Inferring the somatic mutation rate from population sequencing data

A thesis submitted

by

Declan Bennett

to

The Discipline of Bioinformatics,
School of Mathematics, Statistics & Applied Mathematics
National University *of* Ireland, Galway

in partial fulfilment of the requirements for the degree of

M.Sc. in Biomedical Genomics

July 31st 2017

Thesis Supervisor: Prof. Cathal Seoighe

Declaration

I, Declan Bennett, declare that this thesis titled, 'Inferring the somatic mutation rate from population sequencing data' submitted to the Discipline of Bioinformatics, School of Mathematics, Statistics and Applied Mathematics National University of Ireland, Galway in partial fulfilment of the requirements for the degree of M.Sc in Biomedical Genomics in entirely my own work.

I have acknowledged all main sources help. I agree freely that the library may lend or copy this thesis upon request.

Declan Bennett
July 2017

Acknowledgements

To my grandmother, Catherine.

My deepest gratitude is owed to Prof. Cathal Seoighe for his help and guidance throughout this thesis. I could not of progressed this far without his expertise and unyielding encouragement. Dr Pilib Ó Broin for always having time for what at the time may seem like trivial problems and Dr Aaron Golden for uniquely explaining difficult concepts in a way that even I could understand. A special thank you to Shauna Hoey, for always listening to my rants on mutation rate even when it was the last thing she wanted to hear about.

Thanks to my parents, Kevin and Monica, and brothers and sister, Kevin, Stephen and Maryann who have always supported me without fail no matter what the endeavour.

Abstract

Somatic mutation rate has remained one of the most difficult biometrics to measure directly and accurately. The somatic mutation rate has implications in a wide variety of diseases such as cancer and age-related diseases. There is a large variation within both interspecies and intraspecies somatic mutation rates stemming from a large number of factors that can influence and drive mutation. It also plays a significant role in driving the progression of ageing. Over lapping paired-end reads allow for potential calling of somatic variants as the sequencing error is effectively minimized by the overlap.

To test this hypothesis genome-wide association studies were performed on all populations of the 1000 genomes project followed by meta analysis. The phenotype was generated by counting the nucleotides in overlapping paired-end reads from the 1000 genomes deep-exome sequencing project that differed from the reference genome. The association studies were then performed using the low coverage whole genome sequencing genotyping call data and the high density omni genotyping chip data. A third set of GWAS and meta analysis was performed using a log transformed phenotype and the high density genotyping data. To account for the strong correlation between age and number of accumulated somatic mutations Horvath's epigentic clock was used to estimate epigenetic age as a surrogate for chronological age.

The association study returned pathways that are implicated in cancer as well as some DNA damage response pathways, although none were significant after multiple testing correction. As the methylation profiling was performed on cell lines, the resulting DNA methylation age profiles were discarded and not used as a covariate in the GWAS.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Defining the somatic mutation rate

The mutation rate can be defined as the probability of observing a mutation per cell per division or generation [1]. The mathematical definition has been derived from work done by Shapiro (1946) and built on by Armitage (1952) using bacterial cultures [2, 3].

$$\mu = \frac{\ln(2)m}{N_t - 1} \approx \frac{\ln(2)m}{N_t}$$

Where μ is the somatic mutation rate. N_t is the final number of cells in culture. Due to the cells within culture growing asynchronously the average number of cells per generation is $\frac{N}{\ln(2)}$ and the total number of cell divisions is $\frac{N_t}{\ln(2)}$. m is the mean number of normalised mutations. The mutation count distribution from fluctuation assays is non-normally distributed and is normalised using the MSS maximum likelihood method, however, several other methods exist [1]. A major caveat to this method was that mutants were described by a Bernoulli trial based on observable phenotypes, typically by observing loss of an essential gene or by comparison with a reporter construct.

For eukaryotes, it is beneficial to be able to quantify the accumulated mutations without such uncertainty. This poses a greater challenge for accurately and efficiently measuring somatic mutation rate, as opposed to germline mutation rate which can be easily determined from familial trio data. A Somatic mutation that arises in one somatic cell is distributed to a large number of clones, provided the deleterious effect does not over burden the cell. The mutational context is very important with strong deleterious mutations being selected against, positive mutations being selected for and neutral mutations accumulating silently. Somatic mutation rate plays a key role in tumour development and is thought to be a driving force of ageing and age-associated diseases [4, 5]. Mutation rate is the substrate

on which selection acts and is intrinsically associated with genome stability [6]. There is a variety of ways in which mutations arise such as intrinsic polymerase errors, DNA damage(exogenous and endogenous) and vitiated DNA repair pathways which can result in the production of mutator phenotypes. Kimura suggested that the presence of mutator alleles are indirectly selected against due to deleterious mutations in which they cause elsewhere in the genome [7]. Lynch has estimated the base-substitution mutation rate across all species in somatic cells to be $<10^{-7}$ with some species having a mutation rate 1,000 fold below this rate. A middle-aged human will have accumulated $>10^{16}$ mutations by these estimates [8, 9]. Lynch noted, that if 1% of coding mutations impaired fitness, that the number of mutations burdening somatic cells would be in the order of 10^{12} [9]. Selection and apoptotic mechanisms will act if the effect of mutation is detrimental, removing the cell from the environment but where there is a non-loss of function or negligible effect size due to diploidy the mutation will remain possibly being fixed or removed through genetic drift.

In 1937, Sturtevant posed the question as to why the mutation rate did not evolve to zero? Sturtevant concluded that, from the contemporary work done, the differences in mutation rate among species and differences intra-species are direct evidence for the evolvability of the mutation rate and that any accurate measurement of the mutation rate should be a reflection of the optimization of the mutation rate through selective processes. [10, 11, 12, 13]. In the 80 years following Sturtevants paper, many advances have been made through theoretical mathematical derivations. Lynch proposed that the reason the mutation rate did not evolve to zero is not set by cellular mechanism, but by the incapacity of selection to push to biochemical perfection. This viewpoint is the basis of the drift barrier hypothesis [9]. The advent of sequencing technologies has allowed the rapid development of methods to measure mutation rate at an unprecedented level. Even with the advancements of sequencing technology over the past 20 years, the somatic mutation rate has remained one of the most difficult biometrics to measure. Early next generation sequencing (NGS) experiments could only call *de novo* germline variants or early somatic mosaic events i.e, somatic mutations that happened after fertilization. Some methods of inferring somatic mutation rate are detailed in the next section.

1.2 Measuring somatic mutation

1.2.1 Reporter constructs

Before the advent of NGS, estimating the somatic mutation rate in higher eukaryotes relied on the advent of integrating reporter constructs into higher eukaryote genomes. The reporter constructs are typically bacteriophage λ shuttle vectors with integrated selection genes such as the *LacZ* gene. The reporter construct is then purified from the transgenic eukaryote and propagated with bacteria lacking the *LacZ* gene to allow for colony selection of mutated strains. In 1989, Jan Vijg developed the first mouse model for the study of mutagenesis [14]. After the development of higher eukaryote model organisms, there was a surge in advancements in genetic research into biochemical pathways and, in particular, into the genome maintenance pathways. The mouse genome is highly conserved with that of the human genome, with mutant maintenance pathways showing similar phenotypes. Although this method is inapplicable to determining human somatic mutation rate, it is readily applicable to other organisms such *Drosophila melanogaster*. The mouse model is, therefore, seen as an ideal model organism in terms of facilitating our understanding of the biological processes that drive somatic mutation rate.

1.2.2 Mutation accumulation lines

The mutation accumulation (MA) procedure is a well-established protocol for directly estimating the mutation rate of a cell line. The concept itself is quite simple; create several isogenic cell lines from a progenitor cell, then subject each cell line to a series of bottlenecks i.e, isolating a single cell from each cell line and clonally expanding [8]. After several bottlenecks, whole genome sequencing (WGS) can be performed to identify the spectrum of mutations. The average rate of increase in number of mutations per cell line is equal to the mutation rate, under the assumption that the MA procedure is effectively neutral. There are many advantages to MA procedure, such as its ability to be applied to range of species. There is one drawback to MA, however, the procedure is quite time-consuming and WGS is still quite expensive albeit the cost is decreasing rapidly [8]. For this procedure to have clinical relevance it involves transforming tissues into cell lines which, in turn, exacerbates the time and cost restraints. MA does, however, provide an unprecedented and accurate measure of mutation rate.

1.2.3 Phylogenetic metrics

Estimating somatic mutation rate from phylogenetic methods requires first estimating the germline mutation rate and then inferring the somatic mutation rate for different cell types [8]. Sequence comparison is often used to estimate the time since divergence of two species. Some assumptions are necessary, however. The first assumption is that the rate of neutral allele fixation is approximately equal to the mutation rate. Typically *pseudogenes* that are common to the two species are used as sequences that evolved neutrally. Secondly, generation time is required for *Homo sapiens*; 30 years is generally chosen as an average generation time. The effective population size (N_e) as time to fixation is $4N_e$. The final requirement is the time since divergence which may be inferred from the genetic distance between sequence [8, 15].

1.2.4 Botseq

In 2016, *Hoang et al.* devised a method to identify rare somatic mutations in human tissue sample by using a bottleneck sequencing system (fig 1.1). This new method named, botseq, uses multiplexing and a simple dilution step before library amplification. The dilution step allows for random sampling of each set of bar-coded fragments. The advantage of this technology is its ability to determine between clonal variants, polymorphisms and PCR induced damage [16].

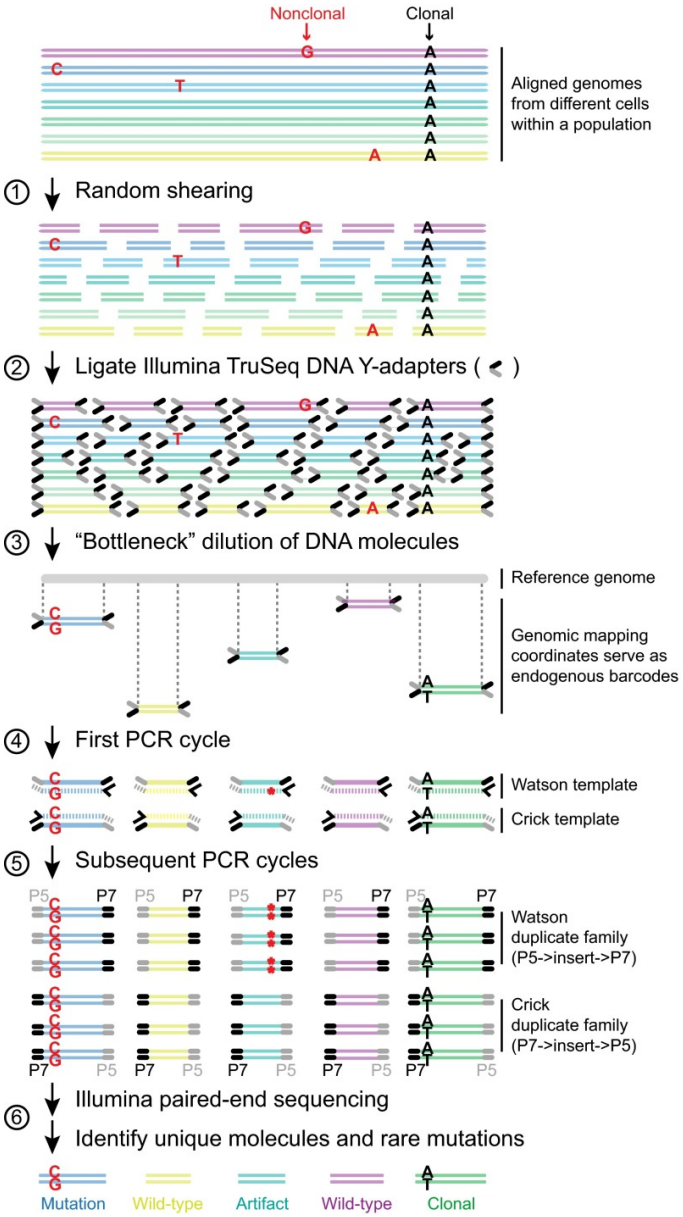

Figure 1.1: Botseq protocol. 1. The aligned genomes represent the variation with in a population of cells. 2. Random fragmentation of each individual fragment. 3 Adapter ligation. 4. The bottleneck serial dilution step acts like a random sampling of specific loci. 4-5. After PCR amplification. 6. Artefacts such as PCR errors are easily discernable from true rare variants and clonal variants after paired-end sequencing. Source: taken from [16].

Botseq allows researchers to obtain and compare rare mutation estimates, as well as mutation prevalence and mutation spectra throughout multiple healthy

tissues and clonal cancer cells. *Hoang et al.* were able to show that the patterns of observed mutations in cells appear to be tissue specific and show similar spectra of mutations to that of cancer in the same tissue type [16]. This observation supports the hypothesis that chromatin structure directly impacts mutation rate throughout the genome leading to a specific mutation footprint per cell type conformation. Botseq carries an advantage over single cell sequencing as it can call private mutations, whereas, in single sequencing the mutation has to be seen in 2 cells to be accurately called as a mutation and not sequencing error.

1.3 Processes driving the somatic mutation rate

1.3.1 DNA replication

Broadly speaking, mutation rate is driven by 3 factors; DNA replication, transcription-associated DNA damage and erroneous DNA damage repair pathways, with replication error providing the largest contribution. Replication fidelity is maintained by 3 mechanisms; Polymerase selectivity, proofreading capabilities and mismatch repair (MMR) mechanisms. Polymerase selectivity is determined by the molecular structure of the polymerase wrapping around the DNA strand. Deviations from the Watson-Crick pairs due to mismatches are subsequently removed via the 3' end exonuclease or through the MMR surveillance protein network [17]. Replication fork stalling due to adducts or crosslinks is dealt with via separate mechanisms; crosslink repair will be discussed in the next section. Adducts cause the disassociation of the high fidelity polymerases in favour for error-prone (low fidelity) polymerases in a process known as translesion synthesis (TLS) [17]. TLS can correctly replicate across the adduct e.g, for T-T dimers, Polη can add A-A opposite the dimer, allowing for successful replication to occur. The dimer is then repaired after replication.

Increased human mutation rate has been shown to be correlated with regions of the genome that are late replicating. This observation had been previously hypothesised due to the increased density of single nucleotide polymorphisms (SNPs) around late replicating genes. *Stamatoyannopoulos et al.* obtained these results from somatic cells concluding that germline and somatic cells have a mirrored mutation rate distribution dependent on replicating time [18]. The increased somatic mutation rate around late replicating genes is often attributed to the depletion of free nucleotides exposing single strand DNA (ssDNA) to endogenous sources of damage for longer periods than early replicating genes [18].

1.3.2 Transcription-associated mutagenesis

Transcription-associated mutagenesis is a significant driver of somatic mutation in the coding region of the genome. As the template strand of DNA is being transcribed, the nontranscribed strand is left in a single strand state. ssDNA is vulnerable to endogenous DNA damage. Formation of R-loops, long ssDNA that have folded back annealing with itself rather than annealing with the template strand, increases the vulnerability of the nontranscribed strand becoming damaged [19]. Transcription can also induce DNA repair, coined transcription coupled repair (TCR). Upon polymerase stalling, the NER pathway is activated removing the aberrant nucleotide. Genes lowly-expressed in cancer show an increased mutation rate over genes highly-expressed suggesting that high fidelity TCR is selected for in cancer. A 3-fold increase in mutation rate is seen in low expressed genes over the highly expressed category. This increase at low expression genes does not, however, capture the full variation in mutation rate seen across the genome [20]. This observation does not hold for the germline. In the testes mutation rate increases with gene expression which is in direct contrast with the observations in the soma [21].

1.3.3 Role of repair

The genome is in constant contact with endogenous and exogenous sources of DNA damage, with each human cell ascertaining approximately 70,000 lesions per day [22, 23]. Several distinct subpathways exist in the DNA repair arsenal to correctly identify, remove and replace lesions or mismatches. Base excision repair (BER) removes damaged bases and single strand DNA breaks, Nucleotide excision repair (NER) removes bulky adducts, such as cyclobutane pyrimidine dimers, either before or after TLS has occured and can be activated by TCR or by the global genome repair pathway. Mismatch repair (MMR) removes incorrectly matched bases that were not removed by polymerase proofreading after replication. For double-strand DNA backbone breaks homologous repair (HR) and nonhomologous end joining (NHEJ) are deployed [24]. Aberrant DNA repair proteins can lead to strong mutator phenotypes; inference of such mutator alleles have identified loci upstream of the DNA repair protein *BRSK2* [6].

1.3.4 Chromatin context affects mutation rate

Regional variation in mutation rates (RViMR) is dictated by early and late replicating genes and the chromatin context within a cell type [25]. Interestingly, highly conserved genes lay in regions of low mutation rate [26]. Not only do the structural features of the genomic landscape, such as nucleosome occupancy, have

an affect on mutation rates but also, epigenetic features, such as methylation at CpG sites. Methylated CpG sites readily undergo deamination in oxidative conditions to form TpG sites. [25]. An elevated mutation rate is also seen at regions of high GC content, such as exonic regions. One would then expect there to be a decreased mutation rate in AT rich regions, such as gene depleted regions of the genome. This, however, is not observed as the AT rich regions are tightly packed in heterochromatin and therefore become inaccessible to repair proteins [25]. *Sabarinathan et al.* found that, DNA binding sites, such as, transcription factor binding sites (TFBS), located within DNase hypersensitive sites (DHS) have an increased mutation rate due to the inaccessibility of NER repair proteins,(fig 1.2) [27].

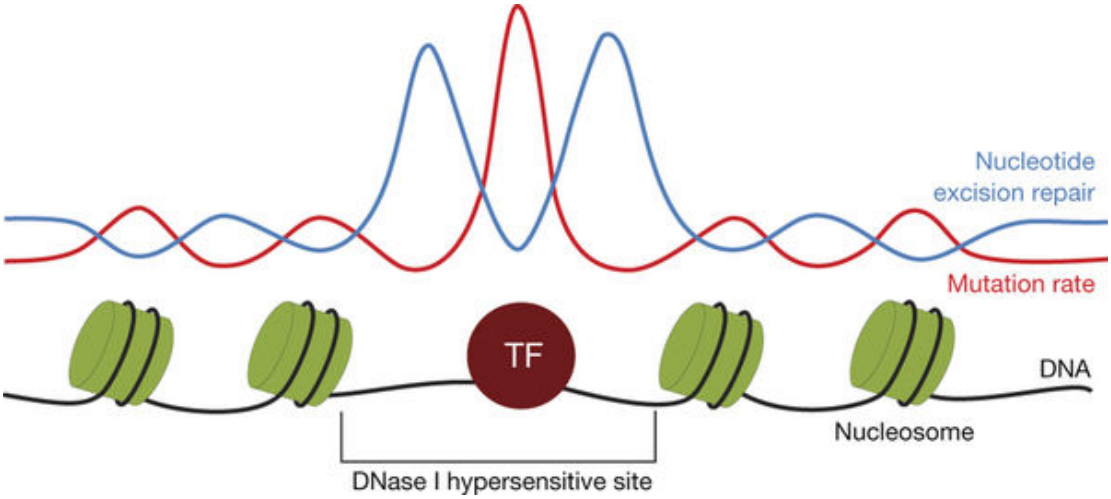

Figure 1.2: Mutation rate is increased at sites of DNA binding. Mutation rate peaks correspond to $\approx$ 170bp. DHS that contain TFBS have a greater mutation rate than DHS with no TFBS. Inaccessibility of NER machinery has is causative of the increase in mutation. Source: taken from [27].

1.4 Implications of the somatic mutation rate

1.4.1 Cancer

Somatic mutations have been associated with cancer for well over 100 years, implicating the somatic mutation rate as a potential driver of tumorigenesis [28, 29]. Although, increased mutation rate is not required for tumorigenesis nor, should it be assumed to be causative, it is responsible for genome instability, which is defined as increased rate of mutation. Genome instability is a well-established hallmark of cancer [30]. Natural selection is expected to play a powerful role in the evolution of tumorigenesis [29]. Increased mutation rate allows increased rate of selection, creating a higher predisposition to tumour development. Mutator alleles, such as, defective DNA damage response proteins, are well established sources of increased mutational load per cell [31]. Tumours (neoplastic cells) effectively evolve at much higher rate over normal tissues due to an increased mutation rate, allowing selection to retain mutations that produce the hallmarks of cancer [32].

1.4.2 Ageing

The second major implication of somatic mutation rate is its role in ageing. The first models of ageing were described in 1959 and 1963 by Orgel and Szilard [33, 34]. Orgel proposed that ageing is caused by 'the error catastrophe' theory. The error catastrophe theory states that errors arising during protein synthesis that affect the accuracy of translating enzymes creates a negative feedback loop that drives ageing, eventually causing death due to increased instability in cellular processes. Although the error catastrophe can explain the latency in ageing, no evidence has been found to support the model [35]. Szilards two-hit model of somatic accumulation states that the somatic mutations accrue linearly, while explaining the non-linearity of ageing. Szilard proposed that human cells have a genetic redundancy (diploid) and that organisms have a cellular redundancy. Once redundancy is exhausted, an exponential increase in mortality is observed [33]. Some assumptions, such as redundancy due to diploidy, do not hold when considering the lifespan triploid or tetraploid organisms, which should theoretically follow a three or four hit model [35]. Although not a perfect model, Szilards two hit model was the first model to accurately predict genetic instability as a cause of senescence [36].

Milholland et al. have proposed a model which merges elements of Orgels and Szilards model, termed 'The somatic mutation catastrophe theory of ageing', (fig 1.3). A key parameter that this model includes is the fact that somatic mutations increase exponentially as the organism ages. To account for the fact that

we do not see a critical amount of erroneous proteins in aged organisms, as Szilard predicted, gene expression dysregulation is factored instead. The regulatory region of the genome constitutes a far greater percentage than the coding region amino acids. Reconciling then, the assumption that the regulatory region of the genome will have a higher burden of mutations and the fact that gene expression dysregulation becomes more apparent in ageing organisms, gene expression can be substituted for altered protein sequences in Szilards model [35]. The somatic mutation catastrophe theory of ageing remains untested and requires a specific experimental design detailed in [35].

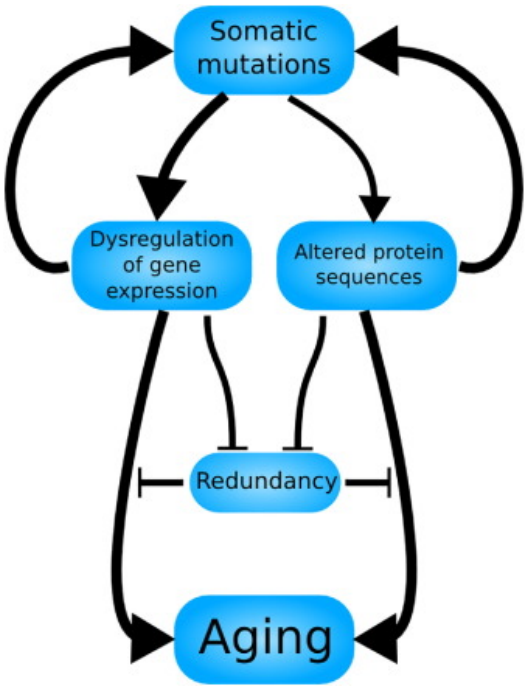

Figure 1.3: The somatic mutation catastrophe theory of ageing combines the error catastrophe theory and the two hit somatic model accumulation. Somatic mutations are increased by a feedback loop dysregulated gene expression and altered protein sequences that lower protein fidelity, weighted by thickness of the arrows. Genetic and cellular redunacy mask the consequences of somatic mutations until exhausted, leading to an exponential increase in ageing. Source: taken from [35]

1.5 Inferring somatic mutation from NGS data

Development of directional adapter molecules has increased the scope of NGS technologies by allowing for ease of mapping reads to the reference genome, as well as reads that span splice junctions in RNA seq experiments. Short fragments

that are paired-end sequenced can be merged based on the concordance in the overlapping region. The increased fragment length facilitates *de novo* assembly. Overlapping reads are fragments of the genome that have been sequenced twice effectively, minimising the probability of a sequencing error. A sequencing error occurring at the same position on read 1 and on read 2 is equal to the square of the probability of a sequencing error occurring. For illumina machines, the average substitution error rate is $\approx 0.1\%$ giving a probability of two sequencing errors at the same position on read 1 and read 2 as 0.01%. This point is the key to this project. Should there be sufficient coverage of reads and overlapping paired-end reads across the genome then novel low frequency variants can be called accurately.

Artefacts, such DNA damage during library preparation and PCR errors, can hinder the detection of real low frequency mutations. DNA damage after fragmentation is a well described phenomenon that leads to an imbalance between paired-end reads. This imbalance is proportional to the damage in the sample. With this fact in mind, algorithms have been developed that can estimate the level of damage within reads [37]. Errors during bridge amplification can be removed by increasing the quality threshold for that nucleotide. PCR amplification errors during library preparation can also be removed by removing reads that begin and end at the same chromosomal positions but have a differing nucleotide(s). PCR artefacts due to polymerase errors are difficult to discern from true mutations. Although population scale WGS projects with sufficient depth are not available there is, however, population scale deep-exome sequencing data publicly available through the 1000 genomes project.

1.6 1000 genomes project

The 1000 genomes project (G1K) was a seven year project that aimed to map and catalogue human genetic variation. The G1K project was made up 2504 individuals from 26 populations around the world and entailed low coverage sequencing, deep-exome sequencing and high-density genotyping, in order to map variants with a frequency of >1%. In total, over 88 million variants were catalogued, including over 99% of common SNPs to multiple ancestries [38]. Derivation of haplotypes from the G1K and also HapMap projects allowed for techniques like imputation to be developed, which statistically infers unobserved genotypes. Genotyping chips need only, then, genotype a subset of reference tag SNPs and through imputation, the remaining genotypes can be inferred. The ability to capture genetic variation by genotyping and imputation has led to an explosion in association studies.

