the set of autosomal SNPs by pairwise LD for maximum pairwise r2 between SNPs (thresholds of 0.999, 0.998, 0.995, 0.99, 0.98, 0.95, 0.9, 0.8, 0.5, 0.2, and 0.1) on h2 estimates from the unweighted GRMs constructed with the resulting set of genome-wide SNPs. The first column shows the LD threshold, the second column shows the h2 estimate for MDD, the third column shows the standard error (SE) of the estimate, and the final column shows the number of SNPs in the SNP set after pruning that went into constructing the GRMs (N).

	LD quartile							
MAF quintile	1		2		3		4	
	h2	SE	h2	SE	h2	SE	h2	SE
1	0.0000	0.0252	0.0036	0.0119	0.0000	0.0081	0.0029	0.0043
2	0.0132	0.0226	0.0000	0.0151	0.0289	0.0109	0.0000	0.0060
3	0.0000	0.0213	0.0117	0.0168	0.0087	0.0124	0.0123	0.0073
4	0.0400	0.0193	0.0276	0.0179	0.0277	0.0139	0.0062	0.0084
5	0.0463	0.0163	0.0681	0.0167	0.0090	0.0132	0.0262	0.0088

Table 3.6 | Estimates of h2 of MDD with whole-genome SNP sets partitioned by LD quartiles and MAF quintiles This table shows the partitioning of whole-genome SNP effects (h2 and SE) by LD quartiles and MAF quintiles. 6M while-genome SNPs are partitioned in to quartiles by their aggregate pairwise LD r2 with all other SNPs in a 100Kb region, and into quintiles based on their MAF. Intersection of LD quartiles and MAF quintiles was performed to generate 20 similarly sized sets of SNPs each representing a LD and MAF category, and each of which is used to generate a unweighted GRM using GCTA (version 1.24.7). A joint REML on all 20 GRMs was performed to estimate the contribution of each set of SNPs to the h2 of MDD.

As the aim of these analyses was to adequately control for uneven LD in the particularly dense set of markers for h2 analysis while leverage information from all high quality SNPs called from sequencing data (as opposed to pruning, which would lead to loss of information), I employed an alternative method LDAK[160] for constructing the GRM, which was able to take into account LD between markers and weigh the SNPs based on their LD with other SNPs in vicinity. A quick check using PCA if SNP weighting was effective (Figure 3.10) showed no signs of domination by structural variants or high-LD regions as shown in Figure 3.10, and projecting PCs 1 to 3 onto the map of China shows the identification of the North-South cline (PC1) as well as other ancestry differences mirroring geographical origins of the samples.

Using this method I arrived at a whole-genome GRM with all SNPs including those in chromosome X, and obtained a whole-genome h2 of 36.0% (SE 5.00%, P-value < 10^{-16}) at the same population prevalence of depression at 8%. This brings it very close to the estimates from twin studies.

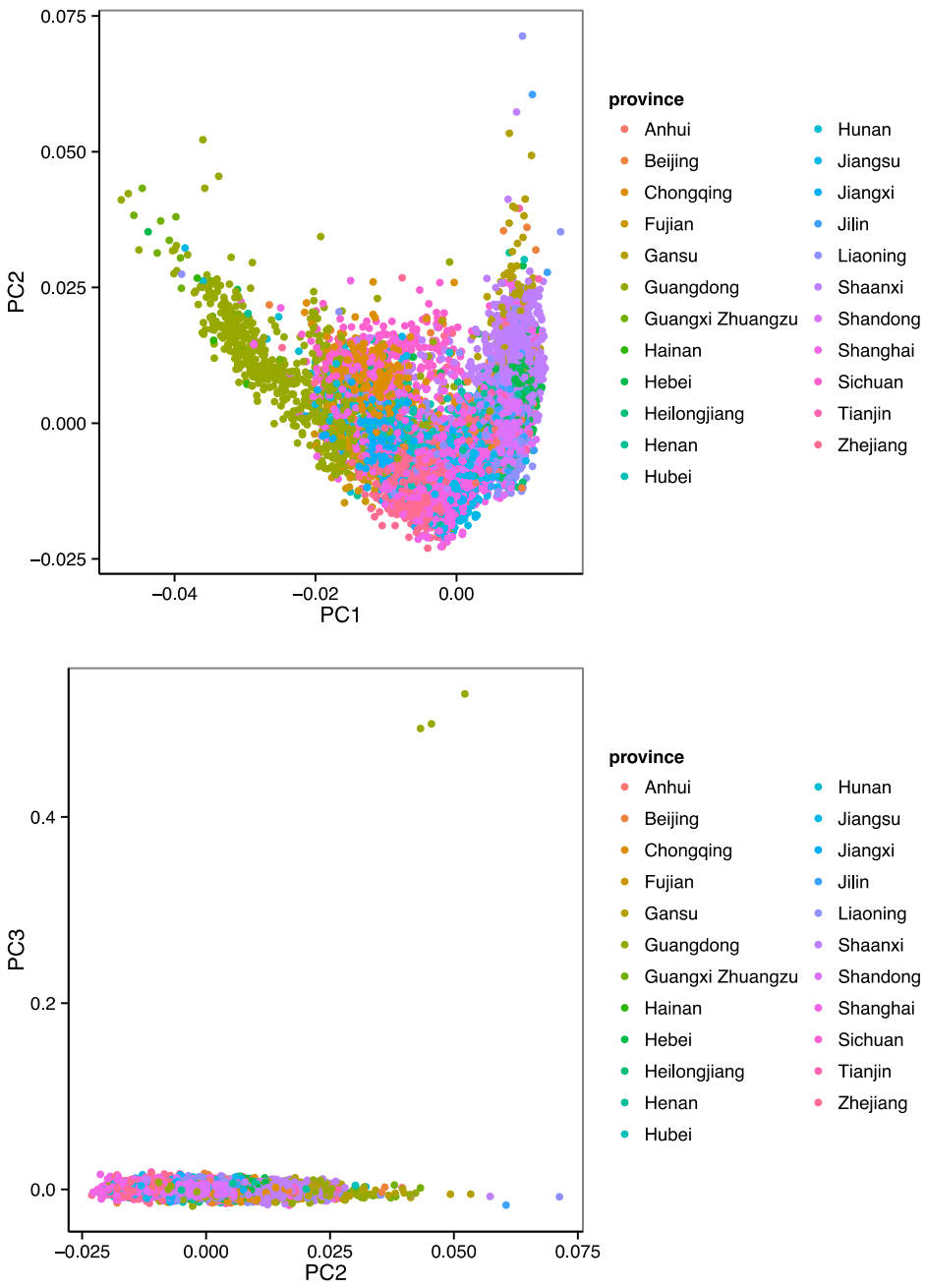


Figure 3.10 | PCA on CONVERGE samples based on GRM constructed with 6M whole-genome imputed SNPs without weighting for local LD This figure plots PC1 against PC2 of a PCA based on a GRM constructed using 6M whole-genome imputed SNPs using LDAK (version 5.9) with weighting for local LD, coloured by provinces where the samples were from, in the upper panel; the lower panel shows the same plot for PCs 2 and 3. Removing the three outliers separated from other samples by PC3 did not cause differences in GWAS results or h2 estimates.

3.3 Polygenic burden of rare deleterious variants

NGS data allows the calling of variants at any position in the genome covered by sequencing reads in more than one sample, and therefore can be used for calling and analysis of rare variants occurring at low frequency in the population or only in single samples. Using this data, I am therefore able to answer the question whether MDD cases have a polygenic burden of rare deleterious coding variants, as previously observed for schizophrenia [164], and if these deleterious variants are enriched in particular gene sets. For this analysis, I restricted analysis to only the exome regions of the genome, and only looked at singleton variants (those detected in a single individual in the CONVERGE cohort), following work on coding variation in schizophrenia where singleton deleterious mutations were the most enriched [164].

3.3.1 Validation of rare variant calls

I called only those variants detected by two or more reads in a single individual, using a custom code that counts the number of high quality reads supporting each allele at every variant position in the exome regions of the genome among all CONVERGE samples. As such, multiallelic sites were also taken into account, though only those alleles supported by two or more reads in a single sample, and occurring only in that sample, would be counted (Methods). Using this approach, I identified 302,838 rare coding SNPs and insertion-deletion polymorphisms (INDELs). Figure 3.11 shows the distribution of number of samples with more than two reads at singleton sites we discovered, and the graph below it shows the cumulative percentage of singleton sites with more than each percentage of samples having 2 reads covering it. The mean number of samples having more than 2 reads at a singleton site is 4,126 (38.8% of the whole sample), corresponding to an allele frequency of 0.024%.

As this analysis was based on hard filters for high quality sequencing reads, and counting of number of reads supporting each private allele, there were concerns regarding the true positive and false positive rates of variant calling, therefore affecting the validity of using these variants for any further analysis. I therefore compared the positions and alleles called at private variants sites in nine samples for whom there were 10X coverage sequence data. This comparison estimated a true positive rate of 96.5% for SNPs and 97.2% for INDELs (Table 3.7), while none of the private variant calls were false positives. The false negative rate, however, is very high, as requiring two reads support of 1x coverage sequencing for a private variant is a very stringent filter that most heterozygous sites, which are expected to make up most of the private variant sites, would not be able to fulfill.

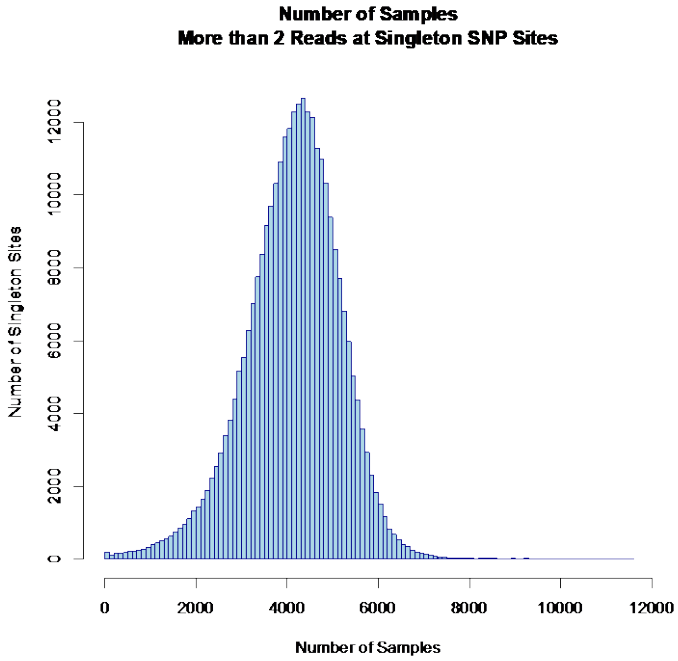


Figure 3.11 | Histogram of number of samples with more than two reads at private variant sites discovered in CONVERGE This histogram shows the number of samples with more than two reads at all 302,838 private variant sites discovered in CONVERGE. The mean number of samples having more than 2 reads at a private variant site is 4,126 (38.8% of the whole cohort), corresponding to an allele frequency of 0.024%.

Sample Number	Variants occurring in only one sample in CONVERGE			
	SNPs		INDELs	
	Discovered	True Positive (%)	Discovered	True Positive (%)
1	77	96.10	6	100.00
2	56	94.64	4	100.00
3	44	88.64	9	88.89
4	57	100.00	6	100.00
5	41	100.00	4	100.00
6	83	96.39	3	100.00
7	88	96.59	12	100.00
8	43	100.00	1	100.00
9	72	95.83	7	85.71

Table 3.7 | True positive rates in the exomes of variant calls occurring only once in the CONVERGE cohort. The table shows the results of comparing singleton variants (defined as occurring on two or more reads in a single individual out of 10,640) from 1.7x coverage sequence data, with genotypes called on the same individuals from 10X coverage data. The first column shows the Barcode of the nine samples for whom we have 10x coverage data. The next two columns show the number of singleton SNPs discovered in the 1.7x coverage sequencing data in each sample (Discovered) and the percentage of it confirmed by the 10x sequencing data (True Positive (%)). The next two columns show the same information for insertions and deletion mutations (INDELs).

To further validate the variant calls by variant type and chromosomal context, I performed a validation experiment for 75 private variants randomly chosen to represent each type of variant (SNPs, and insertions and deletions each of varying length) and each chromosome and chromosomal position (near ends of chromosome, near centromeres, others). Samples these private variants were chosen from were random from the whole CONVERGE cohort, regardless of MDD case status.

I obtained Sanger sequencing data for 64 variants of the 75 variants selected (6 primer design failures, 5 PCR failures). Alleles were called at the private variant sites from manual inspection of the Sanger sequencing traces. Sanger sequencing of 52 randomly chosen variants confirmed with Sanger sequencing for the right variant, mostly heterozygous, showing a high true-positive predictive value (81.25%) of the rare variant calling process. Three examples are shown in Figure 3.12. Of the 12 sites that did not validate, 9 were insertions, while there was no insertion that validated. As such, I disregarded all insertion variant calls for further analysis, since the calling accuracy for insertions were apparently worse than the other types of variants. Removing the insertions from the raw call set would give, based on the Sanger sequencing validation, a 94.5% true positive rate of calling the other types of variants. Based on this validation exercise, the majority of singletons in our data are likely to represent true variant calls.

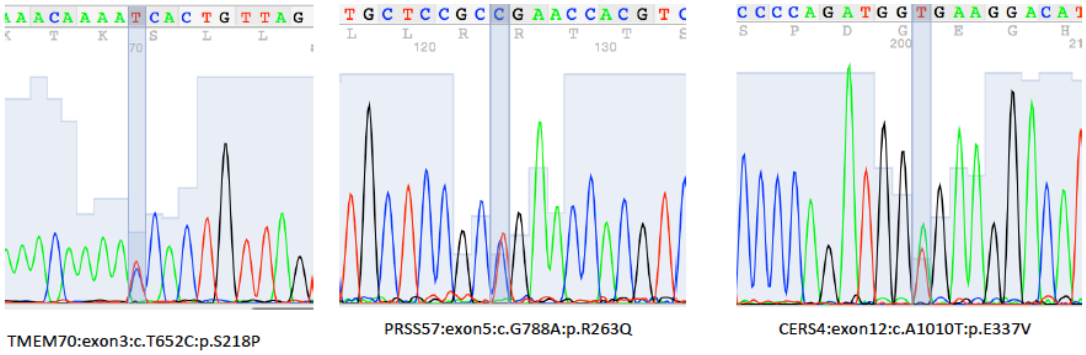


Figure 3.12 | Sanger sequencing read trace of three private variants validated. This figure shows three sanger sequencing traces showing private variants in three samples at three locations in the exome, the leftmost being a synonymous SNP with T/C genotype at the third position of codon 652 in exon3 of *TMEM70* gene, the centre being a non-synonymous SNP with G/A (shown as the reverse strand, T/C) genotype at the third position of codon 788 in exon 5 of *PRSSS7* gene, and the rightmost being a non-synonymous SNP with A/T genotype at the second position of codon 1010 in exon12 of gene *CER54*.

3.3.2 Private variant enrichment in MDD

There are a large number of hypotheses we could potentially test about the role of rare variants in the aetiology of major depression – one could examine rare variants occurring at different frequencies, variants belonging to different annotation classes, classified using different algorithms, occurring in different gene sets and occurring in different diagnostic categories. Given the large sample size in CONVERGE but low sequencing coverage over the exome, and the high false negative rate in detection of private variants as a result, I could only test whether there are differences in the frequency of private variants in cases of MDD compared to controls.

The primary hypothesis was therefore threefold, in order of specificity: a) that private variants would be enriched in cases of MDD, b) private variants that cause coding changes could be enriched in cases of MDD, and c) private deleterious variants would be enriched in cases of MDD. A Bonferroni corrected P-value for this test is 0.05/3 = 0.0167. To classify the private variants based on their effect on coding and potential consequences to protein function, I annotated all the private variants using ANNOVAR (Methods), and defined as 'deleterious singletons' the sum of stop gain, stop loss, nonsynonymous, and frameshift deletion variants [165].

I then tested for differences in the frequency of private variants between cases of MDD and controls for four categories of variants: a) all variants, b) synonymous SNPs, c) non-synonymous SNPs, and d) all deleterious variants (including stop-gains, stop-losses and frameshift deletions). While cases of MDD and controls did not differ significantly in sequencing metrics, including origin, ancestry (based on principal components), experimental batch and technical features of the sequencing (per gene coverage, GC content, and sequence quality), cases had significantly more singleton deleterious mutations than controls (P = 0.003 from logistic regression, Table 3.8).

To confirm the P-values from the enrichment analyses were not inflated due to technical artefacts yet accounted, I ran a series of analyses in which the MDD case status was permuted. For each permutation of the MDD case status I obtained a P value from logistic regression for the number of private variants per sample on the permuted MDD case status. I performed 10,000 such permutations and logistic regressions to generate a distribution of P-values. Quantile-quantile plots were then generated to see if there was any deviation of P-values from the null distribution. These plots indicate no inflation of P-values (Figure 3.13).

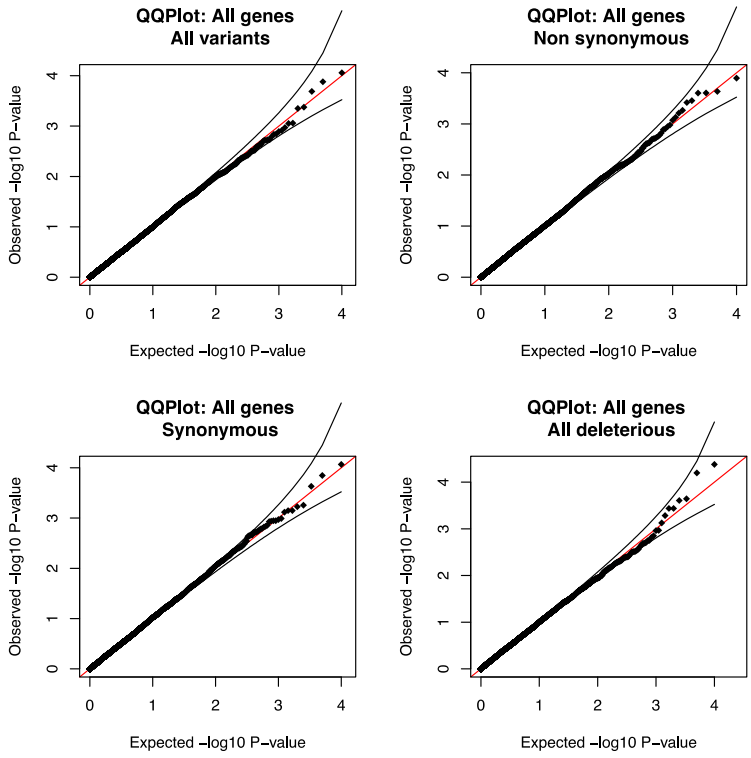


Figure 3.13 | Quantile-quantile plot for the enrichment of coding private variants in cases of MDD Each graph shows the expected association value (negative logarithm base 1) on the horizontal axis and the observed value on the vertical axis. Data are from 10,000 permutations in which MDD case and control status has been randomly assigned. Plots are shown for all variants, and three annotation categories: non synonymous, synonymous and deleterious variants. The latter are defined as the sum of stop gain, stop loss, nonsynonymous, and frameshift deletion variants, using annotation categories from ANNOVAR software.

3.3.3 Private variant enrichment in brain and mitochondrial genes in MDD

Upon finding enrichment in deleterious variants in cases in all genes, I asked if I could narrow the sets of genes driving this enrichment by asking two further questions: a) is the enrichment present in variants occurring in genes expressed in the brain, as opposed to genes not expressed in the brain, and b) is the enrichment present in nuclear encoded genes whose protein products are exported to the mitochondrion?

I therefore counted singleton deleterious mutations occurring in genes expressed in brain, not expressed in the brain, and those whose protein products were localized to the mitochondria. The Bonferroni corrected significant P value for all three hypothesis, when applied on the four gene sets (including the original set containing all genes), is therefore 0.05/12 = 0.0042.

Table 3.8 shows a significant (P = 0.004) excess of private deleterious mutations in cases for genes in expressed in the brain. In contrast, no significant enrichment was seen for associations between MDD and variants in genes not expressed in brain (P = 0.41).

While the choice for genes expressed in the brain is intuitive given MDD is a psychiatric disease likely to have neurological underpinnings in its etiology, I had found GWAS loci near two genes with mitochondrial functions, *SIRT1* and *SLC25A37*. This led to the inquiry of whether singleton deleterious mutations would be enriched in nuclear-encoded genes with mitochondrial localized gene products [166]. As the genes whose products are localized to the mitochondria may have "mitochondrial copies" in the mitochondrial genome similar in sequence, and in some cases encoding similar subunits for the same respiratory complexes, I intersected the list of nuclear encoded genes whose proteins are imported into the mitochondria[166] with the nuclear mitochondrial (NUMT) loci from the hg19 NUMT track from UCSC, and found only three nuclear

encoded genes contained NUMTs (these are: ENSG00000171612, ENSG00000132286 and ENSG00000102738), then excluded all three genes from the analysis.

Table 3.8 shows a modest enrichment in deleterious variants in nuclear-encoded mitochondrial genes (P = 0.009), not passing the significance threshold after Bonferroni correction, but in the direction as expected.

Gene set	Variant type	N	OR	95% CI	P
All (18176)	All variants	302838	1.005	1.001 - 1.009	0.03
	Non-synonymous	179546	1.008	1.001 - 1.014	0.03
	Synonymous	96970	0.997	0.986 - 1.008	0.61
	All deleterious	202374	1.009	1.003 - 1.014	0.003
Brain expressed (10897)	All variants	192137	1.006	0.999 - 1.011	0.07
	Non-synonymous	113136	1.013	1.004 - 1.023	0.006
	Synonymous	62882	0.992	0.977 - 1.004	0.18
	All deleterious	127057	1.011	1.003 - 1.018	0.004
Non brain expressed (7,272)	All variants	110701	0.999	0.992 - 1.009	0.87
	Non-synonymous	66410	0.996	0.983 - 1.009	0.52
	Synonymous	34088	0.988	0.970 - 1.008	0.26
	All deleterious	75317	1.000	0.994 - 1.014	0.41
Nuclear encoded mitochondrial (940)	All variants	11103	1.053	1.001 - 1.096	0.04
	Non-synonymous	6536	1.063	1.001 - 1.129	0.04
	Synonymous	3226	1.004	0.943 - 1.069	0.91
	All deleterious	7414	1.075	1.018 - 1.135	0.009

Table 3.8 | Effect of singleton coding variants on the risk of MDD. The table shows the increased risk of MDD with the number of rare coding variants (where rare means that each variant is found only in a single individual in the CONVERGE sample). Numbers of variants (N) are given for coding variants in all genes, for those expressed in the brain, for nuclear genes that have a role in mitochondrial function, and for brain expressed genes excluding mitochondrial genes. The table gives the odds ratio (OR), the OR 95 % confidence intervals, and the P-values of this analysis for the complete set of variants, and for variants annotated by their predicted function. The category 'all deleterious' includes non-synonymous, stop gain, stop loss, and deletion frameshift variants.

To determine if the increase in the odds ratios attributable to deleterious variants in brain-expressed and nuclear-encoded mitochondrial genes was likely due to chance effects, I randomly selected a set of genes whose total coding length was equal to each gene set, and tested for enrichment in deleterious singletons, keeping case and control status unchanged in 10,000 random selections. The empirical P-value for the odds ratio observed in the brain-expressed gene set was 0.035 and 0.017 for that in the nuclear-encoded mitochondrial genes (Figure 3.14), indicating that the enrichments are unlikely due to chance.

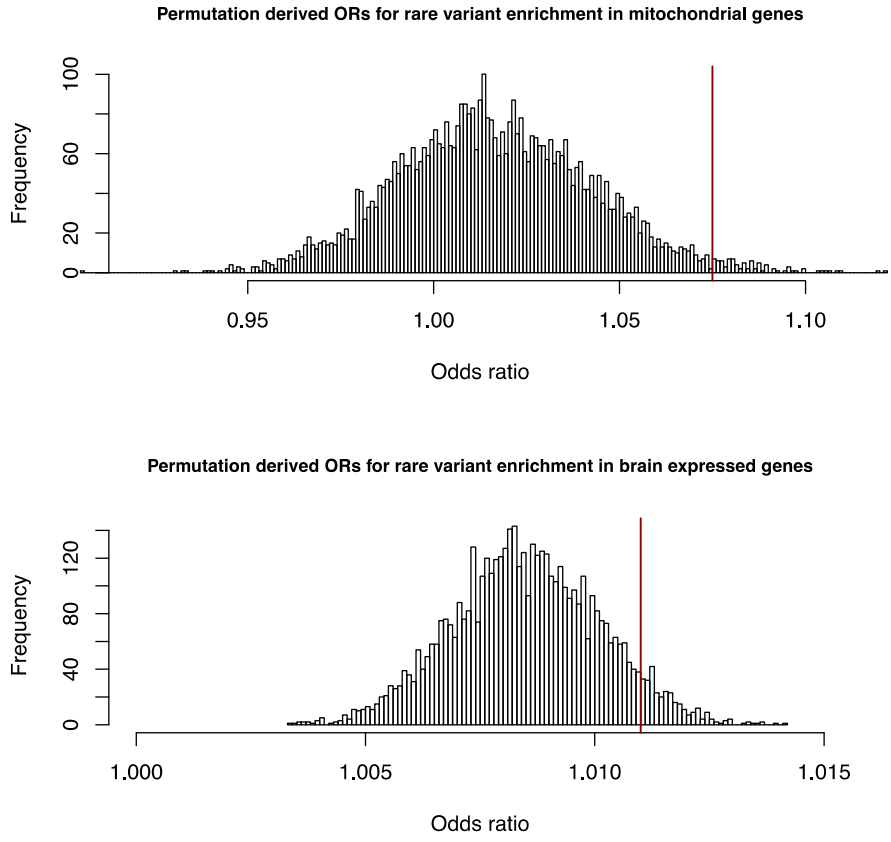


Figure 3.14 | Enrichment of singleton variants in gene subsets. a, Empirical estimation of the odds ratio increase when singleton enrichment analysis is restricted to a subset of genes The graphs show the empirical distribution of odds ratios obtained by randomly selecting an amount of coding DNA equal in length to that used when the analysis is restricted to nuclear encoded mitochondrial genes (upper figure) and genes expressed in the brain (lower figure). Horizontal axes show odds ratios and the vertical axis the frequency with which an odds ratio occurs in 10,000 analyses. The vertical red lines are the observed odds ratios from analyzing enrichment of singleton variants in MD cases.

3.4 Discussion

This chapter described the discovery of two genome-wide significant associations, one implicating *SIRT1* and the other *LHPP*, the bridging of the "missing heritability" using imputed dosages at whole-genome SNPs identified from NGS of all samples, as well as enrichment of rare deleterious mutations in brain-expressed and nuclear-encoded mitochondrial genes in cases of MDD. For MDD, for which there has been limited success in genetic studies, these results are not only encouraging but instructive, as they point towards novel biological pathways that do not coincide with the large number of proposed mechanisms underlying the neurobiology of MDD. The messages from this section of my work, summarized in this chapter, are three-fold.

Firstly, finding a genetic association at SIRT1 opens up new possibilities for research into the etiology of MDD. Reassuringly, restricting MDD cases to only those presenting a severe subtype of MDD melancholia resulted in an increase in genetic signal in association testing at the *SIRT1* locus, and replication in an independent cohort of severe MDD cases and matched controls in China showed significant associations in the same direction of effect. This suggests the association signal represents a real biological signal that is more often present in melancholic MDD cases than other MDD cases and the controls.

This compels more interest in the gene SIRT1. SIRT1 belongs to an evolutionarily well-conserved family of Nicotinamide adenine dinucleotide (NAD^+) - dependent deacetylases and ADP-ribosyltransferases involved in numerous cellular processes, including gene silencing, DNA repair, and metabolic regulation [167-169]. Deletion of *SIRT1* in mice abolishes the well-studied effects of caloric restriction on physical activity [170] and life span extension [171], while overexpression of *SIRT1* mimics many of the positive effects of caloric restriction, including a reduced incidence of cardiovascular and metabolic diseases [172,173], cancer [174] and neurodegeneration [168]. A

systematic analysis of changes in activity of variants of SIRT1 under a range of conditions and in a range of tissues may not have been possible to date. As the variants used in GWAS came from NGS and may be the causal variants instead of merely tagging them, it may be worthwhile to prioritise candidate causal variants from the association results around the SIRT1 GWAS peak for functional testing[175].

Secondly, in agreement with the rationale of reducing phenotypic heterogeneity, maximizing genetic contribution to disease and minimizing differences due to genetic ancestry of samples in the design of CONVERGE, I have found that increasing selectivity on samples used for association studies increases power for finding genetic associations. In addition to the increase in genetic signal with restriction of MDD cases to only those with melancholia, restricting analysis to MDD cases who had not suffered CSA (the most predictive early life adversity of later in life MDD) also increased the genetic signal at *SIRT1*, while causing no change in the signal at *LHPP* and unmasking another associated locus at *SLC25A37*. Results from COVERGE have shown that strict selectivity to reduce heterogeneity and maximizing genetic effects brings the required number of samples down.

Lastly, this chapter presents the first attempt at accounting for the "missing heritability" in MDD by using all variations, common and rare, obtained from whole-genome NGS. It is the first time that SNP-based heritability estimate came near that obtained from twin studies, and the analysis has also shown that it was not just the use of all variants that allowed the capturing of more genetic variation that contributed to the genetic liability to MDD, but the application of methods appropriate for the uneven distribution of causal genetic variants across regions of high and low LD in the genome. This will become more important in future studies using dense sets of genetic markers either directly obtained from sequencing or genotyped on arrays then imputed into haplotypes identified from sequenced reference genomes like those in HapMap and 1000 Genomes Project.

Partitioned analysis showed that variants across all MAF quintiles contributed to liability of MDD, though most of the contribution came from common variants. This was corroborated by evidence of enrichment of deleterious private variants in cases of MDD, especially in particular gene sets. A significant enrichment of private deleterious variants in cases of MDD was found, and restricting to only genes in expressed in the brain or whose products are translocated into the mitochondria preserved the enrichment signal while the complementary gene sets did not, lending credibility to the enrichment signals through potential functional relevance. As brain expressed genes account for more than half the total number of genes encoded by the genome, refinement of definition of gene sets for investigation is needed for better resolution and identification of potential functional networks and pathways. The identification of such finer defined gene sets for private variant studies can be informed by interrogating variants from a wider range of frequencies. Methods combining the information in summary statistics from GWAS on traits of interest and that in expression levels and co-expression patterns obtained from multiple tissue samples[176] are emerging as a means to suggest the involvement of specific tissues and co-expressed genes in the traits. Other methods directly interrogating the additive effect sizes in sets of genomic regions or genes using genotype data, allowing different effect size distributions of variants in different sets, can be applied to estimate the relative contribution of candidate gene sets [177].

One possible criticism for rare variants analysis in CONVERGE is the lack of accuracy in identification of rare variants from low coverage NGS. This was indeed a problem, as false negative rates were very high (>80%, as verified using 10x sequencing on nine samples), and false positive rates were only kept low after very stringent filters over the number and quality of reads supporting the private variants. Targeted deep sequencing of the genes with good statistical evidence to be involved in MDD can be performed as a cost-effect way (as opposed to whole-exome sequencing) of identifying

private or very rare variants contributing to MDD. Enrichment analysis on accurately called deleterious private or very rare variants can be done post-hoc to confirm the validity of gene set identification, and the functional consequences of these variants can be examined to elucidate direction of effect gene function (protective or predisposing to MDD) and disturbances to interactions and pathways.

4. Molecular markers associated with MDD

This chapter discusses the two variable components of the somatic genome suspected I had found to be associated with adverse life experiences: telomeric and mitochondrial DNA (mtDNA). Accelerated shortening of telomeres, the sequence that caps the ends of chromosomes, had been associated with stress[42,44], anxiety [178] and MDD [46] (although not all findings have been replicated[179,180]). Abnormal mitochondrial morphology and altered metabolic activity has been reported in mood disorders [181]. The aim of this chapter is to establish whether telomere length and the levels of mitochondrial DNA represent markers of stress-related illness, and to explore how such molecular signatures might arise.

CONVERGE collected aggregate measures of lifetime adversities, including assessments of childhood sexual abuse [182] and stressful life events [183]. In CONVERGE, as in other populations, both childhood sexual abuse and stressful life events are strongly associated with risk for MDD. More severe forms of abuse are more strongly associated with MDD than milder forms, consistent with a causal relationship [158,183]. Using genomes of subjects in CONVERGE, I was able to look for molecular changes that are associated with disease status and documented adversities experienced by the samples.

4.1 Shortened telomeres and increased mitochondrial DNA (mtDNA) in MDD

I first examined the relationship between mean telomere length and MDD. I estimated the mean length of telomeres (across all chromosomes) using Telseq [184] (version 0.0.1) from all CONVERGE samples, then transformed the distribution to normality using a quantile normal function after controlling for sequencing batch and sample age. I then tested for the association between nTel with MDD in a logistic regression model. MDD was associated with shorter normalized mean telomere length: odds ratio for the standardized measure of mean telomere length's contribution to the risk of MDD is 0.85 (95% confidence interval (CI) = 0.81-0.89, P = 2.84×10^{-14}). Figure 1a shows the distributions of nTel in cases and controls.

Similarly, I estimated for each sample the levels of mitochondrial DNA (mtDNA), for which we had a mean per-sample coverage of 102X. Average read depth per 100 base pairs is calculated for the mtDNA reads mapped to the human mitochondrial reference NC_012920.1 both before and after filtering out poorly mapped reads (Methods). As there are regions in the mitochondrial genome replicated in the nuclear genome commonly known as NUMTs (nuclear copies of mitochondrial DNA), which are most likely present as secondary alignments, I calculated average read depth per segment of 100 base pairs in the mtDNA alignments both before and after filtering and compared the average read depth per 100 base pairs before and after filtering. To reduce errors in estimation of coverage due to NUMTs, segments with big differences in read depth (>5% of the filtered read depth) between the filtered and unfiltered alignments that are more likely to span NUMTs are excluded from the calculation of mtDNA copy number (Figure 4.1).

I arrived at a measure of mtDNA copy number by taking the mean read depth in the filtered alignments across all remaining 100 base pairs segments, then regressing out sequencing batch, sample age and average filtered read depth per site of the whole chromosome 20, and finally transforming the residuals to normality using a quantile

normal function. As with mean telomere length, I tested for the association between amount of mtDNA with MDD. Cases had higher normalized levels of mtDNA than controls: the odds ratio for amount of mtDNA's contribution to the risk of MDD was 1.33 (95% CI = 1.29-1.37, P = $9.00X10^{-42}$). The direction of effect of amount of mtDNA on MDD was opposite to that observed for mean telomere length. Figure 1b shows the distributions of amount of mtDNA for the cases and controls.

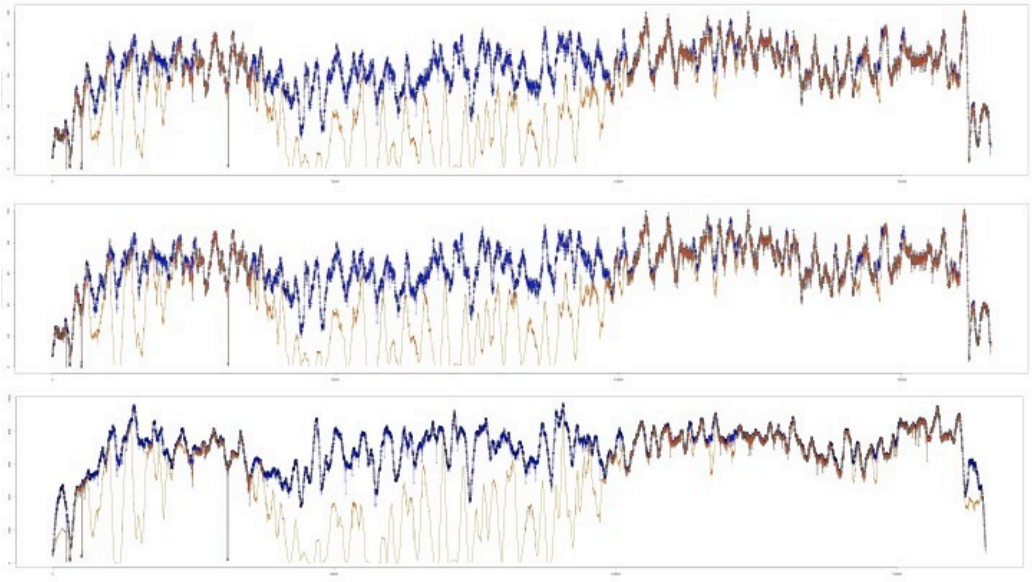


Figure 4.1 | Three samples' sequencing coverage over mitochondrial genome before and after filtering on poorly mapped reads This figure shows the mean read depth per 100bp across the mitochondrial genome from three individuals (top two with high amount of sequencing reads mapping to mtDNA reference, one of which is the a case of MDD and the other a control, the bottom one with low amount of sequencing reads mapping to mtDNA reference), blue lines are unfiltered read depth, red lines are filtered. Filtering implemented in samtools (version 1.08) removed reads that did not passing quality controls, was unmapped, had its paired segment unmapped or mapping to a different chromosome, or was a non-unique or secondary alignment, PCR or optical duplicate. The three traces were representative of other samples, and consistent among themselves in the regions of the mtDNA (from 5000 to 10000bp) where mapping was poor, potentially due to the presence of NUMTs in the nuclear genome were reads originating from mtDNA could map to, causing non-unique mapping to the mtDNA reference.

4.2 Replication of finding in independent cohorts

I was able to replicate the association between a history of MDD and amount of mtDNA quantified through quantitative polymerase chain reaction (qPCR) in two European case-control studies GenDEP and United Kingdom Depression Case-Control study (DeCC). In contrast to the CONVERGE sample, the DNA was extracted from blood, and samples were of both sexes. qPCR measures of relative amount of mtDNA to nuclear DNA was obtained on 216 individuals (108 cases of MDD and 108 controls, 123 women and 93 men). In a logistic model incorporating age as a covariate, the odds ratio for the standardized measure of amount of mtDNA's contribution to the risk of MDD was 1.35 (95% CI = 1.11-2.10, P = 8.3×10^{-5}; Figure 1c).

Despite being a much smaller sample, these data showed much more marked separation between cases and controls in their amount of mtDNA. This might be because the blood from the cases of MDD in these two samples was taken when they were in the middle of episodes of depression. In contrast, DNA was not necessarily taken from all cases of MDD in CONVERGE when they were in episodes of depression, but some when they came into hospitals for health checks, as CONVERGE had included as cases of MDD all those who had history of multiple episodes of severe MDD.

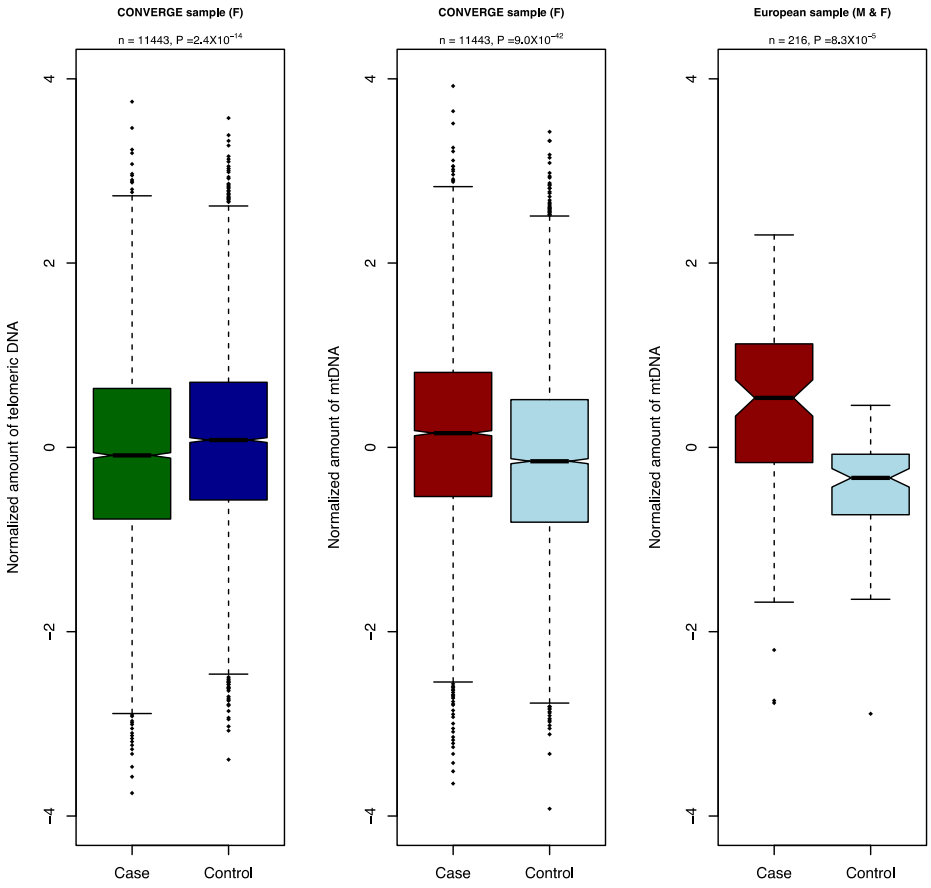


Figure 4.2 | Two molecular markers of depression: mitochondrial DNA and telomere length (A) Boxplot of normalized measure of mean telomere length (vertical axis) for cases and controls in the CONVERGE study. (B) Boxplot of the normalized measure of mtDNA levels (vertical axis) in cases of MDD and controls in the CONVERGE study. (C) Boxplot of the normalized measure of mtDNA levels (vertical axis) in cases of MDD and controls in the GENDEP/DECC studies (labeled European).

To test if this was true, I obtained qPCR measures of relative amount of mtDNA to nuclear DNA in an independent Chinese cohort (54 samples, 27 cases of MDD, 27 controls) of severe MDD, where the source of DNA was blood and samples were of both sexes. All cases of MDD in this cohort were scored on the Hamilton scale (1-30) for the state of their mood at the same time their blood was taken; a high score would mean low mood with 30 being suicidal and above 25 being severely depressed. Cases of MDD selected for qPCR in this cohort all had Hamilton scores above 25 (mean = 29.43, standard deviation = 1.29). If the amount of mtDNA was a state measure of mood, then

selecting cases of MDD with the highest Hamilton scores should allow me to separate out cases of MDD from controls in the Chinese cohort just like I did in the GenDEP and DeCC cohorts. And indeed, the odds ratio for the standardized measure of amount of mtDNA's contribution to the risk of MDD was 1.26 (95% CI = 1.11-1.44, P = 8.16×10^{-4}, Figure 4.3).

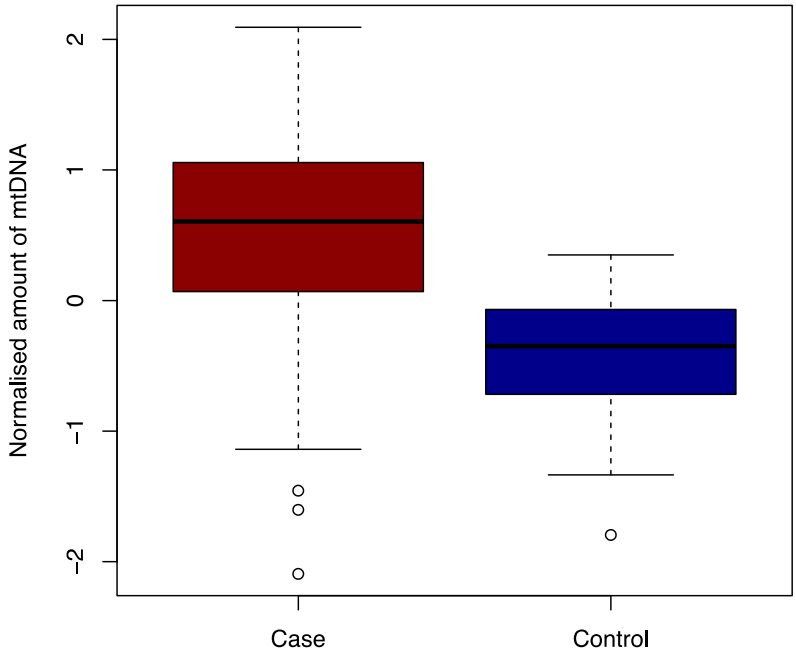


Figure 4.3 | Mitochondrial DNA in MDD cases and controls in a Chinese cohort Boxplot of the normalized measure of mtDNA levels (vertical axis) in cases of severe MDD (n=27) with mean Hamilton score of 29.43 and controls (n=27) in the an independent cohort from Northern China

4.3 Verifying association represent bona fide biological signal

One major concern over the quantification of amount of mtDNA and mean length of telomeric DNA from sequencing or qPCR of DNA samples was there were many points where experimental errors and technical artefacts could be introduced. For example, DNA extraction from different saliva or blood samples were likely to yield different amount of DNA, with different proportions of nuclear and mitochondrial DNA. A higher yield of DNA would more likely give a proportion more representative of the true ratio between nuclear and mitochondria DNA in the samples' cells, while a lower yield would likely contain proportions of DNA more subject to random variations. This is yet one of the many steps in the processing of the samples that could introduce random errors, characterizing errors incurred by each of these steps would be impractical.

4.3.1 Robustness of telomere length and amount of mtDNA measures

It is important to characterize the robustness of amount of mtDNA and mean telomere length measures per sample . If I could show the variances between measures of amount of mtDNA and mean telomere length from different samples of the same person are smaller than that from 18 randomly selected samples from different individuals, then I could be more confident in using the two measures for assessing differences between different individuals.

In the CONVERGE cohort there was one individual who from whom there were 18 saliva DNA samples at different time points during the study, all of which were sequenced separately just like ones collected from different individuals. I therefore compared the variance in amount of mtDNA and mean telomere length measures from these 18 samples to the distributions of variances between the respective measures of 10,000 randomly selected sets of 18 samples from the entire cohort. Figure 4.4 shows the distributions of the randomly sampled data for both measures, where the red

vertical lines show the position of the variance of the 18 samples collected from the same individual. The variance of mtDNA from the 18 samples collected from the same individual lies in the lower 5.9th percentile of the distribution, and that of the mean telomere length measure lie in the lower 1.4th percentile.

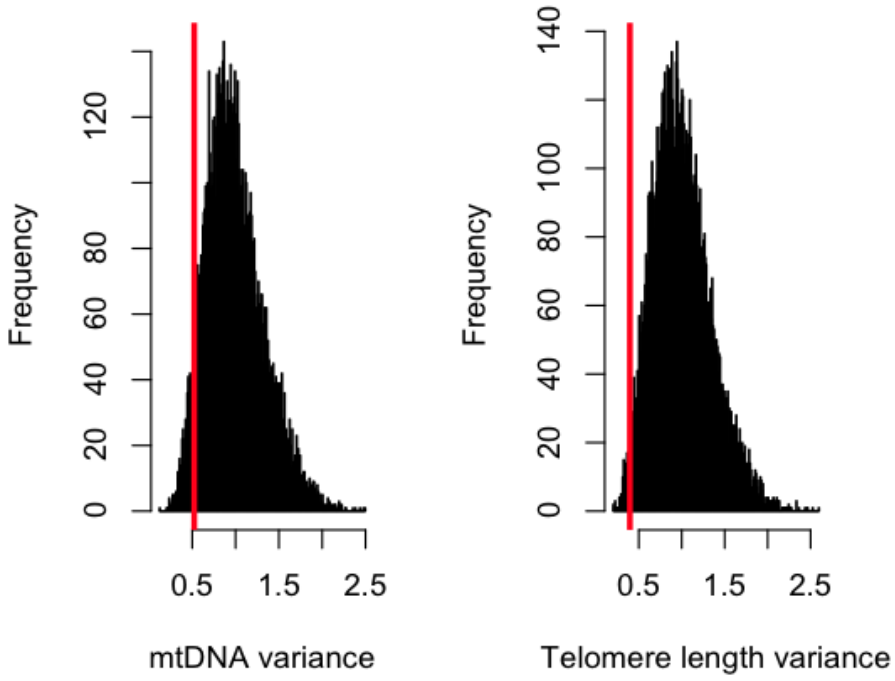


Figure 4.4 | Percentile of variance in measures of amount of mtDNA and mean telomere length from 18 samples from a single individual in an empirical distribution consisting of variance of the same measures obtained from 10,000 randomly selected sets of 18 samples. On the left is a histogram of variance in measures of amount of mtDNA obtained from low coverage sequencing data of 10,000 randomly selected sets of 18 samples from the CONVERGE cohort; the red line marks the variance in measures of amount of mtDNA obtained from 18 samples from the same individual that were collected and sequenced separately lying at the 5.9th percentile of the distribution. On the right the same information is displayed for measure of mean telomere length; the red line marks the variance in measures of mean telomere length obtained from 18 samples from the same individual lying at the 1.4th percentile of the distribution.

In addition to the single individual sequenced 18 times, there were 92 individuals in CONVERGE who were sequenced twice (as they were admitted to two of the participating hospitals). The Pearson correlation coefficient between the mtDNA

measures was 0.28 (p-value = 0.009) and 0.19 for the telomere measure (P-value = 0.018), compared to a non-significant (P = 0.63) correlation of 0.0047 for 10,000 randomly selected pairs of mtDNA measures and a non-significant (P = 0.92) correlation of 0.001 for 10,000 randomly selected sets of 92 pairs of telomere measures. Figure 4.5 shows the correlation coefficients for the 92 duplicate samples (vertical red bar) and correlations obtained from 10,000 randomly sampled 92 pairs.

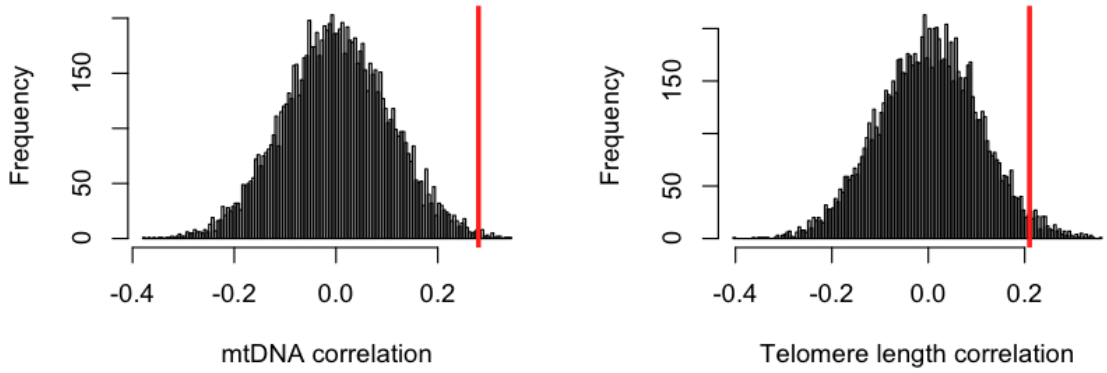


Figure 4.5 | Percentile of correlation coefficient between two measures taken from the same individual among distribution of correlation coefficients obtained from two measures taken from 10,000 randomly selected pairs of individuals. On the left is a histogram of Pearson correlation coefficients between 92 pairs of measures of amount of mtDNA obtained from the same individuals (marked by the red line), in a distribution of correlation coefficients from 10,000 randomly selected sets of 92 pairs of measures from different individuals. On the right the same information is displayed for measure of mean telomere length.

These analyses showed that the measures obtained from the same individual have lower variance than a randomly selected group, indicating the measures obtained from low coverage sequencing were robust and could be seen as a property of the individuals they came from, and therefore could be used for comparison between individuals.

4.3.2 Potential artefacts

Why should both markers be associated with MDD? Were there real links between these molecular markers to disease, or were the association artefacts that arose during data collection or analysis? I therefore explored a number of explanations for the association between the two molecular markers and MDD.

4.3.2.1 Artefacts introduced by data collection process

I first examined the effects of time of day on all samples. The CONVERGE data collection process documented the time of day interviews were started and finished, and DNA samples were always collected at the end of interviews. Using the end times of interviews I could ask if there was any systematic difference between the time of DNA collection between cases of MDD (possibly during specific clinic hours) and controls (possibly a wider range of times), and if any such difference found could be related to the difference in molecular markers quantified from both groups.

Figures 4.6 shows this information for amount of mtDNA (Figure 4.6a) and mean telomere length (Figure 4.6b). Firstly, there was no significant correlation between time of day when DNA was collected and the amount of mtDNA ($P = 0.06$ by linear regression) or telomere length ($P = 0.13$ by linear regression). Secondly, there was no significant difference in the time of day when DNA was collected between cases of MDD and controls (T-test $P = 0.11$).

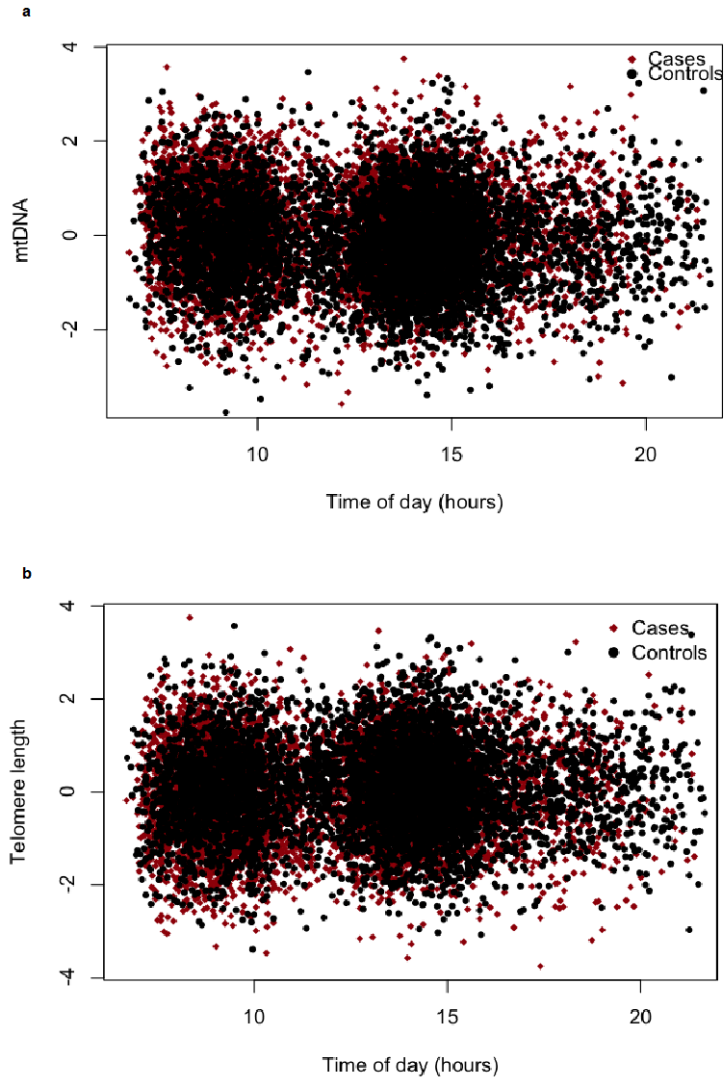


Figure 4.6 | Relationship between time of day of DNA collection for all CONVERGE samples and molecular markers quantified from low coverage sequencing. (a) This plot shows on the horizontal axis the times of day when samples were collected and on the vertical axes the amount of mtDNA quantified from low coverage sequencing in all CONVERGE samples, with cases of MDD coloured in red and controls coloured in black. No significant correlation was found between time of day of DNA collection and the amount of mtDNA (P = 0.06 by linear regression), and no significant difference in time of day of DNA collection was found between cases of MDD and controls (P = 0.11). (b) This plot shows the same information but for mean telomere length instead of amount of mtDNA,; no significant correlation was found between time of day of DNA collection and mean telomere length (P = 0.13 by linear regression).

4.3.2.2 Geographical origin of samples

I examined the effect of geography and DNA sample collection differences between all samples in CONVERGE. Samples in CONVERGE were gathered from 60 hospitals in China across a wide geographical range, and 150 specifically trained interviewers were involved in the saliva DNA collection process. Results for examining the effect of all covariates (by linear regression and analysis of variance (ANOVA)) on the mtDNA and telomere length measures were summarized in the Table 4.1; none of the covariates could account for the differences between cases of MDD and controls in their measures of amount of mtDNA and mean telomere length. None of the covariates were significantly associated with either molecular measure. The inclusion of each covariate in models to explain the association between MDD disease status and both molecular markers showed no significant association between any of the covariates with MDD disease status ($P > 0.1$ in each case) and made no difference between MDD disease status and either molecular markers.

Covariate	F-stat	Degrees of freedom	P-value
Interviewer	1.14	149	0.12
Hospital	1.11	56	0.28
Province	1.51	21	0.063
Time of sample collection	3.584	1	0.058
Date of DNA extraction	0.483	1	0.49

Covariate	F-stat	Degrees of freedom	P-value
Interviewer	1.07	149	0.27
Hospital	0.97	56	0.54
Province	1.38	21	0.12
Time of sample collection	2.24	1	0.13
Date of DNA extraction	0.87	1	0.35

Table 4.1 | Relationship between aspects of DNA collection and molecular markers quantified from low coverage sequencing. This table shows the association by linear regression between five aspects of data collection with amount of mtDNA quantified (upper panel) and mean telomere length (lower panel) from low coverage sequencing of the DNA extracted from the saliva samples. No significant association was found between any of the five covariates and amount of mtDNA or mean telomere length.

4.3.2.3 DNA contamination

Next, I considered artefacts arising due to systematic differences between cases of MDD and controls in the extent of mismapping of sequencing reads to the human mitochondrial DNA reference, which could be caused by heavier contamination of saliva DNA by microbial DNA similar in sequence to mtDNA in either cases of MDD or controls. Before quantification of the amount of mtDNA from sequencing data of all CONVERGE samples I re-mapped all of the sequencing reads that had mapped to the human mitochondrial DNA reference NC_012920.1 to a combined bacterial genome reference

containing 894 complete bacterial genomes, 2024 complete bacterial chromosomes, 154 draft assemblies and 4373 complete plasmids sequences (in total 7390 unique bacterial DNA sequences, see Methods). All reads that were mapped to bacterial DNA sequences were there filtered out using Samtools (v0.1.18) [185] by imposing a mapping quality filter of 59 (Phred-scale probability of being wrongly mapped) and removing reads with FLAGs (-F 1804) that identify unmapped reads, unpaired reads, reads that do not pass quality control, reads that may be PCR or optical duplicates, and reads that are secondary alignments that also map to other areas of the reference. Re-mapping the filtered reads to the combined bacterial reference showed that no filtered read matched bacterial genomes.

To prevent chance mismappings that were missed by the filtering, as well as to avoid including mismappings from nuclear copies of mitochondrial DNA (NUMTs) that also escaped filtering, I selected only non-overlapping 100bp segments in the mtDNA where less than 5% of originally reads were filtered out with the above criteria to be used for quantification of amount of mtDNA. These segments were more likely unique sequences in the human mtDNA. As all of the above results were based on amount of mtDNA quantified in this manner, it was unlikely the measure was affected by contamination, or to an extent that could account for the differences between cases of MDD and controls.

4.3.2.4 Effect of medication in cases of MDD

I considered whether the molecular changes might be due to medication. If elevated amount of mtDNA and decrease mean telomere length in saliva DNA was due to cases of MDD taking antidepressants, I should find a difference between either measures in cases of MDD who did not take antidepressants and those who did, similar to that I did between controls and cases of MDD. In CONVERGE there were 975 cases of MDD who

reported never having taken any anti-depressants. Neither amount of mtDNA in these subjects nor mean telomere length in these cases differed significantly from that assayed in the 4,861 cases of MDD reporting taking anti-depressants (t test P = 0.96 and P = 0.88 respectively). Medication, therefore, cannot explain either molecular changes.

4.3.3 Potential biological explanations

4.3.3.1 Cellular composition of saliva samples

It was possible that both molecular differences between cases of MDD and controls were due to differences in cell type. From literature, most of the cells in saliva are leukocytes, dominated by neutrophils (as in the blood) [186,187]. More recent analyses [188] used cell sorting to show that the white cells in saliva were mostly neutrophils and T-cells in an approximately 2:1 ratio, with smaller numbers of B-cells. For cellular composition to matter in this instance, mtDNA and mean telomere length needed to be different between the different cell types.

I examined the amount of mtDNA quantified from qPCR of flow sorted white cells from peripheral blood mononuclear cells (PBMCs) in culture. For this analysis I had the help of Dr Benjamin Fairfax who kindly performed qPCR quantification of amount of mtDNA was obtained for T-cells (CD3+) B cells (B220+), monocytes, neutrophils and natural killer (NK) cells. The results are shown in Figure 4.7. The amount of mtDNA in T (CD3) and B-cells was almost two orders of magnitude higher than in NK cells. This result means that relatively small changes in the proportion of T and B-cells could therefore have an impact on the total amount of mtDNA in a blood or saliva sample. Therefore it was particularly important to assess the proportion of T and B-cells in the CONVERGE sample.

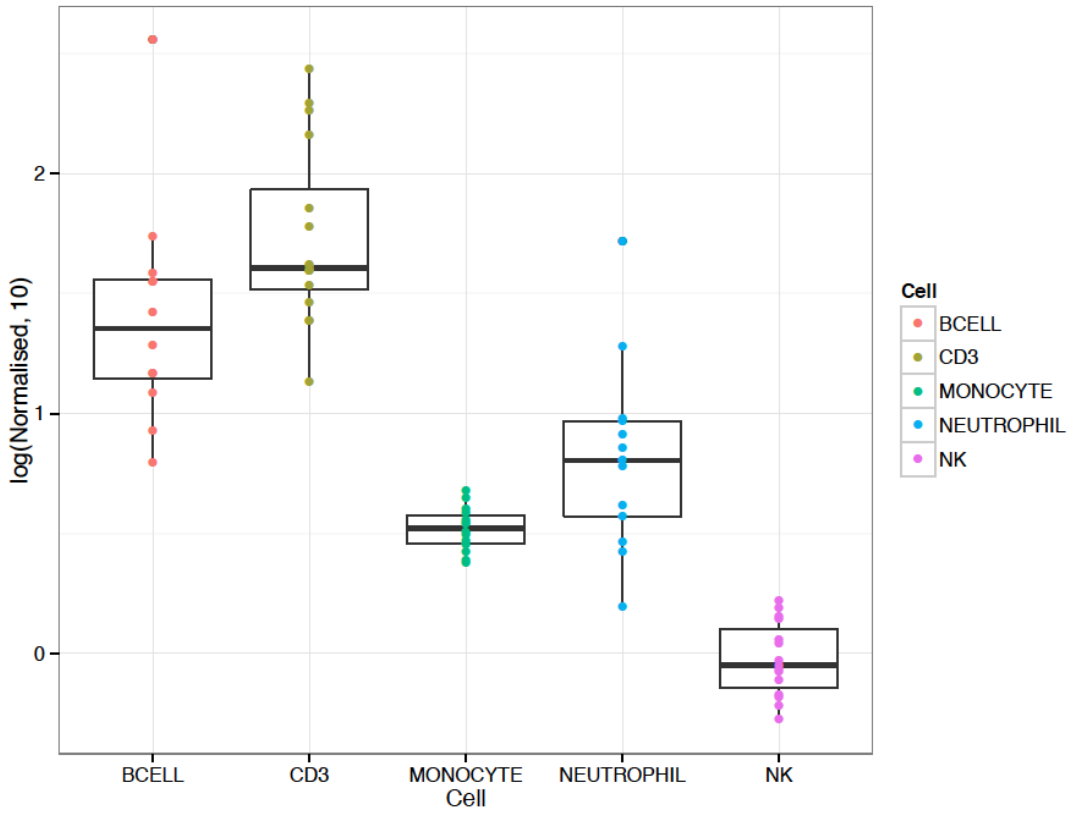


Figure 4.7 | Normalised relative amount of mtDNA to single copy nuclear sequence quantified by qPCR from five flow-sorted white blood cell types, in log scale. The figure shows a boxplot of normalized relative copies of mtDNA to a single-copy nuclear DNA on a log scale in five different white blood cell types separated from cultured PBMCs using flow-cytometry (performed by Dr Benjamin Fairfax); each box represents x replicates of the qPCR.

4.3.3.2 Assessing cellular composition of saliva samples

To assay cellular composition in CONVERGE samples, for which no intact cells remained as DNA had already been extracted from saliva samples, I used two approaches. First, I quantified events of somatic recombinants in the VDJ region of the T-cell receptor and immunoglobulin loci in the genome using insert sizes between paired-end reads, hence assaying the amount of T- and B-cells in the saliva DNA of all CONVERGE samples. Second, I obtained data on methylation state at 20 cytosine-guanine (CpG) dinucleotides sites, each uniquely identifying a particular blood cell type or buccal epithelial cells on

156 samples, 78 of whom were cases of MDD with the highest amount of mtDNA quantified from low coverage sequencing and 78 were controls with the lowest.

4.3.3.3 Detecting somatic recombination using NGS data

I was able to assess the amount of two cell components, T and B cells, by examining the amount of somatic DNA recombination in the DNA. DNA derived from mature T-cells had undergone re-arrangements of the T-cell receptor alpha and beta chains (TRA and TRB respectively); DNA derived from mature B-cells had undergone rearrangement at immunoglobulin heavy chain (IGH) locus. Thus the percentage of reads showing signals of recombination in the IGH locus (chr14:106-107MB) can be used as a proxy for the percentage B cells, and the percentage of reads showing rearrangements at the TRA (chr14:22-23MB) and TRB loci (chr7:141-142MB) can be used as a proxy for the percentage of T-cells.

I examined insert sizes between all paired-end reads in the IGH locus. Most fell in a normal distribution centered around 500 base pairs, as expected from sequencing non-recombined, contiguous DNA segments. A small percentage of read pairs however had insert sizes bigger than 1000 base pairs that map to distinct sections within the IGH locus, indicative of recombination. To show these insert size aberrations were likely indicative of real recombinations instead of random mapping errors, I performed the same analysis in a randomly chosen, similarly sized region on chromosome 2 where recombination was not expected. A comparison between the IGH locus and the region on chromosome 2 is shown in Figure 4.8; as there were more abundant and non-randomly distributed insert sizes in the IGH region, as compared to the much more rare occurrence of abberations in insert sizes from the mean in chromosome 2, I was confident I was detecting somatic recombinations in IGH, and therefore categorized those paired-end reads with insert sizes greater than 1000 base pairs as spanning a recombination.

I counted the number of read pairs whose insert size was indicative of a somatic recombination event (as defined above) and for each sample calculated a ratio of recombined versus non-recombined reads at the IGH (B-cell), TRA or TRB loci (T-cell). I removed those samples whose counts were outliers more than 5 standard deviations away from the mean, and controlled for sequencing batch, whole-genome coverage and age of samples to arrive at the final measure of degree of recombination per sample. I observed no significant differences in the amount of somatic recombination at the IGH (t-test P-value = 0.108), TRA (t-test P-value = 0.175) or TRB (t-test P-value = 0.275) loci between cases and controls (Figure 4.9).

Logisitic regressions of the amount of somatic recombination at all the above loci showed no significant association with MDD and cannot explain the association between amount of mtDNA or mean telomere length and MDD (Table 4.2). For the following logistic regressions, in R notation:

lm0 = lm(MDD ~ 1)

lm1 = lm(MDD ~ molecular marker)

lm2 = lm(MDD ~ somatic recombination)

lm3 = lm(MDD ~ somatic recombination+ molecular marker)

lm4 = lm(MDD ~ molecular marker + somatic recombination)

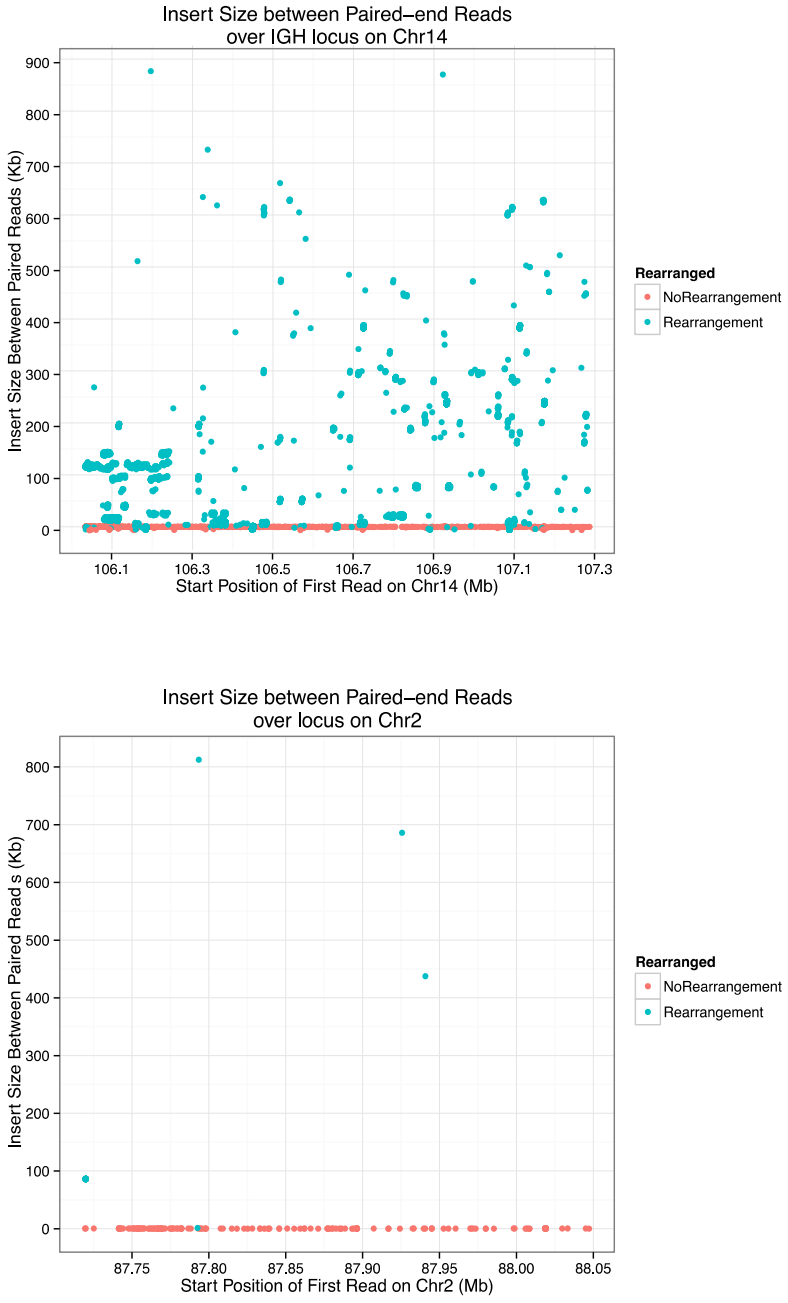


Figure 4.8 | Somatic recombination detected by long insert sizes between paired-end reads in the IGH region. This figure shows the insert sizes in Kb between paired end reads from low-coverage sequencing of all CONVERGE samples over IGH region (chr14:106-107MB, upper panel) expected to display patterns of somatic recombinations in proportion to proportion of B cell contribution to saliva DNA samples, and a randomly selected, similarly sized region (chr2:87-88MB, lower panel) where no such recombination was expected. Abundant high insert sizes were shown in the IGH region, and non-random distributions of the insert were indicative of bona fide recombinations instead of random mapping errors, which would show up as much less abundant and randomly distributed insert sizes as shown in chr2 region.

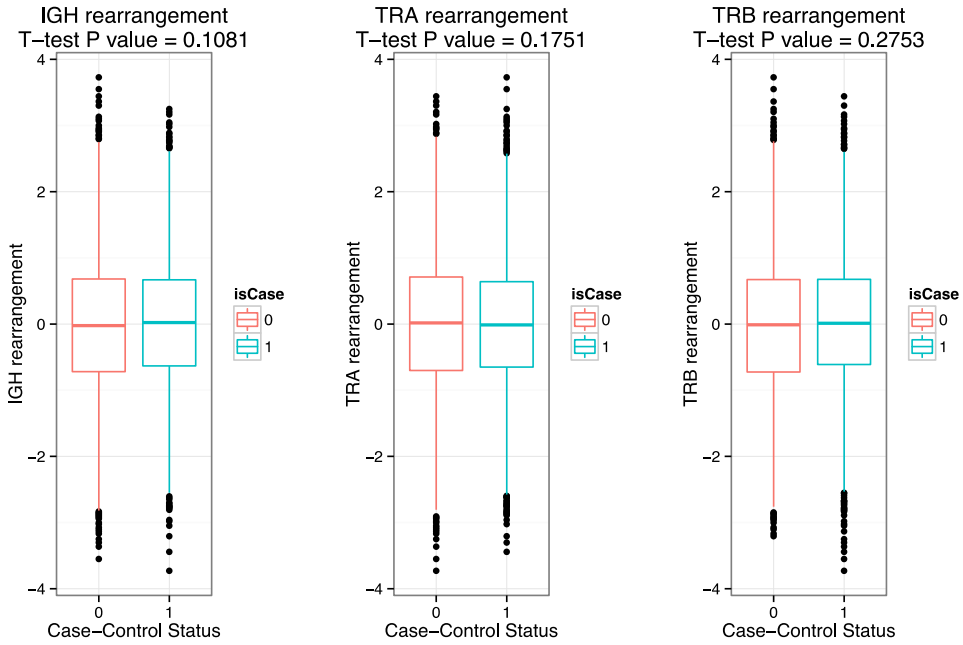


Figure 4.9 | Boxplots and t-test P values of quantile-normalized measures of IGH, TRA and TRB rearrangements in cases of MDD and controls These figures show boxplots of quantile normalized ratios between number of paired end reads with insert sizes over 1000bp, indicating recombination, and those with insert sizes under, in the IGH (chr14:106-107MB), TRA (chr14:22-23MB) and TRB (chr7:141-142MB) regions. Cases of MDD do not differ significantly in amount of recombination at any of the three loci (t-test P-values = 0.108, 0.175 and 0.275 respectively) and hence there was no evidence to conclude that cases of MDD had higher levels of T or B cells in their saliva.