1.6.1 GWAS

Genome-wide association studies (GWAS) probe single common variants for association to a given phenotype. Although not the silver bullet for complex genetics that was once envisaged, large scale GWAS have redefined our understanding of complex disease aetiology. An example of a large scale GWAS meta analysis is the 2014 Psychiatric Genomics Consortium schizophrenia GWAS meta analysis which identified 108 schizophrenia risk loci, 83 of which were novel loci [39]. GWAS can be used for quantitative phenotypes, although phenotypes that do not follow a normal distribution are generally normalised [40]. Some considerations that need to be controlled for in GWAS are the confounding effects of population structure, epistasis between variants and that sources of phenotype variation are caused by common variants with small effect rather large effect rare variants. In this project, a large scale GWAS is performed to assess the performance of the algorithm generating the phenotype by profiling the genetic architecture associated with the phenotype.

1.7 DNA methylation age as a surrogate to chronological age

The number of somatic mutations acquired increases with age in a linear fashion. Other factors such as cell division rate and smoking status are also positively correlated with increased mutation load [4]. There is no phenotype information available for the G1K. This poses the challenge of not being able to reliably call variation in the genetic dataset, versus artefacts in the sequencing files that the algorithm for generating the phenotype has not accounted for i.e. an 80 year old individual is expected to have a far greater number of somatic mutations than a 20 year old individual. One possible method to account for age is to use Horvath's epigenetic clock [41]. Using methylation data it is possible to infer a biological age that is correlated with actual chronological age. This surrogate can then be used to assess correlation between somatic mutation rate, mutation counts, DNA methylation age and mean sample methylation.

The Horvath clock accurately predicts age across all tissues by profiling methylation levels at 353 CpG sites across the genome. In methylation analysis, probe I is often used for scaling/normalization of probe II. Horvath, instead, chose to normalize based on the mean of the largest single study sample coined new gold standard. This was followed by adapting the beta-mixture quantile normalization method (BMIQ), described by *Tesenchendorff et al.*, to rescale the 21k probes overlapping within the 450k and 27k platforms to match the distribution of the new mean gold standard [42]. Horvath chose elastic net regression analysis to

regress a transformed version of chronological age onto 21k beta values of the training dataset. The resulting weighted average of the coefficients from the regression model can then be used to estimate DNA methylation age. Whole blood samples were found to have a Pearson correlation coefficient (r) of 0.98 with an error of 2.7 years [41]. Unfortunately, no population level methylation data is available for the G1K whole-blood samples, however, there is methylation data available for G1K transformed lymphoblastoid cell lines [43]. The major caveat of using transformed cell lines, is that, the precise effects of Epstein-barr virus transformation has on cell methylation profiles, let alone on the 353 CpG DNAm age prediction sites is not fully understood. In a follow up paper, Horvath et al described a 'strong' correlation between 237 lymphoblastoid cell lines and actual age, although the reported Pearson coefficient was 0.59 [44].

The 237 lymphoblastoid cell lines analysed by Horvath et al to assess the correlation between DNAm age and chronological age at time of transformation were part of the human variation panel from 3 cohort; African American, Caucasian-American and Asian-American [45]. Another possible source of error is that the DNAm age of cell lines increases with the number of cell passages, as hypothesized by Horvath and subsequently proven within the same paper [41]. The available methylation data for some of the G1K samples(CEU population) pre-dates the Hapmap project (>1980) whereas the remaining available data (YRI population) was sourced especially for the Hapmap project. Details of the samples is given in the material and methods section.

Chapter 2

Aims

2.1 Project scope

The aim of this project was to assess a novel technique that infers somatic mutation rate from population sized sequencing studies. I will be using the C $\rightarrow$ A substitution rate referred to as somatic mutation rate, phenotype or trait as a quantitative phenotype in multiple genome wide association studies (GWAS) as well as probing the correlation between DNA methylation age and somatic mutation rate for samples with available methylation data. The reasoning behind using C $\rightarrow$ A substitution is to minimize the bias introduced from artefacts such as DNA damage during library preparation stage. Cytosine is readily methylated and mutated to thymine but requires replication to occur before repair. Library preparation creates an oxidative environment in which cytosine has a higher stability over the guanine therefore we expect a C $\rightarrow$ A transversion to be rarely caused by damage during library preparation. The final stage of the analysis will be to assess how the results of the GWAS can help improve the algorithm for calling somatic mutations. The algorithm used to create the phenotype is written pseudocode in the appendix. The overall aim of this project is to evaluate the algorithm as a reasonable proxy for measuring the variation in somatic mutation rate across population sequencing data.

Chapter 3

Methods

The scripts used throughout this project to perform analysis and visualise results can be found at github https://github.com/declan93/MSc-project. This list contains analysis bash scripts as well as R and python scripts for plotting and data manipulation.

3.1 Association study

Plink 1.9 was used for all association studies and meta analysis [46]. The association studies conducted below used the default linear additive model of the plink software package. Three separate association studies were performed on two different sets of SNPs. Firstly, the low coverage whole genome sequencing variant call data for all populations, secondly, the omni platform high density genotyping data available for 19 of 26 populations and finally, a log transformed phenotype with the HD genotype data, as one of the assumptions of quantitative association studies is that the phenotype is normally distributed. Non-normally distributed phenotypes can have significant effects on type I error rates [47].

To account for population structure principal component analysis (PCA) was performed. The 20 top principle components were chosen as covariates in the association study. Combining principal components that capture a small fraction of the variation can increase the statistical power when calling genetic associations, contrary to the popular method of choosing only the top few principal that capture the most variance [48]. For the low coverage data, only the autosomes were analysed; the genotyping data contained 23 chromosomes as well as the mitochondrial chromosome. Of the 26 populations included in the G1K only five are represented in the results section; FIN Finnish in Finland, IBS Iberian population in Spain, ASW Americans of African ancestry in SW USA, TSI Toscani in Italy and ACB African Caribbeans in Barbados.

3.1.1 Low coverage VCF data

Low coverage G1K chromosome VCF file was downloaded from the data portal on the G1K website using either FTP, Aspera or Globus. http://www.internationalgenome.org/. The first step in performing the association was to recode the vcf in plink format and to remove non bi-allelic variants. As each chromosome contained regions where SNPs and indels overlapped, it was necessary to remove these sites to allow merging of chromosomes into one single analysis. The listing of duplicates and subsequent exclusion in the analysis was performed using a mixture of bash and plink. For the quality control step, variants with a sample missingness of 10%, genotype missingness of 10%, minor allele frequency of 5% and variants with a Hardy-Weinberg equilibrium pvalue of below 5% with midp adjustment modifier were excluded from analysis. LD pruning on the variants further diminished the variant count. The parameters chosen were a window size of 50 variants, sliding the window 5 variants at a time and a R^2 of 0.2. A more stringent maf was chosen *post hoc* for two reasons; across all populations chromosome 1 contained $\approx$ 6.5 million variants and chromosome 22 contained $\approx$ 1.1 million variants, and secondly meta analysis on the association results showed excessive significant Cochran's Q results. Population structure has been well documented as a significant confounder in association studies leading to an increase in type I or type II errors [49]. To control for population structure PCA was performed and the top 20 principal components used as covariates in the linear model. The confidence interval flag (–ci), which reports confidence intervals as well as the standard error was also included. The standard error and confidence intervals of each association were required for the meta analysis step. To generate manhattan plots the R package qqman was used [50].

3.1.2 Omni high density genotyping data

There are two genotype data sets available from the G1K project; Phase 1 axiom dataset and Phase 3 HD omni dataset. The Phase 3 HD omni data was downloaded from the G1K in VCF format. The procedure for analysing genotype data mirrors the procedure for low coverage, albeit with the removal of merging of chromosome genotype files and removing overlapping SNP/indels steps. Parameters for filtering genotypes and LD pruning remain the same as well as population stratification methods.

3.1.3 Log transformed phenotype data

In order to normalize the phenotype, the C $\rightarrow$ A mutation rate was log transformed in R and infinite values fixed at 0 and subsequently removed from the analysis.

The procedure described above for the HD genotyping data was then carried out on the transformed phenotype.

3.1.4 Meta analysis

Association studies are often under powered and unable to identify common variants with small effect size. Meta analysis is a statistical approach that can pool independent studies together to boost statistical power and lower false discovery rate. Analysing multiple independent associations as one large meta analysis allows for increased ability to uncover subtle genetic effects. Plink uses a fixed effects model which assumes that the effect of the risk loci is the same across all samples. It does however, have an increased ability to call true positives and true negatives over a random effects model [51]. Heterogeneity is a key issue that must be accounted for when performing a meta analysis. Cochran's Q is typically reported as a metric of heterogeneity and follows a χ^2 distribution with n minus 1 degrees of freedom where n is the number of samples, generally a cut-off of $\alpha = 0.1$ is chosen. Q can be calculated as the sum of the square of the weighted differences in effects in individual studies and the pooled effect across all studies. A second metric, I^2, is generally reported also. I^2 is the percentage of heterogeneity not due to chance [51]. The main sources of heterogeneity are phenotype-based heterogeneity, ancestry-based heterogeneity, population structure, epistasis and gene-environments. Unfortunately, it is impossible to account for all the sources of heterogeneity described. I can however attempt to account for population, phenotype and ancestry-based heterogeneity by using a mixture of strict maf above 5% in the original association analysis and controlling for population structure using PCA. Meta analysis was performed on all three association studies.

3.2 Gene set enrichment analysis

Gene set enrichment analysis was performed on all individual populations across 3 association studies and 3 genome wide meta analyses using magma a tool specifically designed for association results [52]. Analysing the vcf data required that each chromosome be filtered individually before being merged together to create one plink format variant file. Only the results for the meta analysis will be presented in the results section.

Magma uses a multiple regression model that provides better statistical power and is less computationally heavy compared with other available tools. An important parameter to account for is gene size; magma conditions on gene size and gene density allowing for increased true positive rate. There are 3 steps involved in using magma for gene set analysis. Firstly, SNPs are mapped on to the genes

using the human reference build 37 gene location file which can be downloaded from the magma website. http://ctg.cncr.nl/software/magma. The second step in the analysis is the gene analysis step, here already computed p values sample size and genotype data from the association study are used to assign gene p values. Finally, the gene level analysis step uses a general linear model to give some biological meaning to the dataset being analysed. Default parameters for the gene level analysis does not account for multiple testing; the *fwer* flag was added to permute the labels 10,000 times in order to correct for multiple testing. Two gene sets were chosen from the the 5 available curated sets on mSigDB, C2 and C5. C2 contains 4731 general gene sets curated from a wide variety of sources such as online pathway databases and from biomedical literature. C5, on the other hand, contains 5917 gene ontology (GO) gene sets.

3.3 DNA methylation age

Horvath's online calculator was used to compute DNAm age from sample methylation values. The epigenetic clock calculator can be found at https://labs.genetics.ucla.edu/horvat
Methylation data relating to G1K individuals was downloaded from the GEO database under the accession number GSE39672. The full data set of 133 individuals was downloaded. The methylation data for individuals in each population that overlapped with the genotype data from the G1K were separated into 2 cohorts. The central Europeans from Utah (CEU) (n = 60) and the Yuruba tribe in Nigeria (YRI) (n = 73) and each divided into into a full set and a non-warning set. The non-warning sets were the individuals that did not have a low methylation count returned by Horvath's epigentic clock calculator. The full set contained all individuals in that population. The online calculator requires only sample ID, methylation value and CpG ID superfluous columns were removed. Subsequent correlation analysis was conducted in R using the GGally and GGplot2 packages.

Chapter 4

Results

4.1 Association study

4.1.1 Low coverage WGS data

Low coverage whole genome sequencing provided genotype data for the 2533 individuals across the the 26 populations analysed in the G1K project. In order to maintain brevity only a select few populations are presented below. All other analyses have been excluded from this report. FIN was chosen due to the high number of significant p-values returned. ACB was chosen as a more conservative example and the IBS as an example of a population that contains variants associated with the same *pseudogene*.

FIN

After merging of chromosome genotype files and genotype filtering, 440,140 variants and 99 people were carried through to the linear model analysis step. The association study returned 72 significant variants after Bonferroni correction, visualised in figures 4.1& 4.2. The top ten associations are listed in table 4.1.

Table 4.1: FIN population low coverage WGS adjusted association results

	CHR	SNP	UNADJ	GC	BONF	HOLM	SIDAK_SS	SIDAK_SD	FDR_BH	FDR_BY
1	4	rs578009605	4.45E-17	5.89E-17	1.96E-11	1.96E-11	1.96E-11	1.96E-11	1.96E-11	2.66E-10
2	4	rs199795369	2.79E-16	3.66E-16	1.23E-10	1.23E-10	1.23E-10	1.23E-10	6.15E-11	8.34E-10
3	1	rs199809125	2.42E-15	3.14E-15	1.06E-09	1.06E-09	1.06E-09	1.06E-09	3.55E-10	4.81E-09
4	10	rs548717232	4.19E-15	5.42E-15	1.84E-09	1.84E-09	1.84E-09	1.84E-09	4.61E-10	6.26E-09
5	10	rs576691840	6.80E-15	8.77E-15	2.99E-09	2.99E-09	2.99E-09	2.99E-09	5.98E-10	8.12E-09
6	9	rs559799144	8.49E-14	1.08E-13	3.74E-08	3.74E-08	3.74E-08	3.74E-08	5.72E-09	7.77E-08
7	7	rs542320674	9.10E-14	1.16E-13	4.01E-08	4.01E-08	4.01E-08	4.01E-08	5.72E-09	7.77E-08
8	2	rs569983614	1.68E-13	2.13E-13	7.41E-08	7.41E-08	7.41E-08	7.41E-08	9.27E-09	1.26E-07
9	1	rs201130852	4.71E-13	5.93E-13	2.07E-07	2.07E-07	2.07E-07	2.07E-07	2.30E-08	3.12E-07
10	9	rs184202621	9.29E-13	1.17E-12	4.09E-07	4.09E-07	4.09E-07	4.09E-07	3.72E-08	5.04E-07

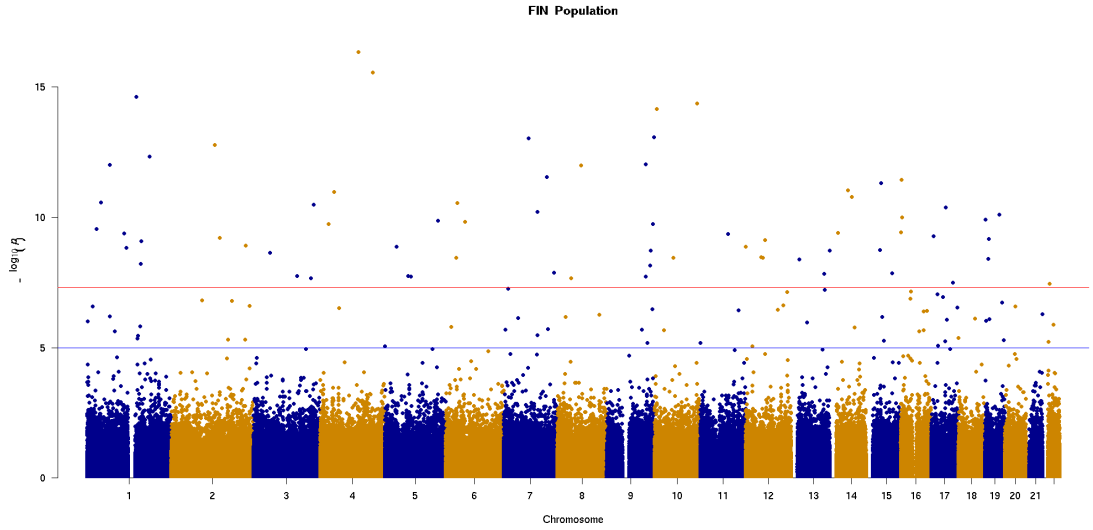

Figure 4.1: Manhattan plot for FIN population VCF data

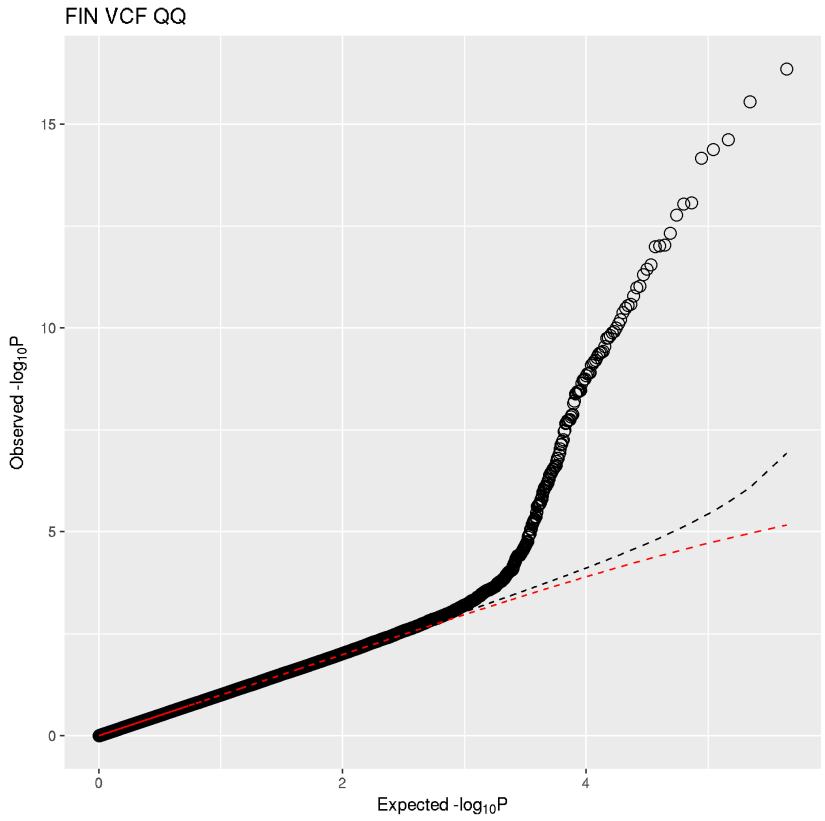

Figure 4.2: Q-Q plot for FIN population VCF data

The genomic inflation estimate λ based on the median χ^2 was 1.01206.λ is the factor of inflation due to systematic differences in the allele frequencies between subpopulations within a population. A λ of 1 indicates that there is no underlying substructure. When plotted, the association result show excessive results greater than the genome wide significance level. The highest association comes from rs578009605 on chromosome 4 with an unadjusted p value of 4.45e−17. rs578009605 is a missense variant in *ZGRF1* gene. The significant hits are distributed evenly across all chromosomes except for chromosomes 20 and 21, neither of which contain variants above the genome-wide significance level. The Q-Q plot shows a a sharp deviation from the expected distribution of p-values.

ACB

After merging the ACB chromosomes, 886,308 variants passed quality control and were carried through for analysis. All 96 individual also passed the filtering stage. The reported genome inflation estimate λ based on the median χ^2 was 1.0057. The association returned 5 significant snps after multiple correction, (table 4.2). The strongest association came from the chromosome 3 missense variant, rs531565115, in the *ZPLD1* gene with an unadjusted p value of 1.3e−6. The polyphen score attributed to rs531565115 is 0.998. *ZPLD1* localises to the cell membrane and has been associated with cerebral cavernous malformation [53]. For the remaining variants, rs587771221 is a intron variant of *NBPF8 pseudogene*, rs201417739 is a missense variant in the thyroid stimulating hormone receptor *TSHR*, rs552557631 is an intronic variant in LINC01598 and esv3646727 a 625 bp inversion located in the intronic region of amyloid presursor protein, *APP*. Association p-values are visualised in figures 4.3 & 4.4. The bulk of the variants follow the null distribution with a sharp upward deviation toward the observed low p-values.

Table 4.2: ACB population low coverage WGS adjusted association results

	CHR	SNP	UNADJ	GC	BONF	HOLM	SIDAK_SS	SIDAK_SD	FDR_BH	FDR_BY
1	3	rs531565115	1.27E-12	1.41E-12	1.13E-06	1.13E-06	1.13E-06	1.13E-06	1.13E-06	1.61E-05
2	1	rs587771221	1.16E-11	1.28E-11	1.03E-05	1.03E-05	1.03E-05	1.03E-05	5.16E-06	7.36E-05
3	14	rs201417739	2.29E-09	2.48E-09	2.03E-03	2.03E-03	2.02E-03	2.02E-03	6.75E-04	9.64E-03
4	20	rs552557631	3.64E-08	3.92E-08	3.23E-02	3.23E-02	3.18E-02	3.18E-02	6.52E-03	9.30E-02
5	21	esv3646727	3.68E-08	3.95E-08	3.26E-02	3.26E-02	3.21E-02	3.21E-02	6.52E-03	9.30E-02

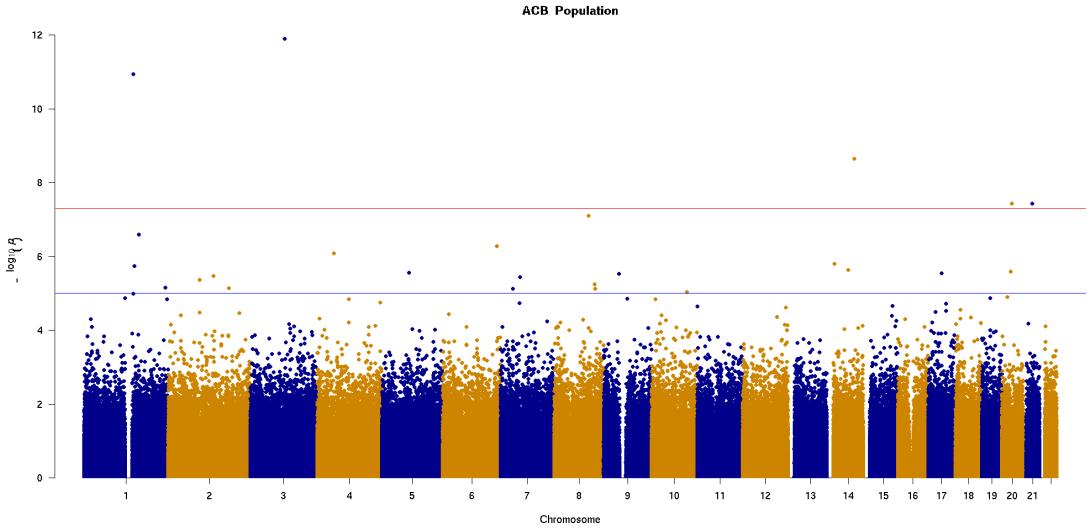

Figure 4.3: Manhattan plot for ACB population VCF data

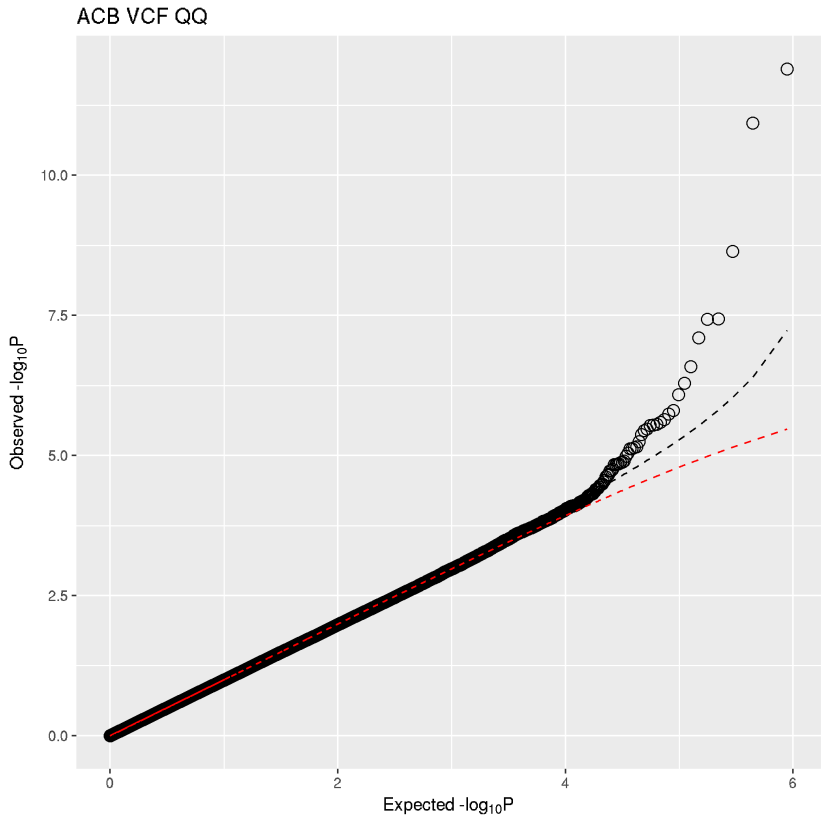

Figure 4.4: Q-Q plot for ACB population VCF data

IBS

The IBS population had a similar number of variants pass the quality control and filtering stages with 434,451 variants carried forward to the analysis stage. All 107 individuals passed quality control. The reported genome inflation estimate λ based on the median χ^2 was 1.00372. Three variants were returned as significant, (table 4.3); rs77353265 which has the highest association with mutation rate is currently not validated, rs80024805 is a missense variant in the processed *pseudogene ANKRD20A19P*. *ANKRD20A9P* is significantly associated with mutation rate in multiple populations in the G1K project and lastly, rs77406825 an intronic variant mapping to *LINC01708* a non coding RNA. The Manhattan plot, figure 4.5, shows the significant results

Table 4.3: IBS population low coverage WGS adjusted association results

	CHR	SNP	UNADJ	GC	BONF	HOLM	SIDAK_SS	SIDAK_SD	FDR_BH	FDR_BY
1	1	rs77353265	0	2.75E-10	1.13E-04	1.13E-04	1.13E-04	1.13E-04	6.91E-05	9.37E-04
2	13	rs80024805	0	3.38E-10	1.38E-04	1.38E-04	1.38E-04	1.38E-04	6.91E-05	9.37E-04
3	1	rs77406825	0	3.33E-08	1.38E-02	1.38E-02	1.37E-02	1.37E-02	4.59E-03	6.23E-02