Effect on MDD	P-values for association with MDD from ANOVA		
	IGH	TRA	TRB
mtDNA anova(lm0, lm1)	$8.38x10^{-43}$	$8.38x10^{-43}$	$8.38x10^{-43}$
Somatic recombination anova(lm0, lm2)	0.108	0.175	0.275
Somatic recombination conditioned on mtDNA anova(lm1, lm4)	0.0882	0.252	0.217
mtDNA conditioned on somatic recombination anova(lm2,lm3)	$7.12x10^{-43}$	$1.09x10^{-42}$	$7.09x10^{-43}$
Effect on MD by covariate	**P Values for association with MD from ANOVA**		
	IGH	**TRA**	**IGH**
Telomere length anova(lm0, lm1)	$2.52x10^{-12}$	$2.52x10^{-12}$	$2.52x10^{-12}$
Somatic recombination anova(lm0, lm2)	0.108	0.175	0.275
Somatic recombination conditioned on telomere length anova(lm1, lm4)	0.0453	0.220	0.122
Telomere length conditioned on somatic recombination anova(lm2,lm3)	$1.22x10^{-12}$	$2.99x10^{-12}$	$1.37x10^{-12}$

Table 4.2 | Table showing P-values of association between somatic recombination and molecular markers on MDD respectively, and conditioned on each other This table shows the P-values of association between somatic recombinations at the IGH, TRA and TRB loci, as well as molecular markers of interest amount of mtDNA (upper panel) and mean telomere length (lower panel), and MDD by analysis of variance (ANOVA) comparing logistic regression models differing by the test term. For each row the test term was highlighted in bold and the logistic regression models compared using ANOVA was shown in the line below. P-values of association for molecular markers amount of mtDNA and mean telomere length without controlling for somatic recombination at any of the three loci were shown in the first line for both tables, as were the same values in all three columns as recombination at none of the three loci were involved in the logistic regression models tested. Both tables show molecular markers were associated with MDD whether conditioned or not on somatic recombination at the three loci (each indicative of proportions of B or T cells in saliva samples), and somatic recombination at none of the three loci were associated with MDD whether on their own or conditioned on molecular markers. This shows the proportion of B and T cells in saliva samples of MDD cases and controls were not significantly different.

4.3.3.4 Assessing cell-type specific methylation states

The second method I used to assess cellular composition in the CONVERGE sample was to assess the degree of methylation of cytosine residues at CpG dinucleotides, which is well known to differ between cell types [189-191]. Methylation of DNA thus contains information about the cellular composition of the tissue from which it was extracted [192], and cell types can be quantified from the methylation profiles [193,194].

I used methylation marks that have been shown to identify leucocytes [193] and epithelial cells in saliva [195] as surrogates for the cellular composition of the saliva samples in the CONVERGE cohort. It has been shown that about 20 methylated CpG can accurately predict the identity of leukocyte subsets: using this approach, estimates of the proportions of cell types are close to 100% accurate [193]. I designed primers for targeted sequencing at 20 CpG dinucleotides, 13 of which were methylated in specific blood cell types that could best distinguish between B lymphocytes, T lymphocytes, eosinophils, basophils, NK cells, monocytes, and granulocytes and neutrophils[193], and 7 CpG sites most significantly differently methylated in buccal epithelial cells and blood cells [192,195], as shown in Table 4.3, on 156 samples after bisulfite conversion of their saliva DNA. It had been shown that about these 20 methylated CpG can accurately predict the identity of leukocyte subsets: using this approach, estimates of the proportions of cell types are close to 100% accurate(Accomando et al., 2014)

The 156 samples were chosen to capture the extremes of the distribution of amount of mtDNA quantified from sequencing data (78 samples were chosen with the from the high end of the distribution and were all cases of MDD, and 78 were chosen from the low end of the distribution and were all controls). Libraries were constructed using the WaferGen SmartChip MyDesign Target Enrichment system, targeted sequencing was performed using Illumina Miseq. The sequencing reads were cleaned for adaptor sequences both at ends and in the middle of the reads before I mapped them to

the hg18 reference genome (to be consistent with the CpG dinucleotide coordinates taken from hg18) that has been processed to the bisulphite converted sequence using Bismark [196] (v0.13.0). Overlapping regions of paired-end reads were excluded to prevent double counting of methylated or unmethylated CpG sites. I then quantified the percentage methylation at all 20 CpG sites using Bismark's methylation extraction utility with the –bedGraph option.

CpG site	Gene	Unmethylated cell type	Methylated cell type
cg02237342	ADORA2A	Blood	Buccal Epithelial
cg20660269	ADORA2A	Blood	Buccal Epithelial
cg06310816	GLI3	Blood	Buccal Epithelial
cg17390350	GLI3	Blood	Buccal Epithelial
cg21472506	OTX1	Blood	Buccal Epithelial
cg14088196	PPFIA1	Blood	Buccal Epithelial
cg18384097	PTPN7	Blood	Buccal Epithelial
cg00226923	FGD2	B cells	All others blood cells
cg00491404	EPS8L3	NK cells	All others blood cells
cg00899659	ZNF22	Eosinophils	All others blood cells
cg02329886	HDC	Basophils	All others blood cells
cg02600394	TXK	T cells and NK cells	All others blood cells
cg02712878	FAIM	All others blood cells	T cells
cg03693099	CEL	Eosinophils	All others blood cells
cg03860768	BLK	B cells	All others blood cells
cg04576021	HLA-DO	B cells	All others blood cells
cg07873128	OSBPL5	CD8+/CD16+ NK cells	All others blood cells
cg08399444	GSG1	Granulocytes + Neutrophils	All others blood cells
cg11206634	SFT2D3	Monocytes	All others blood cells
cg12952132	NCR1	CD16+ NK cells	All others blood cells

Table 4.3 | 20 differentially methulated CpG sites distinguishing between epithelial and different blood cell types This table shows 20 CpG dinucleotides previously shown to be differentially methylated in different blood cell types and buccal epithelial cells. The first column shows the ID of the CpG site, the second column shows the gene it resides in, the third and fourth columns show the unmethylated and methylated cell types at the 20 sites respectively. Methylation was shown to be full or nearly full for all methylated cell types and null or unmethylated cell types for all 20 CpG sites.

Methylation states were readable at 19 sites for more than 90% of the 156 samples tested. One site (cg14088196), an assay for epithelial cells, gave useable data for only 4% of individuals, and was hence excluded from further analysis. Since six other epithelial states interrogated epithelial cells, the loss of this site does not invalidate our ability to quantitate epithelial cell composition.

To verify the assay can distinguish between different cell types, I looked for pairs of CpG sites whose methylation state should reflect the state of methylation of just one cell type, instead of a mixture of cell types. I first looked at the two pairs of redundant CpG sites in the same genes that are methylated in the buccal epithelial cells and unmethylated in blood cells. There were significant positive Pearson correlations between the percentage methylation in each pair using raw data from all samples. I then looked at a pair of CpG sites that distinguishes T cells from other blood cell types through the possession of one methylated and one unmethylated state, and found a significant negative correlation between the different tissues. Results of these association tests are shown in Table 4.4. These results indicate the methylation assay can distinguish between cell types and that we can use this approach to investigate differences in cellular composition in saliva samples taken from cases and controls.

Figure 3.8 shows the percentage methylation at each of the 20 CpG sites (including the excluded site) in cases and controls. There was no significant difference between level of methylation at CpG sites specific for any of the blood cell types, though there was a significant difference between cases of MDD and controls in terms of methylation at CpG sites identifying the buccal epithelial cell types. Cases of MDD had higher levels of methylation at the CpG sites specific for buccal epithelial cell types, indicating therefore saliva samples of cases of MDD contained a higher ratio of buccal epithelial cells than controls.

CpG site	Gene	Methylated cell type	Correlation coefficient	P value of correlation
cg02237342 cg20660269	ADORA2A	Buccal Epithelial	0.713	$< 2x10^{-16}$
cg06310816 cg17390350	GLI3	Buccal Epithelial	0.938	$< 2x10^{-16}$
cg02600394	TXK	All except T and NK cells	-0.242	0.00236
cg02712878	FAIM	T cells		

Table 4.4 | Correlation between raw percentage methylation from sequencing redundant CpG sites conclusively distinguishing the same cell types. This table shows three pairs of CpG dinucleotides each able to identify one particular cell type., the gene they reside in, the cell type in which they are methylated, the Pearson correlation coefficient r between their degree of methylation, and P-value of the correlation. As methylation state at both CpG dinucleotides in each pair should be able to indicate the proportion of the cell type they are both able to identify, their methylation states should be either significantly positively correlated if they were both methylated or unmethylated in the same cell, as in the case of the first two pairs identifying buccal epithelial cells (cg02237342 and cg20660269, cg06310816 and cg17390350), or significantly negatively correlated if one of them identifies the cell type by methylation and the other by lack of methylation, as in the case of the final pair identifying T cells (cg02600394 and cg02712878).

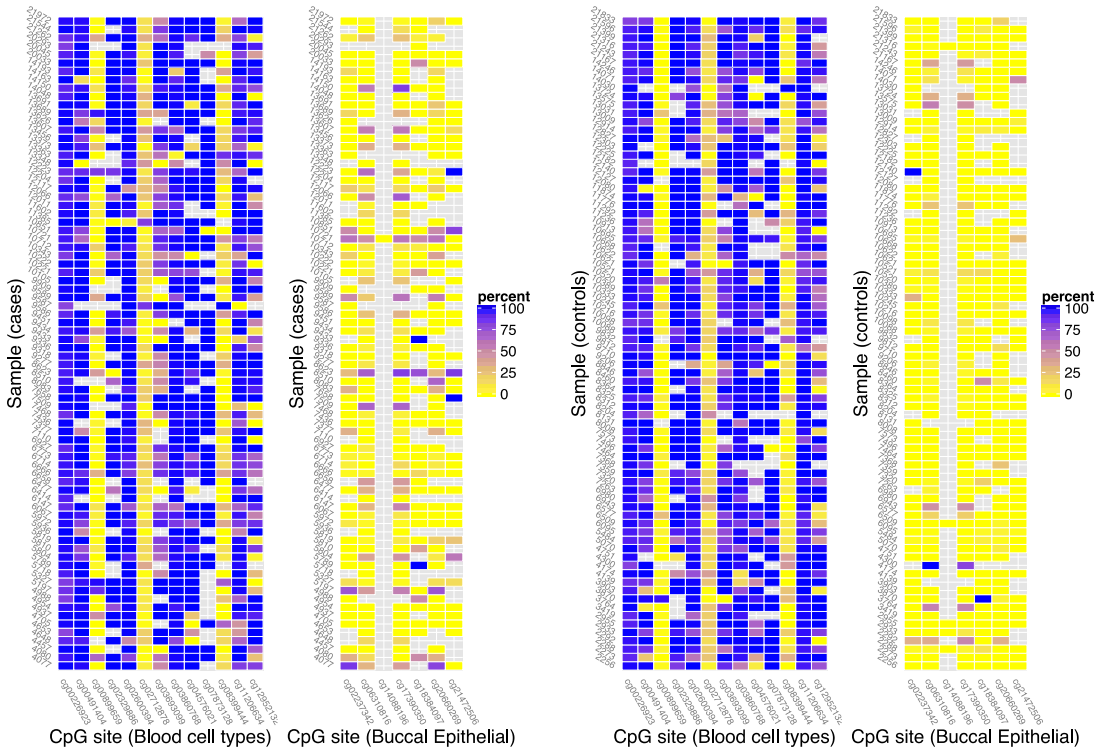


Figure 4.10 | Percentage methylation at CpG sites distinguishing between blood cell types and between buccal epithelial and blood cells in 156 samples. Color indicates the percentage methylation where 100% methylated is blue and 0% methylated is yellow. The horizontal axis lists the sites and the vertical axis the identifiers of the samples. Cases are shown on the left and controls on the right.

4.3.3.5 Cellular composition effect on association signals

As I looked for association between a) MDD, b) amount of mtDNA and c) telomere length and a quantile normalized measure of percentage of methylation at the 19 methylation sites with high quality data, I found a significant difference between cases of MDD and controls in the level of methylation at 3 out of the 7 CpG sites that distinguish between buccal epithelial (methylated) and blood cell types (unmethylated), and at 4 out of the 13 CpG sites that distinguish between different blood cell types.

Methylation differences could not fully account for the mtDNA and telomere length differences between cases of MDD and controls (Table 4.6 for amount of mtDNA, and Table 4.7 for mean telomere length). MDD case-control status remained highly significantly associated with the amount of mtDNA (t test P-value = $5.14X10^{-18}$) and telomere length (t test P-value = $6.83X10^{-5}$) after accounting for the degree of methylation at all of the sites, as shown in Figure 4.11. Expressed as a change in effect size using Nagelkerke's R^2 measure, there is a 6% reduction in the R^2 in a model including methylation and nMT to predict MDD, and a 9% reduction for telomere length. This analysis showed that while the cellular composition of saliva collected from cases of MDD differed slightly from that of controls, it could explain less than 10% of the differences amount of mtDNA and mean telomere length between cases of MDD and controls.

CpG Site	Number of Samples	P values of association			Cell Type
		With CpG methylation		MDD with mtDNA conditioned on CpG methylation	
		MDD	mtDNA		
cg02237342	134	0.002*	0.022	$3.73x10^{-40}$	Buccal
cg06310816	147	0.019*	0.007	$5.21x10^{-45}$	Buccal
cg17390350	147	0.012*	0.001	$7.75x10^{-45}$	Buccal
cg18384097	107	0.612	0.023	$1.01x10^{-33}$	Buccal
cg20660269	134	0.070	0.875	$1.60x10^{-41}$	Buccal
cg21472506	91	0.468	0.098	$4.87x10^{-29}$	Buccal
cg14088196	6	Failed sequencing			Buccal
cg03693099	141	0.005*	0.022	$1.09x10^{-42}$	Blood
cg04576021	130	0.009*	0.025	$1.38x10^{-39}$	Blood
cg11206634	156	0.019*	0.029	$9.21x10^{-48}$	Blood
cg02600394	155	0.035*	0.039	$1.11x10^{-47}$	Blood
cg00491404	148	0.051	0.135	$1.05x10^{-45}$	Blood
cg00899659	144	0.135	0.031	$7.83x10^{-45}$	Blood
cg02329886	121	0.147	0.347	$7.35x10^{-38}$	Blood
cg03860768	154	0.255	0.289	$4.56x10^{-48}$	Blood
cg00226923	156	0.364	0.244	$8.94x10^{-49}$	Blood
cg12952132	147	0.380	0.530	$4.61x10^{-46}$	Blood
cg08399444	151	0.412	0.385	$2.70x10^{-47}$	Blood
cg02712878	156	0.558	0.423	$7.03x10^{-49}$	Blood
cg07873128	114	0.679	0.685	$4.39x10^{-36}$	Blood

Table 4.6 | Association between percentage methylation at CpG sites with MDD and amount of mtDNA, and the association between MDD and amount of mtDNA conditioned on percentage methylation at CpG sties. This table shows the ID of 20 CpG sites for which I had obtained percentage methylation from targeted sequencing of DNA of 156 samples after bisulphite conversion. The first column shows the ID of CpG site, the second column shows the number of samples in which there was sufficient data for determining percentage of methylation, the third and fourth columns show the P-values of association by logistic regression and linear regression respectively between percentage methylation at each CpG site with MDD disease status of the 156 samples, and their amount of mtDNA quantified from low-coverage sequencing. The fifth column shows the association by logistic regression between MDD and amount of mtDNA in the 156 samples after controlling for methylation levels at each of the CpG sites, indicative of abundance the particular cell type they each identify, as shown in column six.

CpG Site	Number of Samples	P values of association			Cell Type
		With CpG methylation		MD with telomere length conditioned on CpG methylation	
		MD	Telomere		
cg02237342	134	0.002*	0.720	$1.69x10^{-4}$	Buccal
cg06310816	147	0.019*	0.808	$5.08x10^{-5}$	Buccal
cg17390350	147	0.012*	0.539	$7.41x10^{-5}$	Buccal
cg18384097	134	0.612	0.855	$1.85x10^{-4}$	Buccal
cg20660269	91	0.070	0.550	0.033	Buccal
cg21472506	107	0.468	0.203	0.011	Buccal
cg14088196	6	Failed sequencing			Buccal
cg03693099	141	0.005*	0.267	$8.64x10^{-5}$	Blood
cg04576021	130	0.009*	0.197	$9.56x10^{-5}$	Blood
cg11206634	156	0.019*	0.166	$3.08x10^{-5}$	Blood
cg02600394	155	0.035*	0.858	$1.45x10^{-5}$	Blood
cg00491404	148	0.051	0.048	$4.06x10^{-4}$	Blood
cg00899659	144	0.135	0.078	$1.31x10^{-4}$	Blood
cg02329886	121	0.147	0.726	$1.06x10^{-3}$	Blood
cg03860768	154	0.255	0.073	$3.55x10^{-5}$	Blood
cg00226923	156	0.364	0.510	$9.55x10^{-6}$	Blood
cg12952132	147	0.380	0.669	$2.31x10^{-5}$	Blood
cg08399444	151	0.412	0.381	$1.93x10^{-5}$	Blood
cg02712878	156	0.558	0.556	$1.44x10^{-5}$	Blood
cg07873128	114	0.679	0.414	$8.59x10^{-4}$	Blood

Table 4.7 | Association between percentage methylation at CpG sites with MDD and mean telomere length, and the association between MDD and mean telomere length conditioned on percentage methylation at CpG sties. This table shows the ID of 20 CpG sites for which I had obtained percentage methylation from targeted sequencing of DNA of 156 samples after bisulphite conversion. The first column shows the ID of CpG site, the second column shows the number of samples in which there was sufficient data for determining percentage of methylation, the third and fourth columns show the P-values of association by logistic regression and linear regression respectively between percentage methylation at each CpG site with MDD disease status of the 156 samples, and their mean telomere length quantified from low-coverage sequencing. The fifth column shows the association by logistic regression between MDD and mean telomere length in the 156 samples after controlling for methylation levels at each of the CpG sites, indicative of abundance the particular cell type they each identify, as shown in column six.

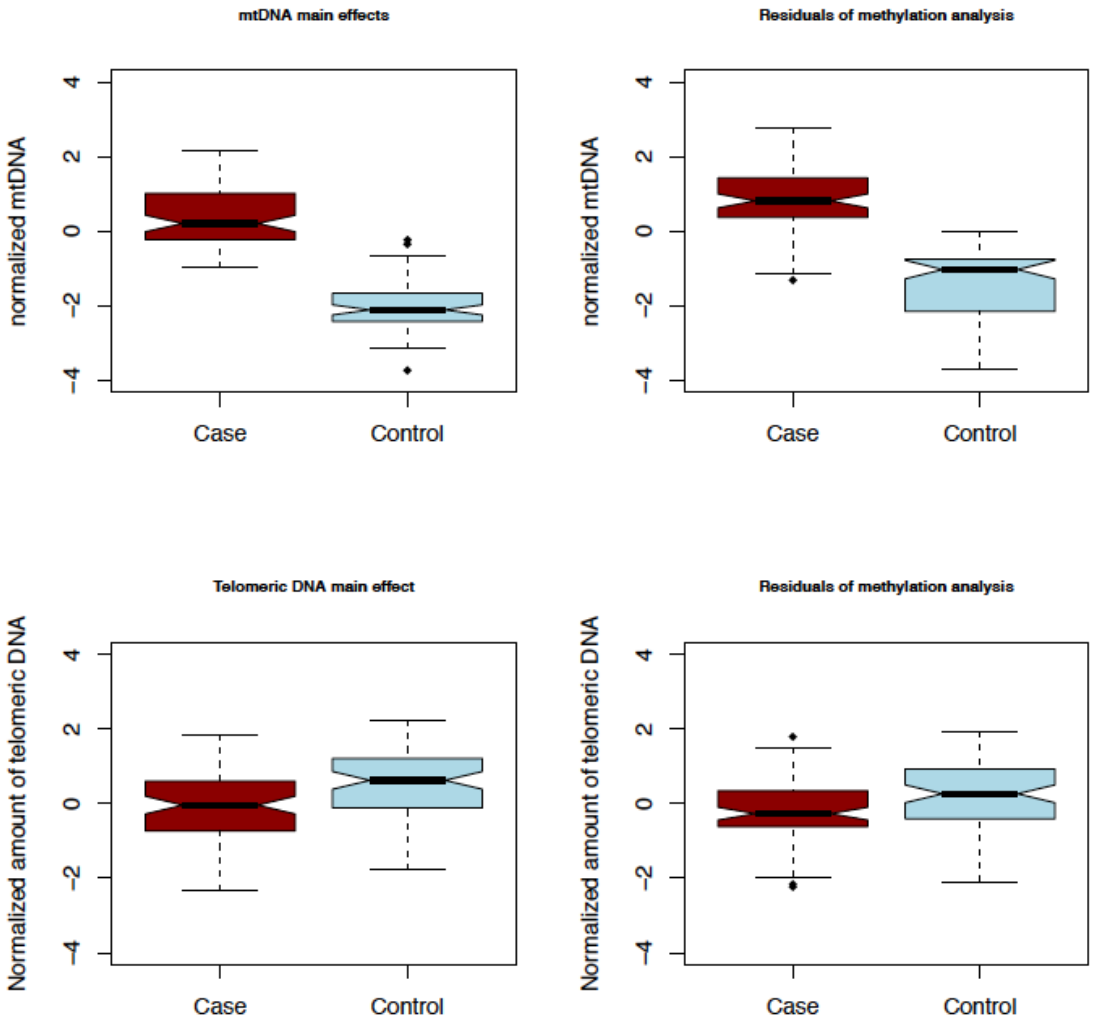


Figure 4.11 | Effect of methylation on the relationship between mtDNA and telomeric DNA and case status. This figure shows boxplots for amount of mtDNA and mean telomeric length measures quantified from low-coverage sequencing in cases of MDD (red) and controls (blue) on the left ("Main effects") and after controlling for percentage methylation at all 19 CpG dinucleotides in the 156 samples ("Residuals of methylation analysis"). As both amount of mtDNA (t test P-value = $5.14X10^{-18}$) and mean telomere length (t-test P-value = $6.83X10^{-5}$) remain significantly different in cases of MDD and controls after controlling for methylation states at 19 CpG dinucleotides indicative of percentage of different blood and buccal epithelial cell types in saliva samples, cellular composition cannot account completely for the differences in either molecular markers in cases of MDD as compared to controls.

4.4 Molecular changes are associated with adversity

4.4.1 Molecular changes are associated with adversity

Having shown the differences in both molecular markers between cases of MDD and controls were unlikely to be due to technical artefacts or biological though irrelevant phenomenon, I started exploring the association in the CONVERGE data between known predisposing factors to MDD and both mean telomere length and mtDNA levels. Mean telomere length was significantly smaller in those who had experienced more stressful life events (P = 0.0018, by linear regression with number of stressful life events, including as covariates sequencing coverage, sequencing batch, body mass index and age), and in those reporting childhood sexual abuse (P = 0.043, by linear regression with childhood sexual abuse including the aforementioned covariates) (Table 4.8).

Amount of mtDNA was significantly correlated with both the total number of stressful life events (linear regression $P = 4.83X10^{-4}$) and with childhood sexual abuse (linear regression $P = 3.65X10^{-5}$). As with the analysis of mean telomere length, the association with childhood sexual abuse was stronger with increasingly severe abuse, with those reporting the most severe form (attempted or completed intercourse) showing up to 1.35 times higher amount of mtDNA compared to those who reported no abuse.

Type	ΔnTel[a]	t- value[b]	P-value[c]	ΔnMT[a]	t- value[b]	P-value[c]	Ncases[d]	Ncontrols[e]	Total[f]
Non-genital									
CSA	0.02	0.35	0.73	0.08	1.37	0.169	186	81	267
Genital									
CSA	-0.08	-1.27	0.20	0.11	2.02	0.045	240	47	287
Intercourse									
CSA	-0.20	-2.45	0.01	0.38	4.67	$3.05x10^{-6}$	159	17	176

Table 4.8 | Relationship between childhood sexual abuse, normalized measure of mean telomere length (nTel) and the normalized measure of mitochondrial DNA levels (nMT). This table shows the results of analysis of variance in which different forms of childhood sexual abuse (CSA) predict normalized measures of mean telomere length (nTel) and the normalized measyre of mitochondrial DNA levels (nMT) (after removing covariates). Non-genital CSA refers to sexual invitation, sexual kissing, and exposing; genital CSA refers to fondling and sexual touching and Intercourse CSA refers to attempted or completed intercourse. [a,b,c] Estimated change in nTel (ΔnTel) or nMT (ΔnMT) over mean nTel or nMT in individuals with no CSA, t-statistic and P-value of tests of hypotheses that corresponding underlying excess is zero. [d,e,f] Number of MDD cases, controls and total individuals in given CSA category. MDD status was not taken into account in the analysis.

4.4.2 Molecular changes are contingent on the depressed state

To investigate a causal relationship between stressful life events, MDD, amount of mtDNA, and mean telomere length, I performed a series of conditional regression analyses, assuming that stressful life events preceded the onset of MDD and the molecular changes.

If stressful life events have independent causal effects on MDD, mtDNA levels and mean telomere length, then the latter two molecular phenotypes should become independent of MDD after conditioning on the number of stressful events. Table 3.8 shows this is not the case because the differences in amount of mtDNA and mean telomere length between cases and controls, when stratified for the number of stressful life events, remained highly significant (t-tests in third column of Table 4.9; amount of

mtDNA P-values range from $1.25x10^{-18}$ to 0.37; mean telomere length P-values range from $4.23x10^{-5}$ to 0.0083). Table 4.10 shows the counts of individuals categorized by MDD disease status and number of stressful life events, with the means and standard errors for amount of mtDNA (Table 4.10a) and mean telomere length (Table 4.10b) within each category.

The next question was to ask whether the effect of stressful life events on MDD is entirely indirect, acting via changes in amount of mtDNA or mean telomere length. I rejected this explanation, because the association between MDD and stressful life events remains highly significant after conditioning on either amount of mtDNA (P-value = $5.60x10^{-99}$) or mean telomere length (P-value =$2.x10^{-100}$) in a logistic regression model. In contrast, the association between stress and amount of mtDNA or mean telomere length disappears when conditioned on MDD (P-value = 0.11 and P-value = 0.11, respectively). In other words, the predictive power of stress on amount of mtDNA and mean telomere length is mediated through a history of MDD. These results are summarized in Table 4.11a for amount of mtDNA and Table 4.11b for mean telomere length.

These conclusions also hold when the number of stressful life events was replaced by a history of childhood sexual abuse (CSA), the most robust predictor of later life MDD. In particular there was no significant difference in the amount of mtDNA or mean telomere length when comparing controls who reported a history of CSA with those who did not. Amount of mtDNA for controls who reported any form of CSA was -0.136 (SE = -0.007), and -0.095 (SE = -0.001) for no such history; t-test P-value = 0.66. Similarly, for mean telomere length for controls who reported any form of CSA was 0.168 (SE = 0.0125), and 0.072 (SE = 0.001) for no such history; t-test P-value = 0.27. These results are summarized in Table 3.10a for amount of mtDNA and Table 3.10b for mean telomere length.

(A) Difference in nMT in cases of MDD and controls per SLE			
SLE	**MDD Control**	**MDD Case**	**T-statistic, P-value**
0	-0.132 (0.019) **2487**	0.142 (0.024) **1689**	-8.86, 1.25×10^{-18}
1	-0.156 (0.026) **1432**	0.0987 (0.027) **1441**	-6.83, 1.01×10^{-11}
2	-0.103 (0.034) **757**	0.165 (0.033) **935**	-5.65, 1.84×10^{-08}
3	-0.068 (0.058) **334**	0.085 (0.044) **507**	-2.12, 0.03
4+	0.062 (0.067) **221**	0.132 (0.040) **666**	-0.89, 0.37

(B) Difference in nTel in cases of MDD and controls per SLE			
SLE	**MDD Control**	**MDD Case**	**T-statistic, P-value**
0	0.078 (0.020) **2542**	-0.053 (0.025) **1722**	4.10, 4.23×10^{-5}
1	0.098 (0.026) **1461**	-0.048 (0.027) **1470**	3.91, 9.42×10^{-5}
2	0.093(0.035) **780**	-0.069 (0.032) **952**	3.36, 8.05×10^{-4}
3	0.042 (0.053) **342**	-0.085(0.045) **517**	1.82, 0.069
4+	0.060 (0.060) **229**	-0.129 (0.038) **677**	2.65, 0.0083

Table 4.9 | Relationship between number of stressful life events (SLE), normalized measures of mtDNA levels (amount of mtDNA), normalized measures of mean telomere length (mean telomere length) and MDD. This table shows the counts of individuals categorized by MDD case status and the number of stressful life events (SLE) with the means and standard errors of the normalised measure of mtDNA levels (amount of mtDNA, A) and normalized mean telomere length (mean telomere length, B) within each SLE category.

(i) glm(MD ~ nMT*SLE)[a]	Df[b]	Deviance[c]	P[d]
NULL	10465		
nMT	1	164.50	$1.18x10^{-37}$
SLE	1	445.91	$5.60x10^{-99}$
nMT *SLE	1	4.96	0.026
(ii) glm(MD ~ SLE*nMT)	**Df**	**Deviance**	**P**
NULL	10465		
SLE	1	460.83	$3.16x10^{-102}$
nMT	1	149.58	$2.14x10^{-34}$
SLE* nMT	1	4.96	0.026
(iii) glm(MD ~ nMT*CSA)	**Df**	**Deviance**	**P**
NULL	11031		
nMT	1	174.91	$6.28x10^{-40}$
CSA	1	288.44	$1.09x10^{-64}$
nMT * CSA	1	0.03	0.86
(iv) glm(MD ~ CSA*nMT)	**Df**	**Deviance**	**P**
NULL	11031		
CSA	1	301.39	$1.64x10^{-67}$
nMT	1	161.95	$4.24x10^{-37}$
CSA * nMT	1	0.03	0.86

Table 4.10a | Effect of normalised measure of normalized amount of mtDNA (nMT), number of stressful life events (SLE) and their interaction on MD disease risk. Shown are the analysis of deviance tables produced by the anova function to compare the fits of nested generalised linear models in the R statistical package. [a]: specification of the model using the notation of R. [b]: The difference in degrees of freedom between the current model and that on the line above. [c]: the difference in deviance (twice the log-likelihood) between the current model and that on the line above. [d]: P-value of test comparing the fit of the current model to the line above. NULL: model with just top 3 principal components calculated from a GRM computed using 561,819 common tagging SNPs as covariates. All fitted models included these covariates.

(i) glm(MD ~ nTel*SLE)	Df	Deviance	P
NULL	10688		
nTel	1	62.23	$3.06x10^{-15}$
SLE	1	451.97	$2.x10^{-100}$
nTel *SLE	1	0.71	0.40
(ii) glm(MD ~ SLE* nTel)	**Df**	**Deviance**	**P**
NULL	10688		
SLE	1	459.67	$5.66x10^{-102}$
nTel	1	54.53	$1.53x10^{-13}$
SLE* nTel	1	0.71	0.40
(iii) glm(MD ~ nTel *CSA)	**Df**	**Deviance**	**P**
NULL	11276		
nTel	1	67.78	$1.83x10^{-16}$
CSA	1	303.47	$5.77x10^{-68}$
nTel * CSA	1	2.36	0.12
(iv) glm(MD ~ CSA* nTel)	**Df**	**Deviance**	**P**
NULL	11276		
CSA	1	307.26	$8.65x10^{-69}$
nTel	1	64.00	$1.25x10^{-15}$
CSA * nTel	1	2.36	0.12

Table 4.10b | Effect of normalised measures of mean telomere length (nTel), number of stressful life events (SLE) and their interaction on MD disease risk. See **Table 3.9a** for description of table columns and contents.

(i) lm(nMT ~ MD *SLE)[a]	Df[b]	Sum Sq[c]	Mean Sq[d]	F[e]	P[f]
MD	1	162.5	162.48	165.84	$1.16x10^{-37}$
SLE	1	2.6	2.57	2.63	0.11
MD *SLE	1	3.4	3.40	3.47	0.062
Residuals	10462	10249.9	0.98		
(ii) lm(nMT ~ SLE* MD)	**Df**	**Sum Sq**	**Mean Sq**	**F**	**P**
SLE	1	17.5	17.48	17.84	$2.42x10^{-5}$
MD	1	147.6	147.57	150.63	$2.18x10^{-34}$
SLE* MD	1	3.4	3.40	3.47	0.062
Residuals	10462	10249.9	0.98		
(iii) lm(nMT ~ MD *CSA)	**Df**	**Sum Sq**	**Mean Sq**	**F**	**P**
MD	1	172.8	172.84	176.32	$6.25x10^{-40}$
CSA	1	4.9	4.93	5.03	0.025
MD *CSA	1	0.4	0.35	0.36	0.55
Residuals	11028	10810.7	0.98		
(iv) lm(nMT ~ CSA* MD)	**Df**	**Sum Sq**	**Mean Sq**	**F**	**P**
CSA	1	18.4	18.41	18.78	$1.48x10^{-05}$
MD	1	159.4	159.36	162.56	$5.68x10^{-37}$
SLE* MD	1	0.4	0.35	0.37	0.55
Residuals	11031	10810.7	0.98		

Table 4.11a | Effect of MD, number of stressful life events (SLE) and their interaction on normalised measure of amount of mtDNA (nMT). Shown are Analysis of Variance (ANOVA) tables, generated by the anova() function applied to comparisons of linear models fitted by the R statistical package. [a] : Specification of the linear model fitted in the R language. [b]: Difference in degrees of freedom between current model and modle above (the null model, not shown, fitted just top 3 principal components calculated from a GRM computed using 561,819 common tagging SNPs as covariates). [c]: Additional Sum of squares (variance explained) by the current model compared the line above. [d] : Mean Square (Sum Sq divided by Df). [e]: Partial F statistic comparing the fit of the current model to that on the line above. [f]: P-value of Partial F-test.

(i) lm(nTel ~ MD *SLE)	Df	Sum Sq	Mean Sq	F	P
MD	1	60.9	60.88	62.38	$3.11x10^{-15}$
SLE	1	2.4	2.44	2.50	0.11
MD *SLE	1	0.3	0.29	0.30	0.59
Residuals	10685	10427.9	0.98		
(ii) lm(nTel ~ SLE* MD)	**Df**	**Sum Sq**	**Mean Sq**	**F**	**P**
SLE	1	9.7	9.67	9.91	0.0017
MD	1	53.7	53.65	54.97	$1.31x10^{-13}$
SLE* MD	1	0.3	0.29	0.30	0.59
Residuals	10685	10427.9	0.98		
(iii) lm(nTel ~ MD *CSA)	**Df**	**Sum Sq**	**Mean Sq**	**F**	**P**
MD	1	65.5	65.47	67.97	$1.85x10^{-16}$
CSA	1	0.5	0.53	0.55	0.46
MD *CSA	1	2.7	2.69	2.79	0.095
Residuals	11273	10859.0	0.96		
(iv) lm(nTel ~ CSA* MD)	**Df**	**Sum Sq**	**Mean Sq**	**F**	**P**
CSA	1	4.0	4.03	4.18	0.041
MD	1	62.0	61.97	64.33	$1.16x10^{-15}$
CSA * MD	1	2.7	2.69	2.79	0.095
Residuals	11273	10859.0	0.96		

Table 4.11b | Effect of MD, number of stressful life events (SLE) and their interaction on normalised measure of mean telomere length (nTel). See **Table 3.10a** for description of table columns and contents

4.4.2 Chronic stress causes molecular changes in mice

To gain a mechanistic understanding of the relationship between stress, mtDNA and telomere length, I undertook the analysis of a few mouse experiments that I had designed together with collaborators Simon Chang and Guo-Jen Huang, who performed the experiments.

4.4.2.1 Stress has dose dependent and reversible effect on molecular markers

The first experiment was designed to establish if chronic stress or adversity could cause changes in amount of mtDNA and mean telomere length. Mice from a single laboratory inbred strain were used so that genetic diversity did not complicate the interpretation of experimental outcome. 16 C57BL/6J mice (eight males and eight females) were stressed for four weeks (for five days a different stressor was administered: tail suspension, force-swim, footshock, restraint and sleep deprivation, followed by two days rest). At zero, two and four weeks of stress, blood and saliva samples were taken from both stressed mice and 16 age-matched non-stressed controls (eight males and eight females), and mtDNA and telomere length were assessed by qPCR (Methods).

Consistent with the findings in humans, in mice stress significantly increases mtDNA levels and decreases mean telomere length in saliva and in blood. After four week of stress, stressed mice showed a mean increase in mtDNA levels in saliva of 210% compared to the unstressed animals (t test P = 0.0036) and in blood of 240% (t test $P=6.1X10^{-5}$). Meanwhile, the amount of telomeric DNA was reduced 28% in saliva (t test P=0.0001) and 30% in blood (t test P =0.0017) in stressed mice as compared to non-stressed. There were no significant differences in the white cell parameters between stressed and non-stressed animals (all P-values > 0.05) indicating that this result was unlikely due to differences in blood cellular composition.

At four weeks, half of the animals (eight stressed mice and eight controls) were kept in home cages without any intervention to model a recovery period of no stress. Molecular markers were again tested in blood and saliva, and results are shown as week eight in Figure 4.12. These tests were carried out to determine whether the molecular changes persisted when animals were no longer subject to stress. Four weeks after the discontinuation of stress, there were no significant differences between control animals and those that had been previously exposed to stress (mtDNA in saliva $P = 0.50$, mtDNA in blood $P = 0.38$, telomere length in saliva $P = 0.85$, telomere length in blood $P = 0.76$, all P-values from t-tests). These results indicate that the molecular changes are, at least in part, reversible.