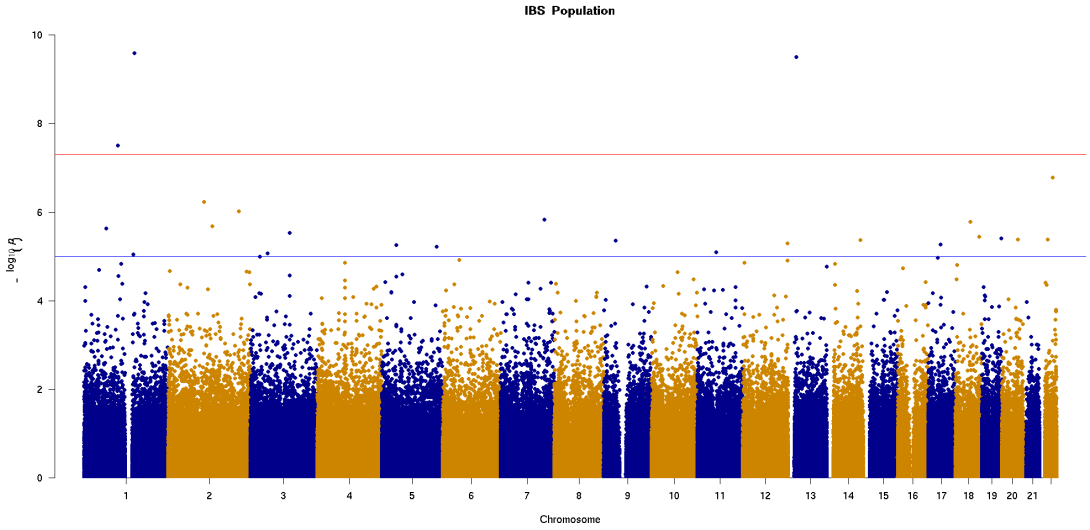

Figure 4.5: Manhattan plot for IBS population VCF data

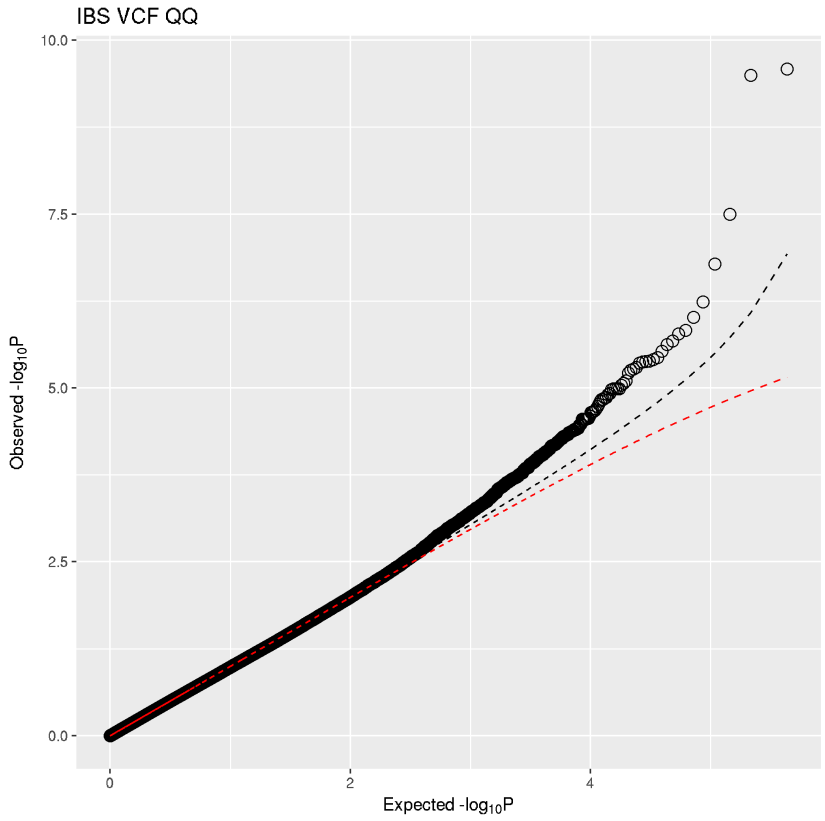

Figure 4.6: Q-Q plot for IBS population VCF data

4.1.2 Omni high density genotyping data

Genotype data was downloaded as one separate file allowing full genome quality control parameters to be represented easily. The 3 populations chosen for low coverage genotyping are presented again to show contrast between the sources of data. Two populations, ASW and TSI, are also presented as they were the only populations to return significant associations after multiple correction. The total number of individuals available for analysis was 2318.

FIN

For the genotyping ≈ 2.5 million variants were analysed. Low genotyping data eliminated 35,900 variants, minor allele threshold removed ≈ 1.125 million and 95,350 variants violated Hardy-Weinberg equilibrium. After quality control and filtering ≈ 1.2 million variants 96 individuals (n=96) were carried through to the association analysis. The reported genome inflation estimate λ based on the median χ^2 was 1.01388. The genotyping rate of the dataset improved from 0.986 to 0.997 after quality control. No variants showed significance after multiple test correcting The p-value distribution was uniform and is not shown below. Not all variants analysed had a reference snp id (rsID)(table 4.4). However it is possible to recover the correct rsID by mapping the variant coordinates to the GRCh37. For example, SNP11-83309647 corresponds to rs1516625, an intron variant in the *DLG2* gene. The Manhattan plot of of the fin population, (fig 4.7), shows an even distribution across all chromosomes. The observed pvalue distribution does not deviate from the expected distribution, (fig 4.8).

Table 4.4: FIN population genotype adjusted association results

	CHR	SNP	UNADJ	GC	BONF	HOLM	SIDAK_SS	SIDAK_SD	FDR_BH	FDR_BY
1	11	SNP11-83309647	4.26E-06	4.86E-06	0.526	5.26E-01	4.09E-01	4.09E-01	0.526	1
2	2	SNP2-13980353	1.78E-05	2.00E-05	1.000	1.00E+00	8.89E-01	8.89E-01	0.658	1
3	4	rs8180202	2.20E-05	2.47E-05	1.000	1.00E+00	9.34E-01	9.34E-01	0.658	1
4	7	rs218148	2.47E-05	2.77E-05	1.000	1.00E+00	9.53E-01	9.53E-01	0.658	1
5	12	SNP12-19617913	2.71E-05	3.03E-05	1.000	1.00E+00	9.65E-01	9.65E-01	0.658	1
6	1	SNP1-22531790	4.00E-05	4.46E-05	1.000	1.00E+00	9.93E-01	9.93E-01	0.658	1
7	12	SNP12-77072647	5.59E-05	6.22E-05	1.000	1.00E+00	9.99E-01	9.99E-01	0.658	1
8	11	rs10501549	5.67E-05	6.29E-05	1.000	1.00E+00	9.99E-01	9.99E-01	0.658	1
9	21	SNP21-24254979	6.40E-05	7.10E-05	1.000	1.00E+00	1.00E+00	1.00E+00	0.658	1
10	1	SNP1-180652274	7.25E-05	8.04E-05	1.000	1.00E+00	1.00E+00	1.00E+00	0.658	1

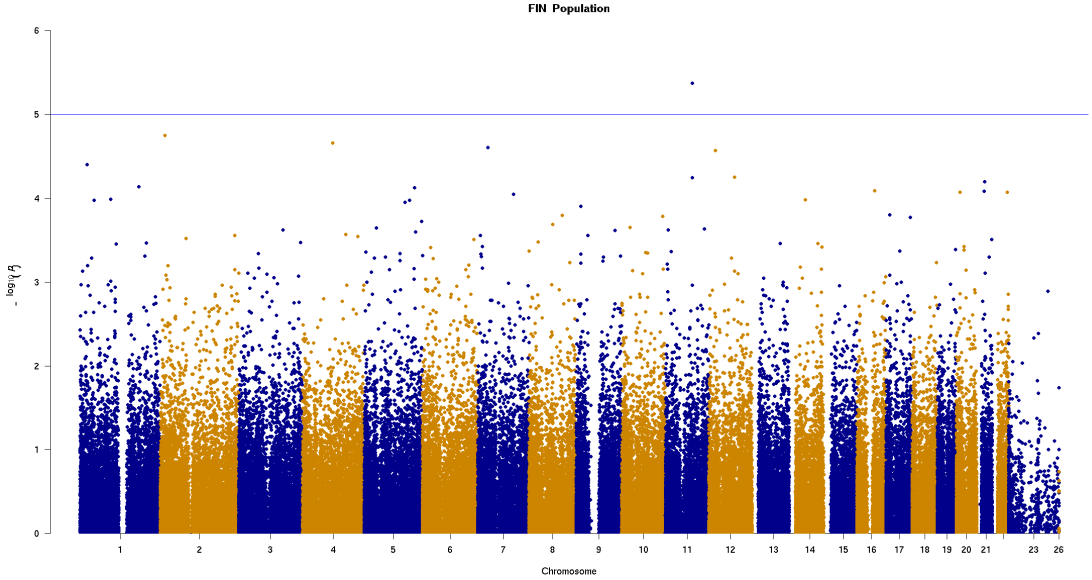

Figure 4.7: Manhattan plot for FIN population genotype data

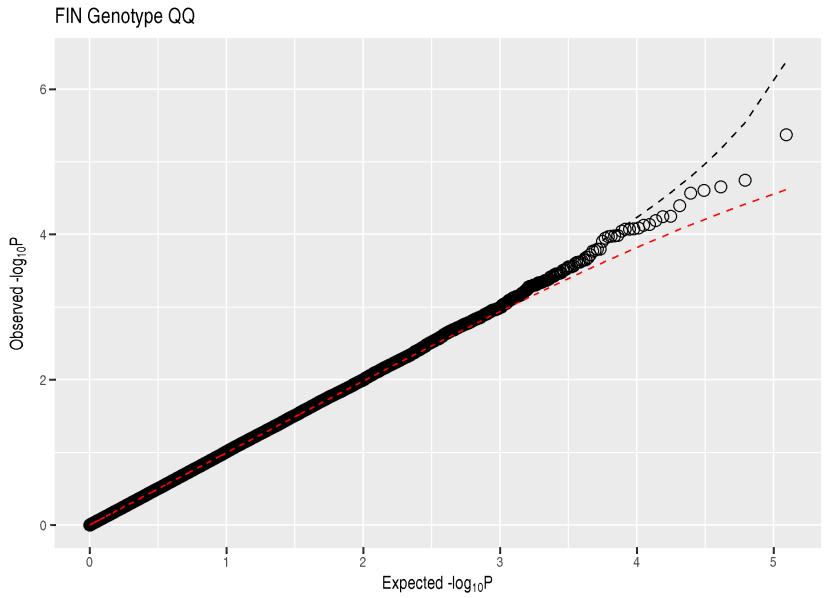

Figure 4.8: Q-Q plot for FIN population geno data

ACB

The genotyping rate for ACB population (n=77) was 0.9844 for ≈ 2.5 million variants. The Hardy-Weinberg equilibrium filter removed 127,613 variants, 855,929

variants were removed due to the minor allele threshold and 38,345 variants were removed due to low genotyping error. The total number of variants carried through to analysis was ≈ 1.4 million variants. The reported genome inflation estimate λ based on the median χ^2 was 1.00331. with a genotyping rate of 0.995935. As with the FIN genotyping data no variants were significant after Bonferroni correction. The distribution of p-values is evenly distributed across the genome, (fig 4.9). All observed p-values fall within the 95% confidence intervals for the expected distribution, (fig 4.10).

Table 4.5: ACB population genotype adjusted association results

	CHR	SNP	UNADJ	GC	BONF	HOLM	SIDAK_SS	SIDAK_SD	FDR_BH	FDR_BY
1	10	SNP10-130688336	1.45E-06	1.50E-06	0.416	4.16E-01	3.40E-01	3.40E-01	0.416	1
2	9	rs1448579	5.46E-06	5.63E-06	1.000	1.00E+00	7.91E-01	7.91E-01	0.710	1
3	18	rs8094791	8.17E-06	8.41E-06	1.000	1.00E+00	9.04E-01	9.04E-01	0.710	1
4	8	SNP8-10047362	9.89E-06	1.02E-05	1.000	1.00E+00	9.41E-01	9.41E-01	0.710	1
5	6	SNP6-130876525	1.49E-05	1.53E-05	1.000	1.00E+00	9.86E-01	9.86E-01	0.770	1
6	15	rs2045325	2.10E-05	2.16E-05	1.000	1.00E+00	9.98E-01	9.98E-01	0.770	1
7	2	rs2256763	3.10E-05	3.18E-05	1.000	1.00E+00	1.00E+00	1.00E+00	0.770	1
8	3	SNP3-6187663	3.14E-05	3.22E-05	1.000	1.00E+00	1.00E+00	1.00E+00	0.770	1
9	3	SNP3-55593916	3.72E-05	3.81E-05	1.000	1.00E+00	1.00E+00	1.00E+00	0.770	1
10	13	rs9805596	3.74E-05	3.83E-05	1.000	1.00E+00	1.00E+00	1.00E+00	0.770	1

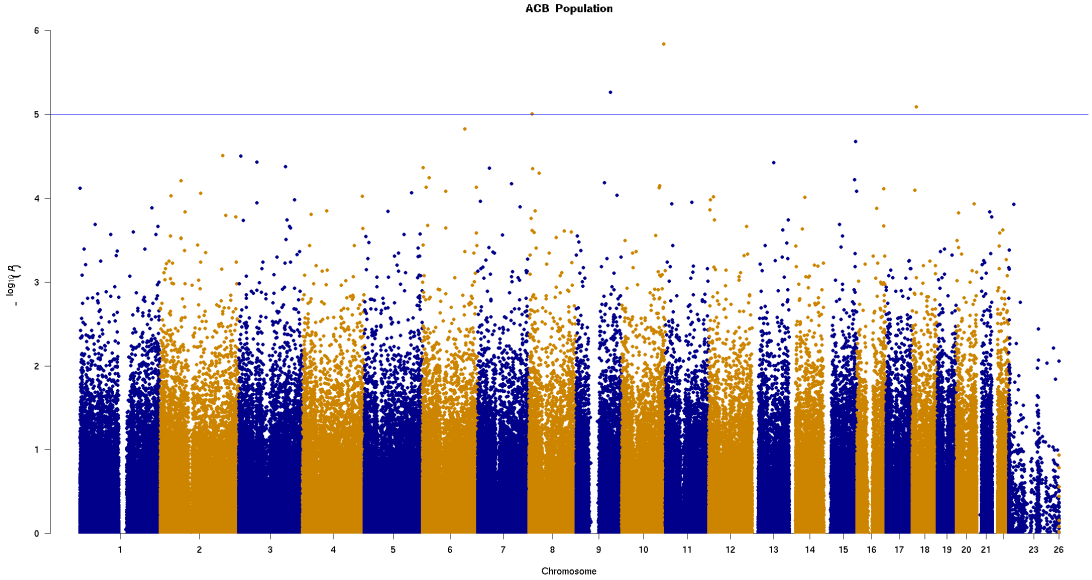

Figure 4.9: Manhattan plot for ACB population genotype data

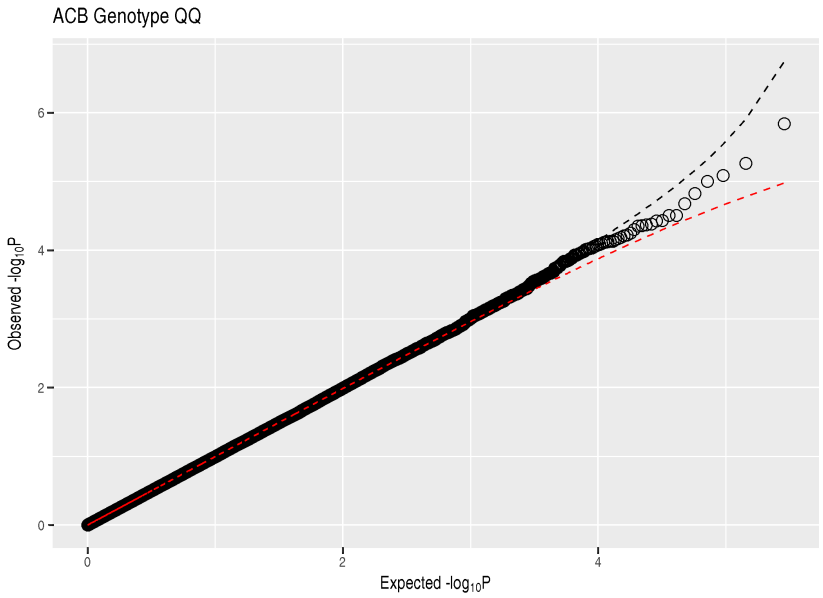

Figure 4.10: Q-Q plot for ACB population geno data

IBS

The total number of variants for analysis was ≈ 2.4 million with a total genotyping rate of 0.985415 for 100 individuals (n=100). The genotype missingness filter

removed 38,163 variants, the Hardy-Weinberg equilibrium filter removed 104,834 variants while the minor allele frequency threshold removed ≈ 1.2 million variants to be carried through for analysis. The genotyping rate improved to 0.996499. The reported genome inflation estimate λ based on the median χ^2 was 1.00964. No variants were significant after multiple testing. The p-value distribution is uniform and evenly spread across the set of chromosomes, (fig 4.11). The observed p-value distribution does not show any significant value enrichment and all values remain inside the 95% confidence interval of the expected p-value distribution, (fig 4.12).

Table 4.6: IBS population genotype adjusted association results

	CHR	SNP	UNADJ	GC	BONF	HOLM	SIDAK_SS	SIDAK_SD	FDR_BH	FDR_BY
1	13	SNP13-97219342	3.82E-06	4.19E-06	0.527	5.27E-01	4.10E-01	4.10E-01	0.512	1
2	7	rs922411	7.83E-06	8.55E-06	1.000	1.00E+00	6.60E-01	6.60E-01	0.512	1
3	2	SNP2-4127309	1.82E-05	1.97E-05	1.000	1.00E+00	9.18E-01	9.18E-01	0.512	1
4	9	rs12351382	2.19E-05	2.38E-05	1.000	1.00E+00	9.51E-01	9.51E-01	0.512	1
5	3	SNP3-50102449	2.79E-05	3.03E-05	1.000	1.00E+00	9.79E-01	9.79E-01	0.512	1
6	14	SNP14-86698288	2.83E-05	3.06E-05	1.000	1.00E+00	9.80E-01	9.80E-01	0.512	1
7	16	rs2908792	2.93E-05	3.17E-05	1.000	1.00E+00	9.83E-01	9.83E-01	0.512	1
8	8	rs7003556	3.12E-05	3.37E-05	1.000	1.00E+00	9.87E-01	9.87E-01	0.512	1
9	21	rs2403729	3.68E-05	3.98E-05	1.000	1.00E+00	9.94E-01	9.94E-01	0.512	1
10	18	SNP18-75310964	3.96E-05	4.27E-05	1.000	1.00E+00	9.96E-01	9.96E-01	0.512	1

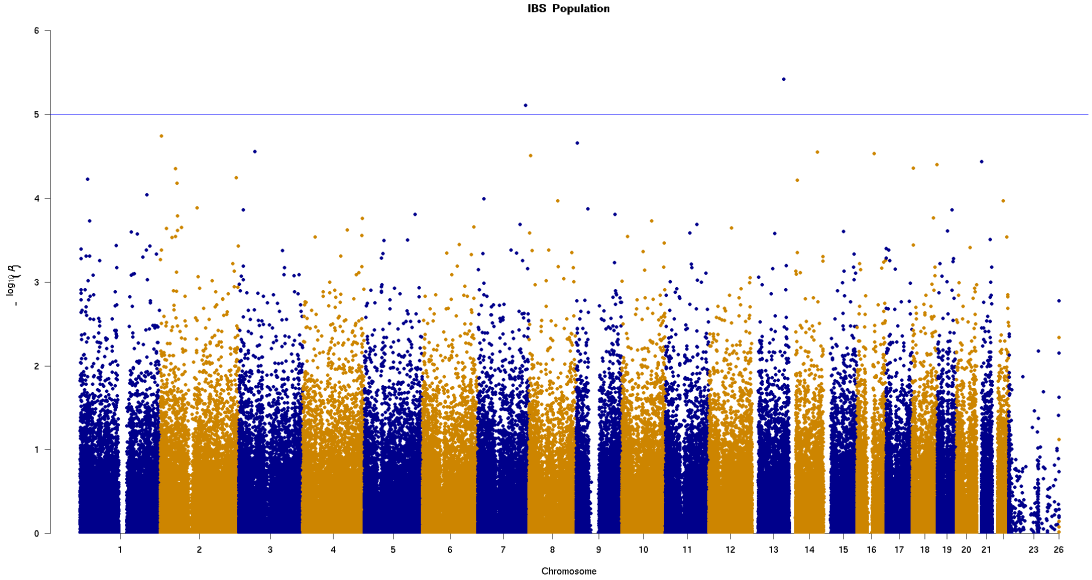

Figure 4.11: Manhattan plot for IBS population genotype data

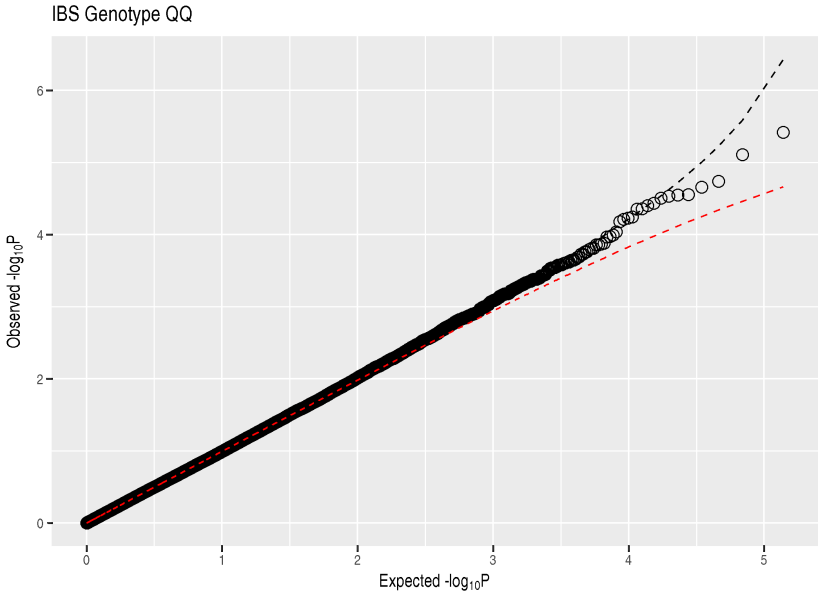

Figure 4.12: Q-Q plot for IBS population geno data

TSI

The genotyping rate of the raw TSI raw genotyping data was 0.975967 with 108 individuals (n=108) and ≈ 2.4 million variants before quality control and filtering.

Ten individuals were removed due to genotype missingness (n=98) > 10%. The Hardy-Weinberg equilibrium threshold removed 102,050 variants, the minor allele frequency threshold removed ≈ 1.1 million variants and 38,500 variants removed due to genotype missingness. The total number of variants carried forward for analysis was ≈ 1.2 million variants. The reported genome inflation estimate λ based on the median χ^2 was 1.01995. After removing 10 individuals and variants that did not pass quality control, the genotyping rate 0.997 SNP2-150826890 corresponds to rs7598384 an intron variant in AC016682.1-001 a long non-coding RNA expressed in the liver and testes. The overall spread of p-values across the genome is uniform, (fig 4.13). The observed p-value distribution does however deviate from the expected distribution showing enrichment for values below 0.005, (fig 4.14).