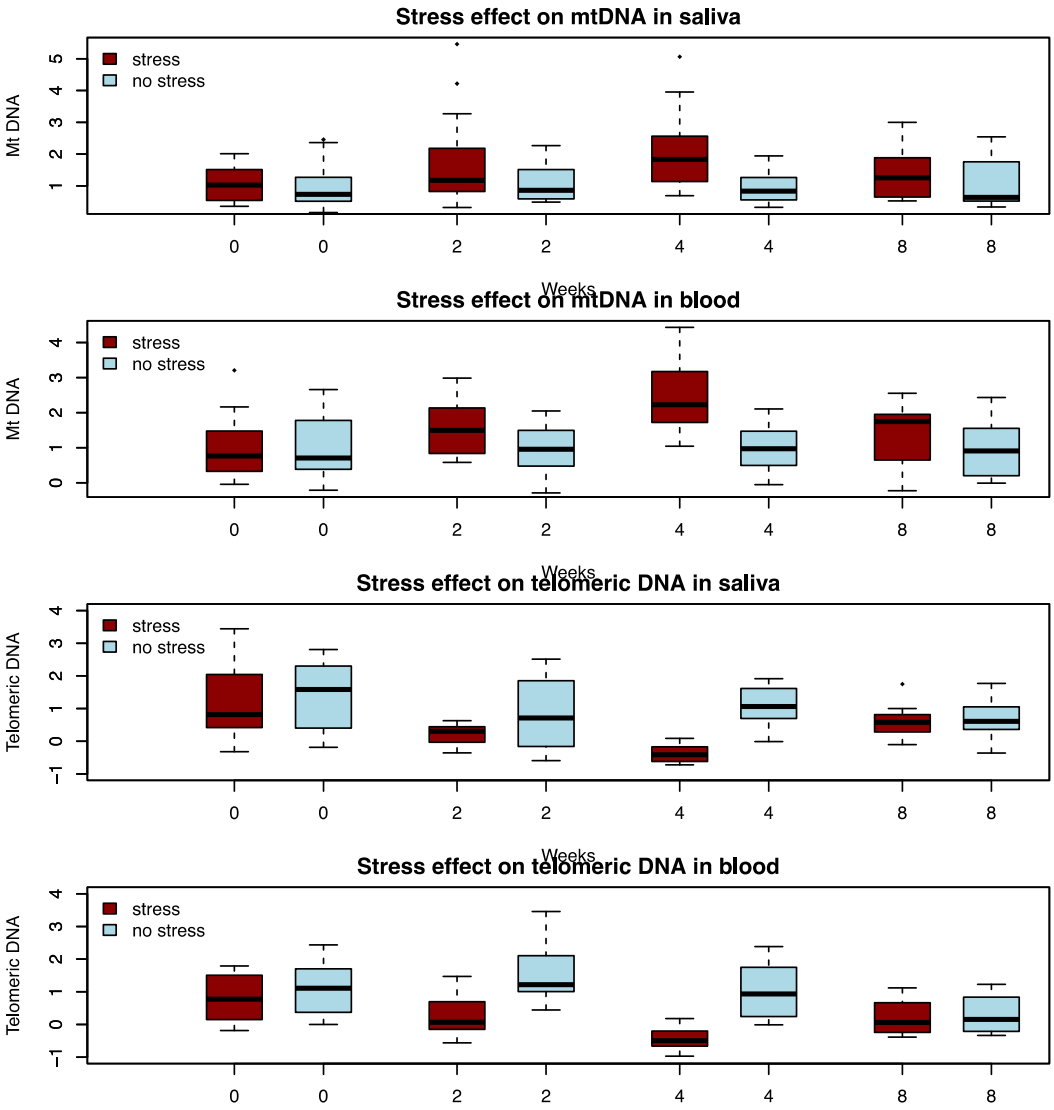


Figure 4.12 | Effect of chronic stress on mtDNA in saliva and blood of mice. Boxplot of relative mtDNA levels and relative mean telomere lengths over time in mice exposed to stress (red) and controls (blue). The vertical axis shows the amount of DNA, assessed by qPCR, relative to the mean of the values obtained before stress was imposed (week 0). The mean of week 0 is set to 1, so that results from subsequent weeks are fold changes relative to pre-stress levels. The horizontal axis is time in weeks from the beginning of the experiment. Stress was discontinued after week 4, so week 8 shows results for previously stressed animals after 4 weeks of living in a home-cage. At the four weeks time point, relative mtDNA levels in blood and saliva were significantly greater in stressed animals (t test $P=6.1X10^{-5}$ and $P = 0.0036$ respectively). Also at the four weeks time point, relative mean telomere lengths in stressed mice were significantly lower in saliva (t test P=0.0001), and blood (t test P =0.0017) as compared to non-stressed mice. Differences between stressed and non-stressed mice in both measures were not significant at the start of the experiment, nor at the 8 week time point.

4.4.2.2 Molecular changes are tissue specific

Immediately after the cessation of stress, multiple tissues from the other 16 animals were assayed for amount of mtDNA and mean telomere length using qPCR (Methods). Figure 4.13 shows results for four tissues: liver, muscle, brain (hippocampus) and ovary (ovary was chosen as we were interested to assess whether the changes might be transmitted to the next generation). For amount of mtDNA, there was a significant increase in liver ($P = 0.005$), a significant decrease in muscle ($P = 0.014$) but no significant alterations in hippocampus ($P = 0.50$) and a suggestive change in ovary ($P= 0.086$). For mean telomere length, there was a significant 54% reduction in liver ($P = 0.03$), but no significant changes in other tissues (muscle $P= 0.23$, hippocampus $P= 0.59$, ovary $P= 0.30$). These results (all P-values from t-tests) reveal tissue specific changes in mtDNA, and possibly in telomere length, as a consequence of stress.

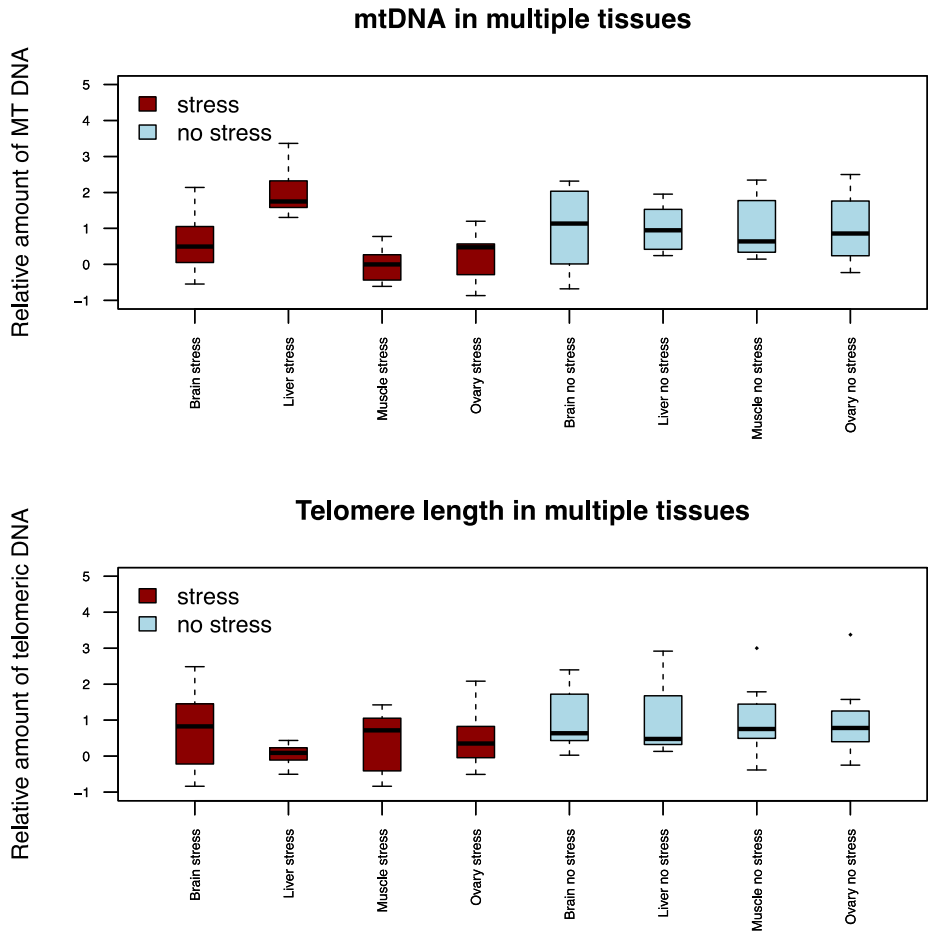


Figure 4.13 | Alterations of mtDNA in different tissues after four weeks of stress. Upper panel shows the assessment of mtDNA in four tissues; lower panel shows the assessment of telomere length in four tissues. The vertical axis shows the amount of mtDNA or telomere length assessed by qPCR, relative to the mean of the values obtained for control animals (no stress). The horizontal axis gives the names of the tissues for the two conditions: stress (red) and no stress (blue).

4.4.2.3 Mitochondrial function is altered in tissues with increased mtDNA

Changes in amount of mtDNA could reflect functional changes in mitochondria in those tissues that underwent changes in amount of mtDNA following chronic stress. To test if oxidative phosphorylation was affected in mice that were stressed, as compared to controls, a second experiment was designed to assay the oxidative phosphorylation capacity of mitochondria-enriched fractions from the liver of stressed and non-stressed mice.

Figure 4.14a shows the mean values for the change in oxygen concentration over time for liver mitochondrial preparations taken from eight stressed and eight control (not stressed) animals. Addition of an equal amount of adenosine diphosphate (ADP, driving force for the electron transport chain after depletion of residual driving force in the fraction with excess respiratory substrates Glutamate and Malate) induced a greater increase in oxygen consumption in mitochondria from the non-stressed animals than stressed ones ($P = 0.038$, from a linear model, Figure 4.14b). The complete quenching of oxidative phosphorylation upon addition of electron transport chain inhibitor KCN in both stressed and non-stressed mice showed oxygen consumption during the experiment was solely due to oxidative phosphorylation. Results of this experiment showed that oxidative phosphorylation efficiency is reduced in liver tissue of mice whose mtDNA levels had increased in response to stress, suggesting either an adaptive switch to glycolysis, or a compromise in mitochondrial function.

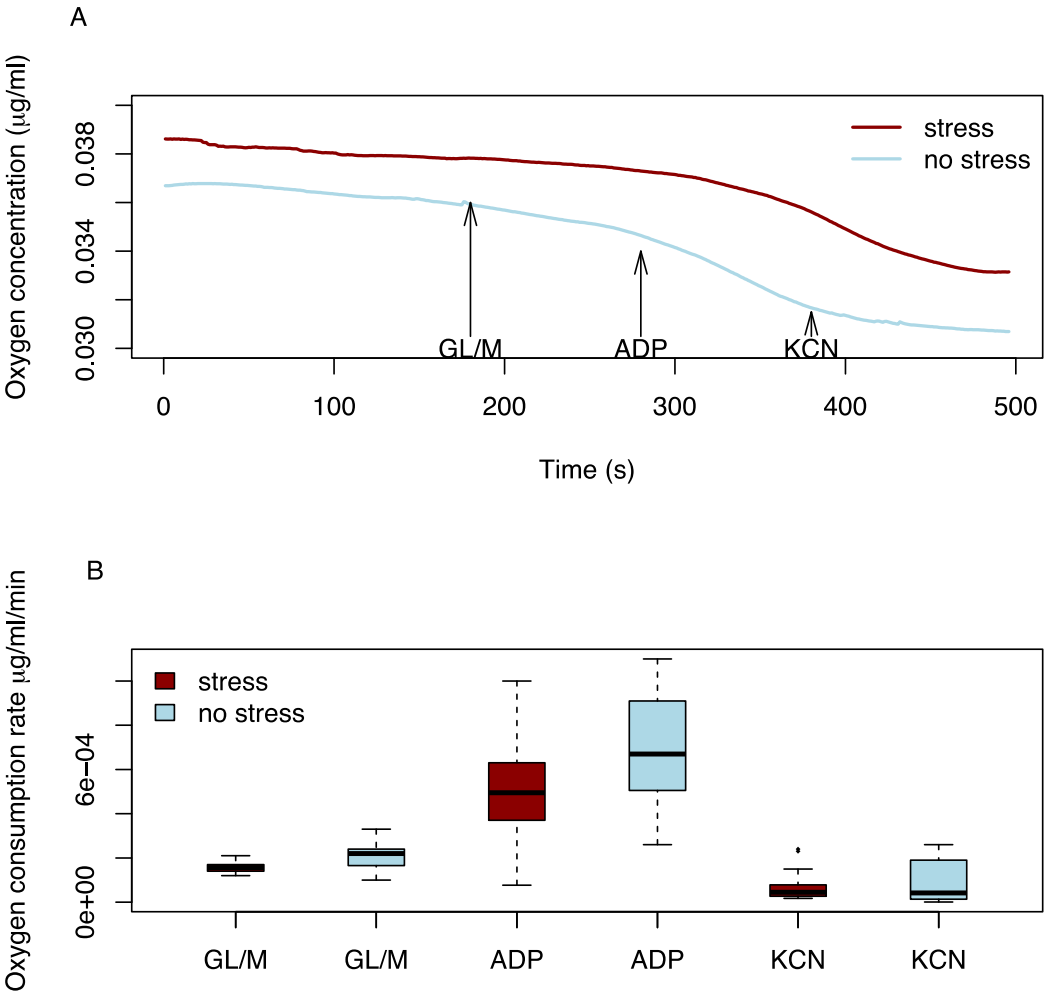


Figure 4.14 | The oxygen consumption of mouse liver after stress administration. **a**. oxygen concentration (vertical axis) detected per second (horizontal axis) per µg of mitochondria. The slope of the curve indicates the rate of oxygen consumption. Glutamate/Malate (GL/MA) is added at 3 minutes after addition of isolated mitochondria, and oxygen consumption was assessed after substrate addition. The addition of adenosine diphosphate (ADP) (100s later) initiates active respiration while potassium cyanide (KCN) (100s later) inhibits all mitochondrial function. **b**. Oxygen consumption rate per µg of mitochondria after the addition of the three compounds, comparing stressed and non-stressed animals.

4.4.2.4 Glucocorticoid administration reproduces the effects of stress

What might be inducing the molecular changes? The final experiment was designed to test one hypothesis: activation of the hypothalamic pituitary adrenal (HPA) axis and the increase in circulation of stress steroids could have elicited the molecular changes [197-201].

In this experiment, corticosterone was administered to eight C57BL/6J female mice over four weeks, and the oil vehicle used for administration of corticosterone was injected in eight control mice over the same period of time at the same intervals. Figure 4.15 shows that at four weeks there was significantly higher mtDNA levels in the saliva (P= 0.011) and in the blood (P= 0.0013) of treated mice compared to controls and that telomere length had significantly reduced in both tissues (in saliva P= 0.0023; in blood P= 0.0016, all P-values from t-tests).

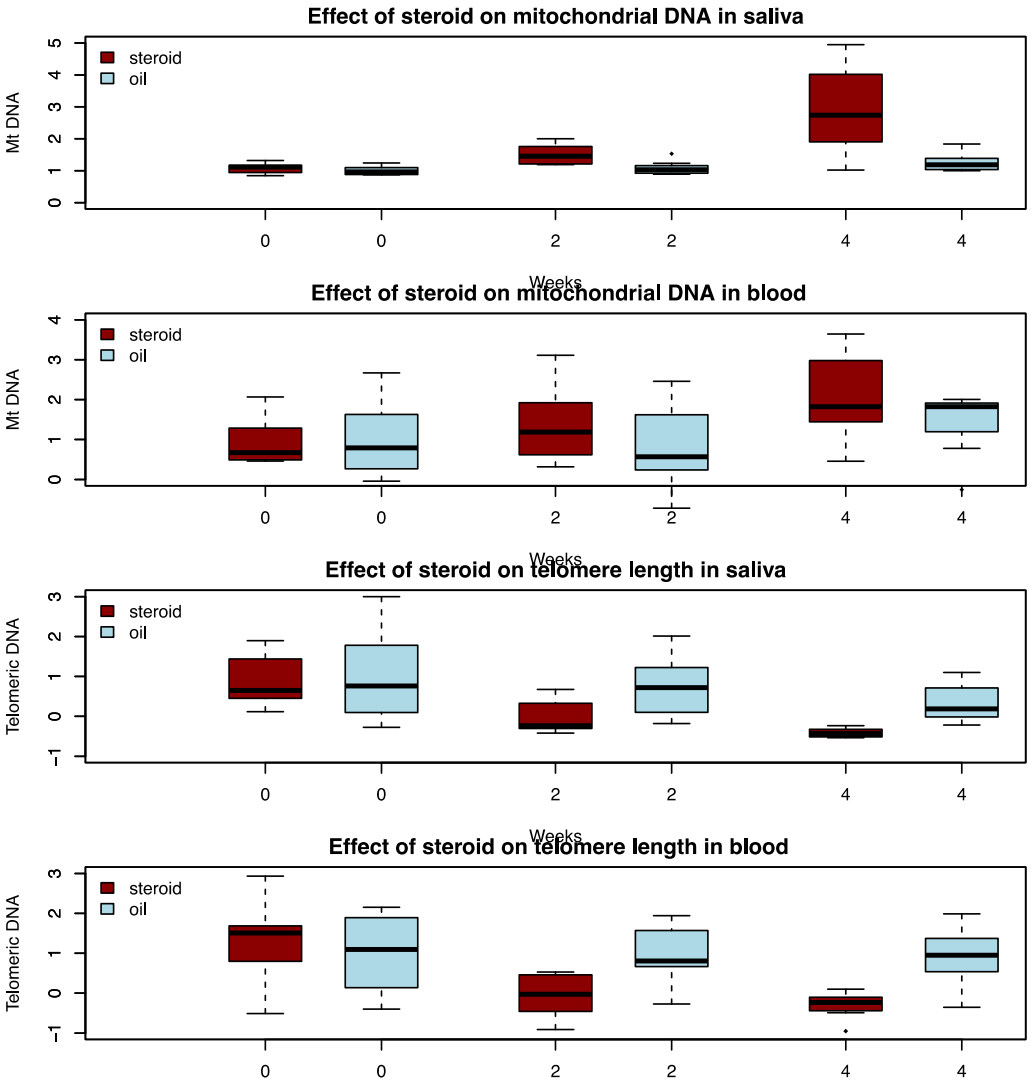


Figure 4.15 | Effect of daily subcutaneous injection of corticosterone on mtDNA and telomere length in saliva and blood in mice. Boxplot of relative mtDNA levels and relative mean telomere lengths over time in mice injected with corticosterone (red) and controls injected the same volumes of oil vehicle(blue). The vertical axis shows the amount of mtDNA assessed by qPCR, relative to the values obtained before corticosterone was injected (week 0). The mean of week 0 is set to 1, so that results from subsequent weeks are fold changes relative to pre-stress levels. The horizontal axis is time in weeks from the beginning of the experiment. After four weeks relative mtDNA levels of mice injected with corticosterone were significantly higher in saliva (t test P= 0.011) and blood (t test P= 0.0013) as compared to mice injected with oil vehicle; relative mean telomere lengths were significantly reduced in both tissues (in saliva: t test P= 0.0023; in blood: t test P= 0.0016) in mice injected with corticosterone as compared to mice injected with oil vehicle.

4.5 Discussion

The first half of this chapter identifies two molecular markers for MDD, increase in amount of mtDNA and decrease in mean telomere length across all chromosomes. To establish the association between the molecular markers and MDD, there is a need to ensure the observations were not due to technical artefacts or biological explanations unrelated to MDD. Is also important to find out which tissues are showing these molecular changes, and what effects these changes might have on cellular and systemic physiology.

There are many possible explanations for the elevated mtDNA levels in cases of MDD as compared to controls. For example, while MDD has been associated with hypercortisolemia[202], the excess secretion of the stress steroid cortisol, it has also long been known that blood cortisol levels fluctuate with the day-night cycle. To attribute molecular differences as due, at least in part, to hypercortisolemia in MDD cases, there is a need to first show there was not a systematic difference in the time of collection of saliva samples for cases of MDD and that for controls.

While careful analysis of detailed documentation of data collection details from CONVERGE showed no such artefacts could explain the difference between cases of MDD and controls in the amount of mtDNA and mean telomere length, lending credibility to these changes as bona fide biological signals, it is still important for any future work to carefully examine all possible sources of artefacts before drawing conclusions. The better one is able to characterize the molecular changes, the closer one gets to their origins and their biological significance.

The second part of the chapter reports two important observations on the relationship between MDD and amount of mtDNA levels and mean telomere length. First, the changes in levels of mtDNA and mean telomere length are contingent on presence of MDD. There were no significant molecular changes in those who report

childhood sexual abuse, but have never been depressed. Second, in a mouse model, while stress over a period of weeks does increase the amount of mtDNA and decrease mean telomere length, both changes were at least partly reversible. While early and cumulative environmental adversity such as CSA and number of SLE are associated with depression in CONVERGE, and have been shown in previous studies to result in permanent changes in physiology and risk of MDD[203], the results presented in this chapter indicate that it is important to recognize two trajectories, one leading to molecular signatures of stress, and one to illness. As the molecular changes were shown to be contingent on MDD, the obvious question is whether there is a genetic link between the control of mtDNA levels and MDD, and if genetic associations for one affect the other. The next chapter describes analyses done in attempt to answer this question, and looks in the mtDNA sequence for other ways MDD, stress and mtDNA levels could be related.

Animal experiments showcased in this chapter were instrumental in establishing the causal relationship between stress and changes in the molecular markers. I was heavily involved in conceptualizing this work, the protocols for qPCR quantification of amount of mtDNA and telomere lengths were results of my own research, and I performed all the statistical analysis on the results. However, Mr Simon Chang and Dr Guo-Jen Huang from Chang Gung Unversity in Taiwan were the ones who maintained the animals and performed behavioural experiments on them. Their expertise in animal models of anxiety and depression and their proficiency in behavioural experiments were indispensible to the success of the analysis. Results from these experiments showed conclusively that mtDNA and telomere changes in chronic stress are time-dependent, irreversible, tissue-specific, and at least in part due to stress steroids produced by the HPA axis.

One perplexing finding was while both phenomenons were striking in saliva, blood and liver, no obvious molecular changes were observed between the single brain

region, hippocampus, assayed in the stressed and control animals. However, the lack of difference in mtDNA levels and mean telomere length between stressed and non-stressed animals in the hippocampus hence cannot be seen as indication that pathways involved in these changes did not have their origins in the central nervous system, as there are many other regions of the brain, each consisting of many different cell types, that these changes could manifest. It is conceivable that the source or components of the signal transduction pathway coordinating mitochondrial changes in different tissues would be elusive to investigations unless close inspection of tissues, especially finely defined brain regions, ideally at the single cell or cell-type level under different environmental conditions both in vivo and in vitro, could be carried out. Further studies aimed at finding the origin of molecular responses following adversity and how it might lead to MDD could benefit from such tissue-based investigations.

5. Genetic and environmental link between mtDNA and MDD

In the previous chapter I have shown that the amount of mtDNA changes in response to external stress: there was significantly more mtDNA in the saliva and blood of people with major depression than in controls. Chronic stress also altered the amount of mtDNA in mouse tissues, and, at least in part, was restored to pre-stress levels following cessation of stress. Changes in cellular composition could not account for these observations, suggesting that the mtDNA alterations reflected changes within cells, and corticosteroid signaling down the hypothalamic-pituitary axis (HPA) may be involved in causing these changes since injection of corticosterone alone can recapitulate effects of chronic stress. Intriguingly, increase in mtDNA in chronically stressed mice was accompanied by a decrease in efficiency in oxidative phosphorylation, instead of an increase, suggesting a reduction in utilization of coding regions of the mtDNA despite the increase in the number of copies.

These observations raise the questions: a) what genetic factors and cellular mechanisms contribute to the increase in mtDNA levels during stress, b) are there changes to the mtDNA sequence, causing the counterintuitive decrease in efficiency of oxidative phosphorylation, and might they constitute a response to stress, and lastly c) do these changes, or the variation in extent of change between individuals, confer different liability to MDD?

I took a genetic-mapping approach to answering these questions, and as in the previous chapter, relied on animal models of chronic stress to explore the consequences of stress *in vivo*. I looked in the nuclear genome for variations that could explain variations in mtDNA levels independently of, or in interaction with, environmental stress. In other words, I wanted to know if there were trans effects from the nuclear genome on the level of mtDNA. Many studies had reported tissue heterogeneity in mitochondrial content. While the range of mitochondrial levels across different tissues

was found to be large and mitochondria could exist in different physical forms in different cell types under different cellular conditions, mitochondrial levels within tissue types and amount of mtDNA per functional unit of mitochondria are relatively constant. Despite evidence pointing to regulatory mechanisms exerting tight and tissue specific control over mitochondrial levels, there had been no report of any genetic basis to that control to date. This apparent gap might have been caused by the lack of a high-throughput means to assay mitochondrial levels while simultaneously genotyping a large number of people at a substantial number of variant sites across the genome. Traditional means, such as a combination of quantitative polymerase chain reaction (qPCR) and whole-genome genotyping array containing a few hundred thousand variant sites, would likely to be both time-consuming and costly. This was overcome by using NGS data in the CONVERGE cohort. In this chapter, I describe the first genome-wide association study (GWAS) on mtDNA levels to find out what might be involved in the molecular pathways controlling mtDNA levels during resting and stressed states.

To answer the second question, I looked into variations in mtDNA sequences in CONVERGE, some of which are clonal (homoplasmic) and likely inherited in uniparental fashion from the mother, some are not (heteroplamic) and likely reflect accumulation of mutations that may have been subject to selection.

That mtDNA mutations could cause disease was first reported in 1988: some patients with mitochondrial myopathy had heteroplasmic deletions in their mtDNA [204]; the maternally inherited Leber hereditary optic neuropathy (LHON), which causes sudden onset blindness, was caused by a homoplasmic missense mutation in the ND4 gene [205]; and myoclonic epilepsy and ragged red fiber disease (MERRF) was caused by a heteroplasmic mutation in the tRNA-Lys gene [206]. These discoveries set the stage for investigating and understanding a broad range of enigmatic familial and age-related diseases. Genetic epidemiological studies quantifying only the most common pathogenic mtDNA mutations have estimated that the incidence of clinical mitochondrial diseases is

about one in 5000[207]. Hundreds of pathogenic mtDNA mutations have now been documented [208,209] and these can affect every tissue in the body, depending on the mutation's severity, nature, and heteroplasmy level. It was found that pathogenic mtDNA mutations can be very common in the population: a survey of newborn cord bloods revealed that one in 200 infants harbored one of 10 common pathogenic mtDNA mutations [210]. The most clinically overt class of mtDNA variants comes in the form of maternally inherited disease mutations.

While heteroplasmy can arise from inheritance, it is also attributable to somatic mutation and is in part tissue specific [211]. Human sequencing shows that low levels of heteroplasmy are relatively common, with at least one heteroplasmy in 90% of individuals [212]. Although mtDNA appears to be particularly exposed to mutation given the relative lack of the DNA repair machinery comparable to that maintaining nuclear DNA integrity [213] there is substantial evidence that mechanisms exist to ensure that the 16Kb mammalian mitochondrial genome is inherited as single, practically clonal, lineage. The mechanism by which this enrichment occurs in either germline or somatic cells remains a mystery.

De novo mtDNA mutations can accumulate at anytime throughout life from the oocyte to the cells of the elderly. The earlier in development the mutation occurs, the more broadly it will be distributed due to later cell divisions – mtDNA mutations that arise in the embryonic period can be dispersed throughout the body, while those that arise in an adult organ will be tissue specific [204]. This genetic mosaicism results in bioenergetic mosaicism and phenotypic complexity, which is compounded by differential sensitivity of different organs to different mitochondrial physiological alterations, depending on their flexibility in terms of metabolic strategy and respiratory substrate utilization. For instance, the brain is the most sensitive to partial bioenergetic defects followed by heart, muscle, kidney, and endocrine systems [214]. Hence, subtle systemic mitochondrial deficiencies can result in organ-specific symptoms.

The mtDNA encodes 37 genes (13 protein components of the electron transport chain, 2 mitochondrial ribosomal RNAs, and 22 tRNAs specific for mitochondrial translation), all of which serve the oxidative phosphorylation pathway in concert with proteins and RNAs encoded by the nuclear genome and imported into the mitochondria. The coding sequences take up 93.23% of the mtDNA (576bp-16023bp), hence any accumulation of mutations in mtDNA, even if completely random, would likely affect gene function. I therefore asked: if amount of mtDNA increases during stress, was there any site-specific accumulation of mutation in response or adaption to the "stressed" or "depressed" state, how may these mutations affect bioenergetics, and could that give a clue to what may be happening in the state of stress at the molecular level?

5.1 Genetic control over mtDNA

The number of mtDNA molecules appears to be tightly regulated, as inferred from the constant amount of mtDNA per mitochondrion in different cells in the presence of marked variation in the number of mitochondria between cell types[51]. The mechanisms involved are largely unknown. One of the components of the mtDNA replication machinery mitochondrial transcription factor A (TFAM) appears to be a key transcription factor controlling the amount of mtDNA [215] but it is still not known how that signal is used by cells to count the amount of mtDNA that they require. A further observation that the number of mitochondria predicts cell division better than other measures such as cell volume [216] or cell size suggest replication of mtDNA and mitochondrial biogenesis, while autonomous, may be regulated by cell cycle machinery and vice versa.

Using NGS data at 6,242,619 SNPs for all CONVERGE samples, I determined that the amount of mtDNA, defined as number of reads mapping to the mitochondrial genome normalized by read coverage of the nuclear genome, is a heritable trait, with an estimated SNP-based heritability [160] (Methods) of 15.6% (standard error (SE) = 5.1%). Two loci were identified to be contributing to variation in the amount of mtDNA exceeding genome-wide significance (genomic control lambda (λ) = 1.017, Manhattan plot in Figure 5.1a, quantile-quantile plot in Figure 5.2).

One locus is within the *TFAM* gene on chromosome 10 (top SNP rs11006126, P-value = 8.73×10^{-28}, Figure 5.1b) and the second over the *CDK6* gene on chromosome 7 (top SNP rs445, P-value = 6.03×10^{-16}, Figure 5.1c). *TFAM* and *CDK6* respectively contribute 1.90% (SE = 1.80%, permutation adjusted Likelihood ratio test (LRT) P-value = 3.71×10^{-10}) and 0.50% (SE = 0.28%, P-value = 1.55×10^{-10}) to the variance in amount of mtDNA (Methods).

The most highly coding associated SNP in the *TFAM* gene lies in exon1 containing the mitochondrial targeting sequence, resulting in a serine to threonine

amino acid substitution at the 12th residue. Conservation of the serine residue between all humans, chimpanzees and gorillas point to this being a derived change which may have functional consequences (Figure 5.3). Dr Tomas Malinauskas performed the structural biology analysis (Methods). The most highly associated SNPs in *CDK6* lie in Intron 1 and do not cause protein changes in *CDK6*.

Both associations have been replicated in an independent cohort of 1753 samples in the Avon Longitudinal Study of Parent and Children[217] (ALSPAC), which forms part of the UK10K study (Methods). Associations at the top two SNPs rs445 and rs11006126 in both separate studies and the joint analysis (Methods) using genotypes from both studies are summarized in Table 5.1.

Cohort	SNP	BETA	SE	P
CONVERGE	rs445	0.119	0.015	4.57E-16
UK10K	rs445	0.130	0.057	2.14E-02
JOINT	rs445	0.120	0.014	3.84E-17
CONVERGE	rs11006126	0.195	0.018	1.17E-27
UK10K	rs11006127	0.179	0.047	1.53E-04
JOINT	rs11006128	0.193	0.017	1.08E-30

Table 5.1 | Replication of association with amount of mtDNA at top GWAS SNPs This table shows the effect size (BETA) and its standard error (SE), and P-values (P) from linear regression for association between the two top SNPs (SNP) in *CDK6* (rs445) and *TFAM* (rs11006127) genes in three cohorts (Cohort): our study CONVERGE, the ALSPAC cohort in UK10K, and a joint cohort containing samples from both CONVERGE and ALSPAC. All associations were significant and the directions of effect at both SNPs in both cohorts are the same.

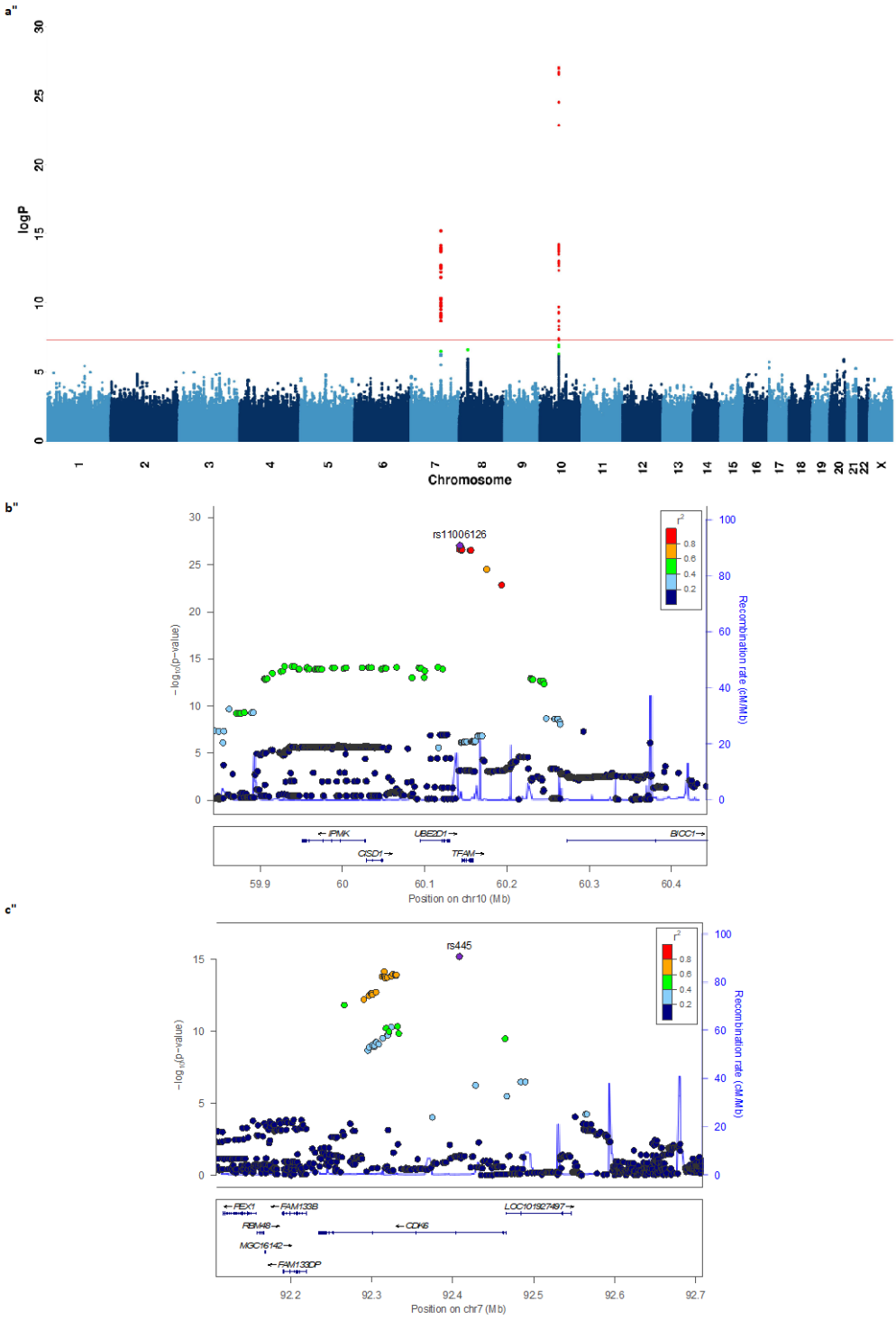


Figure 5.1 | Two loci associated with mtDNA (a) Manhattan plot of genome wide association for amount of mtDNA. Detailed views of two loci associated with amount of mtDNA (b) over the *TFAM* region on chromosome 10 at 60.1 Mb, (c) the *CDK6* gene on chromosome 7 at position 92.4 Mb. The −log10 P-values of imputed SNPs associated with amount of mtDNA are shown on the left y axis. The horizontal axis gives chromosomal position in megabases (Mb). Genes within the regions are shown in the bottom panels. Linkage disequilibrium of each SNP with top SNP, shown in large purple diamond, is indicated by its colour. The plots were drawn using LocusZoom[155].

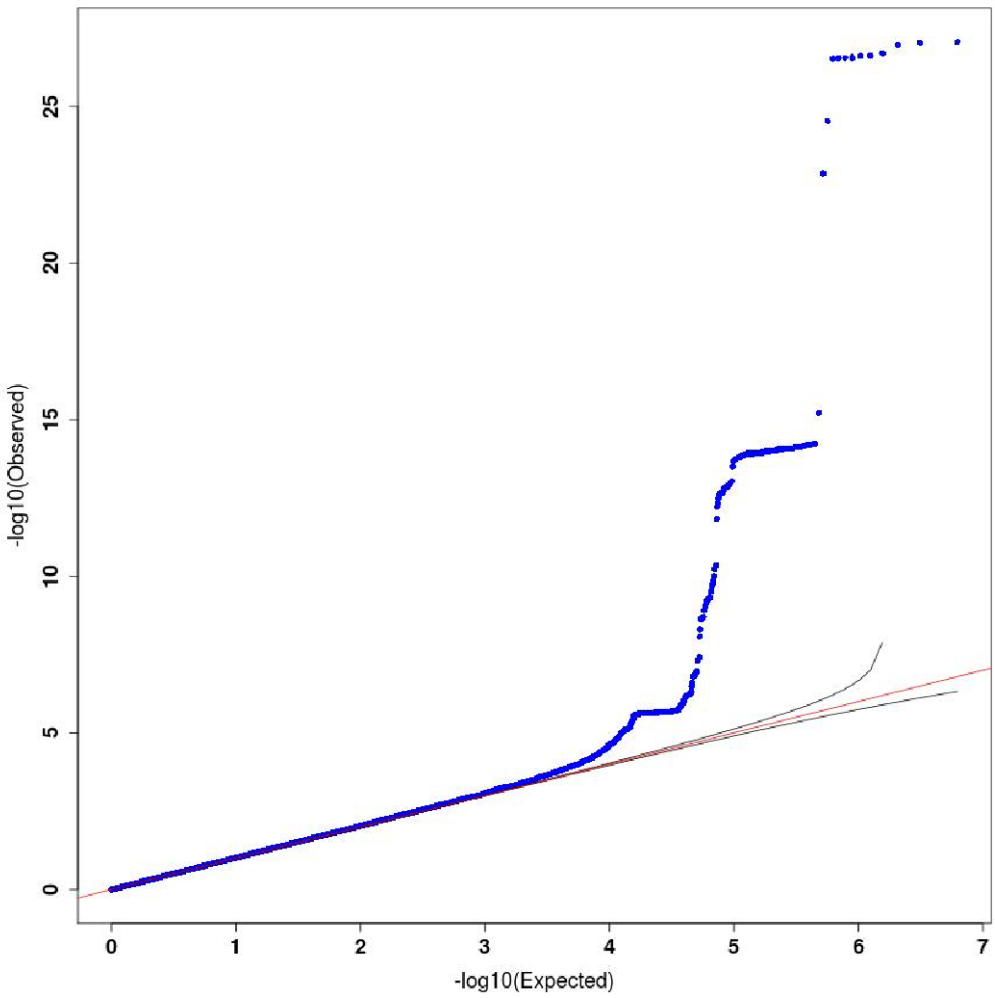


Figure 5.2 | Quantile quantile plots for GWAS on amount of mtDNA Quantile-quantile plot of GWAS for amount of mtDNA using leave-one-chromosome-out (LOCO) linear mixed model implemented in FastLMM on 10,442 samples (5,224 cases of MD, 5,218 controls), genomic control lambda (λ) = 1.017.

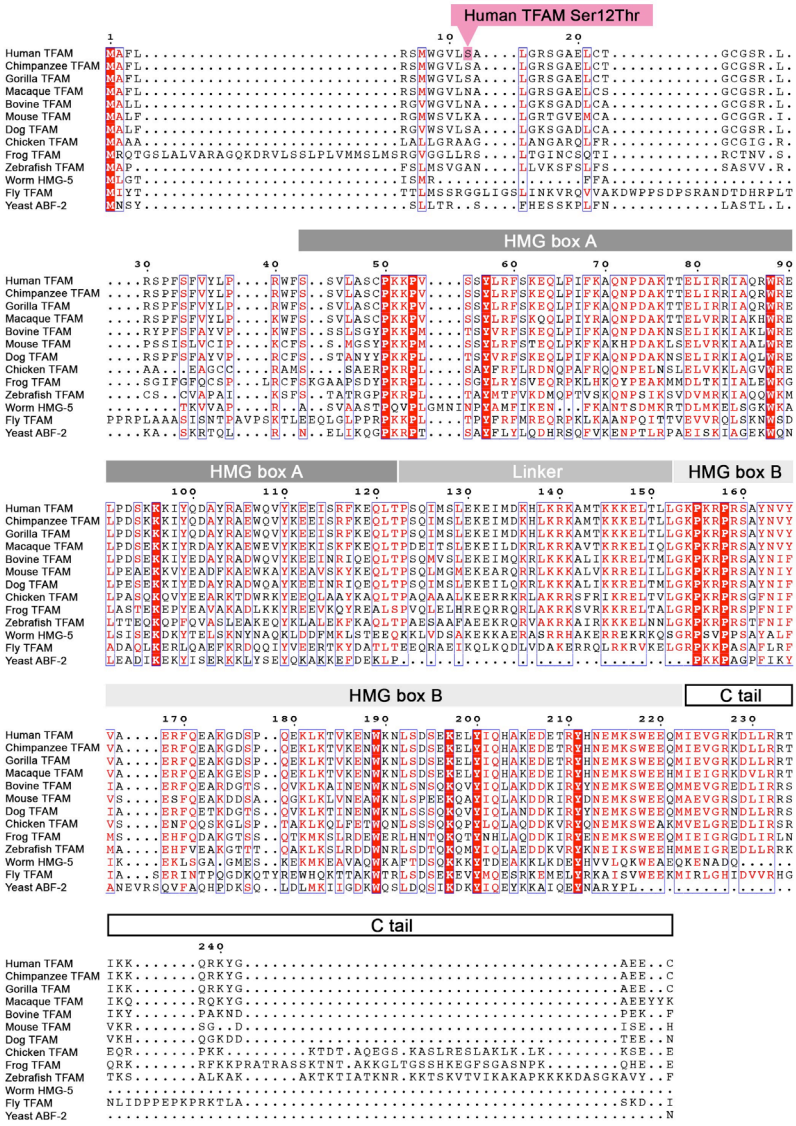


Figure 5.3 | Amino acid sequence alignment of the human TFAM and homologues Amino acid sequence alignment of the TFAM (Transcription factor A, mitochondrial) family members, homologues from worm (HMG-5, high mobility group 5) and yeast (ABF-2; ARS-binding factor 2, mitochondrial). Alignment of human (UniProt Knowledgebase [218] identification accession number Q00059), chimpanzee (*Pan troglodytes*, H2Q1X6), gorilla (*Gorilla gorilla gorilla*, G3QQI3), macaque (*Macaca mulatta*, F7FXP0), bovine (*Bos taurus*, Q0II87), mouse (*Mus musculus*, P40630), dog (*Canis familiaris*, E2RJ35), chicken (*Gallus gallus*, Q8AYT2), frog (*Xenopus laevis*, Q91634), zebrafish (*Danio rerio*, Q08BL2), worm (*Caenorhabditis elegans*, GenBank identification number CCD63785.1), fly (*Drosophila melanogaster*, Q86BR8), and yeast (*Saccharomyces cerevisiae*, Q02486) TFAM-like proteins. Evolutionarily conserved residues are shown in a red background. Numbering corresponds to human TFAM (shown above sequences). Approximate boundaries of domains were derived from the crystal structure of human TFAM [219] (Protein Data Bank (PDB) ID 3TMM) and are shown above sequences. Ser12 of human TFAM is highlighted in pink.

5.2 Genetic overlap between MDD and amount of mtDNA

I then explored the relationship between MDD, genetic control over the amount of mtDNA, and heteroplasmy. I asked first if genetic control over the amount of mtDNA was different in cases of MDD and controls. When we included MDD as a covariate in the GWAS for amount of mtDNA. Peaks at TFAM and CDK6 remained significantly associated with amount of mtDNA (P-value = 6.96×10^{-28}, 6.63×10^{-16} for rs11006126 and rs445 respectively), and there were no further peaks found. Figure 5.4 shows the Manhattan and quantile-quantile plots of this GWAS. Even though MDD is a strong predictor of the amount of mtDNA as discussed in Chapter 4, and even though the genetic correlation between amount of mtDNA and MDD computed from genome-wide SNPs [90] was 46.7% (SE = 14.3%, P-value = 3.53×10^{-4}), SNPs at TFAM and CDK6 were not associated with MDD (P-value = 0.70 for rs11006126, P-value = 0.53 for rs445, from a linear mixed model).

Testing specifically for an interaction with MDD, we found no region of the genome exceeded genome-wide significance. Interaction P-values at the two SNPs that were associated with the amount of mtDNA are 0.70 at rs11006126 and 0.71 at rs445. Therefore, our data indicates that genetic control over the amount of mtDNA is independent of MDD disease status. Figure 5.5 shows the Manhattan and quantile-quantile plots of this analysis.

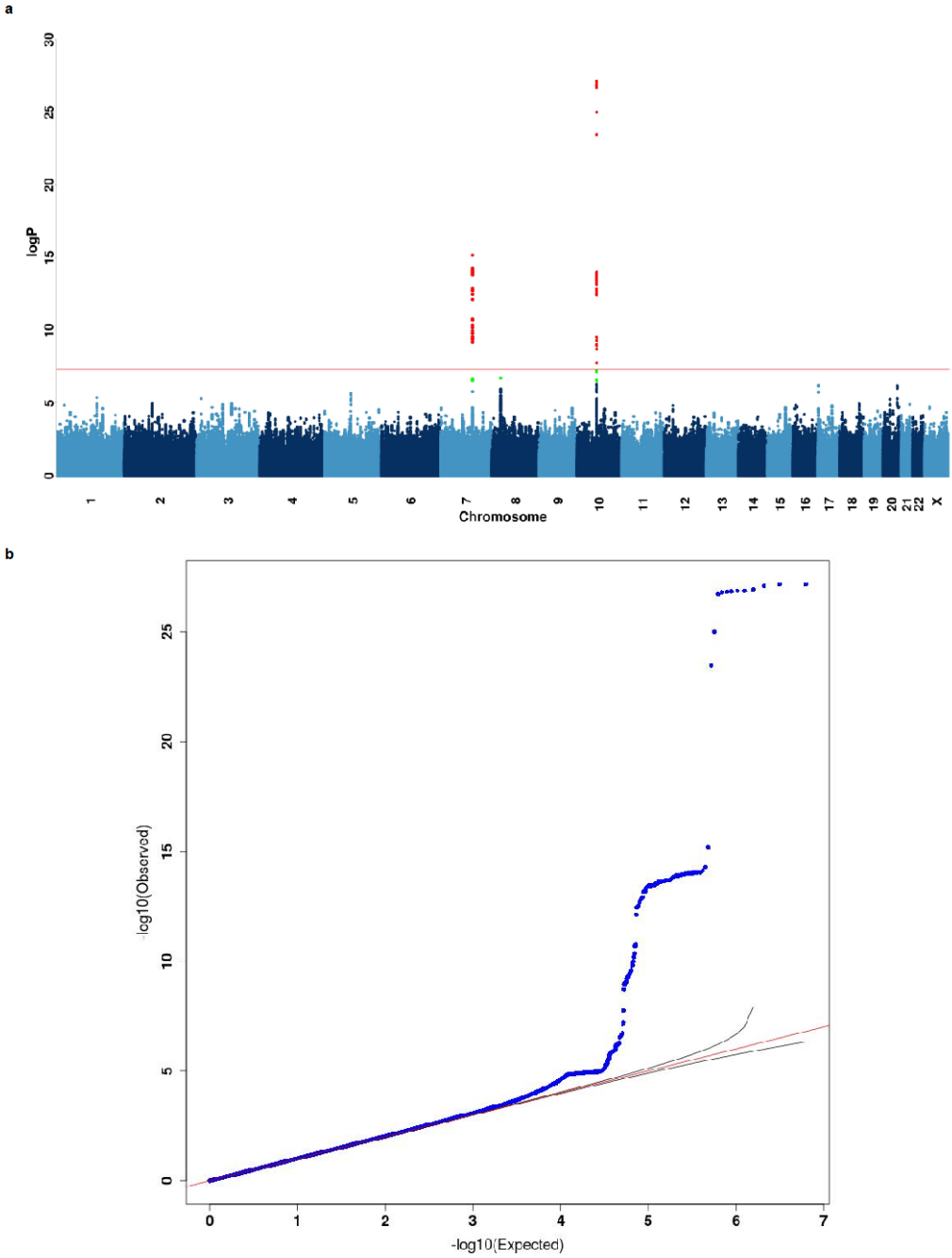


Figure 5.4 | GWAS on amount of mtDNA with MD disease state as covariate (a) Manhattan plot and (b) quantile-quantile plot of genome wide association for amount of mtDNA with MD included as covariate term; genomic control lambda (λ) = 1.018. Both *TFAM* (P-value at rs11006126 = 6.96×10^{-28}) and *CDK6* signal remain (P-value at rs445 = 6.63×10^{-16}).

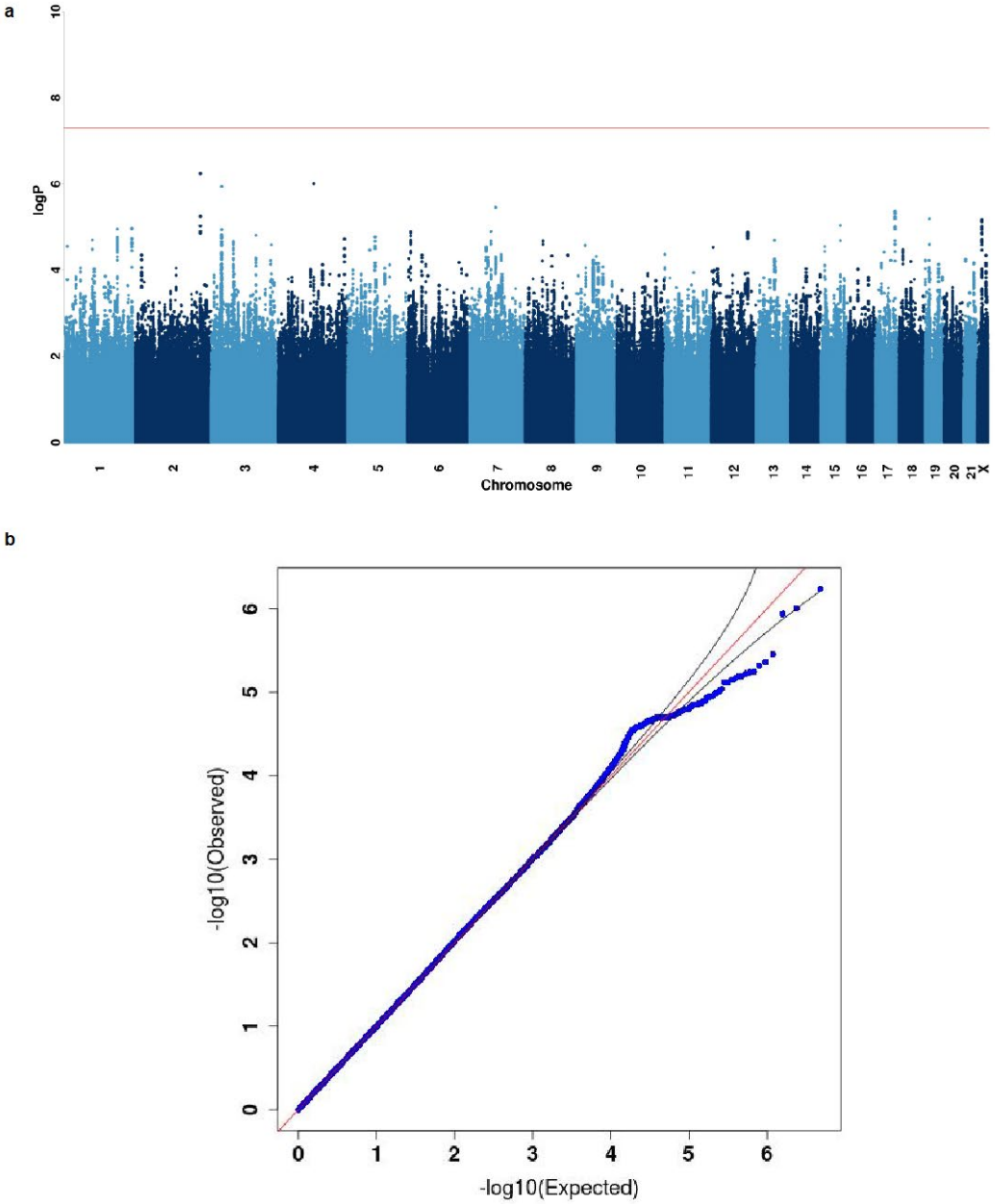


Figure 5.5 | Interaction effect between MDD and genome-wide SNPs on amount of mtDNA
(a) Manhattan plot and (b) quantile-quantile plot of genome wide testing for interaction effect between MDD and SNPs on amount of mtDNA; genomic control lambda (λ) = 1.016. Both *TFAM* (P-value at rs11006126 = 0.702) and *CDK6* signal remain (P-value at rs445 = 0.711).

5.3 Mitochondrial genome control over mtDNA

The finding that the amount of mtDNA is dynamic raised the question whether mitochondrial DNA sequence itself governs the alterations in mtDNA quantity and whether cycles of replication might result in mutations in the mtDNA molecule. The latter was considered likely given the relatively higher mutation rate of mtDNA compared to genomic DNA. Therefore I set out to identify variant sites in the 16Kb mitochondrial genome and to quantify their frequency both between and within individuals (although the sequencing data provided only 1.6X coverage of the nuclear genome, coverage of the mitochondrial genome is approximately 100 fold). I considered two problems that might affect the analysis.

First, I had to ensure that the sequence variants I obtained were in the mitochondrial and not in the nuclear genome, since the latter contains partial copies of the former [220]. After applying stringent quality control filters and criteria (Methods), 89% of the mitochondrial genome was considered sufficiently unique for variant calling (Figure 5.6), consistent with previous reports [221]. Second, I needed to identify and distinguish between two forms of variation in the mitochondrial genome. mtDNA sequence variants can be present in all copies (homoplasmy) or be present only in a fraction of molecules (heteroplasmy). I defined homoplasmic variants as those site were there are two alleles present in the cohort, each supported by more than 90% of sequencing reads in individual samples (Methods).

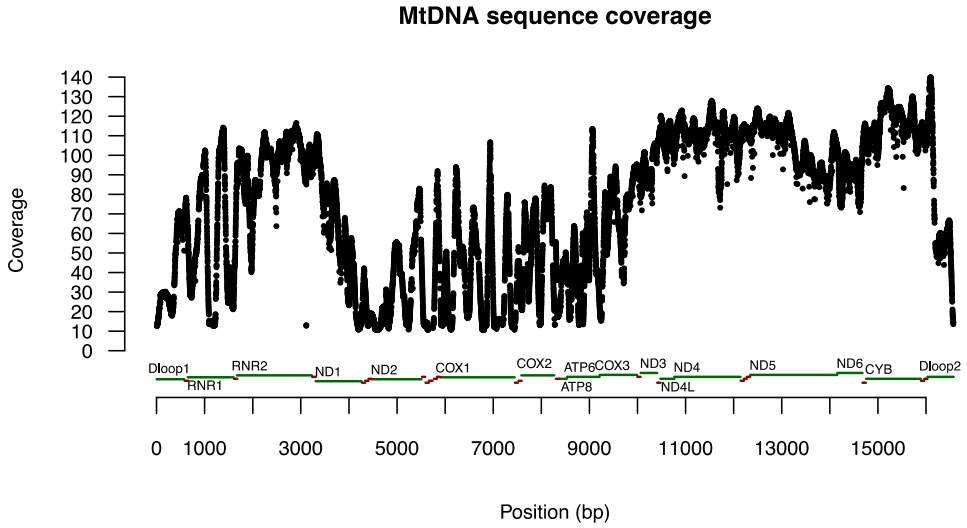


Figure 5.6 | Accessibility of mtDNA for variant calling Mean per site read depth across 10442 samples for all sites on the mtDNA. Reads included in the per site read count have passed the following quality control criteria: a) mapping quality > 59, b) both ends of paired end read map uniquely to the mtDNA reference NC_012920.1, c) do not contain mismatches to the mtDNA reference with total length of five base pairs or above, d) base at site in question has base quality of higher than 20. Sites with more than 10% samples with read depth less than 10 would not be used for calling homoplasmic variants, and sites with more than 10% samples with read depth less than 50 would not be used for calling heteroplasmic variants.

To estimate the sensitivity of our variant calling method, I applied my method of homoplasmic variant calling on the 1000 Genomes Phase 3 whole-genome sequencing data (n=2,598, of which 216 are from Han Chinese populations, CHB and CHS). I compared the set of common homoplasmic variants (>0.1%) I called in CONVERGE with homoplasmic variants called in 1000 Genomes Project Phase 3 samples. 948 out of 1,030 (88.85%) of the common homoplasmic variants I found in CONVERGE were found to be polymorphic in 1000 Genomes Phase 3, and 546 of them (51.17%) were found to be occurring at >0.1% frequency in the Han Chinese populations.

I then checked the allele frequencies of homoplasmic variant sites called in CONVERGE at diagnostic SNPs for haplogroups against previously documented frequencies in Central Asians and East Asians in MitoMap [208], the combined estimates of

allele frequencies from previous studies on Han Chinese populations [222], as well as allele frequencies called from all 1000 Genomes Phase 3 samples and those from Han Chinese populations (Table 5.2). The frequencies of alleles at SNPs diagnostic for the different haplogroups are highly consistent with those previously reported.

Finally, I checked for the presence of four homoplasmic variants previously reported to be present in Asian mitochondrial Haplogroups (A, C, D and X) [223] and associated with colder climates in our sample. Two of the four variants were identified as homoplasmic variants occurring at frequencies above 0.1% in the CONVERGE sample, and were found to be associated with PC1, capturing the North-South cline of geographical origin of our samples, calculated from eigen decomposition of the a GRM calculated from 322,911 common SNPs of minor allele frequency (MAF) > 1% and linkage disequilibrium $r^2 < 0.5$ using GCTA [134] (version 1.24.4, Methods) (Table 5.3). Two variants were not called in our dataset due to lack of coverage at the sites.

Haplogroup	Diagnostic site	Reference Allele	Diagnostic Allele	Frequency (%)					
				Central Asia	East Asia	Han Chinese	1000G	1000G CHSCHB	CONVERGE
A	663	A	G	7.00	7.00	4.00-16.70	5.54	6.02	7.21
B	8280-8289	ACCCCCTCTA	A	5.00	16.00	0.60-3.30	NA	NA	NA
C	13263	A	G	12.00	5.00	2.00-8.00	3.04	3.24	3.98
D	5178	C	A	15.00	26.00	1.40-3.30	5.15	22.55	21.67
E	13626	C	T	0.00	0.00	NA	NA	NA	0.09
F	6392	T	C	5.00	11.00	1.40-6.00	3.71	14.42	14.34
G	4833	A	G	5.00	4.00	NA	0.00	0.00	14.77
H	7028	C	C	15.00	1.00	NA	11.37	0.00	NA
I	4529	A	T	1.00	0.00	NA	0.05	0.00	NA
J	13708	G	A	3.00	1.00	NA	5.27	5.56	4.31
K	12308	A	G	1.00	0.00	NA	NA	NA	NA
L	3594	C	T	0.00	0.00	NA	15.78	0.00	0.02
S	8404	T	C	0.00	0.00	NA	0.19	0.92	0.28
T	4917	A	G	6.00	0.00	NA	0.22	0.00	0.28
U	12308	A	G	10.00	0.00	NA	NA	NA	NA
V	4580	G	A	0.00	0.00	NA	NA	NA	NA
W	11947	A	G	2.00	0.00	NA	0.96	0.00	0.06
X	6371	C	T	0.00	0.00	NA	0.43	0.00	0.02
Y	8392	G	A	1.00	1.00	1.30-3.80	0.19	0.00	NA
Z	9090	T	C	2.00	2.00	1.30-7.10	0.31	1.85	3.24

Table 5.2 | Frequency of diagnostic alleles for mitochondrial haplogroups This table shows in the first four columns the mitochondrial haplogroup, diagnostic diagnostic site for the haplogroup, the reference allele in the human mtDNA reference NC_012920.1, and the diagnostic allele. The next six columns show the frequency (%) of occurrence of the diagnostic allele in Central Asians and East Asians from MitoMap[208], estimates from multiple studies in Han Chinese population[222], all 1000 Genomes Phase 3 samples, 1000 Genomes Phase 3 Chinese Beijing (CHB) and Chinese South (CHS) samples, and CONVERGE study samples.

Variant	Adaptive	Haplogroup	Gene	Amino	Linear	Coefficient	Allele

position	allele			acid change	regression P value with PC1		frequency
4824	G	A	ND2	T112A	NA	NA	NA
8794	T	A	ATP6	H90Y	NA	NA	NA
11969	A	C	ND4	A404T	0.00056	0.0022	0.020
15204	C	C	cytb	I153T	0.00290	0.0025	0.013

Table 5.3 | Correlation between PC1 in CONVERGE and four homoplasmic variants adaptive to arctic climate This table shows in the first three columns position, adaptive allele, haplogroup in which four previously identified homoplasmic variants reported to be present in Asian mitochondrial Haplogroups (A, C, D and X) [223] and associated with colder climates. The next two columns show the genes which the variants reside in, and the amino acid changes they cause. The next two columns show the association P value and regression coefficient between the presence of the adaptive alleles and PC1 in CONVERGE (Methods), which captures the North-South cline. The final column shows the allele frequency of the adaptive alleles in CONVERGE. Two sites at positions 4824 and 8794 were not called in CONVERGE due to lack of coverage.

These quality controls experiments showed that the calling of homoplasmic variants from low-coverage sequencing on the CONVERGE sample was likely to be accurate. Using the set of common homoplasmic variant calls (1,031 sites above 0.1% frequency in CONVERGE) I performed association testing using linear regression for each site with amount of mtDNA. No common homoplasmic variants were significantly associated with the amount of mtDNA after correction for multiple testing.

I next investigated the association between heteroplasmic sites and the amount of mtDNA. As it is easier to detect heteroplasmy in samples with higher mtDNA coverage simply because there are more mtDNA reads, I down-sampled the mtDNA reads such that coverage at all sites in the mtDNA in all samples were exactly 50 reads. I disregarded sites at which more than 10% of individuals had fewer than 50 reads and therefore could not be down-sampled, so estimates of heteroplasmy from all remaining sites were based on equal coverage on an adequate sample size. To be considered a

heteroplasmic site in a sample I required the presence of two alleles, each supported by two or more reads, and to be considered for further analysis we needed the heteroplasmy to be present at higher than 0.1% frequency in the cohort (10 individuals). Using these criteria I identified 26 heteroplasmic sites from low coverage sequencing in all samples.

However introgression of some mitochondrial DNA into the genomes of sequenced subjects might not be present in the reference genome, and would confuse analysis of heteroplasmic variants. Therefore I used additional long-range PCR data on 72 randomly selected samples from the whole cohort (36 cases of MD, 36 controls) to verify the variant sites identified through low-coverage sequencing were indeed of mitochondrial origin and not nuclear sequence variation (Methods). In order to ensure accurate estimates of the degree of heteroplasmy in an individual for testing the association between site-specific degree of heteroplasmy and amount of mtDNA, I checked the degree of heteroplasmy called from each of the 72 randomly chosen samples (number of reads supporting the minor allele out of 50 down-sampled reads) from low-coverage sequencing against that called from high coverage sequencing of long range PCR on their mtDNA. I considered only those sites where the average degree of heteroplasmy among 72 samples were higher than 10% in both datasets, and the Pearson correlation r2 between them were higher than 0.9, leaving therefore only six sites for association testing. One out of the four sites was significantly associated with mtDNA levels (position = 513, P-value = 2.49×10^{-31}, variance explained = 1.86%). This site is located in the D-loop regulatory region, though had not been identified as a transcription factor binding site [221], DNAseI protected sites, or splice cut sites[224]. The association P-values between the six heteroplasmic sites with amount of mtDNA are shown in Table 5.4.

MARKER	REF	ALT	REQ	ANNOTATION	GENE	Association with mtDNA			
						EFFECT	VAREXP	P	LOGP
MT146	T	C	0.005	upstream	RNR1 tRNA-Phe	0.65	0.001	6.14E-01	0.212
MT451	A	I	0.002	regulatory	NA	-0.501	0	4.76E-01	0.322
MT513	G	I	0.416	regulatory	NA	-0.279	0.005	3.27E-09	8.485
MT5894	A	I	0.017	regulatory	NA	-0.617	0.001	3.36E-01	0.473
MT15939	C	I	0.014	regulatory	NA	-0.005	0	9.45E-01	0.025

Table 5.4 | Association between the degree of heteroplasmy at four heteroplasmic sites with the amount of mtDNA This table shows the association between degree of heteroplasmy at four heteroplasmic sites in the mtDNA and amount of mtDNA quantified from low-coverage sequencing in 10,442 samples from CONVERGE. The first four columns show the position of the heteroplasmic site in mtDNA (MARKER), reference allele in the human mitochondrial genome reference NC_012920.1 (REF), the alternative allele (ALT) for which the heteroplasmy is detected, and the frequency of occurrence of this heteroplasmy in the cohort (FREQ). An "I" in the ALT column means the heteroplasmy is for an insertion or deletion mutation. The next two columns show characterisitics of the four heteroplasmic sites: annotation of variant function (ANNOTATION) and nearest gene (GENE). The final four columns show the results of association testing between degree of heteroplasmy at each site with amount of mtDNA by linear regression: direction of effect (EFFECT) from linear regression, variance of amount of mtDNA explained by the site heteroplasmy (VAREXP) from difference in residual sum of squares in analysis of variance (ANOVA) between the model with and without degree of heteroplasmy as the test term in linear regression, P value of association (P) between degree of heteroplasmy and amount of mtDNA, and –log10 of the P value (LOGP). One site position 513 is significantly associated with amount of mtDNA, with a positive effect, and it lies in the D-loop regulatory region in the mtDNA.

5.4. Mitochondria heteroplasmy in stress and MDD

In the previous chapter I observed that the amount of mtDNA increases in response to stress and is higher in cases of MD as compared to controls in the CONVERGE cohort, that these effects were reversible and likely more pronounced in datasets where DNA samples were taken during an active depression episode or state of stress. Could stress be causing alterations in heteroplasmy in mtDNA sequence?

To identify and quantify the number of heteroplasmic sites independent of sequence coverage I down-sampled the mtDNA reads so that each site was covered by 50 reads. I disregarded sites at which more than 10% of individuals did not fulfill this criterion, so that estimates of heteroplasmy from all remaining sites were based on equal coverage on an adequate sample size. I required the presence of two alleles, each supported by two or more reads at sites and occurring at higher than 0.1% frequency in the cohort (10 individuals), for a site to be considered heteroplasmic. Using these criteria we identified heteroplasmy at 26 positions in the mtDNA.

Low-coverage sequencing has a high false-negative rate in the calling of heteroplasmic sites, and some of the heteroplasmies we detected using low-coverage sequencing might be reads originating from NUMTs mapping incorrectly to the mtDNA reference. Therefore I performed the same analysis using the high-coverage sequencing on 72 samples we obtained with long-range PCR, ensuring both high-coverage and mtDNA origin of sequencing reads. I quantified heteroplasmy in the 72 samples, after down-sampling the high-coverage sequencing to 500 reads per site and discarding those samples with fewer than 500 reads covering at each site. I only counted heteroplasmic sites where the fraction of reads supporting the minor heteroplasmic allele was higher than 1% (supported by a minimum of 5 reads). This yielded 408 such sites. Comparison between mean degree of heteroplasmy obtained from low-coverage sequencing on 72

samples and high-coverage sequencing of products of long range PCR on mtDNA on the same samples showed that when the mean degree of heteroplasmy among samples is high (1% or higher) in both low-coverage sequencing and long range PCR data, the correlation between the two estimates is high (Table 5.5). For lower degrees of heteroplasmy, this correlation drops substantially. Therefore I considered further only six sites where heteroplasmy rates are greater than 1% in an individual from both low-coverage sequencing and high coverage sequencing of long range PCR on mtDNA.

I asked whether MDD is associated with mtDNA heteroplasmy in a site specific manner. Using a logistic regression controlling for sequencing batch, age of samples and top 5 PCs I did not find any of the six heteroplasmic sites with verified degrees of heteroplasmies by long-range PCR data to be associated with MDD. This analysis, however, could not rule out whether there were other sites in the mtDNA whose degree of heteroplasmy were directly affected by MD or the state of stress that we could not detect and test with any accuracy. I tested a broader hypothesis: if MDD increases heteroplasmy at some sites in the mtDNA, I should see a higher load of heteroplasmy in cases of MD as compared to controls. Although I was not able to call heteroplasmic sites with high accuracy from low-coverage sequencing data, errors due to calling would be independent of MD disease status and should merely increase noise and mask differences between MD and heteroplasmy load, instead of create such differences.

Position	Mean degree of heteroplasmy		Pearson r2	P-value
	Low coverage	Long range PCR		
MT146	**0.118**	**0.126**	**0.999**	**0.000**
MT451	**0.011**	**0.012**	**0.999**	**0.000**
MT512	0.000	0.006	NA	NA
MT513	**0.374**	**0.411**	**0.984**	**0.000**
MT567	**0.004**	**0.011**	**0.999**	**0.000**
MT955	0.000	0.002	NA	NA
MT1290	0.000	0.000	NA	NA
MT2226	0.000	0.000	NA	NA
MT2487	0.010	0.000	0.013	0.340
MT3167	0.000	0.000	NA	NA
MT3243	0.000	0.000	NA	NA
MT3447	0.000	0.001	NA	NA
MT5894	**0.033**	**0.060**	**0.829**	**0.000**
MT6289	0.000	0.001	NA	NA
MT10283	0.000	0.000	0.003	0.663
MT10306	0.027	0.000	0.000	0.942
MT11031	0.001	0.005	0.000	0.911
MT11866	0.000	0.002	NA	NA
MT12417	0.002	0.005	0.013	0.341
MT12684	0.000	0.000	0.004	0.615
MT15536	0.000	0.000	NA	NA
MT15900	0.000	0.000	0.001	0.806
MT15939	**0.012**	**0.012**	**1.000**	**0.000**
MT16129	**0.183**	**0.183**	**0.993**	**0.000**
MT16179	0.003	0.057	0.003	0.657
MT16496	0.002	0.092	0.017	0.285

Table 5.5 | Correlation between degree of heteroplasmy called in 72 samples from low coverage sequencing and high-coverage sequencing of long range PCR of mtDNA This table shows the degree of heteroplasmy in 72 samples for whom there was both low-coverage sequencing data and high coverage sequencing data from long range PCR of mtDNA. The first column shows the 26 sites in the mtDNA identified to be heteroplasmic at an occurrence of higher than 0.1% in CONVERGE. The next two columns show the mean degree of heteroplasmy among the 72 samples at each of the 26 sites, quantified from low coverage sequencing and high coverage sequencing of long range PCR product on mtDNA respectively. Where the mean degree of heteroplasmy is 0 in either columns, there was no heteroplasmy detected in any of the 72 samples using the relevant dataset. The next two columns show the per site Pearson correlation r2 and between the degree of heteroplasmy called from low coverage sequencing and high coverage sequencing of long range PCR on mtDNA in 72 sample, and the P-value of the correlation. Six sites highlighted in bold showed mean degrees of heteroplasmy of over 1% in both datasets and were highly correlated, and were included in association testing between site-specific degree of heteroplasmy and amount of mtDNA.

Using the 26 heteroplasmic sites I called from low-coverage sequencing on all samples, I computed a heteroplasmic load per sample (number of heteroplasmic sites per sample) and found it to be significantly higher in cases of MDD than controls (mean fold increase = 1.05, P value = 1.40×10^{-4}, Figure 5.7a) after controlling for sequencing batch, age of samples and top 5 principle components (PCs) capturing population structure among the samples (Methods). Using high-coverage sequencing of long-range PCR on mtDNA on 72 samples, I was able to perform the same analysis with higher sensitivity and accuracy of detection of heteroplasmic variants (Methods), and confirm our finding: cases of MD had higher levels of heteroplasmy (mean fold increase = 1.01, P value = 0.032, Figure 5.7b).

The human data alone cannot determine whether the increased rates of heteroplasmy are cause or consequence of the stress-related illness, therefore we turned to an experimental system in which we explored the consequences of exposure to stress in mice. With my collaborators (Mr Simon Chang and Dr Guo-Jen Huang) I generated and analysed long-range PCR of the mtDNA from 12 female C57BL/6J mice, six of which were subject to a four-week chronic stress regime, and six were kept in standard laboratory conditions for the same four-week period. I called heteroplasmy in these mice in the same way I did the 72 CONVERGE samples, down-sampling to 500 reads per site, discarding those samples with fewer than 500 reads covering at each site, and discarding those sites with 8 animals fulfilling the above criterion. I only took into account those heteroplasmic sites where the fraction of reads supporting the minor heteroplasmic allele was higher than 1% (supported by a minimum of 5 reads), yielding a total of 76 sites.

As observed in the previous chapter, I saw higher amount of mtDNA in stressed mice than in non-stressed (mean fold increase = 2.16, P value = 0.004, Figure 5.8a). I expected to see markedly higher heteroplasmic load in the stressed mice compared to the non-stressed, despite the sample size, and it was indeed the case: stressed mice had

higher levels of heteroplasmy (mean fold increase = 1.46, P value = 0.029, Figure 5.8b). This shows that stress could increase heteroplasmic load.