Table 4.7: TSI population genotype adjusted association results

	CHR	SNP	UNADJ	GC	BONF	HOLM	SIDAK_SS	SIDAK_SD	FDR_BH	FDR_BY
1	2	SNP2-150826890	3.16E-08	4.07E-08	0.004	4.33E-03	4.32E-03	4.32E-03	0.004	0.054
2	5	rs2113090	1.40E-06	1.72E-06	0.192	1.92E-01	1.74E-01	1.74E-01	0.096	1.000
3	7	SNP7-154215924	2.98E-06	3.61E-06	0.408	4.08E-01	3.35E-01	3.35E-01	0.111	1.000
4	5	SNP5-176057587	3.24E-06	3.93E-06	0.444	4.44E-01	3.58E-01	3.58E-01	0.111	1.000
5	8	SNP8-52118288	5.04E-06	6.07E-06	0.690	6.90E-01	4.98E-01	4.98E-01	0.138	1.000
6	16	SNP16-26635965	7.84E-06	9.39E-06	1.000	1.00E+00	6.58E-01	6.58E-01	0.179	1.000
7	9	rs290213	1.79E-05	2.12E-05	1.000	1.00E+00	9.14E-01	9.14E-01	0.339	1.000
8	4	rs7677465	1.98E-05	2.34E-05	1.000	1.00E+00	9.34E-01	9.34E-01	0.339	1.000
9	21	SNP21-36983242	2.34E-05	2.76E-05	1.000	1.00E+00	9.59E-01	9.59E-01	0.356	1.000
10	10	SNP10-5975482	3.19E-05	3.75E-05	1.000	1.00E+00	9.87E-01	9.87E-01	0.366	1.000

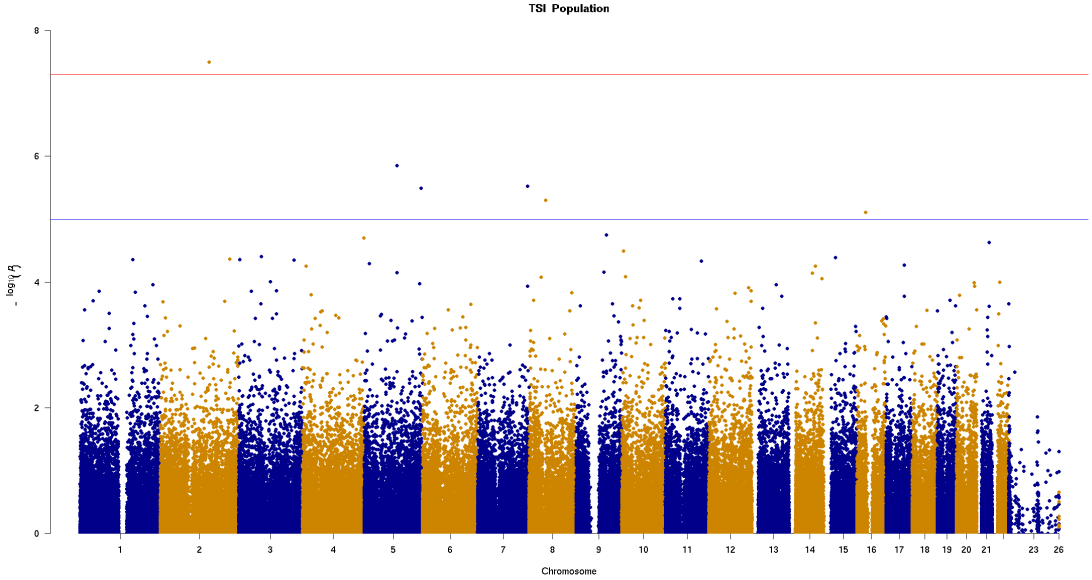

Figure 4.13: Manhattan plot for TSI population genotype data

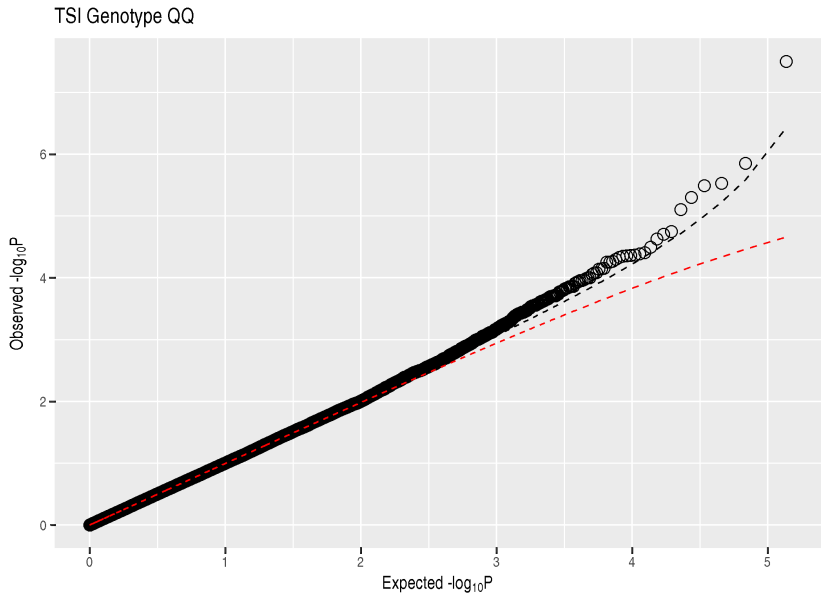

Figure 4.14: Q-Q plot for TSI population geno data

ASW

The total number of variants available for analysis was ≈ 2.45 million with 60 individuals (n=60). The genotyping rate before quality control filtering was 0.984.

One individual was removed due to having a global genotype missingness of >10%, (n=59). Hardy-Weinberg equilibrium excluded 115,985 variants, the minor allele frequency threshold removed a further 829,610 variants while 39,697 variants were removed due to genotype missingness. The total number of variants carried through to analysis was ≈ 1.5 million variants with a total genotype rate of 0.9974. The reported genome inflation estimate λ based on the median χ^2 was 1.0056. rs6512146 was the sole variant to show significance after multiple testing correction, (table 4.8). rs6512146 is an intronic snp in the C3 & PZP like , alpha-2-macroglobulin domain containing PROTEIN 8, *CPAMD8* or *VIP*. Complement proteins are integral to the innate and acquired immune response as well as having function in damage control [54]. *CPAMD8* variants have been previously shown to be associated with anterior segment dysgenesis (ASD) a disorder effecting the anterior segment of the eye [55]. Although the precise function of *CPAMD8* remains unknown, the STRING database provides evidence for interaction between *CPAMD8* and andenylate cyclase activating polypeptide 1 *ADCYAP1*, although evidence for this interaction pre-dates the characterization of *CPAMD8* by 12 years [56]. The distribution of p-values across the genome is evenly spread across the chromosomes, (fig 4.15). The Q-Q plot mirrors that of the TSI population in the there is some enrichment for the significant observed p-values, (fig 4.16).

Table 4.8: ASW population genotype adjusted association results

	CHR	SNP	UNADJ	GC	BONF	HOLM	SIDAK_SS	SIDAK_SD	FDR_BH	FDR_BY
1	19	rs6512146	1.44E-09	1.54E-09	3.97E-04	3.97E-04	3.97E-04	3.97E-04	3.97E-04	5.21E-03
2	2	SNP2-76240545	1.44E-07	1.52E-07	3.96E-02	3.96E-02	3.88E-02	3.88E-02	1.98E-02	2.59E-01
3	4	SNP4-43659731	1.68E-06	1.76E-06	4.62E-01	4.62E-01	3.70E-01	3.70E-01	1.51E-01	1.00E+00
4	8	rs16900312	2.20E-06	2.31E-06	6.05E-01	6.05E-01	4.54E-01	4.54E-01	1.51E-01	1.00E+00
5	19	rs12610057	2.78E-06	2.91E-06	7.65E-01	7.65E-01	5.35E-01	5.35E-01	1.53E-01	1.00E+00

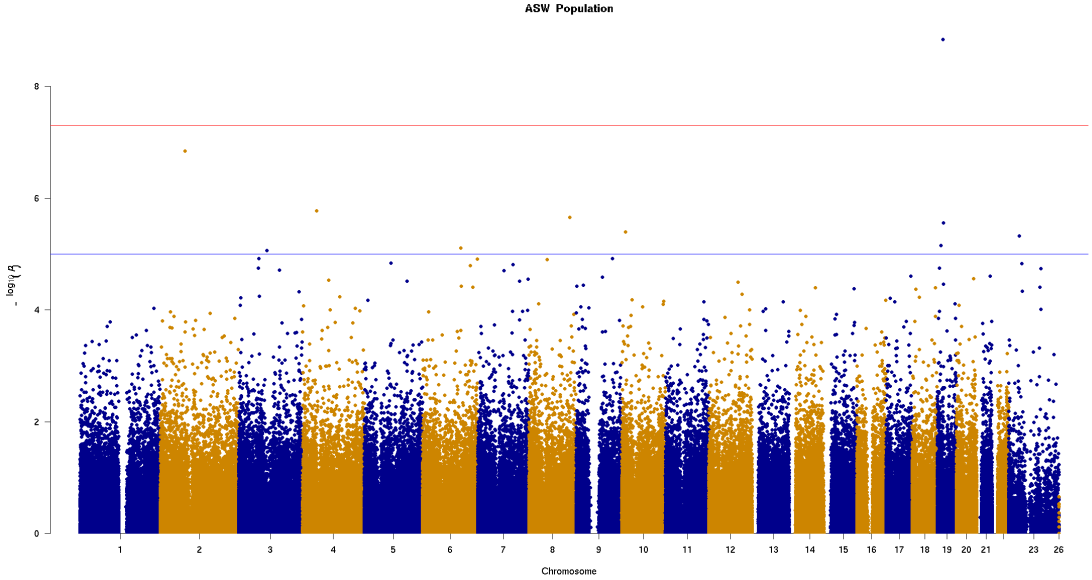

Figure 4.15: Manhattan plot for ASW population genotype data

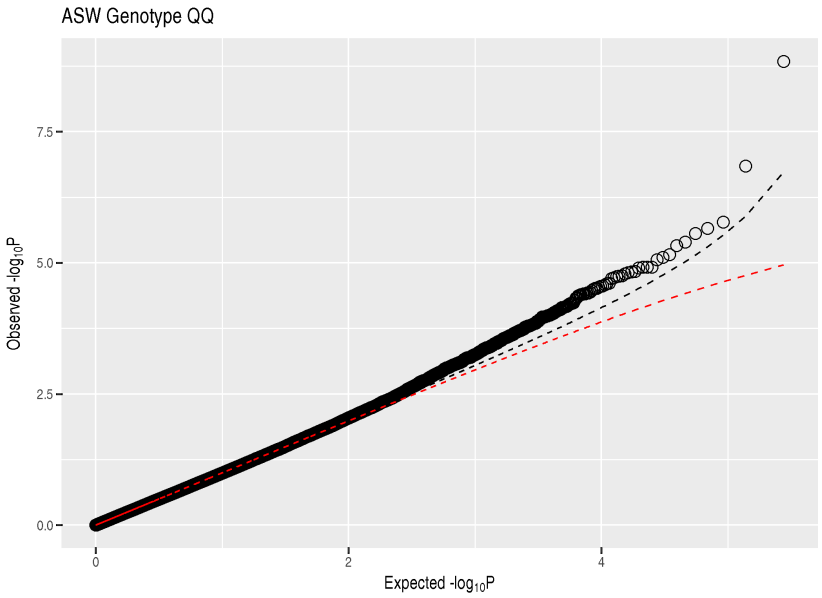

Figure 4.16: Q-Q plot for ASW population geno data

4.1.3 Log transformed genotype data

The initial run of the association study for the log transformed data contained 3 populations CLM, CHB & LWK which showed excessive significant results. Examining the phenotype distribution showed that these populations contained a mutation rate = 0; samples HG00557, NA18552 & NA19473 respectfully. These 3 individuals caused extreme outliers in their respective datasets and were subsequently removed from the association study.

FIN

The total number of variants available for analysis was $\approx$ 2.45 million with a sample size of 96 (n=96). The initial genotyping rate was 0.986. Hardy-Weinberg equilibrium filtering removed 95,350 variants, the minor allele frequency threshold removed $\approx$ 1.1 million variants while 35,900 variants were removed due to missing genotyping data. The total number of variants carried forward for analysis was $\approx$ 1.2 million. LD pruning removed $\approx$ 1.1 million variants leaving 123,551 variants to be analysed in the association study. The reported genome inflation estimate λ based on the median χ^2 was 1.0133 with a genotyping rate of 0.997. No variants were significant after multiple correction, (table 4.9). The low end of the p-values are distributed uniformly across chromosomes, (fig 4.17), while the observed distribution of p-values does not deviate from the expected distribution, (fig 4.18).

Table 4.9: FIN population log transformed adjusted association results

	CHR	SNP	UNADJ	GC	BONF	HOLM	SIDAK_SS	SIDAK_SD	FDR_BH	FDR_BY
1	11	SNP11-83309647	4.91E-06	5.56E-06	6.06E-01	6.06E-01	4.55E-01	4.55E-01	6.06E-01	1.00E+00
2	9	SNP9-36141164	3.50E-05	3.90E-05	1.00E+00	1.00E+00	9.87E-01	9.87E-01	9.28E-01	1.00E+00
3	4	rs8180202	3.59E-05	4.00E-05	1.00E+00	1.00E+00	9.88E-01	9.88E-01	9.28E-01	1.00E+00
4	7	rs218148	4.01E-05	4.45E-05	1.00E+00	1.00E+00	9.93E-01	9.93E-01	9.28E-01	1.00E+00
5	3	SNP3-38717393	5.32E-05	5.89E-05	1.00E+00	1.00E+00	9.99E-01	9.99E-01	9.28E-01	1.00E+00
6	14	SNP14-86945790	6.41E-05	7.09E-05	1.00E+00	1.00E+00	1.00E+00	1.00E+00	9.28E-01	1.00E+00
7	22	rs1984519	6.72E-05	7.43E-05	1.00E+00	1.00E+00	1.00E+00	1.00E+00	9.28E-01	1.00E+00
8	5	rs6580215	7.06E-05	7.80E-05	1.00E+00	1.00E+00	1.00E+00	1.00E+00	9.28E-01	1.00E+00
9	7	SNP7-14159197	8.55E-05	9.42E-05	1.00E+00	1.00E+00	1.00E+00	1.00E+00	9.28E-01	1.00E+00
10	11	SNP11-18659575	1.05E-04	1.16E-04	1.00E+00	1.00E+00	1.00E+00	1.00E+00	9.28E-01	1.00E+00

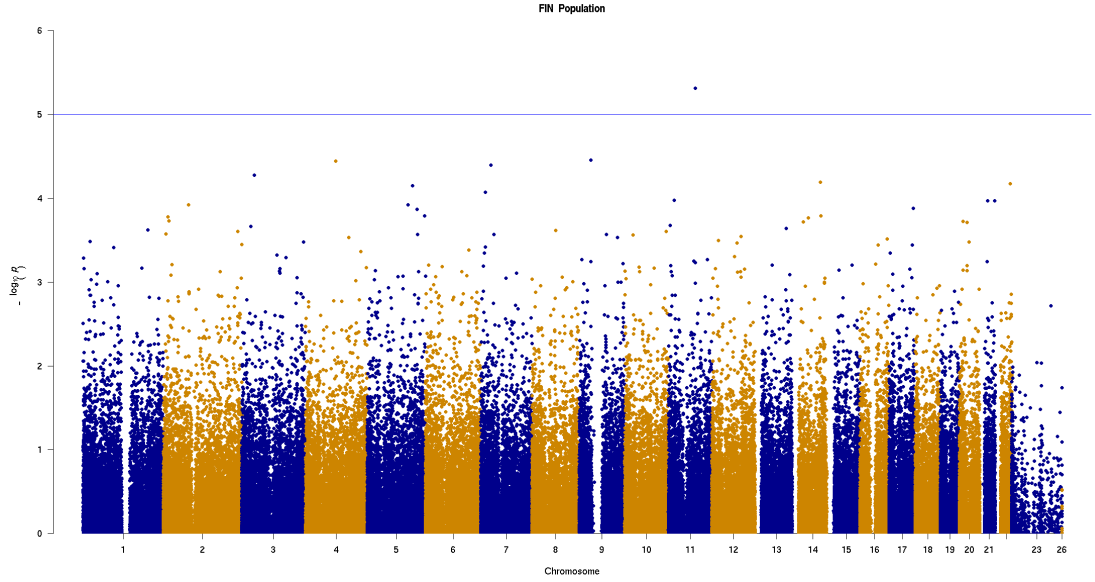

Figure 4.17: Manhattan plot for FIN population log transformed genotype data

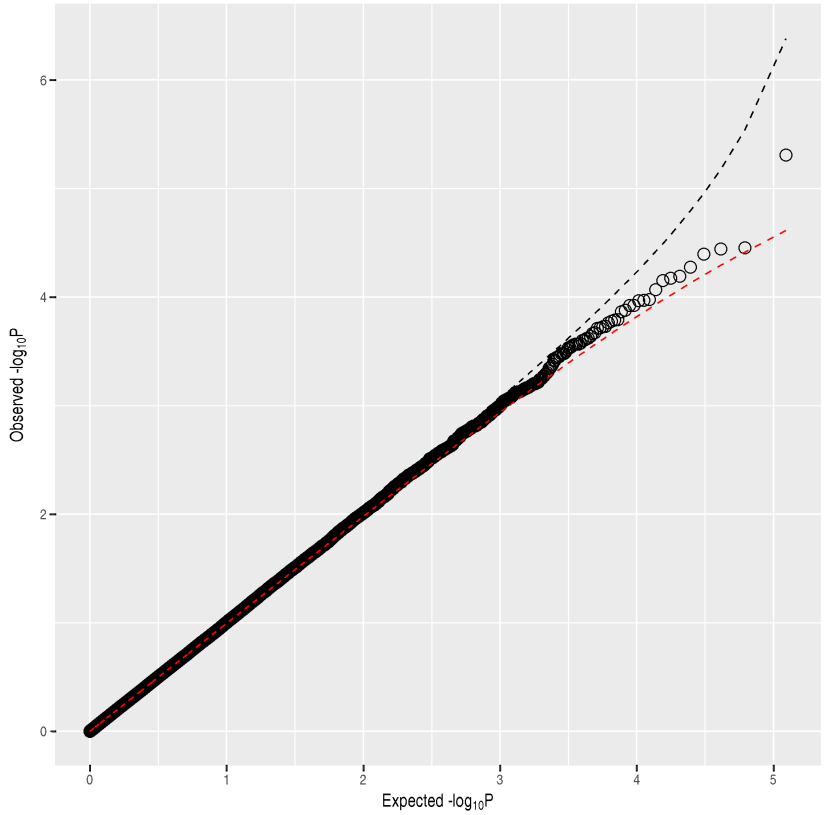

Figure 4.18: Q-Q plot for FIN population log transformed genotype data

ACB

The number of variants available for processing was ≈ 2.45 million with a genotyping rate of 0.984. The minor allele frequency threshold removed 855,929 variants, the Hardy-Weinberg equilibrium constraint filtered 127,613 variants while 38,345 variants were removed due to missing genotype data. The total number of variants carried forward after quality control was ≈ 1.4 million variants and of these, 286,790 remained after LD pruning and were analysed. The reported genome inflation estimate λ based on the median χ^2 was 1.00256. The genotyping rate improved to 0.995. No variant was below the significance threshold after Bonferroni correction, (table 4.10). The variant showing the highest association was rs1448579, a chromosome 9 intron variant in the long non coding RNA *LINC01492*. The spread of variants is uniform across all chromosomes as shown in (fig 4.19). The Q-Q plot shows all variant remain within the 95% confidence limits of the expected distribution, (fig 4.20).

Table 4.10: ACB population log transformed adjusted association results

	CHR	SNP	UNADJ	GC	BONF	HOLM	SIDAK_SS	SIDAK_SD	FDR_BH	FDR_BY
1	9	rs1448579	2.56E-07	2.63E-07	7.34E-02	7.34E-02	7.08E-02	7.08E-02	7.34E-02	9.65E-01
2	13	rs9805596	6.02E-06	6.16E-06	1.00E+00	1.00E+00	8.22E-01	8.22E-01	5.82E-01	1.00E+00
3	10	SNP10-130688336	6.08E-06	6.22E-06	1.00E+00	1.00E+00	8.25E-01	8.25E-01	5.82E-01	1.00E+00
4	8	SNP8-10047362	8.30E-06	8.49E-06	1.00E+00	1.00E+00	9.07E-01	9.07E-01	5.95E-01	1.00E+00
5	4	rs1454218	2.04E-05	2.09E-05	1.00E+00	1.00E+00	9.97E-01	9.97E-01	7.93E-01	1.00E+00
6	6	SNP6-130876525	2.17E-05	2.21E-05	1.00E+00	1.00E+00	9.98E-01	9.98E-01	7.93E-01	1.00E+00
7	15	rs2045325	2.39E-05	2.44E-05	1.00E+00	1.00E+00	9.99E-01	9.99E-01	7.93E-01	1.00E+00
8	22	SNP22-22255335	2.51E-05	2.57E-05	1.00E+00	1.00E+00	9.99E-01	9.99E-01	7.93E-01	1.00E+00
9	8	SNP8-3256254	2.76E-05	2.81E-05	1.00E+00	1.00E+00	1.00E+00	1.00E+00	7.93E-01	1.00E+00
10	1	SNP1-115178922	3.03E-05	3.09E-05	1.00E+00	1.00E+00	1.00E+00	1.00E+00	7.93E-01	1.00E+00

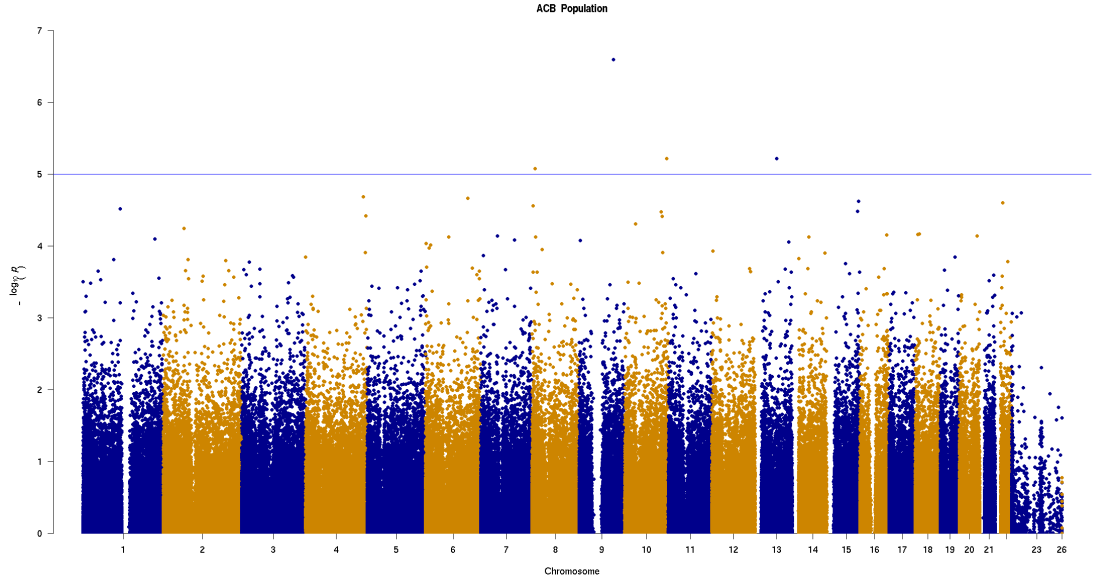

Figure 4.19: Manhattan plot for ACB population log transformed genotype data

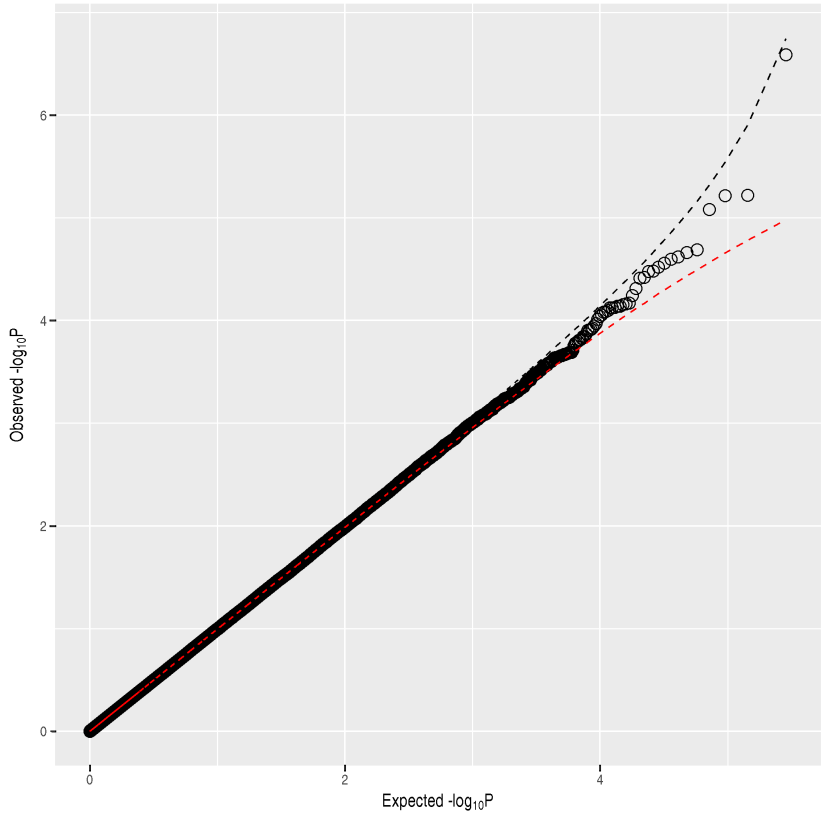

Figure 4.20: Q-Q plot for ACB population log transformed genotype data

IBS

The number of available variants was ≈ 2.45 million with 100 individuals (N=100). The initial genotyping rate was 0.98. The Hardy-Weinberg equilibrium filter removed 104,834 variants, the genotype missingness filtered 38,163 while ≈ 1 million variants. LD pruning removed a further ≈ 1 million variants leaving 138,034 variant to be carried through to analysis. The reported genome inflation estimate λ based on the median χ^2 was 1 with a genotyping rate of 0.996. As with the above population, no variants were deemed to be significantly associated with the phenotype, (table 4.11). Variants were distributed evenly across the genome and the observed p-value distribution did not deviate from the expected distribution of p-values, (figs 4.21 & 4.22).