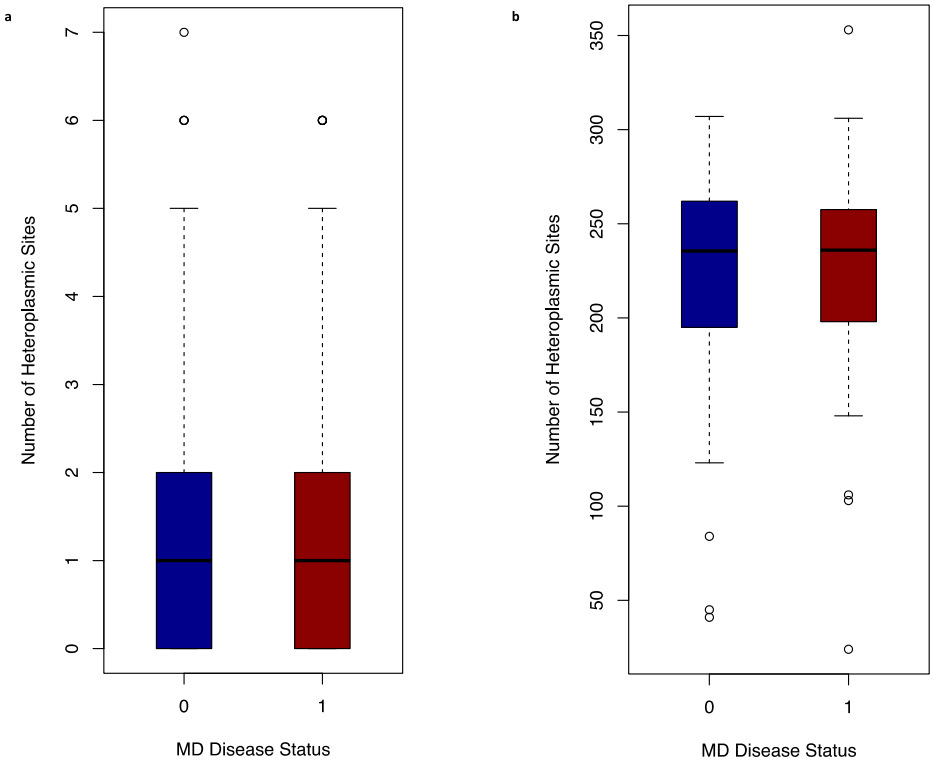


Figure 5.7 | Heteroplasmy counts from low-coverage sequencing of all CONVERGE samples and high coverage sequencing of long range PCR products of 72 samples The figures shows (a) boxplot of the number of heteroplasmic sites quantified per samples for cases of MD (red, 5224 samples) and controls (blue) from low coverage sequencing; cases of MD had higher number of heteroplasmic site per sample than controls mean fold increase = 1.05, P value = 1.40×10^{-4}); (b) boxplot of the amount of mtDNA quantified from low-coverage sequencing of long range PCR of mtDNA from 72 samples (36 cases of MD, shown in red, 36 controls, shown in blue); cases of MD had higher number of heteroplasmic site per sample than controls (mean fold increase = 1.01, P value = 0.032)

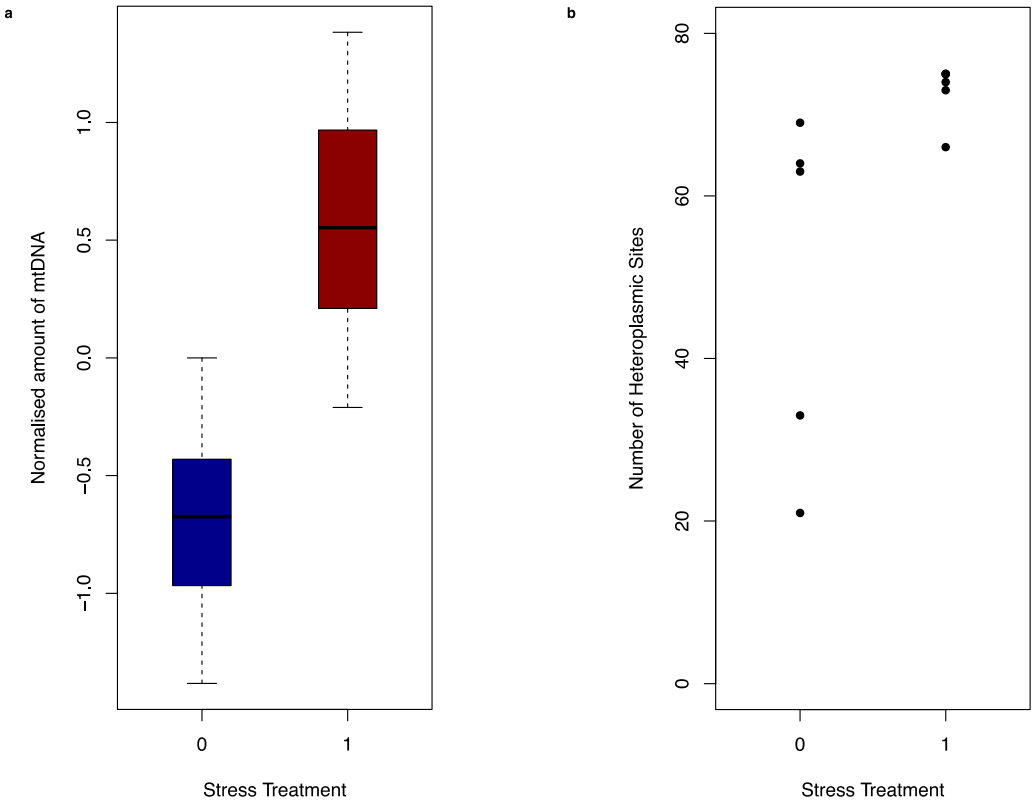


Figure 5.8 | Heteroplasmy counts in stressed and control mice The figures shows (a) boxplot of the amount of mtDNA quantified from high-coverage sequencing of long range PCR of mtDNA from liver samples (controlled for starting mass of genomic DNA extracted from mice livers, then quantile normalized among all twelve mice) of six female mid-age C56BL/6J mice exposed to a four-week chronic stress protocol (shown in red), compared to that from the liver samples of six female, mid-age C56BL/6J control mice (shown in blue). Normalised amount of mtDNA from liver samples of 6 mice was significantly higher than that in non-stressed control mice mean (fold increase = 2.16, P value = 0.0040), and (b) the number of heteroplasmic variant sites found in high-coverage sequencing of long range PCR of mtDNA (downsampled to equal coverage of 500 reads per site) from liver samples of six female mid-age C56BL/6J mice exposed to a four-week chronic stress protocol, compared to that from the liver samples of six female, mid-age C56BL/6J control mice, where each point represents one mouse; stressed mice had significantly higher number of heteroplasmic variants in mtDNA per sample (mean fold increase = 1.46, P value = 0.029).

5.5 Discussion

The findings detailed in this chapter identified two nuclear genomic loci that contribute to amount of mtDNA and one position in the mitochondrial genome. Many previous studies have shown that mtDNA replication is transcription primed [225,226] and therefore comes under tissue-specific regulation through signaling processes [227] responsive to environmental and physiological demands mediated by nuclear respiratory factors (NRFs), nuclear hormone receptors like the PPARs, THRs, ERRs and other transcription factors like the CREB and YY1, which may partly or mostly be influenced by transcription co-activators such as PGC-1 family of co-activators and Sirtuin proteins. Results presented in this chapter show that in spite of the range of regulatory pathways converging on the control of mtDNA levels, the heritable component of mtDNA levels is low and just two genes, TFAM and CDK6, can explain a large proportion of the genetic effect on mtDNA levels.

TFAM is known to interact extensively with mtDNA and play key roles in its transcription, replication and organization into nucleoids. Crystal structures of TFAM bound mtDNA shows that TFAM imposes U-turns on mtDNA [219,223] when bound in both specific and non-specific manners, and this bending of mtDNA at the light strand promoter (LSP) in particular is indispensible to transcription initiation at the LSP and replication of mtDNA [228] since the truncated transcript at LSP is used to prime the asymmetrical mtDNA synthesis [229,230]. TFAM knockout mice suffer severe depletion of mtDNA and are embryonic lethal while overexpression of TFAM in mice causes dose-dependent increases in mtDNA levels [231], while mere import of TFAM into isolated mitochondria from rats increase mtDNA replication [232].

CDK6, by contrast, has not previously been implicated in mitochondrial function (unlike CDK5, which also associates with D-type cyclins[233]). CDK6 belongs to a family of serine-threonine kinase that is part of the core cell cycle machinery by phosphorylating the D-type cyclins (cyclins D1, D2 and D3), which are unique among cyclins to drive cells

into the S phase of the cell cycle by phosphorylating the retinoblastoma protein [234]. Recent findings suggest that CDK6, despite sequence similarity and known functional redundancy with CDK4, has distinct, kinase-independent[235] transcription activation activities in promoting cell-proliferation[236] and angiogenesis[235] and inhibiting differentiation [236]. To my knowledge there is no report of CDK6 being imported into the mitochondria and interacting directly with mtDNA, nor has it been associated with regulation of mitochondrial activities. The novel genetic association between CDK6 and mtDNA levels compels future interest on direct or indirect effects of CDK6 activities on mtDNA levels and mitochondrial functions that may shed light on new links between metabolism and cell proliferation and differentiation, both in health and disease.

I did not find an interaction between the TFAM and CDK6 loci on mtDNA levels, but that cannot preclude any convergence in the molecular pathways that link them to mtDNA regulation. One clue lies in the most highly associated variants in TFAM being in the mitochondrial targeting sequence, which is cleaved before import of TFAM, potentially affecting the translocation of TFAM into the mitochondria through the protein import machinery. Though CDK6 has never been implicated in regulation of mitochondrial protein import, it has been found in yeast that CDK1-cyclinB and CDK3-cyclinA complexes phosphorylate Tom6 [237], a small assembly factor that has a indispensible role in maintaining stable interactions between the receptor and channel proteins within the translocase of the mitochondrial outer membrane (TOM) complex [238]. As most of the more than 1000 proteins necessary for mitochondrial function, maintenance and replication are imported from the cytosol or endoplasmic reticulum through protein import complexes, it is conceivable that changes in protein import would impact mtDNA transcription, replication and mitochondrial function, and genetic factors affecting mitochondrial protein import would cause variations in the dynamic range through which the mitochondria are able to respond to cellular metabolic and biosynthetic demands. Though Tom6 and other small assembly factors Tom5 and Tom7

have been identified in humans to be less crucial in maintaining the assembly of TOM complex [239], explicit investigation of CDK6 phosphorylation of TOM components and TFAM and other protein entry into the mitochondria may yield mechanistic understanding on regulation of mitochondrial function.

This chapter also explored the contribution of homoplasmic and heteroplasmic mitochondrial sequence variations to control over amount of mtDNA. While there were no homoplasmic variants associated with amount of mtDNA, the association between the amount of mtDNA and a heteroplasmic variant in a regulatory region of the mitochondrial genome suggests that the allele alters replication rate. While this may be completely independent of the environment, variants present at this locus could also be responding to environmental stresses and causing preferential replication of mtDNA molecules containing it under particular environmental cues, thereby causing an increase in the population of mtDNA molecules containing this variant. If this was so, mechanisms for such environmental selection and the consequence of preferential clonal copy of mtDNA molecules containing an allele on a variant site over others would be interesting topics for future research as they may be integral to metabolic responses to stress.

It is also possible that the association may represent an example of adaptive mutation [240], in which genetic variation occurs in response to the environment, rather than independently of it. Instances of environmentally induced mutations have been documented in bacteria [241] and a similar process may affect the mitochondrial genome. Likewise, if this was true, then how this adaptive mutation occurs and what it does to cellular metabolism should not be neglected when trying to understand metabolic changes during stress.

The observation that heteroplasmy load increases during stress adds one more piece to this puzzle. It can be seen as merely a case of increase in random mutations in the mtDNA during increased replication, which as the previous two paragraphs explored

may or may not have been driven by particular sequence variants in the mtDNA, which may or may not have arisen due to adaptive mutation. It could however, just as well be adaptive mutation. The question then lies what would be the adaptive advantage of having a higher heteroplasmic load.

Intriguingly, none of the variants, either in the nuclear or mitochondrial genome, contribute to the risk of MDD, despite the fact that the genetic correlation between MDD and the amount of mtDNA is 46%, suggesting that loci exist that exert pleiotropic influence on both risk to depression and mitochondrial function. MDD is a response to anticipated or perceived stress. Over evolutionary time, one of the most common chronic stresses has been famine. In this context the vegetative features of MDD, alterations in appetite and sleep, may in fact be adaptive and driven in part by the cellular changes in metabolism that were suggested by the molecular markers of stress that were found in the mtDNA.

6. Conclusion

This thesis has described an investigation into molecular markers of stress, identified from a genetic study on major depressive disorder (MDD), a psychiatric disease that has eluded previous genetic studies as well as investigations of its mechanistic underpinnings. In the course of this work I have asked: what happens during stress, how may it contribute to the risk of developing MDD, and why does it increase the risk of MDD in some people but not others? Through a GWAS on MDD and the close examination of one molecular marker – increase in amount of mtDNA in cases of MDD – I now have glimpses at what the answer to these questions may be. In this final chapter, I summarise my findings, reflect on the study design and analysis approaches, and hypothesize the biology in stress and MDD.

6.1 Key findings

My thesis presented three main findings using data from the CONVERGE project on MDD, each of which was followed up with a series of secondary and validation analyses.

The first finding came from the GWAS on MDD conducted on 10,640 samples (5,303 cases and 5,337 controls) in the CONVERGE dataset – I found two genome wide significant associations with MDD, one lying at the 5' side of *SIRT1*, and the other in an intron of *LHPP*. Both signals have been replicated in a completely independent cohort of severe MDD cases and matched controls from Northern China, making them the first replicated association loci for MDD to date. Secondary analyses involving restricting cases of MDD to only those with the more severe, melancholic subtype, and removing samples with the known environmental risk factor for MDD, CSA, both increased the association signal at *SIRT1*, and the latter also discovered a new association locus at *SLC35A37*. Both *SIRT1* and *SLC25A37* have known functions involving the mitochondria – the former is a known sensor of cellular energy states and regulator of mitochondrial biogenesis, the latter is a iron transport channel in the inner mitochondrial membrane essential for the production of iron containing oxygen carrying complexes in red blood cells. Both point to mitochondrial involvement in MDD and implicate aberrant metabolic strategies in MDD.

The second finding came from directly interrogating the mtDNA from NGS data on all CONVERGE samples. I found there are more copies of mtDNA in cases of MDD than controls, even after controlling for artefacts incurred during data collection and processing, using all information pertaining to both processes documented in CONVERGE. The results also persist after testing for possible biological explanations unrelated to MDD, such as difference in cellular composition of saliva in cases of MDD and controls. I then found that while mtDNA levels increase with history of life

adversities, especially those in childhood, it does not fully account for the differences in mtDNA levels between cases of MDD and controls. These results all suggest the increase in mtDNA levels in cases of MDD is related to the disease itself, though it can also be caused by stress. Further analyses of results from animal experiments showed stress increases mtDNA levels in a dose-dependent, reversible and tissue specific way that is mediated partly by stress steroids. The changes in mtDNA are accompanied by a change in metabolic efficiency, at least in some tissues. This set of findings suggests genetic overlap between control over amount of mtDNA and MDD, and suggest changes in the mitochondria accompanying the increase in its copy number during stress. It was also found that telomere length decreases with MDD, though its significance has not been as fully explored in this study as that of increase in mtDNA levels in MDD.

The third set of findings came from explicitly testing these hypotheses. I found two genome-wide significant loci associated with amount of mtDNA, both of which have been replicated in an independent cohort of depression samples from the UK. Both associations are independent of MDD status, and neither show interaction with MDD. Nevertheless, genetic correlation between amount of mtDNA and MDD was high at 46%, suggesting undiscovered shared genetic contributions. No significant association was found between the homoplasmic variants in the mtDNA and either MDD or the amount of mtDNA, but one heteroplasmic variant was found to be associated with amount of mtDNA, I also found the total amount of heteroplasmy increases with increasing mtDNA levels, and therefore is higher in cases of MDD than controls. This phenomenon was found also in mice were stressed for four weeks, suggesting the increase in mtDNA heteroplasmy accompanies the increase in mtDNA levels. This is in agreement with the previous finding that increase in mtDNA levels can be accompanied by changes in metabolic efficiencies at least in some tissues – as mtDNA is mostly coding and the gene products are all involved in the electron transport chain needed for aerobic respiration, accumulation of mutation in the mtDNA could compromise respiratory rate.

All three findings suggest stress causes changes in mtDNA, and this change may be larger in cases of MDD than controls. This difference between cases and controls may be due to differences in their regulation of mtDNA levels and sequence mutation during stress, and this may be genetically determined. I have not, however, found what this process may be, and what genetic factors contribute to it.

6.2 Study design and analyses

6.2.1 GWAS study design

Results from COVERGE have shown that stringent criteria in sample recruitment and documentation of known and putative environmental risk factors increase homogeneity in the disease phenotype and maximize the shared genetic contribution to disease in cases of MDD recruited. Without such selectivity for cases of MDD and screening of controls to prevent inclusion of those who may develop MDD in the future, the number of cases of MDD needed for a 80% chance at finding just one GWAS hit was predicted to be 50,000[96]. The two genome-wide significant associations with MDD found in just 10,640 samples at *SIRT1* and *LHPP*, as described in Chapter 3, show the substantial increase in power to detect genetic associations using such an approach from previous predictions. Two findings corroborate the importance of careful sample selection and the value of 'deep' phenotyping: (i) restricting analysis to the melancholic subtype leads to an increase in genetic signal at the *SIRT1* locus, and (ii) excluding cases with a severe environmental cause of MDD led to the discovery of a third genome-wide significant locus on the *SLC25A37* gene.

While the finding of the first genetic associations with MDD through reducing heterogeneity in the disease phenotype studied is indeed fruitful and bears the potential of many future investigations on mechanisms of MDD, it can be argued that the results may be specific to the group of people recruited. Could the GWAS associations be specific to melancholic MDD in Chinese women? I have no answer to that question, but think it is unlikely. While the replication cohort was from North China, the samples selected for the replication were all Han Chinese, and cases were defined as those who had severe, recurring MDD (hence very similar to CONVERGE), it contains both men and women. This shows the *SIRT1* and *LHPP* effects may not be restricted to women. Both *SIRT1* and *LHPP* had also been implicated in MDD by previous studies conducted in populations different from that in CONVERGE.

Moreover, genetic effects specific to one cohort still point to the involvement of biological pathways in MDD, the other elements in which can be implicated in other cohorts. Genetic signals at *SIRT1* and *SLC25A37*, as well as enrichment of rare deleterious variants in MDD cases in genes whose products are localized to the mitochondria, point to the involvement of mitochondria in the etiology of MDD. As shown in Chapter 4, changes in mtDNA in MDD were found to be consistent across sexes, ethnic groups, tissue types, and even species. This speaks to the utility of GWAS results found even in the most specific cohorts – while the association signal at specific variants and genes may not replicate exactly in all other cohorts, they can potentially shed light on biological pathways that are more universal in the etiology of the disease.

6.2.1 Use of NGS data

The availability of NGS data has been central to the analyses and many findings in CONVERGE, and its utility in this study and potentially many future ones has been brought up many times in all chapters of this thesis. My thesis presents but a partial utilization of all the information present in whole-genome NGS data, and more innovative ways of making use of this information can make the same data go a long way further in providing biological insights.

The first and most obvious advantage of using NGS data in genetic studies is its ability to capture all genetic variants present in the study cohort, making any association testing direct and identification of causal variants possible. Even with low-coverage sequencing, which precludes identification of all variants with certainty and necessitates the use of imputation methods to generating complete set of genotypes at all identified variant sites in the cohort, could provide a much denser set of variants for GWAS than a genotyping array. This increases association-mapping resolution, and increases the proportion of genetic contribution to phenotype of interest that can be captured. As

shown in Chapter 3, assessing h2 of MDD using 6 million SNPs across the genome obtained from NGS and appropriate methods that take into account local density of markers have bridged the "missing heritability" of MDD that has perplexed many researchers for a long time. This is a reassuring result, as it shows previously "missing" part of the heritability as found through twin studies was in a large part not due to errors in the assumption that genetic variants contribute to MDD mainly through sum of their independent effects, as can be captured in the SNP-based h2 estimates, rather than interactions between them and with the environment, but both inadequate representation of genetic variation compounded with inadequacies of previous methods. It also suggests that as use of variant information directly identified from NGS becomes more established and prevalent, it will be possible to obtain more accurate estimates of genetic contribution to traits or diseases of interest.

Second, NGS enables the identification of rare, private variants in individual samples that can be used for association with disease status. Chapter 3 shows the calling and validation of private variants in cases of MDD from NGS of 1x coverage in CONVERGE and the enrichment of private deleterious variants in MDD cases, especially in particular functional gene sets. As the average coverage of NGS in CONVERGE is low, very few private variants could be identified with confidence. While high coverage whole-genome NGS on a large sample remains very expensive and therefore may not be feasibly in general, it is possible to narrow down candidate gene sets or regions of the genome for targeted sequencing. As the enrichment of private deleterious variants in MDD cases in particular functional gene sets has shown, rare variants contribute to common diseases like MDD, and may reside in genes of the same biological pathways as implicated by common variant associations.

Third, whole genome NGS in CONVERGE has allows the examination of parts of the genome neglected in conventional genetic analysis, such as the mtDNA and telomeric DNA. NGS contains a lot more information than nucleotide sequences – read

depth, number of mismatches in a read, read mapping quality, singular or multiple mapping alignments, insert sizes between paired-end reads and direction of reads in a pair can all be used to identify artefacts and biological phenomena.

The quantification of the amount of mtDNA and telomeric DNA in Chapter 4 made use of the relative read depth over both regions and that over the rest of the genome; mismatches in a read in the mtDNA was used to identify contaminant DNA among sample DNA as shown in Chapter 2; in Chapter 4, unique mapping to the mtDNA reference was used as a criteria to filter out reads potentially originating from nuclear copies of mtDNA (NUMTs) from those originating from mtDNA, and for identifying regions of the mtDNA where reliable quantification of read depth could be carried out; also in Chapter 4, insert sizes between reads in the VDJ region of the immunoglobulin loci and T cell receptors A and B were used to identify presence of such recombination in the sample's DNA and quantify the proportion of mature B and T cell in the original saliva sample, forensically, when all the cells have been lysed. These are but examples of what can be done using NGS data – other possibilities include looking at copy number variants on a whole-genome scale using read depth information, structural variants using direction of reads in a read pair and insert sizes, perhaps motifs for "fragile" points in the DNA where fragmentation are more prone to occur in cell death, or if the coverage is high enough, *de novo* mutations may be identified, just like heteroplasmic mutations can be identified in mtDNA. While these should not impede or direct efforts away from developing better methods to analyse conventional SNP data, they may be harnessed in useful ways to complement results from SNP analyses. In some cases, such as mtDNA elevation in MDD discussed in this thesis, they can reveal previously unknown phenomenon.

6.3 Biology of MDD and stress

Findings in described in my thesis did not fall into the existing theories of biology of MDD, but introduced a new outlook on MDD etiology. In this section I discuss the new insights findings in this study provide to the field of MDD research.

6.3.1 SIRT1 contribution to MDD

Identification of the association between a locus near *SIRT1* and MDD opens up new possibilities for research into the etiology of MDD. While the definitive experiments to prove the involvement of SIRT1 need to be carried out, there are tantalizing hints that this is indeed the correct gene. There is evidence from many studies, though they do not all point in the same direction.

One extensive study used mice in which the catalytic domain of SIRT1 (exon 4) was deleted in the nervous system [242]. This paper demonstrated that disruption of SIRT1 function in the brain reduces anxiety and depression-related behavior. The authors suggested that *SIRT1* affects behavior by altering serotonin levels in the brain. Fluoxetine, a serotonin reuptake inhibitor commonly prescribed to MDD patients, reversed the effects of the knockout, and it was found that expression of monoamine oxidase A (*MAO-A*), which encodes an enzyme that breaks down and therefore regulates the level of serotonin in the brain, was decreased in SIRT1 knockouts[243]. This seems to suggest increased *SIRT1* expression causes more anxiety and depression.

Another study found that SIRT1 could exert neuro-protective effects through directly deacetylating transducers of regulated CREB activity (TORC1) and promoting its interaction with transcription factor cAMP response element-binding protein (CREB) [244]. This enhances the CREB-mediated transcription at the BDNF promoter IV, which accounts for most of the BDNF gene expression under conditions of neuronal activity [245].

As BDNF is needed for maintaining hippocampal neurogenesis, and impairment of neuronal growth through the BDNF pathway by chronic increase in stress steroid levels had been long postulated as a hypothesis for the etiology of MDD, increase in BDNF activity mediated by SIRT1 activity seem to suggest increased *SIRT1* expression promotes neuroprotection and neurogenesis, therefore offering protection against MDD.

SIRT1 may also be involved in MDD through a more general and widespread mechanism that involves the regulation of mitochondrial biogenesis, likely in a tissue dependent manner [154]. SIRT1 activation by NAD^+ triggers nuclear transcriptional programs that enhance metabolic efficiency and upregulate mitochondrial oxidative metabolism and the accompanying resistance to oxidative stress[169]. NAD+ is an oxidative agent in anaerobic glycolysis, a reducing agent in aerobic respiratory pathways, and lies at the centre of many other biochemical reactions vital for cellular maintenance[246]. SIRT1 functions as both an energy sensor in the cell and a switchboard for control of metabolic strategy.

Cytosolic NAD^+ levels are increased during fasting[247], which will indicate a lack of respiratory substrate for energy production, and cause greater activation of SIRT1. One consequence is mitochondrial biogenesis, through activation of peroxisome proliferator-activated receptor γ (PPARγ) coactivator 1α(PGC-1α). Genetic variation in SIRT1 causing differences in metabolic relation among people could lead to dysregulation of these activities and cause vegetative symptoms such as loss of sleep and appetite in MDD on one end of that spectrum of variations. As will be detailed in Chapter 4, mtDNA levels in cases of MDD were found to be higher than that in controls in both saliva and blood, and mouse put under chronic stress showed tissue specific changes in mtDNA. Given that in animals, behaviour is central to adapting to environmental resources, the idea that *SIRT1* might also impact neural function and in turn behaviour is very attractive.

6.3.2 Relationship between MDD, stress and mtDNA

The finding that there mtDNA increases are contingent on MDD, even though they can be directly induced by chronic stress, suggest two intersecting pathways: one leading to molecular changes, and one to illness.

For the first trajectory, leading from adversity to molecular changes, one possible pathway is through the endocrine system, particularly the activation of the hypothalamic pituitary axis, since changes in both can be reproduced by administration of corticosterone. Release of glucocorticoids is known to increase in response to stress. Severe stressors, such as childhood sexual abuse [248,249], alter pituitary-adrenal and autonomic reactivity. In some circumstances the consequences may be deleterious rather than adaptive: glucocorticoids have been implicated in the pathophysiology of post traumatic stress disorder [250], and it has been known for many years that some patients with MDD exhibit hypersecretion of cortisol [251,252], in part due to corticotrophin releasing factor (CRF) hypersecretion [250]. At least partial involvement of the HPA axis in the molecular changes has been confirmed by mtDNA increases in mice injected with glucocorticoid for four weeks as described in Chapter 4.

For the second trajectory leading to illness, these results could lead to the hypothesis that while adversity may on its own have an effect on levels of mtDNA, the extent and persistence of these molecular changes depend on an individual's susceptibility to MDD, either from genetic or additional environmental predisposing factors. In many individuals the molecular signatures will be small and transitory, but in those with MDD the effects might be larger or last for a longer period of time, causing more physiological changes than smaller, transitory molecular changes. Subjects who have never been depressed, yet suffered severe adversity, may have had detectable alterations in mtDNA levels in particular tissues at the time they experienced stressful

life events, but these changes would have reversed, and no longer be detectable, by the time they were studied. In contrast, recurrent MDD may prevent the reversal of mtDNA levels to normal levels, even though the subjects might not have been experiencing an episode when they were recruited to the study.

The disease-state dependence of the measures is important when considering the potential use of the changes as biomarkers. It is noteworthy in this regard that when DNA was taken from currently severely ill cases of MDD (29 cases from Northern China), a robust difference in the amount of mtDNA could be detected between cases and control samples (25 controls). This suggests that there may be circumstances where the amount of mtDNA could serve as a useful biomarker. The relatively larger increases seen in the mouse experiment (up to 4 fold), suggest that controlling for inter-individual variation would improve the chances of the biomarkers having a clinical application. One potential application would be to use mtDNA levels or mean telomere length from the same patient before and after treatment, assessed by qPCR or other molecular methods, to access recovery from an episode of depression.

Experiments on lab mice were performed to functionally characterize the molecular changes using mouse models. These experiments asked what the molecular changes mean to cellular physiology, and whether such changes are wide-spread , or specific to one tissue or to some with particular metabolic function.

The answers to the first question suggest changes in amount of mtDNA, as demonstrated in chronically stressed mice, could reflect altered metabolic strategies in times of stress. The mtDNA encodes 37 genes (13 protein components of the electron transport chain, 2 mitochondrial ribosomal RNAs, and 22 tRNAs specific for mitochondrial translation), all of which serve the oxidative phosphorylation pathway in concert with proteins and RNAs encoded by the nuclear genome and imported into the mitochondria. Experiments assessing OXPHOS efficiency in mice showed a decrease in

OXPHOS energy production in stressed mice with elevated mtDNA levels. This may indicate increased rates of energy consumption through a switch to glycolysis, a faster though less carbon efficient method of generating energy. More detailed characterization of the molecular changes in mitochondria during stress needs to be performed before there can be an answer. Either way, the consequent decrease in OXPHOS efficiency and output is reminiscent of the well-known lack of motivation and changes in appetite and sleep occurring during the state of depression, and could partly explain the vegetative symptoms of MDD.

The answers to the second question compel further investigations with wider range and better resolution over tissue types. The molecular analyses described in this thesis were restricted to peripheral tissues in humans (blood and saliva) and mice (blood, saliva and liver). The different effects of stress on mtDNA levels and mean telomere length in different tissues in stressed mice (increase in liver, blood and saliva, decrease in muscles, and little change in the hippocampal brain tissue) suggest different, or possibly sequential, pathways governing tissue-specific changes. Interestingly, many studies on the effect of *SIRT1* gene on mitochondrial biogenesis pointed not only to tissue specific effects of *SIRT1* regulation of mitochondrial biogenesis, but signaling between tissues such that *SIRT1* activities in one is sufficient to induce expression of genes involved in mitochondrial biogenesis of another [253]. Whether the signal is transduced through expression of genes whose products are intermediates in the biological pathway, or genes directly involved in mitochondrial biogenesis, or a "mitokine" originating from the affected mitochondria themselves [254,255], there was evidence the starting points of these signaling pathways lie in the central nervous system.

6.3.3 mtDNA heteroplasmy in stress and MDD

The final finding of my thesis is that mtDNA heteroplasmy increases with mtDNA levels during stress and in MDD. Is this change a sign of dysfunction, or might it be an adaptive stress response. The dysfunction hypothesis is very well substantiated by a large body of work on mitochondrial heteroplasmy from model organisms. . Selection for mtDNA containing particular variants has been shown to occur in animals and human alike, and is demonstrably tissue dependent. Artificial introduction of two different mtDNA molecules into mice (cybrid mice) showed that this level of mitochondrial sequence diversity (heteroplasmy) is genetically unstable [256], and tissues systematically eliminate one mtDNA variant in preference over another [257]. During germline transmission mutant molecules are selectively removed, a process which occurs at the level of the organelle [258], possibly through autophagy (whereby cells degrade dysfunctional mitochondria containing mutant molecules) or through inefficient replication of mtDNA in the most dysfunctional units of mitochondria. It is not known whether similar processes occur within somatic cells.

It has been suggested that this phenomenon, especially when observed in the coding regions of the mtDNA, could be partly due to incompatibilities with nuclear genome encoded counterparts for respiratory complexes, where mutant mtDNA encoded subunits may have altered affinities to either the nuclear subunits, respiratory substrates, or complementary factors, therefore causing changes in the ability of respiratory complexes in the production of ATP as well as their generation of reactive oxygen species (ROS). Studies using mouse cybrid cell lines (generated by transferring mitochondria from different sources to mtDNA-less receptor cell lines) showed that while mtDNA variants do not cause differences in the coupling between electron transport activity and ATP production, they do cause significant differences in the production of reactive oxygen species (ROS). Observed ROS differences in cybrid cell

lines may be an underestimation of true differences between ROS production because the increase in measured H2O2 was accompanied by increase catalase activities[259].

The adaptive ROS hypothesis is more radical, though not completely shunned by existing research. Instances of environmentally induced mutations have been documented in bacteria[241] and a similar process may affect the mitochondrial genome. Evidence for this explanation would be the association between mtDNA levels and a heteroplasmic variant lying in a regulatory region of the mitochondrial genome, suggesting that the allele alters replication rate. If this variant, and other heteroplasmic variants like it, causes preferential replication of mtDNA molecules carrying it under only under particular environmental cues such as stress, then it would explain why it might be associated with mtDNA levels in a cohort of MDD cases and controls where mtDNA levels are elevated in cases.

How might this be adaptive? If increased mtDNA heteroplasmy indeed acts through increasing the production of ROS, the adaptive function may be mediated by ROS signaling. ROS is generally known to have toxic effect on the cell, but has been proposed to have regulatory roles in the control of cell functions and particularly in mitochondrial biogenesis[260]. If mechanisms exist to regulate the heteroplasmic load in the mtDNA in an environment dependent manner, either in general or at particular points in the mtDNA (which this study had not yet been able to resolve), one possible function of this regulation would be the control of ROS production and its downstream signaling effects in, non-exhaustively, the control of respiratory substrate utilization and metabolic output. If one was able to monitor the amount of ROS and other metabolites produced by various tissues during chronic stress while quantifying both site specific and general load of heteroplasmy in the mtDNA involved, then tease out the relationship between them, one would likely get closer to understanding what happens during stress.

Taken together, all the findings in this, and the previous chapter, raise questions about how cells manage the turnover of mtDNA sequence – were the increases in amount of mtDNA due to inefficiencies in getting rid of the mutant mtDNA, or a preferential accumulation of mtDNA containing adaptive alleles that may have arisen to meet the demands of a stressful environment, perhaps through causing switches in metabolic strategies? Reminding ourselves of the mouse cybrid experiments suggesting heteroplasmy is "bad" - the increase in heteroplasmy we observe in cases of MDD, likely due to stress (as the mouse experiment implies) might appear detrimental to cell survival. However, it has also been shown that mtDNA mutations need not damage an organism: indeed mutations in genes with mitochondrial function can increase lifespan[240]. Intriguingly, over-expression of TFAM also leads to increased cell survival, with improvements at an organ and system level[241,261,262]. These effects are due in part, if not entirely, to increases in the number of mitochondria[263], suggesting concerted action between alterations in copy number and sequence to protect cells, and organisms, exposed to stress.

6.4 Final thoughts

My thesis revolves around the genetics of MDD, yet it is entitled "Molecular markers of stress". This is because all results presented in this thesis can be interpreted as pointing to molecular changes involving the mitochondria. The multiplicity of mitochondrial functions among different cell types and the universality of its importance, together with the finding that its changes during MDD and chronic stress are widespread, yet specific among tissues, demonstrates that something as elusive as the depressed psychological state may have its physical basis in the most fundamental processes in cellular physiology.

7. Materials and methods

Sample collection

CONVERGE collected cases of recurrent major depression from 58 provincial mental health centres and psychiatric departments of general medical hospitals in 45 cities and 23 provinces of China. Controls were recruited from patients undergoing minor surgical procedures at general hospitals (37%) or from local community centres (63%). A sample size of 6,000 cases and 6,000 controls was chosen on the basis of evidence available when the study was designed (in 2007) of the likely existence of genetic loci with odds ratio of 1.2 and above. All subjects were Han Chinese women with four Han Chinese grandparents. Cases were excluded if they had a pre-existing history of bipolar disorder, psychosis or mental retardation. Cases were aged between 30 and 60 and had two or more episodes of MDD meeting DSM-IV criteria[264] with the first episode occurring between 14 and 50 years of age, and had not abused drugs or alcohol before their first depressive episode. All subjects were interviewed using a computerized assessment system. Interviewers were postgraduate medical students, junior psychiatrists or senior nurses, trained by the CONVERGE team for a minimum of one week. The diagnosis of MDD was established with the Composite International Diagnostic Interview (CIDI) (WHO lifetime version 2.1; Chinese version), which utilized DSM-IV criteria. The interview was originally translated into Mandarin by a team of psychiatrists in Shanghai Mental Health Centre, with the translation reviewed and modified by members of the CONVERGE team. The study protocol was approved centrally by the Ethical Review Board of Oxford University (Oxford Tropical Research Ethics Committee) and the ethics committees of all participating hospitals in China. All interviewers were mental health professionals who are well able to judge decisional capacity. The study posed minimal risk (an interview and saliva sample). All participants provided their written informed consent.

DNA sequencing

DNA was extracted from saliva samples using the Oragene protocol. A barcoded library was constructed for each sample. Sequencing reads obtained from Illumina Hiseq machines were aligned to Genome Reference Consortium Human Build 37 patch release 5 (GRCh37.p5) with Stampy (v1.0.17) [265] using default parameters after filtering out reads containing adaptor sequencing or consisting of more than 50% poor quality (base quality <= 5) bases. Samtools (v0.1.18) [185] was used to index the alignments in BAM format[185] and Picardtools (v1.62, http://picard.sourceforge.net) was used to mark PCR duplicates for downstream filtering. The Genome Analysis Toolkit's (GATK, version 2.6) [116-118] BaseRecalibrator was then run on the BAM files to create base quality score recalibration tables, masking known SNPs and INDELs from dbSNP (version 137, excluding all sites added after version 129). Base quality recalibration (BQSR) was then performed on the BAM files using GATKlite (v2.2.15) while also removing read pairs that did not have the "properly aligned segment" bit set by Stampy (1-5% of reads per sample).