Table 4.11: IBS population log transformed adjusted association results

	CHR	SNP	UNADJ	GC	BONF	HOLM	SIDAK_SS	SIDAK_SD	FDR_BH	FDR_BY
1	2	rs6744195	4.00E-06	4.00E-06	5.52E-01	5.52E-01	4.24E-01	4.24E-01	5.52E-01	1.00E+00
2	8	SNP8-2779994	1.10E-05	1.10E-05	1.00E+00	1.00E+00	7.81E-01	7.81E-01	6.87E-01	1.00E+00
3	7	SNP7-130565354	3.27E-05	3.27E-05	1.00E+00	1.00E+00	9.89E-01	9.89E-01	6.87E-01	1.00E+00
4	17	SNP17-9102114	3.53E-05	3.53E-05	1.00E+00	1.00E+00	9.92E-01	9.92E-01	6.87E-01	1.00E+00
5	22	rs7289667	3.71E-05	3.71E-05	1.00E+00	1.00E+00	9.94E-01	9.94E-01	6.87E-01	1.00E+00
6	18	rs11873533	3.73E-05	3.73E-05	1.00E+00	1.00E+00	9.94E-01	9.94E-01	6.87E-01	1.00E+00
7	10	SNP10-17723667	4.09E-05	4.09E-05	1.00E+00	1.00E+00	9.97E-01	9.97E-01	6.87E-01	1.00E+00
8	16	rs2908792	4.67E-05	4.67E-05	1.00E+00	1.00E+00	9.98E-01	9.98E-01	6.87E-01	1.00E+00
9	6	SNP6-152242332	4.89E-05	4.89E-05	1.00E+00	1.00E+00	9.99E-01	9.99E-01	6.87E-01	1.00E+00
10	10	rs7088963	4.98E-05	4.98E-05	1.00E+00	1.00E+00	9.99E-01	9.99E-01	6.87E-01	1.00E+00

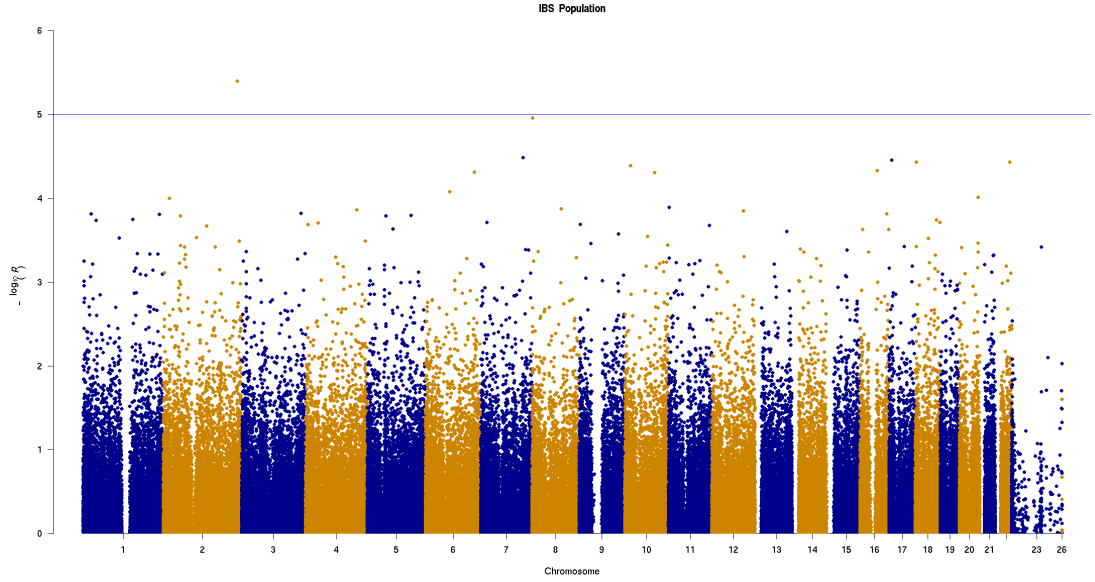

Figure 4.21: Manhattan plot for IBS population log transformed genotype data

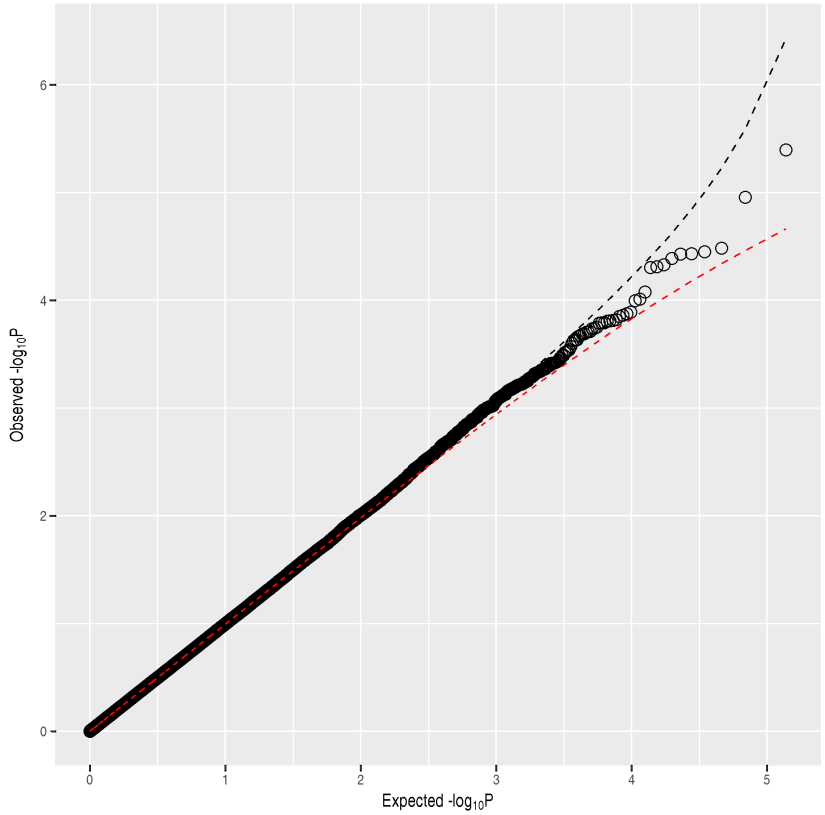

Figure 4.22: Q-Q plot for IBS population log transformed genotype data

4.2 Meta analysis

4.2.1 Low coverage WGS meta analysis

The total number of variants analysed in the meta analysis was 2544719. Two parameters were chosen to deem each variant significant. Firstly, a genome wide p value cut off of $1e-08$ and a Cochran's Q value > 0.1. Of the total number of variants 24 satisfied the above criteria, table 4.12. The Beta value shows the regression coefficients. Negative Beta values are caused by the effect being caused by the minor allele. A list of biologically relevant associations is given below.

The variant with the highest association was rs547104732, a non coding transcript exon variant in *PTCD3*, a gene that plays a role in mitochondrial translation. rs541262090 is a 5' UTR variant in *RALGAPA1* which activates Ras-like small GTPases RALA and RALB. rs201936201 is an intron variant in the *ATF1* gene. *ATF1* plays several roles in positive regulation of transcription and has been implicated in the sarcoma, fibrous histiocytoma [57]. The chromosome 19 variant, rs200114728, encodes a missense variant in the *ZNF845* gene. *ZNF845* may play a role in transcription regulation. This however, has only been inferred from sequence similarity and has not been proven experimentally. rs575061631 is an intron variant in the checkpoint kinase2 pseudogene *CHEK2P2*. Checkpoint kinase proteins have been previously associated with DNA damage repair [58]. A BLASTN query of the *CHEK2P2* nucleotide sequence returned a 96% similarity score to *CHK2*. Both *CHEK2P2* and *ZNF845* show high expression to over expression in the testis. rs568341410 is an intron variant in the *TCF25* transcription factor. *TCF25* has been shown to act as transcriptional repressor.

The majority of variants below the genome wide significance threshold are evenly spread across the genome. The extreme values are located on chromosomes 1, 4, 10, 11 & 13, (fig 4.23).

Table 4.12: Meta analysis for G1K 26 populations VCF dataset

	CHR	BP	SNP	N	P	P.R.	BETA	BETA.R.	Q	I
1	2	86360172	rs547104732	16	3.376E-14	1.67E-09	0.000	0.000	0.149	27.250
2	14	36278054	rs541262090	8	1.359E-13	2.26E-07	0.000	0.000	0.110	40.240
3	6	137147541	rs532273900	8	1.034E-12	2.42E-08	0.000	0.000	0.194	29.330
4	16	21982721	rs551318071	10	1.201E-12	7.18E-09	0.000	0.000	0.186	28.110
5	18	18518431	rs545537217	9	7.316E-12	3.38E-09	-0.000	-0.000	0.246	22.110
6	12	51189603	rs201936201	13	1.517E-10	6.80E-07	0.000	0.000	0.138	30.670
7	19	53856702	rs200114728	13	2.245E-10	5.35E-09	-0.000	-0.000	0.274	16.820
8	5	49432256	rs144073434	4	4.060E-10	4.06E-10	0.000	0.000	0.869	0.000
9	15	20488634	rs575061631	8	4.976E-10	5.85E-06	0.000	0.000	0.120	38.930
10	17	22253615	rs9635795	3	1.357E-09	1.36E-09	-0.000	-0.000	0.503	0.000
11	3	166579814	rs559068488	6	1.657E-09	1.66E-09	-0.000	-0.000	0.506	0.000
12	17	28678955	rs547282804	9	1.676E-09	1.68E-09	0.000	0.000	0.513	0.000
13	1	148756664	rs587742217	16	7.336E-09	7.34E-09	-0.000	-0.000	0.466	0.000
14	17	7259355	rs546814018	9	1.383E-08	1.38E-08	0.000	0.000	0.474	0.000
15	19	16611504	rs566396315	11	1.485E-08	1.49E-08	0.000	0.000	0.617	0.000
16	7	64255019	rs539883638	9	1.577E-08	1.58E-08	0.000	0.000	0.617	0.000
17	11	2224709	rs550199192	15	1.631E-08	4.16E-06	0.000	0.000	0.105	32.950
18	11	489672	rs533395899	16	2.286E-08	1.71E-07	0.000	0.000	0.321	11.560
19	7	50229186	rs17133728	3	2.603E-08	2.03E-04	0.000	0.000	0.118	53.140
20	22	21700177	rs200285191	4	3.322E-08	3.32E-08	-0.000	-0.000	0.624	0.000
21	17	71080806	rs542147764	16	3.995E-08	1.81E-05	0.000	0.000	0.131	29.190
22	16	89973821	rs568341410	19	4.015E-08	8.85E-08	-0.000	-0.000	0.382	6.020
23	1	155296518	rs571889708	10	4.049E-08	1.11E-06	0.000	0.000	0.249	21.080
24	10	97388319	rs562117646	8	4.197E-08	3.98E-05	0.000	0.000	0.140	36.160

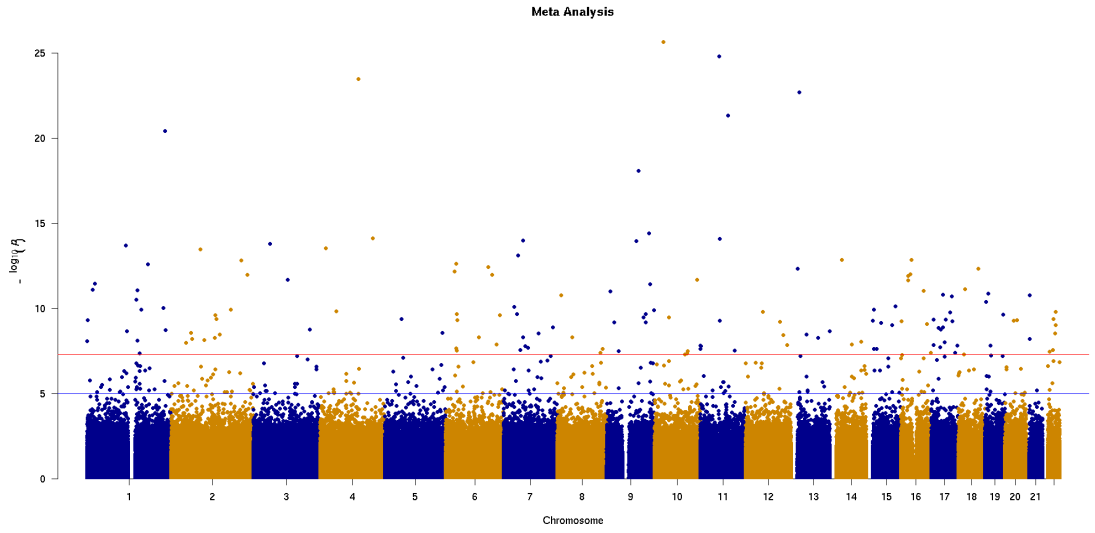

Figure 4.23: Manhattan plot of G1K meta analysis VCF dataset

4.2.2 Omni high density genotyping meta analysis

The total number of variants available for meta analysis of the genotype data was 629,898. A single variant satisfied the Cochran's Q and genome-wide significance thresholds, (table 4.13). The co-ordinates for SNP19-11067585 map to rs745167707, an intron variant, in the hg19 build. The *LDLR* gene encodes low-density lipoprotein receptor. LDLR is responsible for uptake of low-density lipoproteins into the cell.

Table 4.13: Genome wide meta-analysis of genotyped populations

	CHR	BP	SNP	N	P	P.R.	BETA	BETA.R.	Q	I
1	19	11206585	SNP19-11067585	3	4.082E-08	4.08E-08	0.000	0.000	0.565	0.000

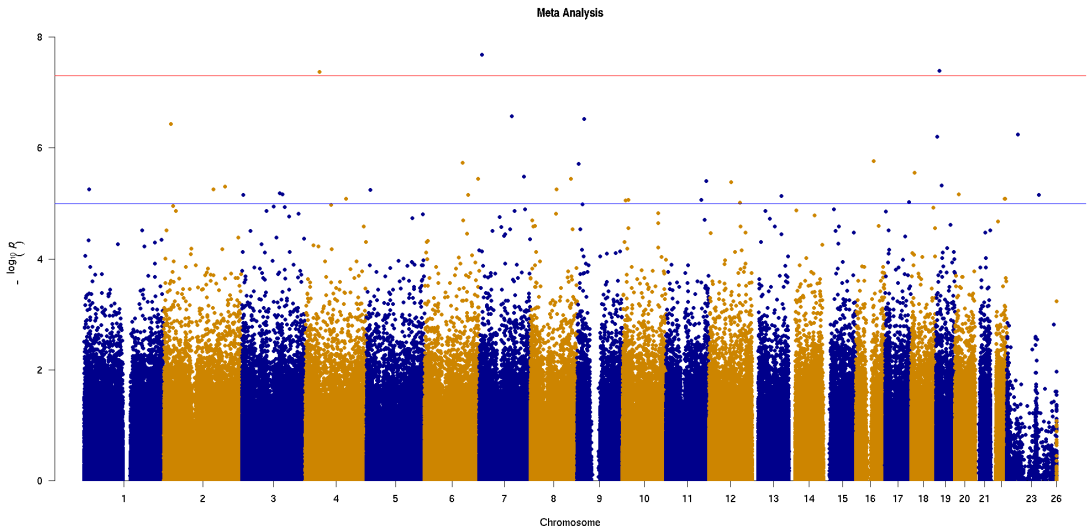

Figure 4.24: Manhattan plot of G1K meta analysis genotyping dataset

4.2.3 Log transformed phenotype meta analysis

For the log transformed genotype data 629,898 variants were analysed. As with the genotype dataset above a sole variant was significant after filtering with Cochran's Q and genome-wide significant threshold. SNP3-135070038 maps to rs13077612 on the hg19 build of the human reference genome. rs13077612 is an intron variant in Ras-related protein Rab-6B (*RAB6B*) a protein highly expressed in brain tissue and localised to the Golgi apparatus.

Table 4.14: Genome wide meta-analysis of log transformed genotyped populations

	CHR	BP	SNP	N	P	P.R.	BETA	BETA.R.	Q	I
1	3	133587348	SNP3-135070038	4	1.65E-08	1.65E-08	3.18E-01	3.18E-01	1	0

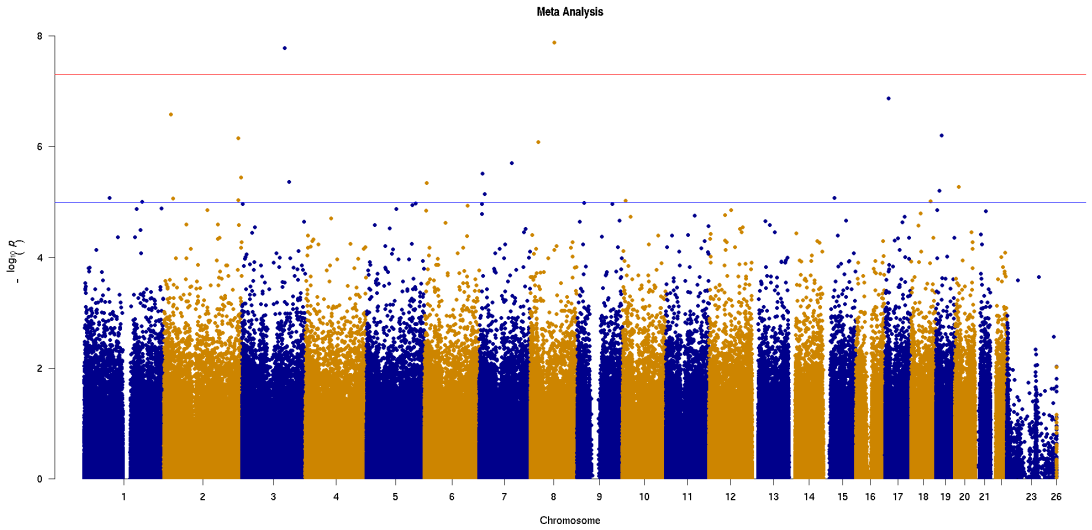

Figure 4.25: Manhattan plot of G1K meta analysis log transformed genotyping dataset

4.3 Gene set enrichment analysis

4.3.1 GO gene sets

low coverage WGS GSEA

The top 10 enriched pathways are shown in table 4.15. P_*CORR* refers to the multiple corrected p-value calculated from the permutation empirical p-value. While no pathway is significantly enriched, there are some biologically relevant pathways included in the top 10. 'DNA damage response signal transduction resulting in transcription', 'ADA2 GCN5 ADA3 transcription activator complex', 'phosphatidylinositol 3 kinase activity' and 'regulation of PI3K signalling pathways' are potential candidates for influencing somatic mutation rate, see Discussion.

Table 4.15: Gene set enrichment analysis for vcf dataset

	NGENES	BETA	BETA_STD	SE	P	P_CORR	FULL_NAME
1	7	1.21	0.03	0.306	4.01E-05	0.186	GO_OUTER_MITOCHONDRIAL_MEMBRANE_PROTEIN_COMPLEX
2	62	0.44	0.03	0.116	6.69E-05	0.282	GO_PHOSPHATIDYLINOSITOL_3_KINASE_ACTIVITY
3	10	1.06	0.03	0.305	2.55E-04	0.687	GO_DDR_SIGNAL_TRANS_RESULT_IN_TRANSCRIPTION
4	116	0.29	0.03	0.086	3.06E-04	0.750	GO_REGULATION_OF_PI3K_SIGNALING
5	11	1.04	0.03	0.304	3.29E-04	0.773	GO_POS_REG_OF_NO_SYNTHASE_PROCESS
6	7	1.21	0.03	0.369	5.06E-04	0.890	GO_PROTEIN_CHANNEL_ACTIVITY
7	644	0.12	0.03	0.038	5.56E-04	0.908	GO_PEPTIDYL_AMINO_ACID_MODIFICATION
8	108	0.28	0.03	0.090	8.74E-04	0.973	GO_INOSITOL_LIPID_MEDIATED_SIGNALING
9	10	0.89	0.02	0.292	1.13E-03	0.990	GO_PROTEIN_IMPORT_INTO_MITOCHONDRIAL_MATRIX
10	10	1.02	0.03	0.338	1.33E-03	0.995	GO_ADA2_GCN5_ADA3_TRANSCRIP_ACT_COMPLEX

Omni high density genotyping GSEA

The gene set enrichment analysis for GO terms returned ontologies relating to reproduction, development and reproductive processes; acrosome reaction, female sex differentiation and ovulation cycle process are all shown to be enriched, (table 4.16). The multi cellular organism reproduction gene set was remained significant after multiple correction with a corrected p-value of 0.02.

Table 4.16: GO enrichment analysis for genotype dataset

	NGENES	BETA	BETA_STD	SE	P	P_CORR	FULL_NAME
1	668	1.36E-01	2.64E-02	0.03	3.81E-06	0.02	GO_MULTICELLULAR_ORGANISM_REPRODUCTION
2	1150	9.59E-02	2.40E-02	0.02	3.32E-05	0.16	GO_REPRODUCTION
3	777	1.14E-01	2.37E-02	0.03	3.94E-05	0.18	GO_MULTI_ORGANISM_REPRODUCTIVE_PROCESS
4	16	7.34E-01	2.24E-02	0.19	4.15E-05	0.19	GO_REGULATION_OF_ACROSOME_REACTION
5	110	2.89E-01	2.31E-02	0.07	4.76E-05	0.21	GO_FEMALE_SEX_DIFFERENTIATION
6	628	1.18E-01	2.22E-02	0.03	8.37E-05	0.34	GO_SEXUAL_REPRODUCTION
7	30	5.60E-01	2.34E-02	0.15	9.59E-05	0.37	GO_REG_OF_VIRAL_RELEASE_HOST_CELL
8	82	3.08E-01	2.13E-02	0.08	1.01E-04	0.39	GO_OVULATION_CYCLE_PROCESS
9	36	4.36E-01	2.00E-02	0.12	1.22E-04	0.45	GO_MULTIVESICULAR_BODY
10	143	2.41E-01	2.19E-02	0.07	1.37E-04	0.49	GO_HORMONE_MED_SIGNAL_PATH

Log transformed phenotype GSEA

The top GO terms enriched for the log-transformed association study returned a greater diversity than the genotype dataset above. Although no pathway was significantly enriched for, the top enrichment was ER calcium homeostasis. Development pathways also appeared in the top ten; segmemtation, oogenesis, somitogenesis and female sex differentiation pathways, (table 4.17). Steroid hormone pathway and regulation of interferon signal pathway also featured in the top ten.

Table 4.17: GO enrichment analysis for log transformed genotype dataset

	NGENES	BETA	BETA_STD	SE	P	P_CORR	FULL_NAME
1	19	6.04E-01	2.01E-02	0.170	1.89E-04	0.60	GO_ENDOPLASMIC_RETICULUM_CA_HOMEOSTASIS
2	41	4.37E-01	2.14E-02	0.123	1.99E-04	0.62	GO_CENTROSOME_CYCLE
3	29	5.22E-01	2.15E-02	0.151	2.79E-04	0.73	GO_GALACTOSYLTRANSFERASE_ACTIVITY
4	85	3.02E-01	2.12E-02	0.089	3.23E-04	0.77	GO_SEGMENTATION
5	110	2.53E-01	2.02E-02	0.074	3.31E-04	0.78	GO_FEMALE_SEX_DIFFERENTIATION
6	18	6.54E-01	2.12E-02	0.194	3.73E-04	0.82	GO_REG_OF_URINE_VOLUME
7	65	3.17E-01	1.95E-02	0.094	3.90E-04	0.83	GO_STEROID_HORMONE_RCPT_SIGNAL_PATH
8	31	4.92E-01	2.09E-02	0.148	4.48E-04	0.87	GO_REG_TYPE_I_INTERFERON_SIGNAL_PATH
9	59	3.19E-01	1.87E-02	0.097	4.88E-04	0.88	GO_OOGENESIS
10	59	3.45E-01	2.02E-02	0.106	5.77E-04	0.92	GO_SOMITOGENESIS

4.3.2 General gene sets

low coverage WGS GSEA

For the general gene sets analysed for the low coverage WGS dataset association study, no pathways were significantly enriched after multiple testing correction, (table 4.18). Pathways associated with cell signalling, proliferation and cancer are the most frequent in the top ten. The gene set Liu TOPB1 targets are the genes that are up-regulated due to knockdown of the p53 repressor TOPB1. Innate immune response to viral recognition is also enriched, KEGG RIG I like receptor signalling pathway. The highest association is the cell signalling response to copper toxicosis.

Table 4.18: Gene set enrichment analysis for vcf dataset

	NGENES	BETA	BETA_STD	SE	P	P_CORR	FULL_NAME
1	11	1.17	0.03	0.295	3.47E-05	0.145	VANDESLUIS_COMMD1_TARGETS_GROUP_4_DN
2	38	0.58	0.03	0.159	1.51E-04	0.474	KEGG_RIG_I_LIKE_RECEPTOR_SIGNALING_PATHWAY
3	9	1.08	0.03	0.327	4.92E-04	0.863	LIU_TOPBP1_TARGETS
4	17	0.73	0.03	0.232	8.20E-04	0.964	BIOCARTA_RAS_PATHWAY
5	65	0.37	0.03	0.117	8.61E-04	0.967	KEGG_TOLL_LIKE_RECEPTOR_SIGNALING_PATHWAY
6	5	1.34	0.03	0.439	1.11E-03	0.987	MOTAMED_RESPONSE_TO_ANDROGEN_DN
7	77	0.33	0.02	0.111	1.57E-03	0.997	LEIN_CHOROID_PLEXUS_MARKERS
8	59	0.35	0.02	0.123	2.01E-03	1.000	KEGG_MELANOMA
9	20	0.63	0.02	0.219	1.98E-03	1.000	SHI_SPARC_TARGETS_UP
10	50	0.40	0.02	0.141	2.31E-03	1.000	BEGUM_TARGETS_OF_PAX3_FOXO1_FUSION_UP

Omni high density genotyping GSEA

Gene set enrichment analysis for the genotyping data showed frequent enrichment for cancer associated pathways, (table 4.19). After multiple testing correction, no pathways were deemed significant. Of the 10 pathways shown, 2 are not cancer related; 'Goering blood and HDL cholesterol qtl trans' and 'Yang MUC2 targets duodenum 6mo up'. The top enrichment 'Zucchini metastasis up' represents the top significantly differentiated genes between metastastic breast cancer and normal breast epithelium.

Table 4.19: Gene set enrichment analysis for genotype dataset

	NGENES	BETA	BETA_STD	SE	P	P_CORR	FULL_NAME
1	40	4.93E-01	2.38E-02	0.13	4.72E-05	0.20	ZUCCHI_METASTASIS_UP
2	124	2.74E-01	2.32E-02	0.07	5.27E-05	0.22	WELCSH_BRCA1_TARGETS_DN
3	301	1.69E-01	2.22E-02	0.04	8.59E-05	0.32	BLUM_RESPONSE_TO_SALIRASIB_DN
4	58	3.74E-01	2.17E-02	0.10	1.01E-04	0.36	WEIGEL_OX_STRSS_BY_HNE_AND_TBH
5	320	1.55E-01	2.09E-02	0.04	2.70E-04	0.68	RHEIN_ALL_GLUCOCORTICOID_THERAPY_DN
6	55	3.55E-01	2.01E-02	0.10	3.73E-04	0.79	ZHAN_MULTIPLE_MYELOMA_UP
7	13	7.00E-01	1.93E-02	0.21	4.02E-04	0.82	GOERING_BLD_HDL_CHOL_QTL_TRANS
8	10	8.81E-01	2.13E-02	0.26	4.05E-04	0.82	YANG_MUC2_TARGETS_DUODENUM_6MO_UP
9	5	1.23E+00	2.09E-02	0.37	4.09E-04	0.82	ZIRN_TRETINOIN_RESPONSE_WT1_DN
10	39	4.26E-01	2.03E-02	0.13	4.24E-04	0.83	LUI_THYROID_CANCER_PAX8_PPARG_DN

Log transformed phenotype GSEA

No gene set was enriched after multiple correction, (table 4.20). As with the genotype gene set enrichment analysis, cancer associated pathways dominated the enrichment analysis for the log transformed data set. The top 2 gene sets for the genotype dataset (Zucchi metastasis up & Welch BRCA1 targets) set overlapped with the log transformed data set.