Quality control of sequencing data

Raw sequencing reads were filtered for reads that contain more than 50% bases with Phred-scaled base quality <= 5. The filtered reads from each sequencing lane were mapped to the Genome Reference Consortium Human Build 37 patch release 5 (GRCh37.p5) with Stampy (v1.0.17) with default settings and --solexa for selecting Solexa read qualities and --bwaoptions=–q10 for quality threshold for trimming reads down to 35bp, run with bwa (v0.5.9-r16) [266]. Mapped sequencing reads from the same sample with the same read group were then merged, coordinate-sorted and indexed with SAMtools (v0.1.18) into one BAM file per sample. Read group names were then

changed with SAMtools to a convention that reflects both sequencing batch and sample ID. MarkDuplicates in Picardtools (v1.62) was then run to mark PCR duplicates for downstream filtering. Base quality score recalibration (BQSR) was applied to the mapped sequencing reads using BaseRecalibrator in Genome Analysis Toolkit (GATK, basic version 2.6) with the known insertion and deletion (INDEL) variations 1000 Genomes Projects Phase 1 and known single nucleotide polymorphisms (SNPs) from dbSNP (v137, excluding all sites added after v129) excluded from the empirical error rate calculation. GATKlite (v2.2.15) [118] was then used to write out sequencing reads with the recalibrated base quality scores while removing reads that did not have the "proper pair" flag bit (meaning both reads in the mate-pair are correctly oriented, and their separation is within 5 standard deviations from the mean insert size between mate-pairs) set by Stampy (1-5% of reads per sample) using the --read_filter ProperPair option.

Variant calling from low-coverage sequencing on all 11,670 samples

Variant calling (for both SNPs and INDELs) was performed on the CONVERGE dataset using sequencing reads from all samples with UnifiedGenotyper in GATK (v2.7-2-g6bda569) using option --genotype_likelihood_model BOTH and default annotation outputs for variant calls. dbSNP v137 rsids were used to fill in the variant ID column of the result variant call format (VCF) files using the --dbSNP option. Variant quality score recalibration (VQSR) was performed with GATK's VariantRecalibrator (v2.7-4-g6f46d11) in SNP variant calls using the SNPs in 1000 Genomes Phase 1 ASN Panel[91] as the known, truth and training sets with a prior of 15.0, and the following default annotations: -an QD -an HaplotypeScore -an MQRankSum -an ReadPosRankSum -an FS -an MQonly. By modeling the distribution of these annotations at all called biallelic SNP sites relative to that at just the sites present in the training set of known biallelic SNPs

from 1000 Genomes Phase 1 ASN Panel using a Gaussian mixed model, the raw call set was clustered and each site assigned VQSLOD scores based on their distance to the centre of the cluster. As such, outliers would be assigned low scores and excluded from further analysis as potential variant calling errors. VQSR was only performed on raw SNP calls because they could have different sequencing and variant discovery error distributions for the default GATK variant annotations from INDELs. Sets or "tranches" of SNPs with increasing VQSLOD scores (increasing stringency in exclusion of outliers and concomitant decreasing sensitivity to known SNPs) were generated at the specified sensitivities to known SNPs using the –tranche option: -tranche 100.0 -tranche 99.9 -tranche 99.0 -tranche 98.0 -tranche 95.0 -tranche 90.0 -tranche 89.0 -tranche 87.0 -tranche 85.0 -tranche 80.0 -tranche 79.0. Transition-to-transversion (TiTv) ratios were calculated for each tranche. A sensitivity to known SNPs of 90.0% was chosen for the inclusion of novel SNPs in downstream analyses as it balances optimization of TiTv ratio in novel SNPs with retaining as many potential novel SNPs as possible so as not to lose variant information. All known sites identified as polymorphic in CONVERGE were included in all further analyses regardless of their VQSLOD scores.

Variant calling on 10x coverage sequencing on nine samples

Variant calling (for both SNPs and INDELs) was performed on the high coverage dataset using sequencing reads from all nine samples with UnifiedGenotyper in GATK (v2.7-2-g6bda569) using option --genotype_likelihood_model BOTH and default annotation outputs for variant calls. dbSNP v137 rsids are used to fill in the variant ID column of the result variant call format (VCF) files using the --dbSNP option. Variant quality score recalibration (VQSR) was performed using two different training sets. The first VQSR was performed using SNP variant calls using the SNPs in 1000 Genomes Phase 1 ASN Panel as the known, truth and training sets with a prior of 15.0 and the default

annotations just like it was done for the raw variant calls from the low coverage data, but with two maximum Gaussians for building the positive model instead of one, with the --maxGaussian option. The second VQSR was performed using SNPs from the HumanOmniZhongHua-8 (v1.0) BeadChip. Sets or "tranches" of SNPs with increasing VQSLOD scores (increasing stringency in exclusion of outliers and concomitant decreasing sensitivity to known SNPs) were generated at the specified sensitivities to known SNPs using the –tranche option: -tranche 100.0 -tranche 99.9 -tranche 99.0 -tranche 98.0 -tranche 95.0 -tranche 90.0 -tranche 89.0 -tranche 87.0 -tranche 85.0 -tranche 80.0 -tranche 79.0. Transition to Transversion (TiTv) Ratio was calculated for each tranche. A sensitivity to known SNPs in the Zhonghua chip of 90.0% was chosen for the inclusion of novel SNPs in downstream analyses as it balances optimization of TiTv ratio in novel SNPs while retaining as many potential novel SNPs as possible to not lose variant information.

Sample filtering for GWAS

A range of quality control filters were imposed on all 11,670 samples and only those samples with high-quality sequencing results and imputed genotype probabilities were selected for downstream analysis. I looked into both the nuclear genome and mitochondrial genome for an excess of variants called, since this would indicate cross-sample contamination due to technical issues during sequencing. I quantified the number of singleton variants called in the genic regions of the nuclear genome and found a mean of 71.55 singletons variants per sample supported by more than 2 sequencing reads passing sequencing quality controls. I excluded 117 samples with a number of singletons greater than the 99th percentile. I asked if there were any samples with an excess of mtDNA heteroplasmies per sample. Coverage of the mitochondrial genome is on average 102X, allowing us to obtain high quality sequence for this part of

the genome. I found a mean of 15.70 heteroplasmic sites per sample, and 116 samples were found to have greater than the 99th percentile of number of heteroplasmic sites. Of these 116 samples, 26 were already discarded for having excess nuclear genome singletons; I excluded the additional 90. I checked imputation quality based on the certainty of genotypes imputed (maximum genotype probability > 0.9) and identified 29 individuals who had fewer than 90% of their sites with maximum genotype probabilities > 0.9. I excluded these samples from further analysis. I checked the 11,434 remaining samples for relatedness. Although being unrelated to other individuals recruited for the CONVERGE study was a clear criteria in our data collection process, there were instances when the same patient or a relative of the patient visited multiple hospitals and was thus recruited more than once. To exclude duplicates and first degree relatives from our sample for GWAS, I estimated pairwise genome-wide Identity by Descent (IBD) using Identity by State (IBS) information in hard-called genotypes from imputed genotype probabilities at 399,211 common, tagging SNPs across all autosomes (MAF >1%, LD <0.5, all known in 1000 Genomes Phase 1). I implemented this in PLINK (v1.07)[123] with the option --genome. I excluded a total of 392 samples (duplicates and first degree-relatives) from our final set of samples for GWAS. I retained second-degree relatives and beyond, correcting for the relatedness between them using a linear-mixed model. I also excluded 402 samples with incomplete phenotype information, giving a final set of 10,640 samples (5,303 cases of MDD, 5,337 controls) for GWAS on MDD.

GWAS on MDD using linear mixed model and liability score estimates

I performed GWAS one chromosome at a time, with a mixed-linear model including a genetic relationship matrix (GRM) constructed from SNPs in the tagging set other than those on the chromosome being studied. This method is implemented in Factored Spectrally Transformed Linear Mixed Models[146] (FastLMM version 2.06.20130802).

Manhattan plots and quantile-quantile plots of the log10 of P-values of the GWAS were generated with custom code in R. Genomic-control inflation factor lambda was calculated using custom code in R.

Replication and joint analyses

I genotyped the replication sample on a MassARRAY™ system mass spectrometer. TYPER4.0 was used to assess the reliability of genotype calls generated by SpectroREAD from the mass spectra. Default genotype call inclusion criteria were used. To perform the association analysis with MDD case-control status at these 12 sites in the replication sample, I obtained effect sizes for discovery from logistic regression with PC correction, and then for replication from logistic regression, and then performed fixed-effects meta-analysis.

Estimation of whole-genome and single gene SNP-based heritability for MDD

For the unweighted estimation of h2, I used GCTA[134] (version 1.24.4) to construct a genetic relatedness matrix (GRM) using all SNPs from the autosomes, and another GRM using all SNPs from the X chromosome, before combining them into a GRM with contributions from all 6,242,619 SNPs used in GWAS. I estimated h2 of MDD using restricted maximum likelihood (REML). For the partitioned analysis with different chromosomes, different LD bins, and differernt LD and MAF bins, I used the GREML framework [267,268] where many GRMs are fitted simultaneously. To generated the different LD bins, I calculated "LD scores" which are sums of LD r2 values of each SNP with all other SNPs in a 100kb window using LDSC[162], then ranked them and took quartiles. For the weighted estimation of h2, LDAK (version 4.9) was used to calculation local pairwise correlations between SNPs and generating weightings of each SNP in the calculation GRM adjusted for local LD between all samples. A GRM was generated using

all 6,242,619 SNPs from all chromosomes, on which restricted maximum likelihood (REML) is used for estimating the proportions of variance explained by all SNPs in the GRMs. For all analyses prevalence of MDD was set at 8% to correct for the ascertainment bias in the CONVERGE cohort, and PCs 1 to 5 from eigen decomposition of the GRMs used in respective analyses were used as covariates.

Estimation of gene-based heritability for MDD

Gene-based heritability was obtained for all genes using LDAK (version 4.9}. Gene boundaries were obtained from RefSeq for human reference genome Build 37, hg19 (Feb 2009). Genes with significant heritabilities were identified as those with post-permutation Likelihood Ratio Test (LRT) P value $<5x10^{-6}$.

Calling rare exonic variants

All sequenced reads mapping to exonic regions of the human genome reference GRCh37.p5 and passing a series of quality control were interrogated using SAMtools mpileup. Read-pairs with both reads uniquely mapped, with mapping quality of or above 55, were included in this analysis. The exon coordinates were downloaded from the Ensembl release 67 *Homo sapiens* gtf file (ftp.ensembl.org/pub/release-67/gtf/homo sapiens). The coordinates contained 96,130,824 base pair positions in 254,986 exons in 21,946 genes. Singletons (both SNPs and INDELs) were called when two reads supported the same alternative in a single sample. I chose a set of genes expressed in the brain (eGenetics/SANBI EST) [269] downloaded from biomart[270] for Ensembl release 75 February 2014. The expression data comes from ESTs in dbEST which were annotated with eVOC terms by SANBI[271] and sequence mapped to Ensembl gene predictions. A list of nuclear encoded mitochondrial genes was obtained from SANBI, excluding three genes that overlapped with nuclear mitochondrial sequences

from the hg19 NUMT track from UCSC. Gene sizes were calculated using data from Ensembl release 77 October 2014 downloaded from ftp://ftp.ensembl.org. I calculated true positive rates for the called singletons by comparing with genotypes called in nine individuals for whom I also had 10X coverage. I assumed the singletons found in the same sample in both the low and 10X coverage data were true positives (Extended Data Table 3). I successfully amplified DNA at 64 deleterious variant sites, including 12 insertions. I obtained Sanger sequencing of each amplicon to validate the next generation sequencing calls.

Gene annotation and enrichment analysis

All exonic SNPs were annotated using ANNOVAR[165]. Variants of each annotation category and in each exon were aggregated for every individual and used in logistic regression as predictors of disease status, controlling for measures associated with sequencing runs, including sequencing batch, read mapping quality, sequence coverage over the genome, and GC content. In addition I included principal components from the common variant analysis, and city of origin of the individual. To confirm that P-values in the rare exonic variant analyses were not inflated, I permuted the case/control status within each annotation category and calculated the contribution to the risk of MDD of all variants, non-synonymous, synonymous and deleterious variants (defined as the sum of stop gain, stop loss, nonsynonymous, and frameshift deletion variants). The procedure was repeated 10,000 times to generate a distribution of P-values. To obtain P-values of the increased odds-ratios seen when enrichment was sought only in a subset of genes, an empirical distribution was generated by randomly selecting segments of coding DNA equal in length to that in the gene set, and testing for enrichment (but not permuting case/control status). This procedure was repeated 10,000 times and the rank of the odds ratios seen in the real data set obtained from the empirical distributions.

The GENDEP and DeCC studies, samples and Quantitative PCR

Cases and control samples were drawn from the United Kingdom Depression Case–Control (DeCC) study[272], and the Genome-Based Therapeutic Drugs for Depression (GENDEP) study[273]). mtDNA copy number was estimated from DNA extracted from blood samples by quantitative PCR, using a TaqMan® Universal PCR MasterMix on an ABI StepOnePlus Real-Time PCR System (Life Technologies, Inc.). The pre-designed TaqMan® assay Hs02596867_s1 was used to amplify a fragment of the MT-CYB gene on the mitochondrial chromosome in duplex with the TaqMan® RNaseP Copy Number Reference Assay (Life Technologies, Part Number 4403326) as an internal control.

Extracting and quality control of mitochondrial reads from low coverage whole-genome sequencing data

All reads mapped to the human mitochondrial genome NC_012920.1 are extracted from the whole genome BAM files mapped to GRCh37.p5 using Samtools (v0.1.18). The mitochondrial reads extracted are then converted to the FASTQ format using Picard (v1.108) and mapped to a combined reference containing 894 complete bacterial genomes, 2024 complete bacterial chromosomes, 154 draft assemblies and 4373 complete plasmids sequences (in total 7390 unique bacterial DNA sequences) available on NCBI using BWA (v0.5.6). All reads mapped to bacterial DNA sequences can be filtered out using using Samtools (v0.1.18) by imposing a mapping quality filter of 59 (Phred-scale probability of being wrongly mapped) and removing reads with FLAGs (-F 1804) that identify unmapped reads, unpaired reads, reads that do not pass quality control, reads that may be PCR or optical duplicates, and reads that are secondary alignments that also map to other areas of the reference. No reads from filtered BAM files of any sample map onto the combined bacterial reference.

Estimation of mtDNA copy number

Average read depth per 100 base pairs is calculated for the mtDNA reads mapped to NC_012920.1 both before and after filtering out poorly mapped reads including those potentially from bacterial genomes using SAMTOOLs (v0.1.18). As there are regions in the mitochondrial genome replicated in the nuclear genome commonly known as NUMTs (nuclear copies of mitochondrial DNA), which are most likely present as secondary alignments, I calculated average read depth per segment of 100 base pairs in the mtDNA alignments both before and after filtering and compared the average read depth per 100 base pairs before and after filtering. To reduce errors in estimation of coverage due to NUMTs, segments with big differences in read depth (>5% of the filtered read depth) between the filtered and unfiltered alignments that are more likely to span NUMTs are excluded from our calculation of mtDNA copy number. I arrived at a measure of mtDNA copy number by taking the mean read depth in the filtered alignments across all remaining 100 base pairs segments, then regressing it with sequencing batch, sample age and average filtered read depth per site of the whole chromosome 20, and transforming the residuals to normality using a quantile normal function.

Estimation of mean telomere length

Mean telomere length is quantified from low-coverage whole genome sequencing data mapped to Genome Reference Consortium Human Build 37 patch release 5 (GRCh37.p5) with the Telseq v0.0.1[184]. The estimated mean telomere length output from Telseq is already corrected for whole genome coverage and GC content by the programme; it is then corrected for batch and sample age before transforming the residuals to normality using a quantile normal function.

Animal experiments

All experiments were carried out in strict accordance with the recommendations in the Guide for Laboratory Animals Facilities and Care as promulgated by the Council of Agriculture, Executive Yuan, ROC, Taiwan. The protocol was approved by the Institutional Animal Care and Use Committee of Chang Gung University (Permit Number: CGU13-067). Mouse stress experiment: mice (strain C57BL6/J, female n = 8, male n = 8, aged 12 months) were stressed over five days followed by two days rest, repeated for 4 weeks. On the first day animals were suspended from their tails for 10 minutes. This was repeated three times, with 5 minutes rest between tail suspensions. On the second day animals were placed in a cylinder of deep water from which there was no escape for 10 minutes. The forced swim was repeated twice with a 10 minute rest. On the third day, a foot shock was administered three times (0.75 mA for 10 second with 10 second rest). On the fourth day animals were restrained in a cylindrical tube (12 cm in length and 3 cm in diameter) for 3hrs. On the fifth day animals were sleep deprived for 24hrs (mice were put in water tank, containing multiple and visible platforms (4.5 cm in height and diameter) surrounded by water for 24 hours). For the glucorticorticoid experiment mice (strain C57BL6/J, female n = 8, aged 12 months) underwent daily subcutaneous injection of 30 mg/kg corticosterone (Sigma) or vehicle (oil) for 28 days.Association between mtDNA, telomere length and stress was performed in a linear mixed model using the lme4 package in the statistical software language R [274]. The null model included only weight. Variation in the amount of mtDNA between different tissues was assessed by a t test, comparing values between controls and experimental animals.

Mouse DNA extraction

DNA was extracted from mouse tissues using a QIAamp DNA Investigator Kit (Qiagen). Saliva was collected from mice by inserting a disposable inoculation loop into the animal's mouth, allowing the animal to chew for a few seconds, before rinsing the loop in QIAamp DNA Investigator Kit ATL buffer. Blood was collected from a superficial tail vein.

Quantification of mtDNA levels by Quantitative PCR (qPCR)

Quantitative PCR was carried out using the SYBR Green kit supplied by Roche Molecular Biochemicals (Pleasanton, CA). A nuclear genomic fragment of 160 bp was amplified from the mouse *Gapdh* gene (forward 5′ TGACGTGCCGCCTGGAGAAAC 3′, reverse 5′ CCGGCATCGAAGGTGGAAGAG 3′. A 117 base pair fragment of the mitochondrial genome (positions 13603-13719) was amplified with primers described [275] (forward 5′ CCCAGCTACTACCATCATTCAAGT 3′, reverse 5′ GATGGTTTGGGAGATTGGTTGATGT 3′). Quantitative PCR was performed under the following conditions: denaturation 95^0C for 10 minutes followed by 50 cycles of 15 secs 95^0C and 1 minute at 60^0C. An estimate of the mtDNA copy number was calculated using the mean of Gapdh as a control[276].

Quantification of telomere length by Monochrome Multiplex qPCR (MMQPCR)

Average telomere length was measured from mouse DNA using a previously described monochrome multiplex qPCR (MMQPCR) method[277] with the following conditions: denaturation at 95^0C for 15 minutes followed by 2 cycles of 15 secs 94^0C and 60 secs 49^0C, 4 cycles of 15 secs at 94^0C and 30 secs at 59^0C, 20 cycles of 15 secs at 85^0C and 30 secs at 59^0C, and 27 cycles of 15 secs at 94^0C, 10 secs at 84^0C and 15 secs at 85^0C. Forward and reverse telomeric primers were 5′ ATACCAAGGTTTGGGTTTGGGTTTGGGTTTGGGTTCATGG 3′ and 5′ GAGGCAATATCCCTATCCCTATCCCTATCCCTATCCCTAACC 3′. Average telomere length

ratio was estimated from the ratio of telomere product to that of a single copy nuclear gene albumin, forward and reverse primers for which were 5′ CGGCGGCGGGCGGCGCGGGCTGGGCGGAAACGCTGCGCAGAATCCTTG 3′ and 5′ GCCCGGCCCGCCGCGCCCGTCCCGCCGCTGAAAAGTACGGTCGCCTG 3′.

Mitochondrial oxygen consumption

I measured oxygen consumption from a liver mitochondrial preparation over time using a Clark electrode. After the addition of respiratory substrates (glutamate and malate) oxygen consumption was monitored for 100s, after which ADP was added, and oxygen consumption measured for a further 100 seconds. Potassium cyanide (KCN) was added 100 seconds later to inhibit all mitochondrial oxygen consumption.

Estimation of whole-genome and single gene SNP-based heritability for amount of mtDNA

LDAK (version 4.9) was used to estimate local linkage disequilibrium (LD) by calculation of local pairwise correlations between SNPs and generating weightings of each SNP in the calculation of a genetic relatedness matrix (GRM) adjusted for local LD between all samples. A GRM was generated using all 6,242,619 SNPs from all chromosomes, on which restricted maximum likelihood (REML) is used for estimating the proportions of variance explained by all SNPs in the GRMs for amount of mtDNA. Gene-based heritability was obtained for all genes using LDAK. Gene boundaries were obtained from RefSeq for human reference genome Build 37, hg19 (Feb 2009).

Nuclear DNA GWAS on mtDNA using linear mixed model

I performed GWAS on mtDNA levels one chromosome at a time, with a mixed-linear model including a genetic relationship matrix constructed from SNPs other than those

on the chromosome being studied. This method is implemented in Factored Spectrally Transformed Linear Mixed Models (FastLMM version 2.06.20130802). Manhattan plots and quantile-quantile plots of the log10 of P-values of the GWAS were generated with custom code in R. Genomic-control inflation factor lambda was calculated in R.

Replication cohort for mtDNA GWAS results

A measure of mtDNA was obtained from the Avon Longitudinal Study of Parents and Children (ALSPAC) [217] (www.alspac.bris.ac.uk) whole genome-sequencing cohort[278], and normalized for the amount of nuclear DNA (using chromosome 20 read depth). Genotype dosages for two SNPs (rs445 and rs1100612) were obtained from ALSPAC, analysed for association with mtDNA by linear regression and the joint sample (CONVERGE and ALSPAC) analysed with a fixed-effects meta-analysis. ALSPAC data are available through a searchable data dictionary http://www.bris.ac.uk/alspac/researchers/dataMaccess/dataMdictionary/. Ethical approval for the study was obtained from the ALSPAC Ethics and Law Committee and the Local Research Ethics Committees.

Amino acid sequence alignment for TFAM gene

Alignment of human UniProt Knowledgebase [218] identification accession number Q00059) TFAM amino acid sequences with several other organisms was performed using MSAProbs [279], minor manual adjustments were based on the results from HHpred, alignments were formatted using ESPript and ENDscript [280]. Approximate boundaries of domains were derived from the crystal structure of human TFAM [219] (Protein Data Bank (PDB) ID 3TMM).

Estimation of genetic correlation between amount of mtDNA and MDD

Genetic correlation between amount of mtDNA and MDD was estimated using bivariate REML in GCTA [90] (version 1.24.4), where the input GRM was the same GRM used for estimating whole-genome SNP-based heritability, generated with LDAK using all 6,242,619 SNPs from all chromosomes. The binary MDD status was converted to liability scale with a population prevalence of disease of 8% for this estimation.

Calling of homoplasmic and heteroplasmic variants in the mtDNA from low-coverage sequencing

I obtained total and per allele read depth for only sequencing reads that map uniquely to the mitochondria using Samtools mpileup (v0.1.18). Filtering criteria also include a base quality threshold of 20 (Phred-scaled) for bases, and mapping quality of 50 (Phred-scaled) and no more than 4 mismatches to the reference for reads. To filter out potential mismappings due to nucleotide homopolymer runs on the mtDNA, I only interrogated 14,796 out of 16,569 sites in the mtDNA sequence, masking out 1,773 sites in the mtDNA for being in or beside homopolymer runs of 4 or more nucleotides. For calling of homoplasmic variants, I first discarded any site in the mtDNA in any sample that was covered by fewer than 10 sequencing reads, and any site that is discarded by this criterion in more than 10% of the samples. A site is considered a homoplasmic variation in the cohort if there were two alleles present in the cohort, and each of them was supported by more than 90% of reads in more than 0.1% of the cohort (10 samples among 10,442). I identified 1,031 homoplasmic sites with the above criteria. For calling of heteroplasmic variants, sequencing coverage at each site in the mtDNA from each sample was independently and randomly downsampled to 50 reads; sites in any sample covered by fewer than 50 reads were discarded from the analysis, and sites with more than 10% of the samples discarded by the above criterion were then completely

disregarded for further analysis due to lack of evidence. A sites was considered potentially heteroplasmic if there was presence of 2 or more alleles each supported by more than or equal to 2 reads (4%, out of 50 reads). I then calculated both a) the degree of heteroplasmy per potential heteroplasmic site in a single individual and b) the frequency of occurrence of heteroplasmy at a each site in the population. I only considered those sites where more than 0.1% (10 samples among 10,442) of the sample show any degree of heteroplasmy.

Association testing with mtDNA variants using linear regression model

Testing of association between mtDNA levels and homoplasmic and heteroplasmic variants in the mtDNA above 0.1% in occurrence in CONVERGE were carried out with a linear regression model using a custom script in R. Homoplasmic mtDNA variation among samples in the were coded as binary measures representing the reference and alternative alleles relative to the human mitochondrial genome reference NC_012920.1. Heteroplasmic mtDNA variation among samples were coded as residuals from a linear regression of number of reads supporting the alternative allele with site-specific read depth, whole-genome sequencing coverage and sequencing batch. Significant thresholds are determined with Bonferroni correction of multiple testing on the standard p-value significant threshold of 0.05.

Long range PCR on mtDNA and high-coverage sequencing of PCR product

For long-range polymerase chain reaction (PCR) on 72 human DNA samples from CONVERGE I designed a pair of primers on D-loop sequences on the mitochondrial DNA that did not show sequence similarity with nuclear DNA (Forward primer: 5'-TGAGGCCAAATATCATTCTGAGGGGC-3'; Reverse primer: 5'-TTTCATCATGCGGAGATGTTGGATGG-3') for the mtDNA. Long range PCR using these

primers covered the whole of the 16,569 bp of the mitochondrial genome. I performed the PCR on 72 samples with thermal cycling conditions as follows: one 1 minute cycle at 98℃, 30 cycles of 10s at 98℃ then 8 minutes 15 seconds at 72℃, one 10 minutes cycle at 72℃, then storing of PCR products at 4℃. For long-range polymerase chain reaction (PCR) on 12 mouse DNA samples I designed a pair of primers on D-loop sequences on the mitochondrial DNA that did not show sequence similarity with nuclear DNA (Forward primer: 5'- CCCAGCTACTACCATCATTCAAGT-3'; Reverse primer: 5'-GAGAGATTTTATGGGTGTAATGCGG-3') for the mtDNA. Long range PCR using these primers covered the whole of the 16,229 bp of the mitochondrial genome. I performed the PCR on 72 samples with thermal cycling conditions as follows: one 1 minute cycle at 98℃, 30 cycles of 10s at 98℃, 30s at 68℃ then 8 minutes 15 seconds at 72℃, one 10 minutes cycle at 72℃, then storing of PCR products at 4℃. PCR mix for both reactions contain 5µl 5xGC buffer, 1µl 10mM dNTPs, 5µl Primer cocktail, 0.5µl 2U/µl High-Fidelity DNA Polymerase, 100ng sample DNA. The PCR products were then sheared to approximately 500bp fragments and used to construct libraries for sequencing using Illumina Hiseq4000, yielding 100bp paired-end reads with average per site read depth of 1757 for human samples and 1368 for mouse samples. Sequencing products from this process contains only mitochondrial DNA.

Calling of heteroplasmic variants in human and mouse mtDNA from high-coverage sequencing of long-range PCR product

Sequencing reads from human samples were mapped to the human mtDNA reference NC_012920.1 using BWA (v0.5.6) and Samtools (v0.1.18), while sequencing reads from mice samples were mapped using the same softwares to mtDNA reference sequence in

mouse reference genome build GRCm38. I obtained total and per allele read depth for only sequencing reads that map to the human and mouse mitochondria using Samtools mpileup (v0.1.18). Filtering criteria also include a base quality threshold of 20 (Phred-scaled) for bases, and mapping quality of 50 (Phred-scaled) and no more than 4 mismatches to the reference for reads. To filter out potential mismappings due to nucleotide homopolymer runs on the mtDNA, I only interrogated 14,796 out of 16,569 sites in the human mtDNA sequence and 14,717 out of 16,229 site in the mouse mtDNA sequence, masking out 1,773 and 1512 sites in the mtDNA for being in or beside homopolymer runs of 4 or more nucleotides. Sequencing coverage at each site is then independently and randomly downsampled to 500 reads; sites in any sample covered by fewer than 500 reads were discarded from the analysis, and sites with more than a third of the human and mice samples discarded, respectively, were then completely disregarded for further analysis due to lack of evidence. A site is considered heteroplasmic in a sample if there was presence of 2 or more alleles each supported by more than or equal to 1% of sequencing reads (5 reads out of 500 reads). I then calculated the degree of heteroplasmy per sample per heteroplasmic site.

Reference:

1. Alonso, J. *et al.* Prevalence of mental disorders in Europe: results from the European Study of the Epidemiology of Mental Disorders (ESEMeD) project. *Acta Psychiatr Scand Suppl*, 21-7 (2004).
2. Kessler, R.C. *et al.* The epidemiology of major depressive disorder: results from the National Comorbidity Survey Replication (NCS-R). *JAMA* **289**, 3095-105 (2003).
3. Liao, S.C. *et al.* Low prevalence of major depressive disorder in Taiwanese adults: possible explanations and implications. *Psychol Med* **42**, 1227-37 (2012).
4. Fava, M. & Kendler, K.S. Major depressive disorder. *Neuron* **28**, 335-41 (2000).
5. Cuijpers, P., van Straten, A., Andersson, G. & van Oppen, P. Psychotherapy for depression in adults: a meta-analysis of comparative outcome studies. *J Consult Clin Psychol* **76**, 909-22 (2008).
6. Cuijpers, P., Andersson, G., Donker, T. & van Straten, A. Psychological treatment of depression: results of a series of meta-analyses. *Nord J Psychiatry* **65**, 354-64 (2011).
7. Weissman, M.M. *et al.* Cross-national epidemiology of major depression and bipolar disorder. *JAMA* **276**, 293-9 (1996).
8. Kendler, K.S., Gardner, C.O., Neale, M.C. & Prescott, C.A. Genetic risk factors for major depression in men and women: similar or different heritabilities and same or partly distinct genes? *Psychol Med* **31**, 605-16 (2001).
9. Kendler, K.S., Gatz, M., Gardner, C.O. & Pedersen, N.L. A Swedish national twin study of lifetime major depression. *Am J Psychiatry* **163**, 109-14 (2006).
10. Kessler, R.C. The effects of stressful life events on depression. *Annu Rev Psychol* **48**, 191-214 (1997).
11. Gladstone, G., Parker, G., Wilhelm, K., Mitchell, P. & Austin, M.P. Characteristics of depressed patients who report childhood sexual abuse. *Am J Psychiatry* **156**, 431-7 (1999).
12. Zlotnick, C., Mattia, J. & Zimmerman, M. Clinical features of survivors of sexual abuse with major depression. *Child Abuse Negl* **25**, 357-67 (2001).
13. Gamble, S.A. *et al.* Childhood sexual abuse and depressive symptom severity: the role of neuroticism. *J Nerv Ment Dis* **194**, 382-5 (2006).
14. Kendler, K.S. *et al.* Childhood sexual abuse and adult psychiatric and substance use disorders in women: an epidemiological and cotwin control analysis. *Arch Gen Psychiatry* **57**, 953-9 (2000).

15. Nelson, E.C. *et al.* Association between self-reported childhood sexual abuse and adverse psychosocial outcomes: results from a twin study. *Arch Gen Psychiatry* **59**, 139-45 (2002).
16. Bennett, A.J. *et al.* Early experience and serotonin transporter gene variation interact to influence primate CNS function. *Mol Psychiatry* **7**, 118-22 (2002).
17. Caspi, A. *et al.* Influence of life stress on depression: moderation by a polymorphism in the 5-HTT gene. *Science* **301**, 386-9 (2003).
18. Hariri, A.R. *et al.* Serotonin transporter genetic variation and the response of the human amygdala. *Science* **297**, 400-3 (2002).
19. Kendler, K.S., Kuhn, J.W., Vittum, J., Prescott, C.A. & Riley, B. The interaction of stressful life events and a serotonin transporter polymorphism in the prediction of episodes of major depression: a replication. *Arch Gen Psychiatry* **62**, 529-35 (2005).
20. Lesch, K.P. *et al.* Association of anxiety-related traits with a polymorphism in the serotonin transporter gene regulatory region. *Science* **274**, 1527-31 (1996).
21. Munafo, M.R., Durrant, C., Lewis, G. & Flint, J. Gene X environment interactions at the serotonin transporter locus. *Biol Psychiatry* **65**, 211-9 (2009).
22. Munafo, M.R. *et al.* 5-HTTLPR genotype and anxiety-related personality traits: a meta-analysis and new data. *Am J Med Genet B Neuropsychiatr Genet* **150B**, 271-81 (2009).
23. Murphy, D.L. *et al.* Genetic perspectives on the serotonin transporter. *Brain Res Bull* **56**, 487-94 (2001).
24. Risch, N. *et al.* Interaction between the serotonin transporter gene (5-HTTLPR), stressful life events, and risk of depression: a meta-analysis. *JAMA* **301**, 2462-71 (2009).
25. Salichon, N. *et al.* Excessive activation of serotonin (5-HT) 1B receptors disrupts the formation of sensory maps in monoamine oxidase a and 5-ht transporter knock-out mice. *J Neurosci* **21**, 884-96 (2001).
26. Young, E.A., Carlson, N.E. & Brown, M.B. Twenty-four-hour ACTH and cortisol pulsatility in depressed women. *Neuropsychopharmacology* **25**, 267-76 (2001).
27. Nemeroff, C.B. *et al.* Elevated concentrations of CSF corticotropin-releasing factor-like immunoreactivity in depressed patients. *Science* **226**, 1342-4 (1984).
28. Raadsheer, F.C., Hoogendijk, W.J., Stam, F.C., Tilders, F.J. & Swaab, D.F. Increased numbers of corticotropin-releasing hormone expressing neurons in the hypothalamic paraventricular nucleus of depressed patients. *Neuroendocrinology* **60**, 436-44 (1994).

29. Nemeroff, C.B. The role of corticotropin-releasing factor in the pathogenesis of major depression. *Pharmacopsychiatry* **21**, 76-82 (1988).
30. Axelson, D.A. *et al.* Hypercortisolemia and hippocampal changes in depression. *Psychiatry Res* **47**, 163-73 (1993).
31. Suri, D. & Vaidya, V.A. Glucocorticoid regulation of brain-derived neurotrophic factor: relevance to hippocampal structural and functional plasticity. *Neuroscience* **239**, 196-213 (2013).
32. Gluckman, P.D., Hanson, M.A., Cooper, C. & Thornburg, K.L. Effect of in utero and early-life conditions on adult health and disease. *N Engl J Med* **359**, 61-73 (2008).
33. Nanni, V., Uher, R. & Danese, A. Childhood maltreatment predicts unfavorable course of illness and treatment outcome in depression: a meta-analysis. *Am J Psychiatry* **169**, 141-51 (2012).
34. Rich-Edwards, J.W. *et al.* Abuse in childhood and adolescence as a predictor of type 2 diabetes in adult women. *Am J Prev Med* **39**, 529-36 (2010).
35. McEwen, B.S. Protective and damaging effects of stress mediators. *N Engl J Med* **338**, 171-9 (1998).
36. McEwen, B.S. Central effects of stress hormones in health and disease: Understanding the protective and damaging effects of stress and stress mediators. *Eur J Pharmacol* **583**, 174-85 (2008).
37. Miller, G.E., Chen, E. & Parker, K.J. Psychological stress in childhood and susceptibility to the chronic diseases of aging: moving toward a model of behavioral and biological mechanisms. *Psychol Bull* **137**, 959-97 (2011).
38. Graham, J.E., Christian, L.M. & Kiecolt-Glaser, J.K. Stress, age, and immune function: toward a lifespan approach. *J Behav Med* **29**, 389-400 (2006).
39. Kendler, K.S., Karkowski, L.M. & Prescott, C.A. Causal relationship between stressful life events and the onset of major depression. *Am J Psychiatry* **156**, 837-41 (1999).
40. Blackburn, E.H. Telomeres and telomerase: their mechanisms of action and the effects of altering their functions. *FEBS Lett* **579**, 859-62 (2005).
41. Price, L.H., Kao, H.T., Burgers, D.E., Carpenter, L.L. & Tyrka, A.R. Telomeres and early-life stress: an overview. *Biol Psychiatry* **73**, 15-23 (2013).
42. Epel, E.S. *et al.* Accelerated telomere shortening in response to life stress. *Proc Natl Acad Sci U S A* **101**, 17312-5 (2004).
43. Tyrka, A.R. *et al.* Alterations of Mitochondrial DNA Copy Number and Telomere Length with Early Adversity and Psychopathology. *Biol Psychiatry* (2015).