Table 4.20: Gene set enrichment analysis for log transformed genotype dataset

	NGENES	BETA	BETA_STD	SE	P	P_CORR	FULL_NAME
1	124	2.84E-01	2.40E-02	0.071	3.01E-05	0.13	WELCSH_BRCA1_TARGETS_DN
2	17	7.34E-01	2.31E-02	0.184	3.38E-05	0.15	WANG_RECURRENT_LIVER_CANCER_UP
3	4	1.52E+00	2.32E-02	0.394	6.01E-05	0.24	CASTELLANO_HRAS_TARGETS_UP
4	124	2.60E-01	2.20E-02	0.069	8.67E-05	0.33	JOHNSTONE_PARVB_TARGETS_2_UP
5	40	4.67E-01	2.26E-02	0.126	1.09E-04	0.39	ZUCCHI_METASTASIS_UP
6	34	4.72E-01	2.10E-02	0.134	2.26E-04	0.63	MARZEC_IL2_SIGNALING_DN
7	20	5.97E-01	2.04E-02	0.173	2.77E-04	0.70	BILANGES_SERUM_SENSITIVE_VIA_TSC1
8	25	4.99E-01	1.91E-02	0.146	3.14E-04	0.74	PID_IL27_PATHWAY
9	68	3.14E-01	1.97E-02	0.092	3.22E-04	0.75	LIN_APC_TARGETS
10	751	9.35E-02	1.91E-02	0.028	4.90E-04	0.88	GRADE_COLON_CANCER_UP

4.4 DNA methylation age

The online calculator returns the predicted DNAm based on the methylation levels of 353 CpG sites in the genome as well as gender, mean methylation scores and information on the quality of the input methylation data. Of the 60 CEU samples, 41 overlapped with the G1K project with 22 of these having sufficient methylation levels. For the 73 YRI samples, 62 samples overlapped with G1K samples and 46 had sufficient methylation levels. Four correlation datasets were created; 2 contained the full overlapping samples with low quality samples included (YRI full & CEU full) and 2 contained the overlapping samples with low quality samples removed (YRI non-warning & CEU non-warning)[1], (figs 4.26, 4.27, 4.28 & 4.29). For the CEU full population DNAm counts and DNAm age showed the highest correlation (r=-0.714) although methylation counts and DNAm age are expected to be intrinsically linked. Somatic mutation rate did not correlate with DNAm (r=-0.0486). For the CEU non-warning subset somatic mutation counts and DNAm counts showed the highest, albeit moderate to low, correlation followed by somatic mutation rate and DNAm age with a weak correlation (r=-0.272).

Directionality of some of the correlations was subject to change between CEU and CEU non-warning datasets i,e DNAm counts and somatic mutation rate, (figs 4.26 & 4.27). Both YRI subsets showed and increased number of positive correlations. For the full YRI set, somatic mutation counts and DNAm counts had the highest correlation r=0.333 with somatic mutation rate and DNAm age maintaining negative direction and a correlation of r=-0.196. In the YRI non warning subset there is a minor increase in the correlation between somatic mutation rate and DNAm age r=-0.253 the remaining correlations remain relatively constant with the exception somatic mutation counts and DNAm which differs from r=0.0488 in the YRI full dataset to r=0.00343 in the YRI non warning dataset. Linear models with confidence intervals describing the association between somatic mutation rate and DNAm age can be seen in Figs. 4.30 to 4.33.

Linear models for somatic mutation counts and DNAm age have been included in supplementary data. The CEU population showed the largest deviance from the mean DNAm age. The mean DNAm age of the CEU full data set is 73.88 years with a range of 5.46-125.18 years while the mean DNAm age for the non warning subset is 76.56 years with a range of 50.56-108.36 years. For the full YRI dataset, the mean DNAm age is 57.3 years with a range of 15.64-109.27 years whereas the mean for the non-warning subset is 59.68 years with a range of 16.65-109.27 years. The distribution of DNAm age varies substantially between CEU and YRI populations, (figs 4.34 & 4.35). The YRI population follows a normal distribution of DNAm whereas the CEU population is non-normal distribution.

[1]Non-warning refers to samples with high quality methylation data after Epigenetic clock estimation

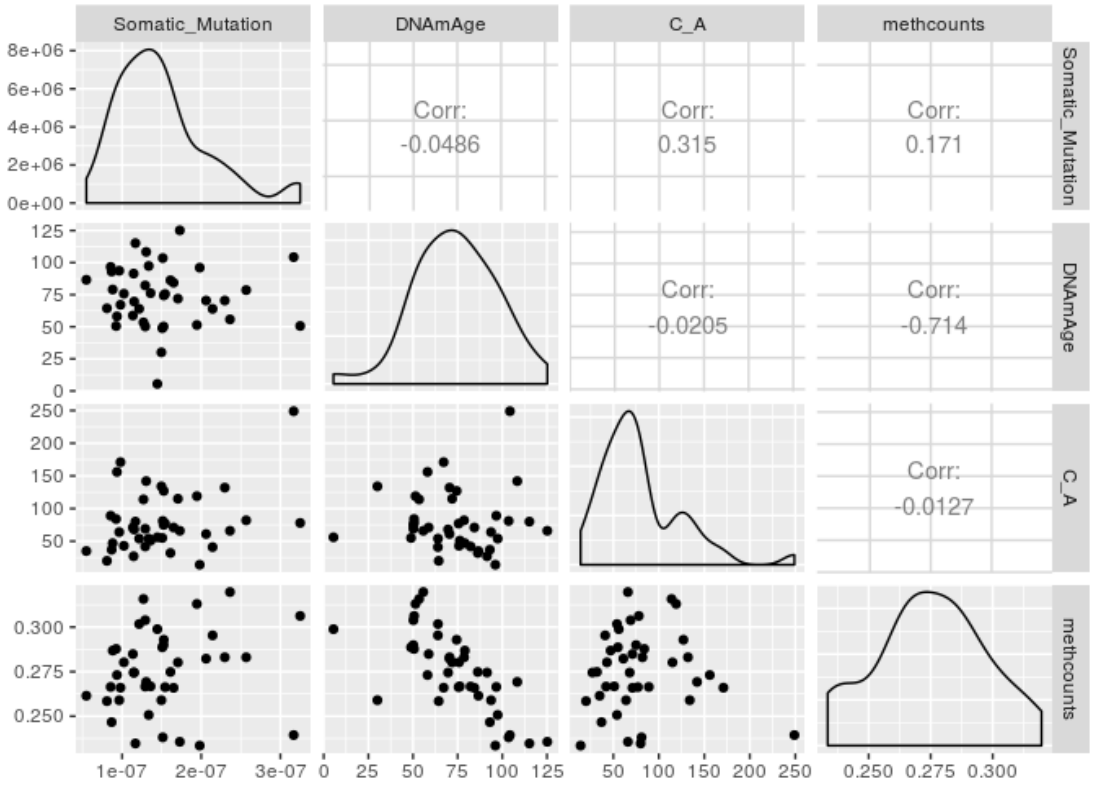

Figure 4.26: Correlation plot for the full CEU overlapping set of samples

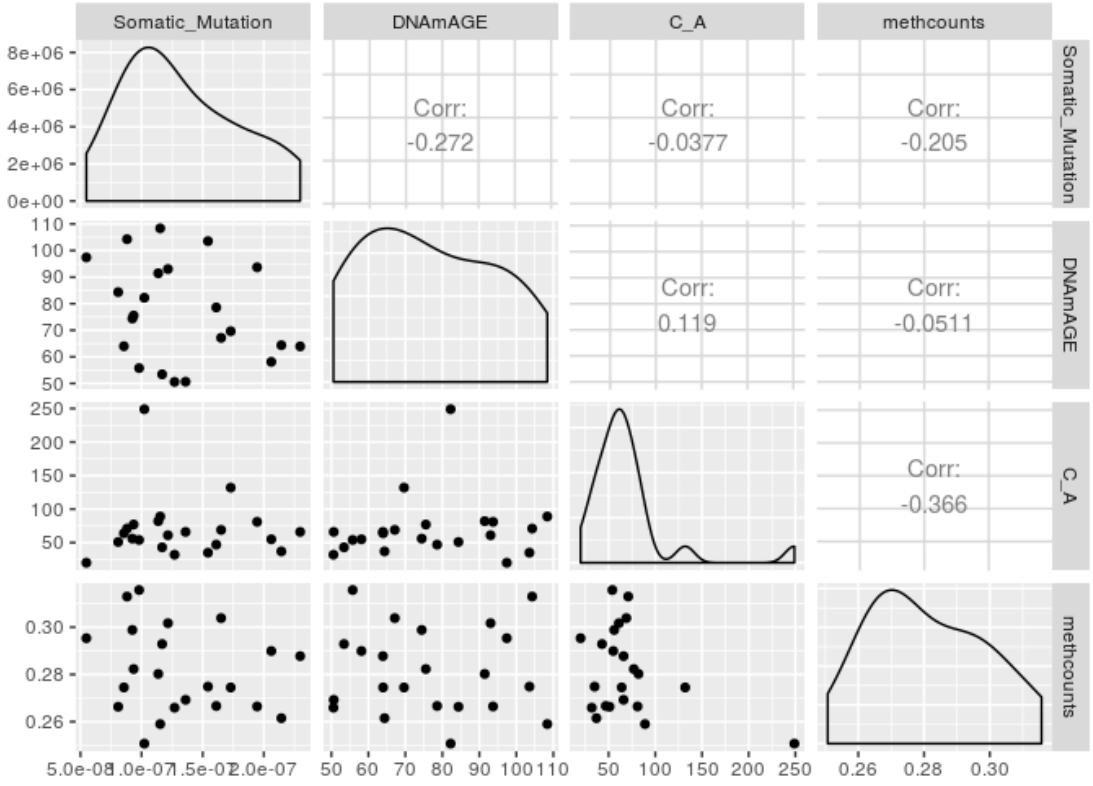

Figure 4.27: Correlation plot for the non warning CEU overlapping G1K sample

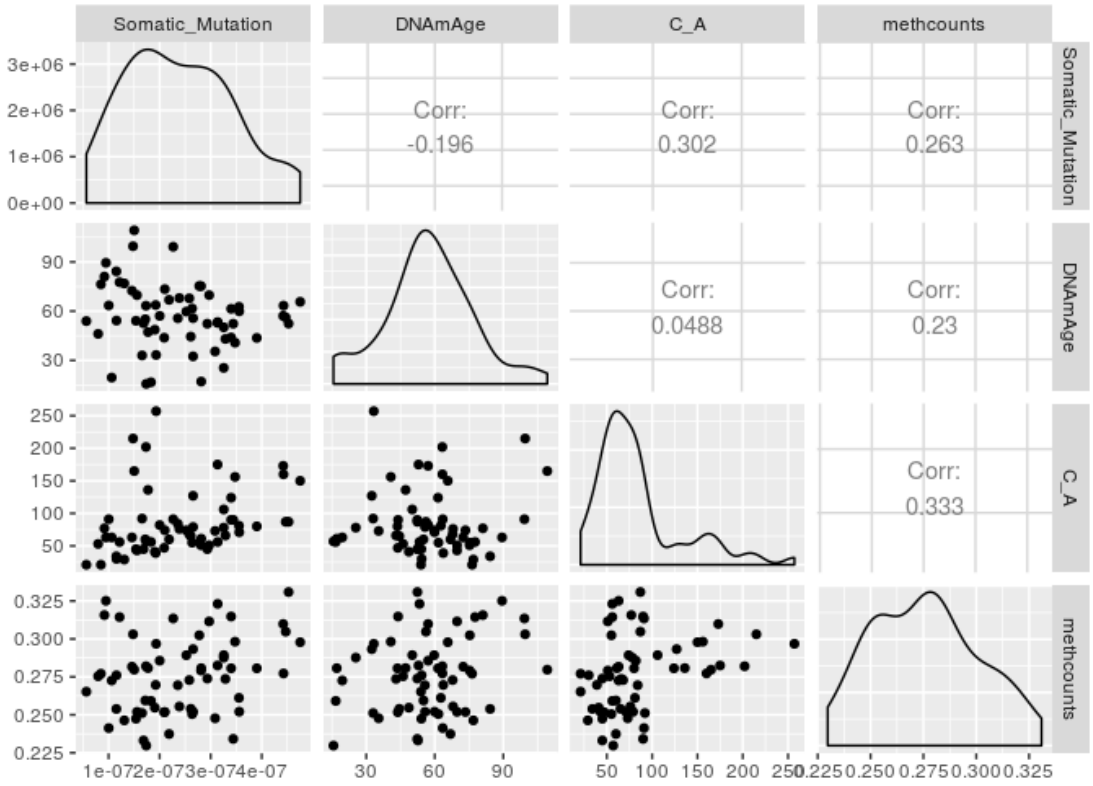

Figure 4.28: Correlation plot for the full YRI overlapping set of samples

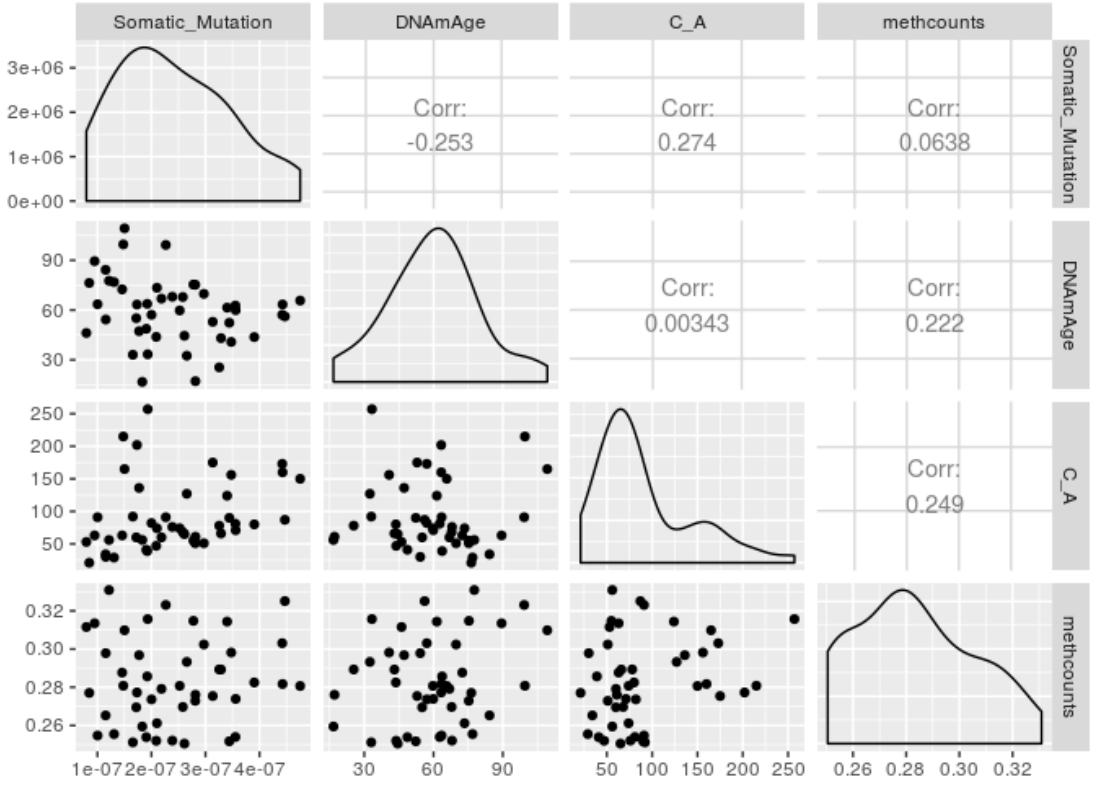

Figure 4.29: Correlation plot for the non warning YRI overlapping G1K sample

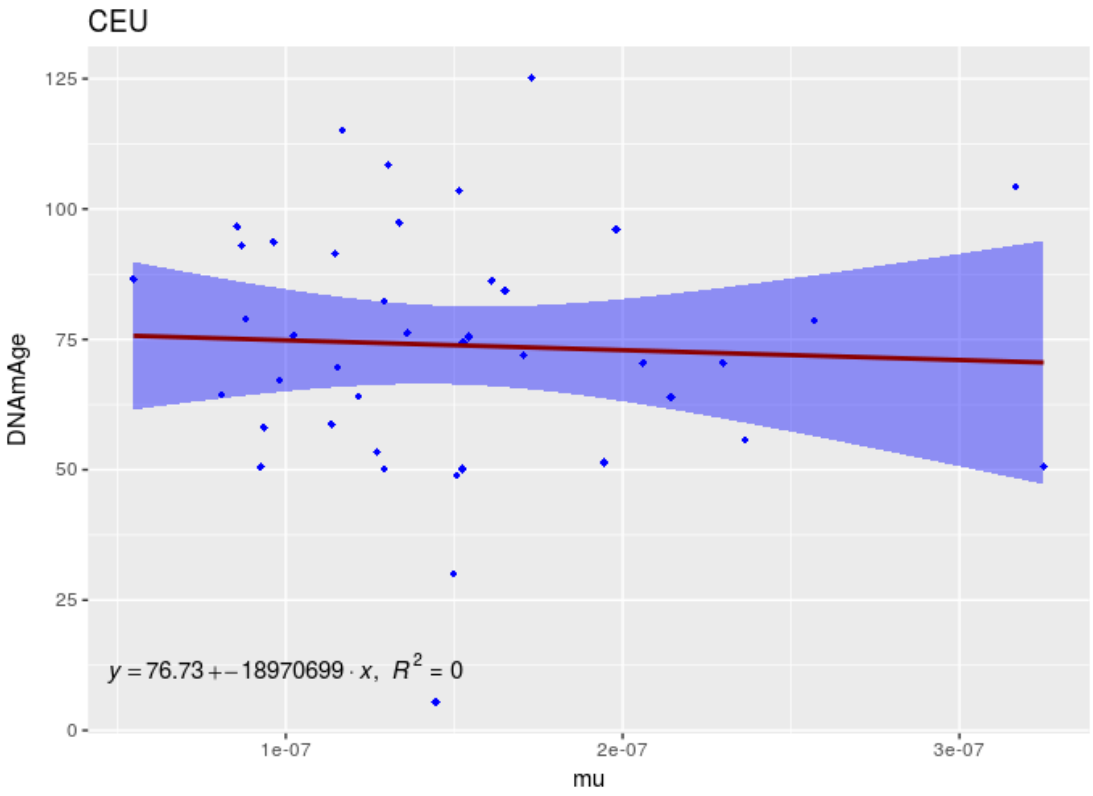

Figure 4.30: Plotted linear model for the full CEU overlapping set of samples

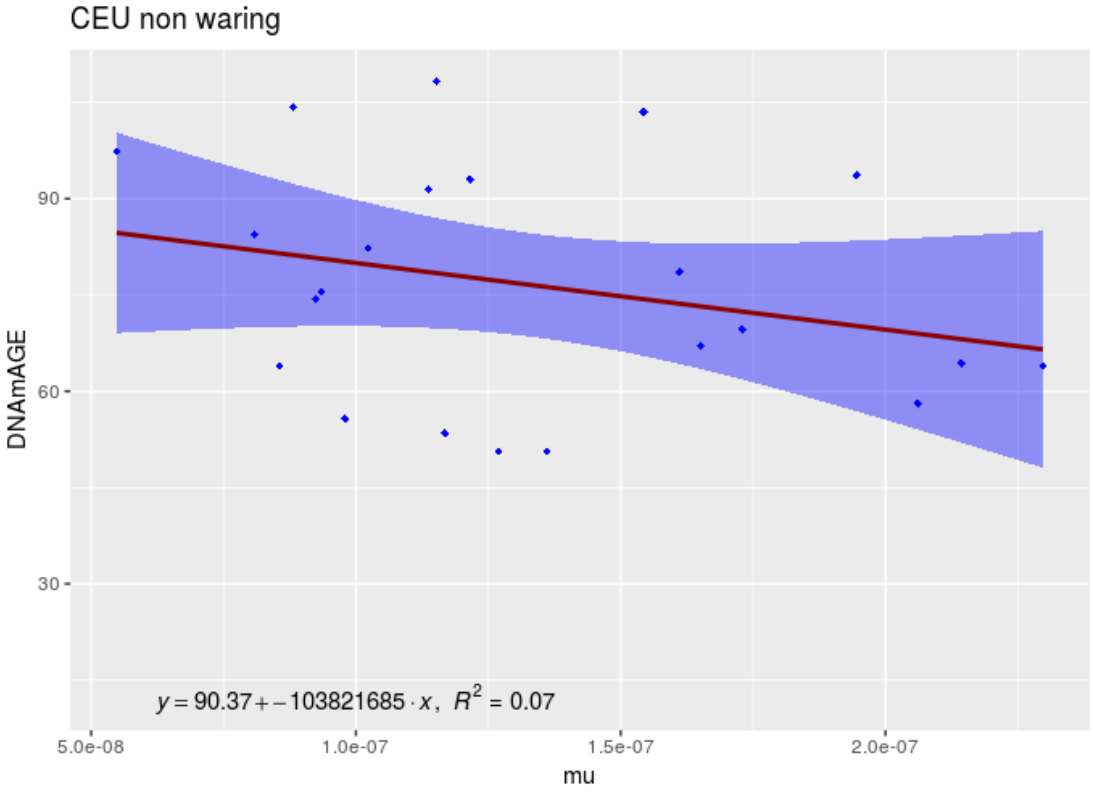

Figure 4.31: Plotted linear model for the non warning CEU overlapping G1K sample

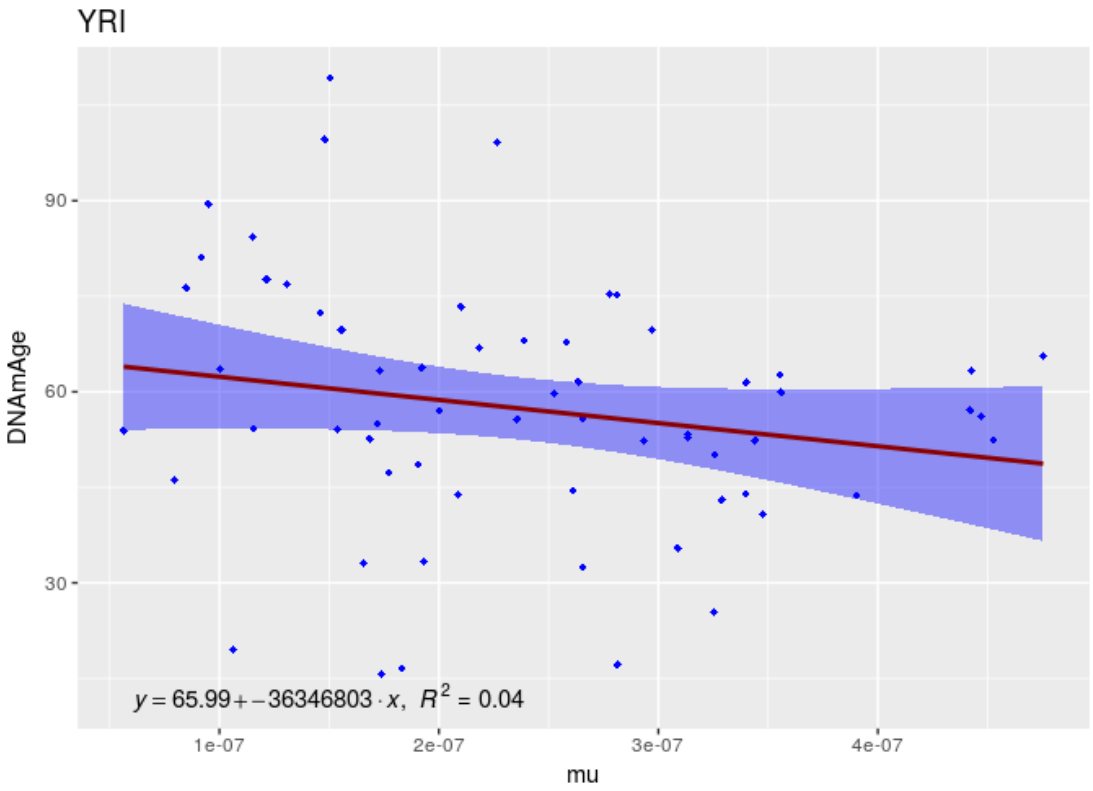

Figure 4.32: Plotted linear model for the full YRI overlapping set of samples

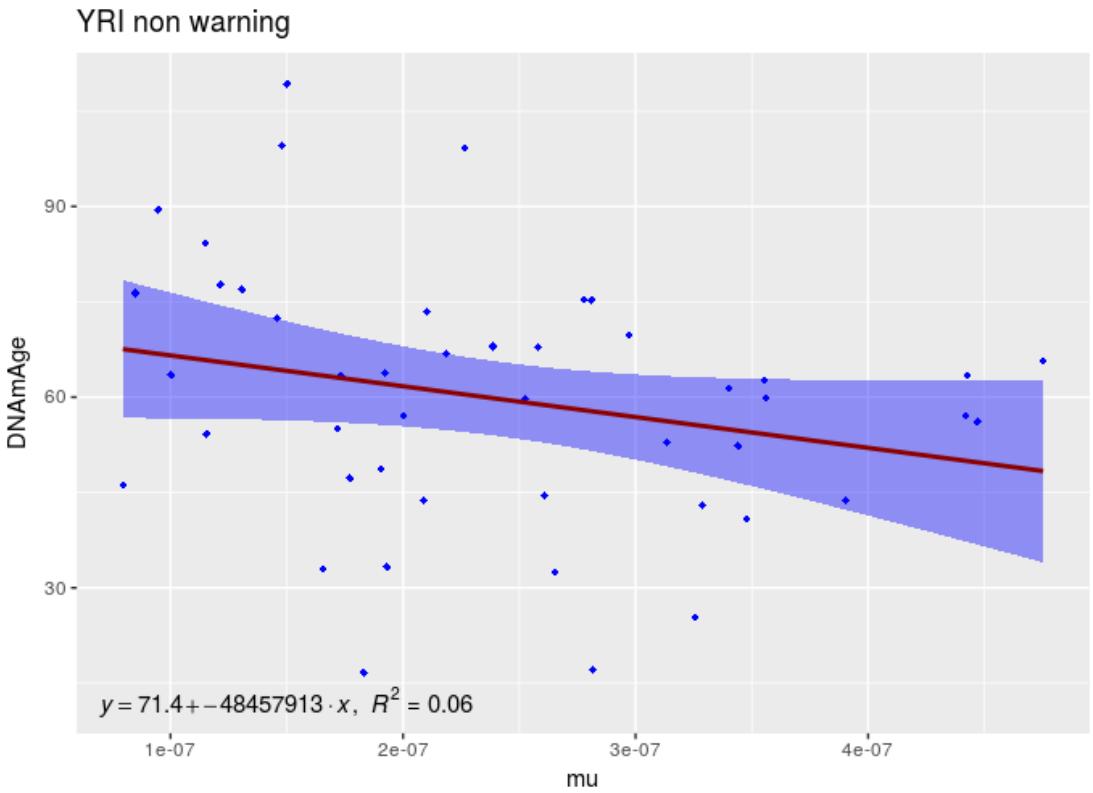

Figure 4.33: Plotted linear model for the non warning YRI overlapping G1K sample

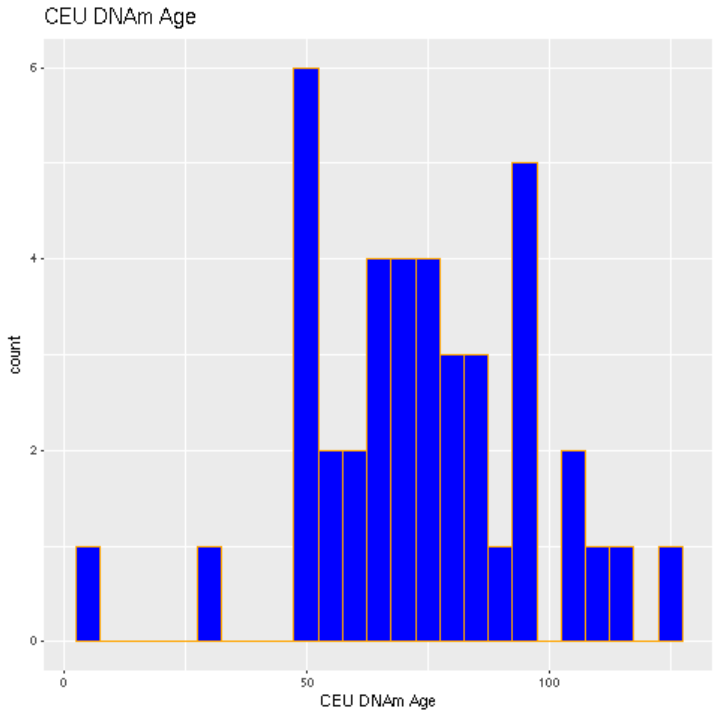

Figure 4.34: DNAm age distribution for the CEU population

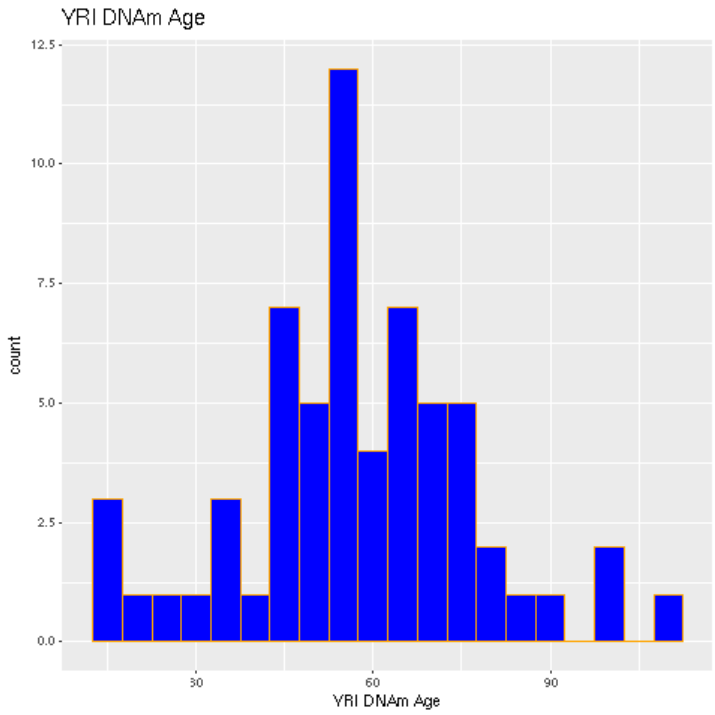

Figure 4.35: DNAm age distribution for the YRI population

Chapter 5

Discussion

5.1 Disparity between low coverage and genotyped data

The first point of discussion is the clear disparity between WGS low coverage genotyping and snp based array association results. Of the 26 populations, 3 populations; southern han Chinese (CHS), Colombians from Medellin (CLM) and Finnish in Finland (FIN), showed clear differences between the low coverage data and the genotyped data association results. In the presence of dominant or epistatic interactions, the additive genetic variance, LD and deviations from the Hardy Weinberg proportions are increased after a population undergoes a bottleneck [59]. Considering only an additive model, the genetic variation would be expected to decrease by Wrights inbreeding coefficient during a bottleneck event such as the bottleneck event that the Finnish population underwent approximately 4000 years ago [59, 60]. There is evidence of population bottlenecks occurring as humans migrated throughout Asia during the last ice age and as a result, 54.2% of the Han Chinese population share 3 Asian specific haplotypes all containing a T $\rightarrow$ C mutation at the M122 locus [61]. For the Colombian population it has been well documented that the effective population size was greatly reduced due to the Spanish and Portuguese colonization of the Americas, as well as an influx of African slaves [62].

Given the effect that each of these events would have had on LD within each population, it is plausible that there is a correlation between LD and the amount of genome wide significant associations. As LD is decayed at a fixed rate

$$D_n = (1 - r)^n \cdot D_0$$

. Where r is the rate of recombination, D_0 is the LD at the initial generation and n is the n-th generation. Given that LD breaks down at a fixed rate it is

plausible that there is a correlation between the time since the bottleneck and the association results. The Finnish population bottleneck happened most recently $\approx 4,000$ years ago and showed the greatest enrichment of significant associations followed by the CHS bottleneck $\approx 100,000$ years ago and the decreased effective population size of the CLM population ≈ 500 years ago. This correlation does not, however, imply causation. It is also possible that the low coverage WGS data has a much higher error rate when calling genotypes than the genotyping chip. In some cases the reported error in WGS genotyping has been reported as high as 15% [63]. The more snps that an index variant is in LD with i.e the genetic variation captured by the index snp, the greater probability that the index snp will be representative of the causal variant [64]. As the quantitative phenotype in question here is derived from the genetic variance in the population and modelled using plinks default additive model, there is a probable increase in type I errors. The percentage of homozygous alleles being called heterozygous by WGS in the G1K Finnish population is 0.2669% compared with the percentage of a heterozygous allele being called homozygous by WGS in the same population is 0.0402% [63]. Although not reported in the results section, the number of variants that failed HWE was in the region of 300,000 whereas the the genotyping data had 95,350 variants violate HWE. Although an interesting topic, comprehending and probing what is driving spurious associations in these populations falls outside of the scope of this project.

5.2 Association study results

The low coverage association study results varied greatly among populations; the Finnish population contained 72 variants above the genome wide significance threshold, $1e-08$, whereas Mexican ancestry in LA and Peruvians from Lima did not contain variants below $1e-05$. The omni genotyping data returned consistent conservative results when compared to the owcoverage data. Two populations, TSI and ASW, showed statistically significant associations. The log transformed phenotype association study returned no significant results for any population. For the low coverage data *pseudogenes* were highly representative throughout the majority of populations. The function of the *pseudogene ANKRD20A9P* remains unknown. It has, however, previously been associated with longevity in Han Chinese samples [65]. *Pseudogenes* are a product of mutation events such as gene duplications that acquire loss of function mutations (non-processed) and loss of function mutations that occur in non duplicated genes(unitary). The third type of *pseudogene* is the retrotransposon (processed) *pseudogene* which can transcribed and re-integrate back into the genome increasing the mutational burden on the organism. Another interesting mechanism in which *pseudogenes* can influence

mutation rate is P-element-induced wimpy testis in *Drosophila*- Piwi-interacting non-coding RNA (Piwi-piRNAs). The Piwi-piRNA pathway is involved in maintaining genomic stability by inhibiting the activity of retrotransposons in non-aging, high-proliferating tissues such as somatic stem cells in the testis and in some forms of cancer [66]. The Piwi-piRNA pathway could potentially be a factor in germline tissues having a mutation rate 100-fold lower than the somatic mutation rate, assuming that the energetics involved in genome maintenance are the same between somatic and germline tissues. An alternative and more likely reasoning behind the enrichment of *pseudogenes* is their ubiquitous nature.

The second most frequent class of biomolecule to show repeated statistical association is the long non-coding RNAs (LINCs). Unfortunately the vast majority of LINCs are uncharacterised. From the characterised LINCs a wide array of biological functions has been determined such as epigenetic regulation, alternate splicing, RNA removal and translation. The pathogenicity LINCs have also been implicated in a host cancers and rare genetic diseases [67]. rs578009605 is a missense variant in *ZGRF1*, a DNA binding protein with presumed helicase activity. Previous genome-wide screens have shown that *ZGFR1* plays a role in the positive regulation of crosslink repair [68]. The polyphen score for rs578009605 is 0.999 and SIFT score is 0 indicating that this variant is deleterious. Although the index variant is representative variant of all the variants in the associated loci, there are multiple consequences for rs578009605 depending on transcript splicing. In one such transcript the consequence of rs578009605 is non-sense mediated decay of the RNA. The full mechanism of action is not understood; loss of *ZGFR1* could increase the mutation rate of an organism by forcing erroneous repair pathways to be used rather than HR. The final individual result of interest is rs6512146 an intronic snp in *CPAMD8*. *CPAMD8* regulates peptidases as well as having a function in eye development. It is important to note that the causal variant may be located within 100 Kb window of the index variant and may be functional variant or regulatory variant for genes located else where.

5.2.1 Meta analysis

CHEK2P2 is a *pseudogene* that is highly expressed in the testis. Although *CHEK2P2* has been previously reported to be implicated in the DNA damage response due to hypoxia by *Viaggi et al* [69].The cited paper made no reference to *CHEK2P2* but rather *CHEK2*, a well-documented DNA damage response protein with high sequence similarity to *CHEK2P2* [58]. This does not, however, rule out *CHEK2P2* playing a role in DNA damage response. What is interesting is that *CHEK2P2* is almost solely expressed in the testis. It is well documented that the germline has a mutation rate much lower than that of the soma. *CHEK2P2* may function or be implicated in the dampening of the mutation rate in the testis.

An interesting hypothesis is that the variation in mutation rate across human genomes is correlated with SNP density [70]. Distinct mutational patterns can be found throughout the genome, such as an increased SNP density around late replicating genes as well as at sequences that surround transcription start sites. The latter effect is only seen at genes that are actively transcribed in the tissue type [70]. Several variants across the low coverage genotype meta GWAS showed significant association with genes involved with transcription processes. The transcription factor *ATF1* has a host of targets from proteins involved in cell survival to cell cycle kinases. *ATF1* fused with *EWSR1* has been shown to be implicated in rare cancers such as angiomatoid fibrous histiocytoma, a rare soft tissue tumour occurring frequently in children and adolescents [57]. *TCF25* or *Nulp1* is a transcription repressor which, when expressed, binds to the X-linked inhibitor of apoptosis protein (XIAP). This interaction allows *TCF25* to possibly play a role in tumour progression as well as transcription regulation [71]. *ZNF845* has been identified as a protein which potentially could be involved in transcription.

RALGAPA1 acts as a GTPase activator for Ras-like small GTPases RalA and RalB. RalA activation required for Ras-induced tumorigenesis. The rate of cell proliferation is also correlated with tumorigenesis. Selection plays a crucial role in eliminating cells that have accumulated deleterious mutations allowing for selection of cells that confer an advantage. This phenomena is frequently observed in cancer i.e, A progenitor cell that acquires a driver mutation (*TP53*) will clonally expand through the population out-competing cells that do not have the conferred advantage [28]. It then makes sense to assume that individuals with a higher mutation rate may have an increased susceptibility to developing cancer. Mutation rates that can overwhelm cancerous cells should still, in theory be selected against and removed from the population. This observation poses the question 'In the prodromal phase of cancer, is there greater pressure for positive selection for a higher mutation rate but when the illness has progressed does selection favour more conservative mutation rate?' Assuming that a carcinogenic mutation satisfying Knudsons two hit hypothesis can in fact develop in cells with a normal low mutation rate [1].

For gene set enrichment analysis, only one gene set was returned significant after multiple testing correction, 'GO multicellular organism reproduction' gene set which contained 668 genes. A gene set this large is guaranteed not to be mutually exclusive with the pathway in question and many pathways may be represented by subsets of this gene set. As expected we see cancer gene sets represented in a high number but none significant after multiple correction. The breast cancer data sets, Zucchi metastasis and Welch *BRCA1* targets are represented in both genotyping datasets (log transformed pheno & untransformed), suggesting a true

[1] This idea is a kin to a mutator phenotype being initially selected for but then being selected against after disrupting some vital process

enrichment. The gene set with the strongest association with somatic mutation to be returned was the gene set 'DNA damage response signal transduction resulting in transcription' from the low coverage meta analysis. Due to the high number of *pseudogenes* associated with the mutation rate phenotype it would be interesting if a PIWI-piRNA gene set were available to test whether there was an enrichment for PIWI-piRNAs implicating them in increased mutation rate.

5.3 DNAm age non concordant with somatic mutation rate

Using Horvath's epigentic clock it is possible to infer chronological age from methylation data. Using chronological age as a covariate would allow for a greater explanation of the variation within the data set. The epigentic clock shows a highest correlation with fresh samples. Unfortunately no biopsy methylation data is available for the G1K project. There was, however, transformed cell lines available for the CEU and YRI populations. The ability to transform whole blood samples into lymphoblastoid cell lines has had a profound effect on the progression of human genetics. However, the effect that Epstein-Barr virus (EBV) immortalization has on the methylation patterns at CpG sites is not well understood. Research has indicated that the higher the number of cell passages in immortalized cell lines the greater the difference in methylation pattern from the white blood cell control [72]. The changes in methylation pattern are randomly distributed across the genome and *de novo*, increased or decreased methylation at a particular site may be observed relative to the control. With this information, it seemed unlikely that using methylation data for transformed G1K project samples would be representative of the true methylation pattern of the individual at the the time of transformation. Horvath's DNAm age clock uses 353 CpG sites and is potentially vulnerable to distortion or misrepresentation of methylation at a particular site. In a follow up paper, Horvath described a correlation between biological age and chronological age in cell lines transformed by EBV (r=0.59) [44]. This warranted the DNAm age correlation analysis.

The first observation of note was the range of DNAm age returned for the CEU was $\approx$ 120 years with mean age of 73.88 while the YRI population had more reasonable DNAm age (mean= 57.3) return, albeit still improbable for a population sequencing project. The second observation was to look at how the DNAm correlates with somatic mutation rate, mutation count, and methylation level as well as fit linear models to DNAm age as a function of somatic mutation rate. The direction of the correlation between DNAm and mutation rate is inverse to what is expected with the the older DNAm individuals showing a lower mutation rate whereas a higher mutation rate would be expected. A strong correlation is

also expected for DNAm age and mutation counts, whereas we see a correlation coefficient close to zero for all data sets. The linear models mirror the correlation results with somatic mutation rate being a poor predictor of DNAm age.

Finally, the distribution of DNAm was examined. The CEU population was not normally distributed while the YRI population did show a normally distributed DNAm. This observation is concordant with methylation patterns changing significantly with the increased number of passages [73]. Both YRI and CEU cell lines were created for the HapMap project, although the CEU cell line collection began in 1980 and were frozen whereas the YRI samples were collected especially for the HapMap project. [74]. For the G1K project some of the CEU and YRI cell line samples did not pass quality control and were excluded from the project [75]. The date of each transformation is unavailable, however, the deviance in DNAm age would suggest that CEU cell lines were older or at least underwent passaging far more frequently than the YRI cell lines as increased passaging has been shown to influence methylation patterns at CpG sites across the genome [73].

Chapter 6

Conclusion

6.1 A plausible network of proteins affecting somatic mutation rate?

Although we do see associations that may have an affect on the somatic mutation rate, it is also necessary to find alternative and often contrary explanations for the results of the association study. As DNA replication proteins are expected to be the main contributors to mutation rate and subsequently were not shown to be associated with the C $\rightarrow$ A mutation rate, it does not necessarily mean that the algorithm generating the mutation rate is not capturing the true mutation rate nor does it mean that results are incorrect. Several plausible protein pathways have been shown to be associated with C $\rightarrow$ A mutation rate. Unfortunately, no smoking gun was observed but many of the processes that did show an association with the C $\rightarrow$ A mutation rate are intertwined with the cancer phenotype. Directly implying causation due to mutation rate is not possible as it cannot be determined whether mutation rate is a cause or the effect of the associated cancer pathways.

6.2 Capturing the phenotypic variantion

There is a high likelihood that the true phenotypic variance is not fully captured by one segment of nucleotide substitution. From Fishers work *'The Genetical Theory of Natural Selection'* in 1930, the phenotypic variance (V_P) can be explained as the sum of the total genetic variance (V_G) and variance due to environmental factors (V_E)[76].

$$V_P = V_G + V_E$$

The total genetic variance can be partitioned into; the additive variance (V_A), the dominance variance (V_D) and the epistatic variance (V_I).

$$V_G = V_A + V_D + V_I$$

If the C $\rightarrow$ A mutation rate is not representative of the actual phenotypic variance then the association analysis will perform poorly. Conversely if the phenotype variance is captured fully the causative loci may be too small of effect and be undetectable with a sample size of ≈ 100.

Replication and transcription errors may be more associated with the stochastic nature of DNA replication and transcription processes, such as the higher mutational load at the 5' end of the later replicating Okazaki fragments [77]. Later replicating regions of the genome also show a much higher mutation rate than early replicating regions which may implicate the cellular environment as a contributor to increased mutation at these sites. Should this be the case then, the phenotypic variance intraspecies could potentially show a larger variance due to the environment and not be effectively identified in an under-powered association study [18]. In terms of the genetic variantions affect on phenotypic variance.

Sampling mutations

As we can only call mutations that lay in regions where overlapping paired-end reads are mapped, the counting of mutations is like a random sampling of mutations from the exonic region of the genome. Therefore, we cant assume that the frequency of each base is going to be $\approx 25\%$. With the knowledge that exonic regions are GC rich and that guanine nucleotides are readily oxidised after fragmentation leading to G $\rightarrow$ T transversions, cytosine is also readily methylated, followed by deamination leading to C $\rightarrow$ T transversions. We should then expect to see a higher number of X $\rightarrow$ T mutations, thus, creating difficulty in measuring the rate at which substitutions are occurring. In practice, this may not necessarily be the case, however, selection is expected to be more active on the coding region of the genome. If mutations were to decrease the fitness of the cell, they would be removed via apoptotic processes or be out competed. Neutral substitutions acting as passengers would then be expected to make up the bulk of the total mutations. Deep exome sequencing will therefore miss the full set of mutations, giving a misrepresentation of the phenotypic variance. To capture the full phenotypic variance due to the genetic component further work is required to unify the individual substitution rates. DNA damage after library preparation, the main caveat is detailed below.

DNA damage artefacts

Sequencing artefacts can have a profound effect on the generation of the phenotype The fragmentation step of library preparation can induce an oxidative environment leading to an increase in DNA damage resulting in amplification errors. In particular, guanine is readily oxidised to 8-oxo-dG which results in the G $\rightarrow$ T transversion. Deamination of methylated cytosine can also lead to sequencing artefacts. DNA damage artefacts can be observed by measuring the substitution and reverse complement imbalance between reads 1 and 2 from paired-end sequencing data. Of the 4 million FASTQ files sampled from the G1K project, 41% of sequences are estimated to contain DNA damage artefacts from post library preparation [37]. For G $\rightarrow$ T transversions in the G1K project we see a global imbalance of greater than 1.5 for 41% of reads. This corresponds to 41% of reads containing 1.5 times more variants on read 1 than on read 2. True variants show no imbalance between read 1 and 2 [37]. It therefore seems highly probable that the G1K reads have inflated the mutation count beyond that of what would be expected. C $\rightarrow$ A substitutions constitute $\approx$ 20% of the total substitutions in the soma although the effect of sequencing artefacts in single cell sequencing is still unknown [78].

6.3 Future work

As the results of the association are not conclusive, they have highlighted some key aspects aspects that must be addressed. The next step in the development of this project is to account for DNA damage in the G1K project as a way of unifying the substitution rates and potentially capturing the full phenotypic variation due to the genetic component. Two avenues can be explored to do this. Firstly, use the damage estimates as a covariate in the GWAS analysis by using pre-existing perl scripts which are available online. The second avenue involves incorporating the imbalance estimator into the algorithm for generating the phenotype, then filtering out the reads that show excessive imbalance. Other artefacts that can hinder the inference of the somatic mutation rate are PCR errors and PCR over-amplification. For the latter, these reads can be easily dealt with by matching reads with the same start and stop sites and removing them. PCR errors on the other hand arise during the library amplification step and can not be easily identified, although, tools have been developed to remove PCR artefacts in pyrosequencing experiments [79]. As a work around, it may be possible to use a non-overlapping read to verify, whether or not, the mutation is a true mutation or an error carried forward from library preparation.

The mutation rate is calculated using

$$\mu = \frac{N_m}{T_s}$$

where μ is the mutation rate, N_m is the total mutation counts and T_s is the total number of sites where the mutation could have of arisen. Should population data become available, along with age or methylation data it would be interesting to see the difference in mutation rate by changing T_s to an inferred number of cell divisions N_d

$$\mu = \frac{N_m}{N_d}$$

here N_d is inferred directly from chronological age or epigenetic age by using Horvath's epigenetic clock on the methylation data.

To conclude, should these issues be addressed, namely; DNA damage during library preparation, PCR errors and PCR over amplification, capturing the variation in human somatic mutation rate amongst populations is attainable through the use of over lapping paired-end reads.

Bibliography

[1] P. L. Foster, "Methods for determining spontaneous mutation rates," in *DNA Repair, Part B*, vol. 409 of *Methods in Enzymology*, pp. 195 – 213, Academic Press, 2006.

[2] P. Armitage, "The statistical theory of bacterial populations subject to mutation," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 14, no. 1, pp. 1–40, 1952.

[3] A. Shapiro *et al.*, "The kinetics of growth and mutation in bacteria.," *The kinetics of growth and mutation in bacteria.*, 1946.

[4] B. Milholland, A. Auton, Y. Suh, and J. Vijg, "Age-related somatic mutations in the cancer genome," *Oncotarget*, vol. 6, no. 28, p. 24627, 2015.

[5] J. Vijg, "Somatic mutations, genome mosaicism, cancer and aging," *Current Opinion in Genetics & Development*, vol. 26, pp. 141 – 149, 2014. Molecular and genetic bases of disease.

[6] C. Seoighe and A. Scally, "Inference of candidate germline mutator loci in humans from genome-wide haplotype data," *PLoS genetics*, vol. 13, no. 1, p. e1006549, 2017.

[7] M. Kimura, "On the evolutionary adjustment of spontaneous mutation rates," *Genetical Research*, vol. 9, no. 01, pp. 23–34, 1967.

[8] M. Lynch, M. S. Ackerman, J.-F. Gout, H. Long, W. Sung, W. K. Thomas, and P. L. Foster, "Genetic drift, selection and the evolution of the mutation rate," *Nat Rev Genet*, vol. 17, pp. 704–714, Nov. 2016.

[9] M. Lynch, "Evolution of the mutation rate," *Trends in Genetics*, vol. 26, no. 8, pp. 345–352, 2010.

[10] A. Sturtevant, "Essays on evolution. i. on the effects of selection on mutation rate," *The Quarterly Review of Biology*, vol. 12, no. 4, pp. 464–467, 1937.

[11] M. Demerec, "Frequency of spontaneous mutations in certain stocks of drosophila melanogaster," *Genetics*, vol. 22, no. 5, p. 469, 1937.

[12] N. Timofeeff-Ressovsky, "Qualitativer vergleich der mutabilität vondrosophila funebris unddrosophila melanogaster," *Zeitschrift für Induktive Abstammungs-und Vererbungslehre*, vol. 71, no. 1, pp. 276–280, 1936.

[13] C. F. Baer, M. M. Miyamoto, and D. R. Denver, "Mutation rate variation in multicellular eukaryotes: causes and consequences," *Nature reviews. Genetics*, vol. 8, no. 8, p. 619, 2007.

[14] J. A. Gossen, W. De Leeuw, C. Tan, E. C. Zwarthoff, F. Berends, P. Lohman, D. L. Knook, and J. Vijg, "Efficient rescue of integrated shuttle vectors from transgenic mice: a model for studying mutations in vivo," *Proceedings of the National Academy of Sciences*, vol. 86, no. 20, pp. 7971–7975, 1989.

[15] M. Kimura, *The neutral theory of molecular evolution*. Cambridge University Press, 1983.

[16] M. L. Hoang, I. Kinde, C. Tomasetti, K. W. McMahon, T. A. Rosenquist, A. P. Grollman, K. W. Kinzler, B. Vogelstein, and N. Papadopoulos, "Genome-wide quantification of rare somatic mutations in normal human tissues using massively parallel sequencing," *Proceedings of the National Academy of Sciences*, p. 201607794, 2016.