44. Shalev, I. *et al.* Exposure to violence during childhood is associated with telomere erosion from 5 to 10 years of age: a longitudinal study. *Mol Psychiatry* **18**, 576-81 (2013).
45. Verhoeven, J.E. *et al.* Major depressive disorder and accelerated cellular aging: results from a large psychiatric cohort study. *Mol Psychiatry* **19**, 895-901 (2014).
46. Wolkowitz, O.M. *et al.* Leukocyte telomere length in major depression: correlations with chronicity, inflammation and oxidative stress--preliminary findings. *PLoS One* **6**, e17837 (2011).
47. Lindqvist, D. *et al.* Psychiatric disorders and leukocyte telomere length: Underlying mechanisms linking mental illness with cellular aging. *Neurosci Biobehav Rev* **55**, 333-64 (2015).
48. Berman, S.B., Pineda, F.J. & Hardwick, J.M. Mitochondrial fission and fusion dynamics: the long and short of it. *Cell Death Differ* **15**, 1147-52 (2008).
49. Chen, H. & Chan, D.C. Mitochondrial dynamics--fusion, fission, movement, and mitophagy--in neurodegenerative diseases. *Hum Mol Genet* **18**, R169-76 (2009).
50. Bogenhagen, D. & Clayton, D.A. Mouse L cell mitochondrial DNA molecules are selected randomly for replication throughout the cell cycle. *Cell* **11**, 719-27 (1977).
51. Robin, E.D. & Wong, R. Mitochondrial DNA molecules and virtual number of mitochondria per cell in mammalian cells. *J Cell Physiol* **136**, 507-13 (1988).
52. Fragouli, E. *et al.* Altered levels of mitochondrial DNA are associated with female age, aneuploidy, and provide an independent measure of embryonic implantation potential. *PLoS Genet* **11**, e1005241 (2015).
53. Poulton, J. *et al.* Variation in mitochondrial DNA levels in muscle from normal controls. Is depletion of mtDNA in patients with mitochondrial myopathy a distinct clinical syndrome. *J Inherit Metab Dis* **18**, 4-20 (1995).
54. Wang, L. *et al.* Plasma nuclear and mitochondrial DNA levels in acute myocardial infarction patients. *Coron Artery Dis* **26**, 296-300 (2015).
55. Mi, J. *et al.* The relationship between altered mitochondrial DNA copy number and cancer risk: a meta-analysis. *Sci Rep* **5**, 10039 (2015).
56. Zhou, W. *et al.* Peripheral blood mitochondrial DNA copy number is associated with prostate cancer risk and tumor burden. *PLoS One* **9**, e109470 (2014).
57. Zhang, Q. *et al.* Circulating mitochondrial DAMPs cause inflammatory responses to injury. *Nature* **464**, 104-7 (2010).
58. Oka, T. *et al.* Mitochondrial DNA that escapes from autophagy causes inflammation and heart failure. *Nature* **485**, 251-5 (2012).

59. Bratic, A. & Larsson, N.G. The role of mitochondria in aging. *J Clin Invest* **123**, 951-7 (2013).
60. Finkel, T. & Holbrook, N.J. Oxidants, oxidative stress and the biology of ageing. *Nature* **408**, 239-47 (2000).
61. Maurya, P.K. *et al.* The role of oxidative and nitrosative stress in accelerated aging and major depressive disorder. *Prog Neuropsychopharmacol Biol Psychiatry* **65**, 134-144 (2015).
62. Barrientos, A. *et al.* Reduced steady-state levels of mitochondrial RNA and increased mitochondrial DNA amount in human brain with aging. *Brain Res Mol Brain Res* **52**, 284-9 (1997).
63. Miller, M.W. & Sadeh, N. Traumatic stress, oxidative stress and post-traumatic stress disorder: neurodegeneration and the accelerated-aging hypothesis. *Mol Psychiatry* **19**, 1156-62 (2014).
64. Kim, M.Y., Lee, J.W., Kang, H.C., Kim, E. & Lee, D.C. Leukocyte mitochondrial DNA (mtDNA) content is associated with depression in old women. *Arch Gerontol Geriatr* **53**, e218-21 (2011).
65. He, Y. *et al.* Leukocyte mitochondrial DNA copy number in blood is not associated with major depressive disorder in young adults. *PLoS One* **9**, e96869 (2014).
66. Sullivan, P.F., Neale, M.C. & Kendler, K.S. Genetic epidemiology of major depression: review and meta-analysis. *Am J Psychiatry* **157**, 1552-62 (2000).
67. Holmans, P. *et al.* Genetics of recurrent early-onset major depression (GenRED): final genome scan report. *Am J Psychiatry* **164**, 248-58 (2007).
68. Zubenko, G.S. *et al.* A collaborative study of the emergence and clinical features of the major depressive syndrome of Alzheimer's disease. *Am J Psychiatry* **160**, 857-66 (2003).
69. McGuffin, P. *et al.* Whole genome linkage scan of recurrent depressive disorder from the depression network study. *Hum Mol Genet* **14**, 3337-45 (2005).
70. Abkevich, V. *et al.* Predisposition locus for major depression at chromosome 12q22-12q23.2. *Am J Hum Genet* **73**, 1271-81 (2003).
71. Levinson, D.F. The genetics of depression: a review. *Biol Psychiatry* **60**, 84-92 (2006).
72. Bosker, F.J. *et al.* Poor replication of candidate genes for major depressive disorder using genome-wide association data. *Mol Psychiatry* **16**, 516-32 (2011).

73. Wellcome Trust Case Control, C. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661-78 (2007).
74. Todd, J.A. *et al.* Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. *Nat Genet* **39**, 857-64 (2007).
75. Sladek, R. *et al.* A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* **445**, 881-5 (2007).
76. McPherson, R. *et al.* A common allele on chromosome 9 associated with coronary heart disease. *Science* **316**, 1488-91 (2007).
77. Rioux, J.D. *et al.* Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis. *Nat Genet* **39**, 596-604 (2007).
78. Gudmundsson, J. *et al.* Two variants on chromosome 17 confer prostate cancer risk, and the one in TCF2 protects against type 2 diabetes. *Nat Genet* **39**, 977-83 (2007).
79. Easton, D.F. *et al.* Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature* **447**, 1087-93 (2007).
80. Weedon, M.N. *et al.* Genome-wide association analysis identifies 20 loci that influence adult height. *Nat Genet* **40**, 575-83 (2008).
81. Frayling, T.M. *et al.* A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science* **316**, 889-94 (2007).
82. Kooner, J.S. *et al.* Genome-wide scan identifies variation in MLXIPL associated with plasma triglycerides. *Nat Genet* **40**, 149-51 (2008).
83. Kathiresan, S. *et al.* Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. *Nat Genet* **40**, 189-97 (2008).
84. Psychiatric, G.C.B.D.W.G. Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. *Nat Genet* **43**, 977-83 (2011).
85. Schizophrenia Working Group of the Psychiatric Genomics, C. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421-7 (2014).
86. Tenesa, A. & Haley, C.S. The heritability of human disease: estimation, uses and abuses. *Nat Rev Genet* **14**, 139-49 (2013).

87. Shyn, S.I. *et al.* Novel loci for major depression identified by genome-wide association study of Sequenced Treatment Alternatives to Relieve Depression and meta-analysis of three studies. *Mol Psychiatry* **16**, 202-15 (2011).
88. Middeldorp, C.M. *et al.* Suggestive linkage on chromosome 2, 8, and 17 for lifetime major depression. *Am J Med Genet B Neuropsychiatr Genet* **150B**, 352-8 (2009).
89. Lewis, C.M. *et al.* Genome-wide association study of major recurrent depression in the U.K. population. *Am J Psychiatry* **167**, 949-57 (2010).
90. Lee, S.H., Yang, J., Goddard, M.E., Visscher, P.M. & Wray, N.R. Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. *Bioinformatics* **28**, 2540-2 (2012).
91. Genomes Project, C. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56-65 (2012).
92. Rietschel, M. *et al.* Genome-wide association-, replication-, and neuroimaging study implicates HOMER1 in the etiology of major depression. *Biol Psychiatry* **68**, 578-85 (2010).
93. Kohli, M.A. *et al.* The neuronal transporter gene SLC6A15 confers risk to major depression. *Neuron* **70**, 252-65 (2011).
94. Kendler, K.S., Aggen, S.H. & Neale, M.C. Evidence for multiple genetic factors underlying DSM-IV criteria for major depression. *JAMA Psychiatry* **70**, 599-607 (2013).
95. Hemani, G. *et al.* Inference of the genetic architecture underlying BMI and height with the use of 20,240 sibling pairs. *Am J Hum Genet* **93**, 865-75 (2013).
96. Flint, J. & Kendler, K.S. The genetics of major depression. *Neuron* **81**, 484-503 (2014).
97. Kendler, K.S., Gardner, C.O. & Prescott, C.A. Clinical characteristics of major depression that predict risk of depression in relatives. *Arch Gen Psychiatry* **56**, 322-7 (1999).
98. Rabbee, N. & Speed, T.P. A genotype calling algorithm for affymetrix SNP arrays. *Bioinformatics* **22**, 7-12 (2006).
99. Barrett, J.C. & Cardon, L.R. Evaluating coverage of genome-wide association studies. *Nat Genet* **38**, 659-62 (2006).
100. Parkes, M., Cortes, A., van Heel, D.A. & Brown, M.A. Genetic insights into common pathways and complex relationships among immune-mediated diseases. *Nat Rev Genet* **14**, 661-73 (2013).

101. Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet* **39**, 906-13 (2007).
102. Lee, S. *et al.* The epidemiology of depression in metropolitan China. *Psychol Med* **39**, 735-47 (2009).
103. McGuffin, P., Katz, R., Watkins, S. & Rutherford, J. A hospital-based twin register of the heritability of DSM-IV unipolar depression. *Arch Gen Psychiatry* **53**, 129-36 (1996).
104. Brown, R.A. *et al.* Depression among cocaine abusers in treatment: relation to cocaine and alcohol use and treatment outcome. *Am J Psychiatry* **155**, 220-5 (1998).
105. Schuckit, M.A. Alcohol and depression: a clinical perspective. *Acta Psychiatr Scand Suppl* **377**, 28-32 (1994).
106. Mardis, E.R. Next-generation DNA sequencing methods. *Annu Rev Genomics Hum Genet* **9**, 387-402 (2008).
107. Genome of the Netherlands, C. Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat Genet* **46**, 818-25 (2014).
108. Clay Montier, L.L., Deng, J.J. & Bai, Y. Number matters: control of mammalian mitochondrial DNA copy number. *J Genet Genomics* **36**, 125-31 (2009).
109. Lupski, J.R. *et al.* Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy. *N Engl J Med* **362**, 1181-91 (2010).
110. Ng, S.B. *et al.* Exome sequencing identifies the cause of a mendelian disorder. *Nat Genet* **42**, 30-5 (2010).
111. Nikopoulos, K. *et al.* Next-generation sequencing of a 40 Mb linkage interval reveals TSPAN12 mutations in patients with familial exudative vitreoretinopathy. *Am J Hum Genet* **86**, 240-7 (2010).
112. Roach, J.C. *et al.* Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* **328**, 636-9 (2010).
113. Pasaniuc, B. *et al.* Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. *Nat Genet* **44**, 631-5 (2012).
114. Li, Y., Sidore, C., Kang, H.M., Boehnke, M. & Abecasis, G.R. Low-coverage sequencing: implications for design of complex trait association studies. *Genome Res* **21**, 940-51 (2011).

115. Ajay, S.S., Parker, S.C., Abaan, H.O., Fajardo, K.V. & Margulies, E.H. Accurate and comprehensive sequencing of personal genomes. *Genome Res* **21**, 1498-505 (2011).
116. Van der Auwera, G.A. *et al.* From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics* **11**, 11 10 1-11 10 33 (2013).
117. DePristo, M.A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43**, 491-8 (2011).
118. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297-303 (2010).
119. Wang, Y., Lu, J., Yu, J., Gibbs, R.A. & Yu, F. An integrative variant analysis pipeline for accurate genotype/haplotype inference in population NGS data. *Genome Res* **23**, 833-42 (2013).
120. Browning, S.R. & Browning, B.L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet* **81**, 1084-97 (2007).
121. Robinson, J.T. *et al.* Integrative genomics viewer. *Nat Biotechnol* **29**, 24-6 (2011).
122. Mount, D.W. Using the Basic Local Alignment Search Tool (BLAST). *CSH Protoc* **2007**, pdb top17 (2007).
123. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559-75 (2007).
124. McClellan, J. & King, M.C. Genetic heterogeneity in human disease. *Cell* **141**, 210-7 (2010).
125. Qin, P. *et al.* A panel of ancestry informative markers to estimate and correct potential effects of population stratification in Han Chinese. *Eur J Hum Genet* **22**, 248-53 (2014).
126. Chen, J. *et al.* Genetic structure of the Han Chinese population revealed by genome-wide SNP variation. *Am J Hum Genet* **85**, 775-85 (2009).
127. Patterson, N., Price, A.L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet* **2**, e190 (2006).
128. Reich, D.E. & Goldstein, D.B. Detecting association in a case-control study while correcting for population stratification. *Genet Epidemiol* **20**, 4-16 (2001).
129. Pritchard, J.K. & Rosenberg, N.A. Use of unlinked genetic markers to detect population stratification in association studies. *Am J Hum Genet* **65**, 220-8 (1999).

130. Price, A.L., Zaitlen, N.A., Reich, D. & Patterson, N. New approaches to population stratification in genome-wide association studies. *Nat Rev Genet* **11**, 459-63 (2010).
131. Devlin, B., Bacanu, S.A. & Roeder, K. Genomic Control to the extreme. *Nat Genet* **36**, 1129-30; author reply 1131 (2004).
132. Price, A.L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**, 904-9 (2006).
133. Seldin, M.F. & Price, A.L. Application of ancestry informative markers to association studies in European Americans. *PLoS Genet* **4**, e5 (2008).
134. Yang, J., Lee, S.H., Goddard, M.E. & Visscher, P.M. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* **88**, 76-82 (2011).
135. Yu, J. *et al.* A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat Genet* **38**, 203-8 (2006).
136. Zhao, K. *et al.* An Arabidopsis example of association mapping in structured samples. *PLoS Genet* **3**, e4 (2007).
137. Kang, H.M. *et al.* Efficient control of population structure in model organism association mapping. *Genetics* **178**, 1709-23 (2008).
138. Kang, H.M. *et al.* Variance component model to account for sample structure in genome-wide association studies. *Nat Genet* **42**, 348-54 (2010).
139. Zhang, Z. *et al.* Mixed linear model approach adapted for genome-wide association studies. *Nat Genet* **42**, 355-60 (2010).
140. Sul, J.H. & Eskin, E. Mixed models can correct for population structure for genomic regions under selection. *Nat Rev Genet* **14**, 300 (2013).
141. Price, A.L., Zaitlen, N.A., Reich, D. & Patterson, N. Response to Sul and Eskin. *Nat Rev Genet* **14**, 300 (2013).
142. Lippert, C. *et al.* FaST linear mixed models for genome-wide association studies. *Nat Methods* **8**, 833-5 (2011).
143. Wang, K., Hu, X. & Peng, Y. An analytical comparison of the principal component method and the mixed effects model for association studies in the presence of cryptic relatedness and population stratification. *Hum Hered* **76**, 1-9 (2013).
144. Tucker, G. *et al.* Two variance component model improves genetic prediction in family data sets. *bioRxiv preprint* (2015).
145. Yang, J., Zaitlen, N.A., Goddard, M.E., Visscher, P.M. & Price, A.L. Advantages and pitfalls in the application of mixed-model association methods. *Nat Genet* **46**, 100-6 (2014).

146. Listgarten, J. *et al.* Improved linear mixed models for genome-wide association studies. *Nat Methods* **9**, 525-6 (2012).
147. Korte, A. *et al.* A mixed-model approach for genome-wide association studies of correlated traits in structured populations. *Nat Genet* **44**, 1066-71 (2012).
148. Segura, V. *et al.* An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nat Genet* **44**, 825-30 (2012).
149. Widmer, C. *et al.* Further improvements to linear mixed models for genome-wide association studies. *Sci Rep* **4**, 6874 (2014).
150. Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nat Genet* **44**, 821-4 (2012).
151. Tucker, G., Price, A.L. & Berger, B. Improving the power of GWAS and avoiding confounding from population stratification with PC-Select. *Genetics* **197**, 1045-9 (2014).
152. Neff, C.D. *et al.* Evidence for HTR1A and LHPP as interacting genetic risk factors in major depression. *Mol Psychiatry* **14**, 621-30 (2009).
153. Guo, X. *et al.* The NAD(+)-dependent protein deacetylase activity of SIRT1 is regulated by its oligomeric status. *Sci Rep* **2**, 640 (2012).
154. Canto, C. & Auwerx, J. PGC-1alpha, SIRT1 and AMPK, an energy sensing network that controls energy expenditure. *Curr Opin Lipidol* **20**, 98-105 (2009).
155. Pruim, R.J. *et al.* LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* **26**, 2336-7 (2010).
156. Sun, N. *et al.* A comparison of melancholic and nonmelancholic recurrent major depression in Han Chinese women. *Depress Anxiety* **29**, 4-9 (2012).
157. Kendler, K.S. The diagnostic validity of melancholic major depression in a population-based sample of female twins. *Arch Gen Psychiatry* **54**, 299-304 (1997).
158. Chen, J. *et al.* Childhood sexual abuse and the development of recurrent major depression in Chinese women. *PLoS One* **9**, e87569 (2014).
159. Fergusson, D.M. & Mullen, P.E. *Childhood Sexual Abuse: An Evidence Based Perspective*, (Sage Publications, Inc, Thousand Oaks, CA, 1999).
160. Speed, D., Hemani, G., Johnson, M.R. & Balding, D.J. Improved heritability estimation from genome-wide SNPs. *Am J Hum Genet* **91**, 1011-21 (2012).
161. Sullivan, P.F., Daly, M.J. & O'Donovan, M. Genetic architectures of psychiatric disorders: the emerging picture and its implications. *Nat Rev Genet* **13**, 537-51 (2012).

162. Bulik-Sullivan, B.K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* **47**, 291-5 (2015).
163. Lee, S.H. *et al.* Estimation of SNP heritability from dense genotype data. *Am J Hum Genet* **93**, 1151-5 (2013).
164. Purcell, S.M. *et al.* A polygenic burden of rare disruptive mutations in schizophrenia. *Nature* **506**, 185-90 (2014).
165. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* **38**, e164 (2010).
166. Pagliarini, D.J. *et al.* A mitochondrial protein compendium elucidates complex I disease biology. *Cell* **134**, 112-23 (2008).
167. Baur, J.A. Biochemical effects of SIRT1 activators. *Biochim Biophys Acta* **1804**, 1626-34 (2010).
168. Donmez, G., Wang, D., Cohen, D.E. & Guarente, L. SIRT1 suppresses beta-amyloid production by activating the alpha-secretase gene ADAM10. *Cell* **142**, 320-32 (2010).
169. Haigis, M.C. & Sinclair, D.A. Mammalian sirtuins: biological insights and disease relevance. *Annu Rev Pathol* **5**, 253-95 (2010).
170. Chen, D., Steele, A.D., Lindquist, S. & Guarente, L. Increase in activity during calorie restriction requires Sirt1. *Science* **310**, 1641 (2005).
171. Boily, G. *et al.* SirT1 regulates energy metabolism and response to caloric restriction in mice. *PLoS One* **3**, e1759 (2008).
172. Bordone, L. *et al.* SIRT1 transgenic mice show phenotypes resembling calorie restriction. *Aging Cell* **6**, 759-67 (2007).
173. Pfluger, P.T., Herranz, D., Velasco-Miguel, S., Serrano, M. & Tschop, M.H. Sirt1 protects against high-fat diet-induced metabolic damage. *Proc Natl Acad Sci U S A* **105**, 9793-8 (2008).
174. Herranz, D. *et al.* Sirt1 improves healthy ageing and protects from metabolic syndrome-associated cancer. *Nat Commun* **1**, 3 (2010).
175. Cho, S.W., Kim, S., Kim, J.M. & Kim, J.S. Targeted genome engineering in human cells with the Cas9 RNA-guided endonuclease. *Nat Biotechnol* **31**, 230-2 (2013).
176. Consortium, G.T. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648-60 (2015).
177. Speed, D. & Balding, D.J. MultiBLUP: improved SNP-based prediction for complex traits. *Genome Res* **24**, 1550-7 (2014).

178. Okereke, O.I. *et al.* High phobic anxiety is related to lower leukocyte telomere length in women. *PLoS One* **7**, e40516 (2012).
179. Jodczyk, S., Fergusson, D.M., Horwood, L.J., Pearson, J.F. & Kennedy, M.A. No association between mean telomere length and life stress observed in a 30 year birth cohort. *PLoS One* **9**, e97102 (2014).
180. Glass, D., Parts, L., Knowles, D., Aviv, A. & Spector, T.D. No correlation between childhood maltreatment and telomere length. *Biol Psychiatry* **68**, e21-2; author reply e23-4 (2010).
181. Shao, L. *et al.* Mitochondrial involvement in psychiatric disorders. *Ann Med* **40**, 281-95 (2008).
182. Martin, J., Anderson, J., Romans, S., Mullen, P. & O'Shea, M. Asking about child sexual abuse: methodological implications of a two stage survey. *Child Abuse Negl* **17**, 383-92 (1993).
183. Tao, M. *et al.* Examining the relationship between lifetime stressful life events and the onset of major depression in Chinese women. *J Affect Disord* **135**, 95-9 (2011).
184. Ding, Z. *et al.* Estimating telomere length from whole genome sequence data. *Nucleic Acids Res* **42**, e75 (2014).
185. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-9 (2009).
186. Raeste, A.M. Degeneration of oral leukocytes. *Scand J Dent Res* **80**, 285-91 (1972).
187. Schiott, C.R. & Loe, H. The origin and variation in the number of leukocytes in the human saliva. *J Periodontal Res Suppl*, 24-6 (1969).
188. Vidovic, A. *et al.* Determination of leucocyte subsets in human saliva by flow cytometry. *Arch Oral Biol* **57**, 577-83 (2012).
189. Christensen, B.C. *et al.* Aging and environmental exposures alter tissue-specific DNA methylation dependent upon CpG island context. *PLoS Genet* **5**, e1000602 (2009).
190. Sehouli, J. *et al.* Epigenetic quantification of tumor-infiltrating T-lymphocytes. *Epigenetics* **6**, 236-46 (2011).
191. Wieczorek, G. *et al.* Quantitative DNA methylation analysis of FOXP3 as a new method for counting regulatory T cells in peripheral blood and solid tissue. *Cancer Res* **69**, 599-608 (2009).

192. Slieker, R.C. *et al.* Identification and systematic annotation of tissue-specific differentially methylated regions using the Illumina 450k array. *Epigenetics Chromatin* **6**, 26 (2013).
193. Accomando, W.P., Wiencke, J.K., Houseman, E.A., Nelson, H.H. & Kelsey, K.T. Quantitative reconstruction of leukocyte subsets using DNA methylation. *Genome Biol* **15**, R50 (2014).
194. Houseman, E.A. *et al.* DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics* **13**, 86 (2012).
195. van Dongen, J. *et al.* Epigenetic variation in monozygotic twins: a genome-wide analysis of DNA methylation in buccal cells. *Genes (Basel)* **5**, 347-65 (2014).
196. Krueger, F. & Andrews, S.R. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**, 1571-2 (2011).
197. Carpenter, W.T., Jr. & Bunney, W.E., Jr. Adrenal cortical activity in depressive illness. *Am J Psychiatry* **128**, 31-40 (1971).
198. Du, J. *et al.* Dynamic regulation of mitochondrial function by glucocorticoids. *Proc Natl Acad Sci U S A* **106**, 3543-8 (2009).
199. Choi, J., Fauce, S.R. & Effros, R.B. Reduced telomerase activity in human T lymphocytes exposed to cortisol. *Brain Behav Immun* **22**, 600-5 (2008).
200. Wikgren, M. *et al.* Short telomeres in depression and the general population are associated with a hypocortisolemic state. *Biol Psychiatry* **71**, 294-300 (2012).
201. Tomiyama, A.J. *et al.* Does cellular aging relate to patterns of allostasis? An examination of basal and stress reactive HPA axis activity and telomere length. *Physiol Behav* **106**, 40-5 (2012).
202. Gillespie, C.F. & Nemeroff, C.B. Hypercortisolemia and depression. *Psychosom Med* **67 Suppl 1**, S26-8 (2005).
203. Gluckman, P.D., Hanson, M.A., Spencer, H.G. & Bateson, P. Environmental influences during development and their later consequences for health and disease: implications for the interpretation of empirical studies. *Proc Biol Sci* **272**, 671-7 (2005).
204. Holt, I.J., Harding, A.E. & Morgan-Hughes, J.A. Deletions of muscle mitochondrial DNA in patients with mitochondrial myopathies. *Nature* **331**, 717-9 (1988).
205. Zhu, D.P., Economou, E.P., Antonarakis, S.E. & Maumenee, I.H. Mitochondrial DNA mutation and heteroplasmy in type I Leber hereditary optic neuropathy. *Am J Med Genet* **42**, 173-9 (1992).
206. Shoffner, J.M.t. & Wallace, D.C. Oxidative phosphorylation diseases. Disorders of two genomes. *Adv Hum Genet* **19**, 267-330 (1990).

207. Schaefer, A.M., Taylor, R.W., Turnbull, D.M. & Chinnery, P.F. The epidemiology of mitochondrial disorders--past, present and future. *Biochim Biophys Acta* **1659**, 115-20 (2004).

208. Lott, M.T. *et al.* mtDNA Variation and Analysis Using MITOMAP and MITOMASTER. *Curr Protoc Bioinformatics* **1**, 1 23 1-1 23 26 (2013).

209. Wallace, D.C. & Chalkia, D. Mitochondrial DNA genetics and the heteroplasmy conundrum in evolution and disease. *Cold Spring Harb Perspect Biol* **5**, a021220 (2013).

210. Chinnery, P.F., Elliott, H.R., Hudson, G., Samuels, D.C. & Relton, C.L. Epigenetics, epidemiology and mitochondrial DNA diseases. *Int J Epidemiol* **41**, 177-87 (2012).

211. He, Y. *et al.* Heteroplasmic mitochondrial DNA mutations in normal and tumour cells. *Nature* **464**, 610-4 (2010).

212. Ye, K., Lu, J., Ma, F., Keinan, A. & Gu, Z. Extensive pathogenicity of mitochondrial heteroplasmy in healthy human individuals. *Proc Natl Acad Sci U S A* **111**, 10654-9 (2014).

213. Liu, P. & Demple, B. DNA repair in mammalian mitochondria: Much more than we thought? *Environ Mol Mutagen* **51**, 417-26 (2010).

214. Wallace, D.C. A mitochondrial paradigm of metabolic and degenerative diseases, aging, and cancer: a dawn for evolutionary medicine. *Annu Rev Genet* **39**, 359-407 (2005).

215. Campbell, C.T., Kolesar, J.E. & Kaufman, B.A. Mitochondrial transcription factor A regulates mitochondrial transcription initiation, DNA packaging, and genome copy number. *Biochim Biophys Acta* **1819**, 921-9 (2012).

216. Johnston, I.G. *et al.* Mitochondrial variability as a source of extrinsic cellular noise. *PLoS Comput Biol* **8**, e1002416 (2012).

217. Boyd, A. *et al.* Cohort Profile: the 'children of the 90s'--the index offspring of the Avon Longitudinal Study of Parents and Children. *Int J Epidemiol* **42**, 111-27 (2013).

218. Magrane, M. & Consortium, U. UniProt Knowledgebase: a hub of integrated protein data. *Database (Oxford)* **2011**, bar009 (2011).

219. Ngo, H.B., Kaiser, J.T. & Chan, D.C. The mitochondrial transcription and packaging factor Tfam imposes a U-turn on mitochondrial DNA. *Nat Struct Mol Biol* **18**, 1290-6 (2011).

220. Hazkani-Covo, E., Zeller, R.M. & Martin, W. Molecular poltergeists: mitochondrial DNA copies (numts) in sequenced nuclear genomes. *PLoS Genet* **6**, e1000834 (2010).
221. Marinov, G.K., Wang, Y.E., Chan, D. & Wold, B.J. Evidence for site-specific occupancy of the mitochondrial genome by nuclear transcription factors. *PLoS One* **9**, e84713 (2014).
222. Yao, Y.G., Kong, Q.P., Bandelt, H.J., Kivisild, T. & Zhang, Y.P. Phylogeographic differentiation of mitochondrial DNA in Han Chinese. *Am J Hum Genet* **70**, 635-51 (2002).
223. Ruiz-Pesini, E., Mishmar, D., Brandon, M., Procaccio, V. & Wallace, D.C. Effects of purifying and adaptive selection on regional variation in human mtDNA. *Science* **303**, 223-6 (2004).
224. Mercer, T.R. *et al.* The human mitochondrial transcriptome. *Cell* **146**, 645-58 (2011).
225. Bonawitz, N.D., Clayton, D.A. & Shadel, G.S. Initiation and beyond: multiple functions of the human mitochondrial transcription machinery. *Mol Cell* **24**, 813-25 (2006).
226. Lee, D.Y. & Clayton, D.A. Initiation of mitochondrial DNA replication by transcription and R-loop processing. *J Biol Chem* **273**, 30614-21 (1998).
227. Dominy, J.E. & Puigserver, P. Mitochondrial biogenesis through activation of nuclear signaling proteins. *Cold Spring Harb Perspect Biol* **5**(2013).
228. Ngo, H.B., Lovely, G.A., Phillips, R. & Chan, D.C. Distinct structural features of TFAM drive mitochondrial DNA packaging versus transcriptional activation. *Nat Commun* **5**, 3077 (2014).
229. Holt, I.J., Lorimer, H.E. & Jacobs, H.T. Coupled leading- and lagging-strand synthesis of mammalian mitochondrial DNA. *Cell* **100**, 515-24 (2000).
230. Holt, I.J. & Reyes, A. Human mitochondrial DNA replication. *Cold Spring Harb Perspect Biol* **4**(2012).
231. Ekstrand, M.I. *et al.* Mitochondrial transcription factor A regulates mtDNA copy number in mammals. *Hum Mol Genet* **13**, 935-44 (2004).
232. Gensler, S. *et al.* Mechanism of mammalian mitochondrial DNA replication: import of mitochondrial transcription factor A into isolated mitochondria stimulates 7S DNA synthesis. *Nucleic Acids Res* **29**, 3657-63 (2001).
233. Ekholm, S.V. & Reed, S.I. Regulation of G(1) cyclin-dependent kinases in the mammalian cell cycle. *Curr Opin Cell Biol* **12**, 676-84 (2000).

234. Otto, T. & Sicinski, P. The kinase-independent, second life of CDK6 in transcription. *Cancer Cell* **24**, 141-3 (2013).
235. Kollmann, K. *et al.* A kinase-independent function of CDK6 links the cell cycle to tumor angiogenesis. *Cancer Cell* **24**, 167-81 (2013).
236. Matushansky, I., Radparvar, F. & Skoultchi, A.I. CDK6 blocks differentiation: coupling cell proliferation to the block to differentiation in leukemic cells. *Oncogene* **22**, 4143-9 (2003).
237. Schmidt, O. *et al.* Regulation of mitochondrial protein import by cytosolic kinases. *Cell* **144**, 227-39 (2011).
238. Dekker, P.J. *et al.* Preprotein translocase of the outer mitochondrial membrane: molecular dissection and assembly of the general import pore complex. *Mol Cell Biol* **18**, 6515-24 (1998).
239. Kato, H. & Mihara, K. Identification of Tom5 and Tom6 in the preprotein translocase complex of human mitochondrial outer membrane. *Biochem Biophys Res Commun* **369**, 958-63 (2008).
240. Rosenberg, S.M. Evolving responsively: adaptive mutation. *Nat Rev Genet* **2**, 504-15 (2001).
241. Hastings, P.J., Bull, H.J., Klump, J.R. & Rosenberg, S.M. Adaptive amplification: an inducible chromosomal instability mechanism. *Cell* **103**, 723-31 (2000).
242. Cohen, D.E., Supinski, A.M., Bonkowski, M.S., Donmez, G. & Guarente, L.P. Neuronal SIRT1 regulates endocrine and behavioral responses to calorie restriction. *Genes Dev* **23**, 2812-7 (2009).
243. Libert, S. *et al.* SIRT1 activates MAO-A in the brain to mediate anxiety and exploratory drive. *Cell* **147**, 1459-72 (2011).
244. Jeong, H. *et al.* Sirt1 mediates neuroprotection from mutant huntingtin by activation of the TORC1 and CREB transcriptional pathway. *Nat Med* **18**, 159-65 (2012).
245. Shieh, P.B. & Ghosh, A. Molecular mechanisms underlying activity-dependent regulation of BDNF expression. *J Neurobiol* **41**, 127-34 (1999).
246. Imai, S. & Guarente, L. NAD+ and sirtuins in aging and disease. *Trends Cell Biol* **24**, 464-71 (2014).
247. Yang, H. *et al.* Nutrient-sensitive mitochondrial NAD+ levels dictate cell survival. *Cell* **130**, 1095-107 (2007).
248. Heim, C. *et al.* Pituitary-adrenal and autonomic responses to stress in women after sexual and physical abuse in childhood. *JAMA* **284**, 592-7 (2000).

249. Sjoberg, R.L. Long-term neuroendocrine effects of childhood maltreatment. *JAMA* **284**, 2321-2 (2000).
250. Arborelius, L., Owens, M.J., Plotsky, P.M. & Nemeroff, C.B. The role of corticotropin-releasing factor in depression and anxiety disorders. *J Endocrinol* **160**, 1-12 (1999).
251. Sachar, E.J., Hellman, L., Fukushima, D.K. & Gallagher, T.F. Cortisol production in depressive illness. A clinical and biochemical clarification. *Arch Gen Psychiatry* **23**, 289-98 (1970).
252. Gibbons, J.L. & Mc, H.P. Plasma cortisol in depressive illness. *J Psychiatr Res* **1**, 162-71 (1962).
253. Price, N.L. *et al.* SIRT1 is required for AMPK activation and the beneficial effects of resveratrol on mitochondrial function. *Cell Metab* **15**, 675-90 (2012).
254. Durieux, J., Wolff, S. & Dillin, A. The cell-non-autonomous nature of electron transport chain-mediated longevity. *Cell* **144**, 79-91 (2011).
255. Woo, D.K. & Shadel, G.S. Mitochondrial stress signals revise an old aging theory. *Cell* **144**, 11-2 (2011).
256. Sharpley, M.S. *et al.* Heteroplasmy of mouse mtDNA is genetically unstable and results in altered behavior and cognition. *Cell* **151**, 333-43 (2012).
257. Boore, J.L. Transmission of mitochondrial DNA--playing favorites? *Bioessays* **19**, 751-3 (1997).
258. Fan, W. *et al.* A mouse model of mitochondrial disease reveals germline selection against severe mtDNA mutations. *Science* **319**, 958-62 (2008).
259. Moreno-Loshuertos, R. *et al.* Differences in reactive oxygen species production explain the phenotypes associated with common mouse mitochondrial DNA variants. *Nat Genet* **38**, 1261-8 (2006).
260. Suliman, H.B., Welty-Wolf, K.E., Carraway, M., Tatro, L. & Piantadosi, C.A. Lipopolysaccharide induces oxidative cardiac mitochondrial damage and biogenesis. *Cardiovasc Res* **64**, 279-88 (2004).
261. Ikeuchi, M. *et al.* Overexpression of mitochondrial transcription factor a ameliorates mitochondrial deficiencies and cardiac failure after myocardial infarction. *Circulation* **112**, 683-90 (2005).
262. Lee, S.S. *et al.* A systematic RNAi screen identifies a critical role for mitochondria in C. elegans longevity. *Nat Genet* **33**, 40-8 (2003).
263. Hayashi, Y. *et al.* Reverse of age-dependent memory impairment and mitochondrial DNA damage in microglia by an overexpression of human mitochondrial transcription factor a in mice. *J Neurosci* **28**, 8624-34 (2008).

264. Association, A.P. *Diagnostic and statistical manual of mental disorders*, (American Psychiatric Association, Washington, D.C., 1994).
265. Lunter, G. & Goodson, M. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Res* **21**, 936-9 (2011).
266. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-60 (2009).
267. Yang, J. *et al.* Genome partitioning of genetic variation for complex traits using common SNPs. *Nat Genet* **43**, 519-25 (2011).
268. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat Genet* **42**, 565-9 (2010).
269. Kruger, A., Hofmann, O., Carninci, P., Hayashizaki, Y. & Hide, W. Simplified ontologies allowing comparison of developmental mammalian gene expression. *Genome Biol* **8**, R229 (2007).
270. Smedley, D. *et al.* The BioMart community portal: an innovative alternative to large, centralized data repositories. *Nucleic Acids Res* **43**, W589-98 (2015).
271. Kelso, J. *et al.* eVOC: a controlled vocabulary for unifying gene expression data. *Genome Res* **13**, 1222-30 (2003).
272. Cohen-Woods, S. *et al.* Depression Case Control (DeCC) Study fails to support involvement of the muscarinic acetylcholine receptor M2 (CHRM2) gene in recurrent major depressive disorder. *Hum Mol Genet* **18**, 1504-9 (2009).
273. Uher, R. *et al.* Genetic predictors of response to antidepressants in the GENDEP project. *Pharmacogenomics J* **9**, 225-33 (2009).
274. R-Development-Core-Team. *A language and environment for statistical computing.*, (R Foundation for Statistical Computing, Vienna, 2004).
275. Furda, A.M., Bess, A.S., Meyer, J.N. & Van Houten, B. Analysis of DNA damage and repair in nuclear and mitochondrial DNA of animal cells using quantitative PCR. *Methods Mol Biol* **920**, 111-32 (2012).
276. Livak, K.J. & Schmittgen, T.D. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods* **25**, 402-8 (2001).
277. Jiao, J., Kang, J.X., Tan, R., Wang, J. & Zhang, Y. Multiplex time-reducing quantitative polymerase chain reaction assay for determination of telomere length in blood and tissue DNA. *Anal Bioanal Chem* **403**, 157-66 (2012).
278. Consortium, U.K. *et al.* The UK10K project identifies rare variants in health and disease. *Nature* **526**, 82-90 (2015).

279. Liu, Y., Schmidt, B. & Maskell, D.L. MSAProbs: multiple sequence alignment based on pair hidden Markov models and partition function posterior probabilities. *Bioinformatics* **26**, 1958-64 (2010).

280. Robert, X. & Gouet, P. Deciphering key features in protein structures with the new ENDscript server. *Nucleic Acids Res* **42**, W320-4 (2014).