[17] W. Yang, "An overview of y-family dna polymerases and a case study of human dna polymerase η," *Biochemistry*, vol. 53, no. 17, pp. 2793–2803, 2014.

[18] J. A. Stamatoyannopoulos, I. Adzhubei, R. E. Thurman, G. V. Kryukov, S. M. Mirkin, and S. R. Sunyaev, "Human mutation rate associated with dna replication timing," *Nature genetics*, vol. 41, no. 4, pp. 393–395, 2009.

[19] S. Jinks-Robertson and A. S. Bhagwat, "Transcription-associated mutagenesis," *Annual review of genetics*, vol. 48, pp. 341–359, 2014.

[20] M. S. Lawrence, P. Stojanov, P. Polak, G. V. Kryukov, K. Cibulskis, A. Sivachenko, S. L. Carter, C. Stewart, C. H. Mermel, S. A. Roberts, *et al.*, "Mutational heterogeneity in cancer and the search for new cancer genes," *Nature*, vol. 499, no. 7457, p. 214, 2013.

[21] C. Chen, H. Qi, Y. Shen, J. Pickrell, and M. Przeworski, "Contrasting determinants of mutation rates in germline and soma," *bioRxiv*, p. 117325, 2017.

[22] H. Ide and M. Kotera, "Human dna glycosylases involved in the repair of oxidatively damaged dna," *Biological and Pharmaceutical Bulletin*, vol. 27, no. 4, pp. 480–485, 2004.

[23] A. Tubbs and A. Nussenzweig, "Endogenous dna damage as a source of genomic instability in cancer," *Cell*, vol. 168, no. 4, pp. 644–656, 2017.

[24] P. A. Jeggo, L. H. Pearl, and A. M. Carr, "Dna repair, genome stability and cancer: a historical perspective," *Nature Reviews Cancer*, 2015.

[25] K. D. Makova and R. C. Hardison, "The effects of chromatin organization on variation in mutation rates in the genome," *Nature Reviews Genetics*, vol. 16, no. 4, pp. 213–223, 2015.

[26] J. H. Chuang and H. Li, "Functional bias and spatial organization of genes in mutational hot and cold regions in the human genome," *PLoS biology*, vol. 2, no. 2, p. e29, 2004.

[27] R. Sabarinathan, L. Mularoni, J. Deu-Pons, A. Gonzalez-Perez, and N. Lopez-Bigas, "Nucleotide excision repair is impaired by binding of transcription factors to dna," *bioRxiv*, p. 028886, 2015.

[28] I. Martincorena and P. J. Campbell, "Somatic mutation in cancer and normal cells," *Science*, vol. 349, no. 6255, pp. 1483–1489, 2015.

[29] I. Tomlinson and W. Bodmer, "Selection, the mutation rate and cancer: ensuring that the tail does not wag the dog.," *Nature medicine*, vol. 5, no. 1, 1999.

[30] A. Tubbs and A. Nussenzweig, "Endogenous dna damage as a source of genomic instability in cancer," *Cell*, vol. 168, no. 4, pp. 644–656, 2017.

[31] W. P. Roos, A. D. Thomas, and B. Kaina, "Dna damage and the balance between survival and death in cancer biology," *Nature Reviews. Cancer*, vol. 16, no. 1, p. 20, 2016.

[32] D. Hanahan and R. A. Weinberg, "Hallmarks of cancer: the next generation," *cell*, vol. 144, no. 5, pp. 646–674, 2011.

[33] L. Szilard, "On the nature of the aging process," *Proceedings of the National Academy of Sciences*, vol. 45, no. 1, pp. 30–45, 1959.

[34] L. E. Orgel, "The maintenance of the accuracy of protein synthesis and its relevance to ageing," *Proceedings of the National Academy of Sciences*, vol. 49, no. 4, pp. 517–521, 1963.

[35] B. Milholland, Y. Suh, and J. Vijg, "Mutation and catastrophe in the aging genome," *Experimental Gerontology*, 2017.

[36] H. Zetterberg, M. Båth, M. Zetterberg, P. Bernhardt, and O. Hammarsten, "The szilard hypothesis on the nature of aging revisited," *Genetics*, vol. 182, no. 1, pp. 3–9, 2009.

[37] L. Chen, P. Liu, T. C. Evans, and L. M. Ettwiller, "Dna damage is a pervasive cause of sequencing errors, directly confounding variant identification," *Science*, vol. 355, no. 6326, pp. 752–756, 2017.

[38] . G. P. Consortium *et al.*, "A global reference for human genetic variation," *Nature*, vol. 526, no. 7571, p. 68, 2015.

[39] S. Ripke, B. M. Neale, A. Corvin, J. T. Walters, K.-H. Farh, P. A. Holmans, P. Lee, B. Bulik-Sullivan, D. A. Collier, H. Huang, *et al.*, "Biological insights from 108 schizophrenia-associated genetic loci," *Nature*, vol. 511, no. 7510, p. 421, 2014.

[40] W. G. Hill, "Understanding and using quantitative genetic variation," *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, vol. 365, no. 1537, pp. 73–85, 2010.

[41] S. Horvath, "Dna methylation age of human tissues and cell types," *Genome biology*, vol. 14, no. 10, p. 3156, 2013.

[42] A. E. Teschendorff, F. Marabita, M. Lechner, T. Bartlett, J. Tegner, D. Gomez-Cabrero, and S. Beck, "A beta-mixture quantile normalization method for correcting probe design bias in illumina infinium 450 k dna methylation data," *Bioinformatics*, vol. 29, no. 2, pp. 189–196, 2012.

[43] E. L. Moen, X. Zhang, W. Mu, S. M. Delaney, C. Wing, J. McQuade, J. Myers, L. A. Godley, M. E. Dolan, and W. Zhang, "Genome-wide variation of cytosine modifications between european and african populations and the implications for complex traits," *Genetics*, vol. 194, no. 4, pp. 987–996, 2013.

[44] S. Horvath, M. Gurven, M. E. Levine, B. C. Trumble, H. Kaplan, H. Allayee, B. R. Ritz, B. Chen, A. T. Lu, T. M. Rickabaugh, *et al.*, "An epigenetic clock analysis of race/ethnicity, sex, and coronary heart disease," *Genome biology*, vol. 17, no. 1, p. 171, 2016.

[45] H. Heyn, S. Moran, I. Hernando-Herraez, S. Sayols, A. Gomez, J. Sandoval, D. Monk, K. Hata, T. Marques-Bonet, L. Wang, *et al.*, "Dna methylation contributes to natural human variation," *Genome research*, vol. 23, no. 9, pp. 1363–1372, 2013.

[46] C. C. Chang, C. C. Chow, L. C. Tellier, S. Vattikuti, S. M. Purcell, and J. J. Lee, "Second-generation plink: rising to the challenge of larger and richer datasets," *Gigascience*, vol. 4, no. 1, p. 7, 2015.

[47] L. Goh and V. B. Yap, "Effects of normalization on quantitative traits in association test," *BMC bioinformatics*, vol. 10, no. 1, p. 415, 2009.

[48] H. Aschard, B. J. Vilhjálmsson, N. Greliche, P.-E. Morange, D.-A. Trégouët, and P. Kraft, "Maximizing the power of principal-component analysis of correlated phenotypes in genome-wide association studies," *The American Journal of Human Genetics*, vol. 94, no. 5, pp. 662–676, 2014.

[49] N. Liu, H. Zhao, A. Patki, N. A. Limdi, and D. B. Allison, "Controlling population structure in human genetic association studies with samples of unrelated individuals," *Statistics and its interface*, vol. 4, no. 3, p. 317, 2011.

[50] S. D. Turner, "qqman: an r package for visualizing gwas results using qq and manhattan plots," *BioRxiv*, p. 005165, 2014.

[51] E. Evangelou and J. P. Ioannidis, "Meta-analysis methods for genome-wide association studies and beyond," *Nature Reviews Genetics*, vol. 14, no. 6, pp. 379–389, 2013.

[52] C. A. de Leeuw, J. M. Mooij, T. Heskes, and D. Posthuma, "Magma: generalized gene-set analysis of gwas data," *PLoS computational biology*, vol. 11, no. 4, p. e1004219, 2015.

[53] F. Gianfrancesco, T. Esposito, S. Penco, V. Maglione, C. Liquori, M. Patrosso, O. Zuffardi, A. Ciccodicola, D. Marchuk, and F. Squitieri, "Zpld1 gene is disrupted in a patient with balanced translocation that exhibits cerebral cavernous malformations," *Neuroscience*, vol. 155, no. 2, pp. 345–349, 2008.

[54] Z.-F. Li, X. hua Wu, and E. Engvall, "Identification and characterization of cpamd8, a novel member of the complement $3/\alpha 2$-macroglobulin family with a c-terminal kazal domain," *Genomics*, vol. 83, no. 6, pp. 1083 – 1093, 2004.

[55] S.-S. Cheong, L. Hentschel, A. E. Davidson, D. Gerrelli, R. Davie, R. Rizzo, N. Pontikos, V. Plagnol, A. T. Moore, J. C. Sowden, *et al.*, "Mutations in cpamd8 cause a unique form of autosomal-recessive anterior segment dysgenesis," *The American Journal of Human Genetics*, vol. 99, no. 6, pp. 1338–1352, 2016.

[56] C. Felley, J. Qian, S. Mantey, T. Pradhan, and R. Jensen, "Chief cells possess a receptor with high affinity for pacap and vip that stimulates pepsinogen release," *American Journal of Physiology-Gastrointestinal and Liver Physiology*, vol. 263, no. 6, pp. G901–G907, 1992.

[57] S. Rossi, K. Szuhai, M. Ijszenga, H. J. Tanke, L. Zanatta, R. Sciot, C. D. Fletcher, A. P. Dei Tos, and P. C. Hogendoorn, "Ewsr1-creb1 and ewsr1-atf1 fusion genes in angiomatoid fibrous histiocytoma," *Clinical Cancer Research*, vol. 13, no. 24, pp. 7322–7328, 2007.

[58] R. A. Freiberg, E. M. Hammond, M. J. Dorie, S. M. Welford, and A. J. Giaccia, "Dna damage during reoxygenation elicits a chk2-dependent checkpoint response," *Molecular and cellular biology*, vol. 26, no. 5, pp. 1598–1609, 2006.

[59] J. Wang, A. Caballero, and W. G. Hill, "The effect of linkage disequilibrium and deviation from hardy–weinberg proportions on the changes in genetic variance with bottlenecking," *Heredity*, vol. 81, no. 2, pp. 174–186, 1998.

[60] A. Sajantila, A.-H. Salem, P. Savolainen, K. Bauer, C. Gierig, and S. Pääbo, "Paternal and maternal dna lineages reveal a bottleneck in the founding of the finnish population," *Proceedings of the National Academy of Sciences*, vol. 93, no. 21, pp. 12035–12039, 1996.

[61] B. Su, J. Xiao, P. Underhill, R. Deka, W. Zhang, J. Akey, W. Huang, D. Shen, D. Lu, J. Luo, *et al.*, "Y-chromosome evidence for a northward migration of modern humans into eastern asia during the last ice age," *The American Journal of Human Genetics*, vol. 65, no. 6, pp. 1718–1724, 1999.

[62] A. Moreno-Estrada, S. Gravel, F. Zakharia, J. L. McCauley, J. K. Byrnes, C. R. Gignoux, P. A. Ortiz-Tello, R. J. Martínez, D. J. Hedges, R. W. Morris, *et al.*, "Reconstructing the population genetic history of the caribbean," *PLoS genetics*, vol. 9, no. 11, p. e1003925, 2013.

[63] F. C. Ceballos, S. Hazelhurst, and M. Ramsay, "Assessing runs of homozygosity: A comparison of snp array and whole genome sequence low coverage data," *bioRxiv*, 2017.

[64] B. K. Bulik-Sullivan, P.-R. Loh, H. K. Finucane, S. Ripke, J. Yang, N. Patterson, M. J. Daly, A. L. Price, B. M. Neale, S. W. G. of the Psychiatric Genomics Consortium, *et al.*, "Ld score regression distinguishes confounding from polygenicity in genome-wide association studies," *Nature genetics*, vol. 47, no. 3, pp. 291–295, 2015.

[65] Y. Zeng, C. Nie, J. Min, X. Liu, M. Li, H. Chen, H. Xu, M. Wang, T. Ni, Y. Li, *et al.*, "Novel loci and pathways significantly associated with longevity," *Scientific reports*, vol. 6, p. 21243, 2016.

[66] Á. Sturm, A. Perczel, Z. Ivics, and T. Vellai, "The piwi-pirna pathway: road to immortality," *Aging Cell*, 2017.

[67] X. Shi, M. Sun, H. Liu, Y. Yao, and Y. Song, "Long non-coding rnas: a new frontier in the study of human diseases," *Cancer letters*, vol. 339, no. 2, pp. 159–166, 2013.

[68] B. Adamson, A. Smogorzewska, F. D. Sigoillot, R. W. King, and S. J. Elledge, "A genome-wide homologous recombination screen identifies the rna-binding protein rbmx as a component of the dna damage response," *Nature cell biology*, vol. 14, no. 3, p. 318, 2012.

[69] C. D. Viaggi, S. Cavani, M. Malacarne, F. Floriddia, G. Zerega, C. Baldo, M. Mogni, M. Castagnetta, G. Piombo, D. Coviello, *et al.*, "First-trimester euploid miscarriages analysed by array-cgh," *Journal of applied genetics*, vol. 54, no. 3, pp. 353–359, 2013.

[70] P. Cui, F. Ding, Q. Lin, L. Zhang, A. Li, Z. Zhang, S. Hu, and J. Yu, "Distinct contributions of replication and transcription to mutation rate variation of human genomes," *Genomics, proteomics & bioinformatics*, vol. 10, no. 1, pp. 4–10, 2012.

[71] H. Steen and D. Lindholm, "Nuclear localized protein-1 (nulp1) increases cell death of human osteosarcoma cells and binds the x-linked inhibitor of apoptosis protein," *Biochemical and biophysical research communications*, vol. 366, no. 2, pp. 432–437, 2008.

[72] D. Grafodatskaya, S. Choufani, J. Ferreira, D. Butcher, Y. Lou, C. Zhao, S. Scherer, and R. Weksberg, "Ebv transformation and cell culturing destabilizes dna methylation in human lymphoblastoid cell lines," *Genomics*, vol. 95, no. 2, pp. 73–83, 2010.

[73] S. Bork, S. Pfister, H. Witt, P. Horn, B. Korn, A. D. Ho, and W. Wagner, "Dna methylation pattern changes upon long-term culture and aging of human mesenchymal stromal cells," *Aging cell*, vol. 9, no. 1, pp. 54–63, 2010.

[74] R. A. Gibbs, J. W. Belmont, P. Hardenbol, T. D. Willis, F. Yu, H. Yang, L.-Y. Ch'ang, W. Huang, B. Liu, Y. Shen, *et al.*, "The international hapmap project," 2003.

[75] . G. P. Consortium *et al.*, "A map of human genome variation from population scale sequencing," *Nature*, vol. 467, no. 7319, p. 1061, 2010.

[76] R. A. Fisher, *The genetical theory of natural selection: a complete variorum edition*. Oxford University Press, 1930.

[77] M. A. Reijns, H. Kemp, J. Ding, S. M. de Procé, A. P. Jackson, and M. S. Taylor, "Lagging strand replication shapes the mutational landscape of the genome," *Nature*, vol. 518, no. 7540, p. 502, 2015.

[78] B. Milholland, X. Dong, L. Zhang, X. Hao, Y. Suh, and J. Vijg, "Differences between germline and somatic mutation rates in humans and mice," *Nature Communications*, vol. 8, 2017.

[79] J. M. Kebschull and A. M. Zador, "Sources of pcr-induced distortions in high-throughput sequencing data sets," *Nucleic acids research*, vol. 43, no. 21, pp. e143–e143, 2015.

Appendix

Algorithm 1 Somatic mutation rate algorithm

```
mask Phase 3 variants in reference
while IN SAM file do
    hash array of chromosomal positions
    if MD:Z in SAM then
        get chromosomal position
        while len(cigarstring) > 0 do
            get mismatches
            update position marker
            if cigar string matches non digits then
                get nucleotide sequence at mismatch position
                if  mismatch occurs more than once then
                    force count as one mismatch
                end if
            end if
        end while
    end if
end while
close SAM file
while IN SAM file do
    map position of read & mate
    filter based on bitwise flag
    if MAP > 60 then
        if read maps to reference & no gaps in alignment then
            if read pos > mate pos then
                get overlapping read size
                for overlapping bases do
                    get read and mate position in read and reference genome and
                    get quality scores
                    if sites above minimum threshold then
                        hash high quality over lapping sites
                    end if
                end for
            end if
```

```
34:         if overlapping region > 0 then
35:           while stepping through MD tag do
36:             remove matching bases
37:             increment position by number of matches
38:             if mismatching base then
39:               record reference position
40:               force only single base variants
41:               if length of mismatch ≯ 1 then
42:                 store mismatching nucleotide
43:                 store quality score
44:               end if
45:               increment by number of mismatching bases
46:             end if
47:           end while
48:           while length of mate MD tag do
49:             step through mate MD tag as described above
50:             if read and mate agree on mismatch and all criteria is satisfied
    then
51:               count number of times mismatch and reference disagree
52:             end if
53:             increment by number of mismatching bases
54:           end while
55:         end if
56:         count total number of overlapping bases
57:       end if
58:     end if
59: end while
60: close SAM file
```

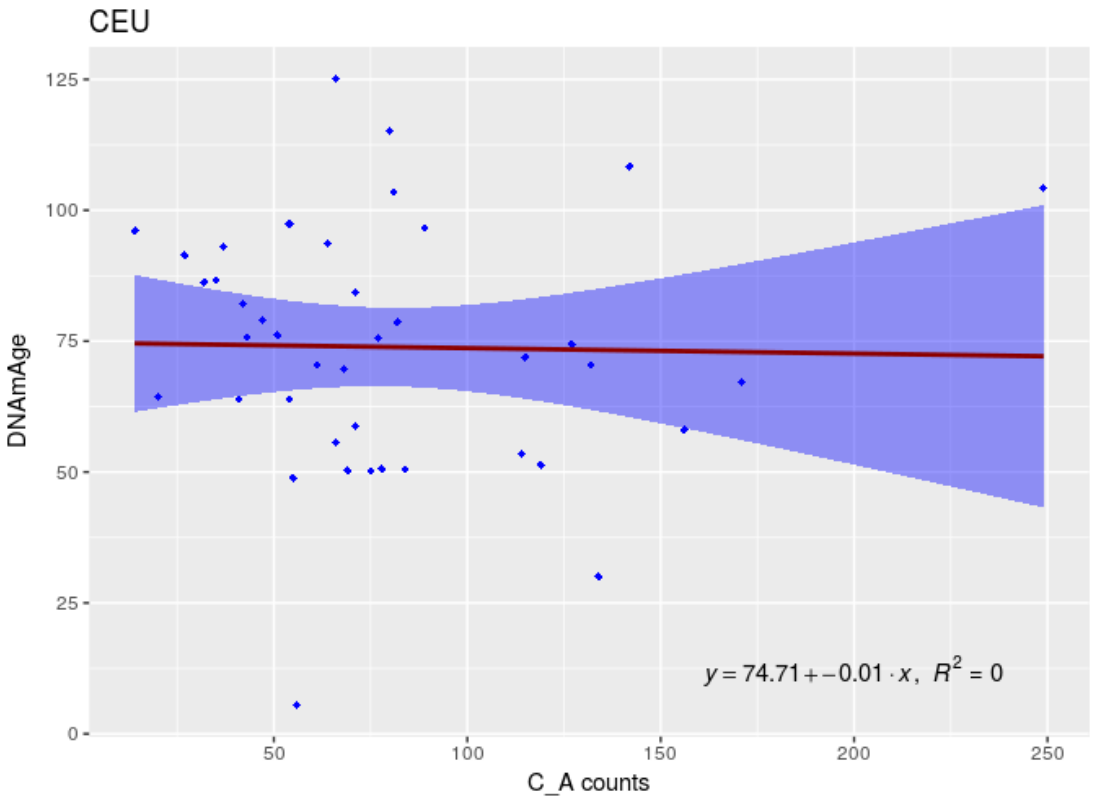

Figure 1: Linear model for the full CEU counts vs DNAm

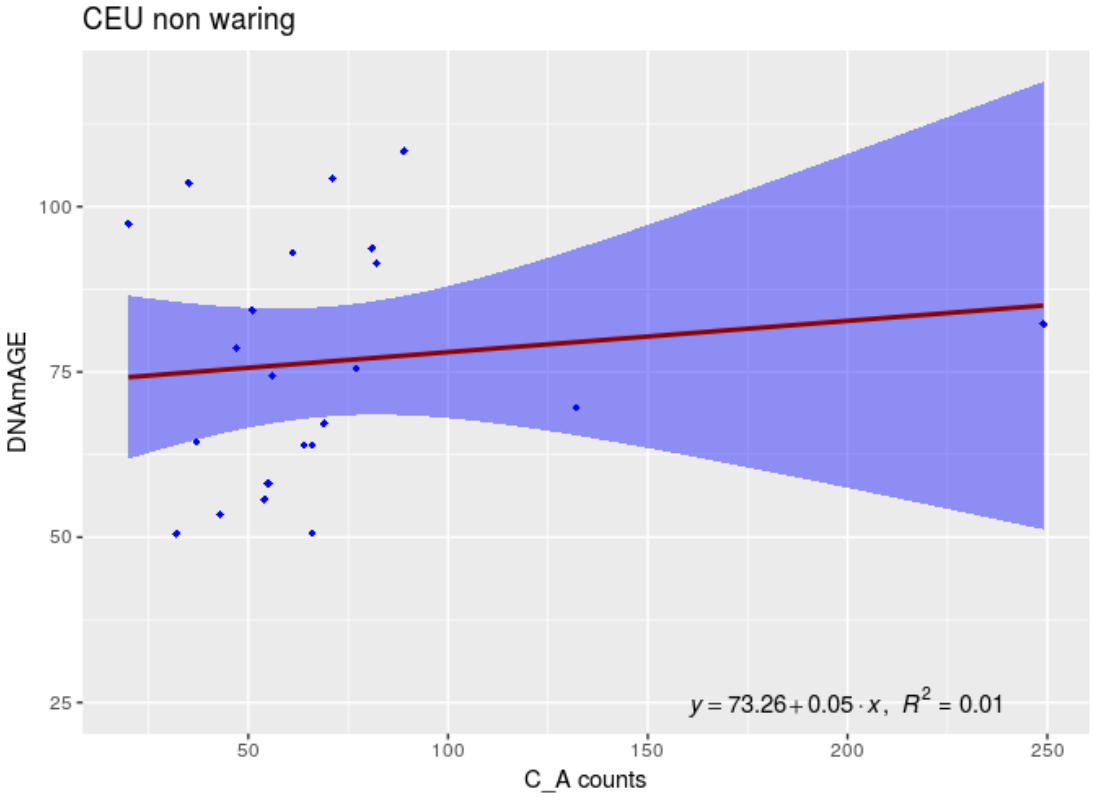

Figure 2: Linear model for the non-warning CEU counts vs DNAm

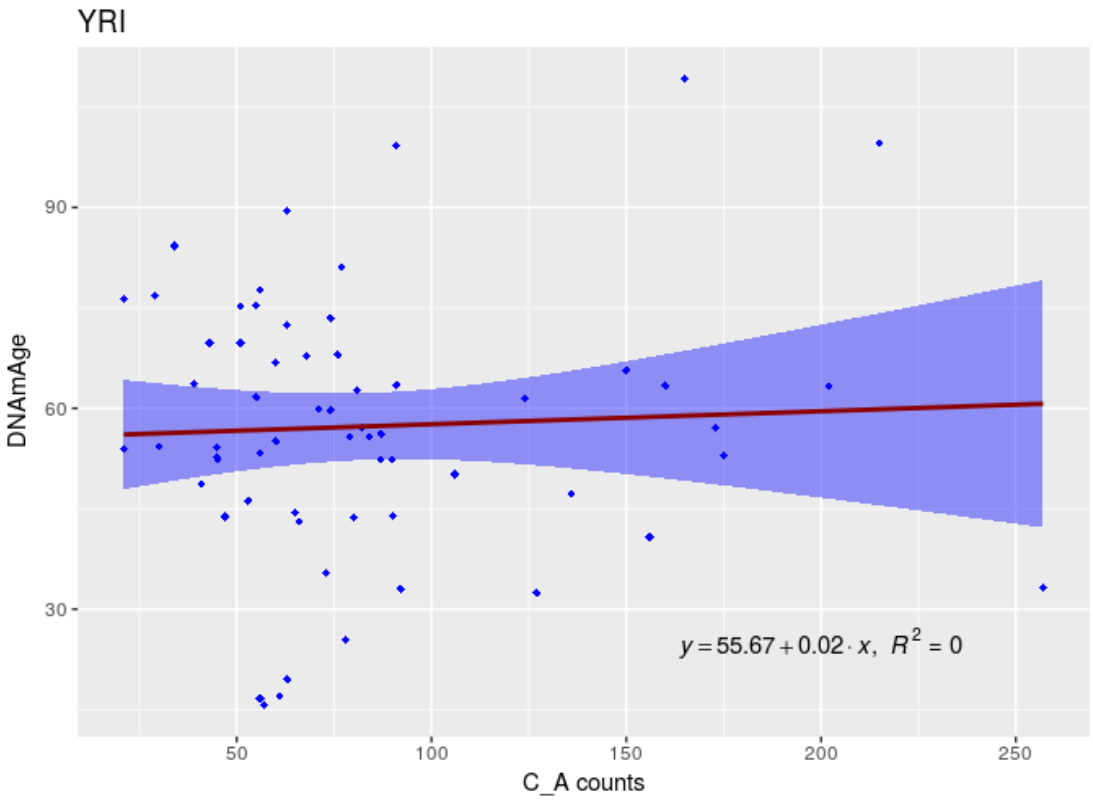

Figure 3: Linear model for the full YRI counts vs DNAm

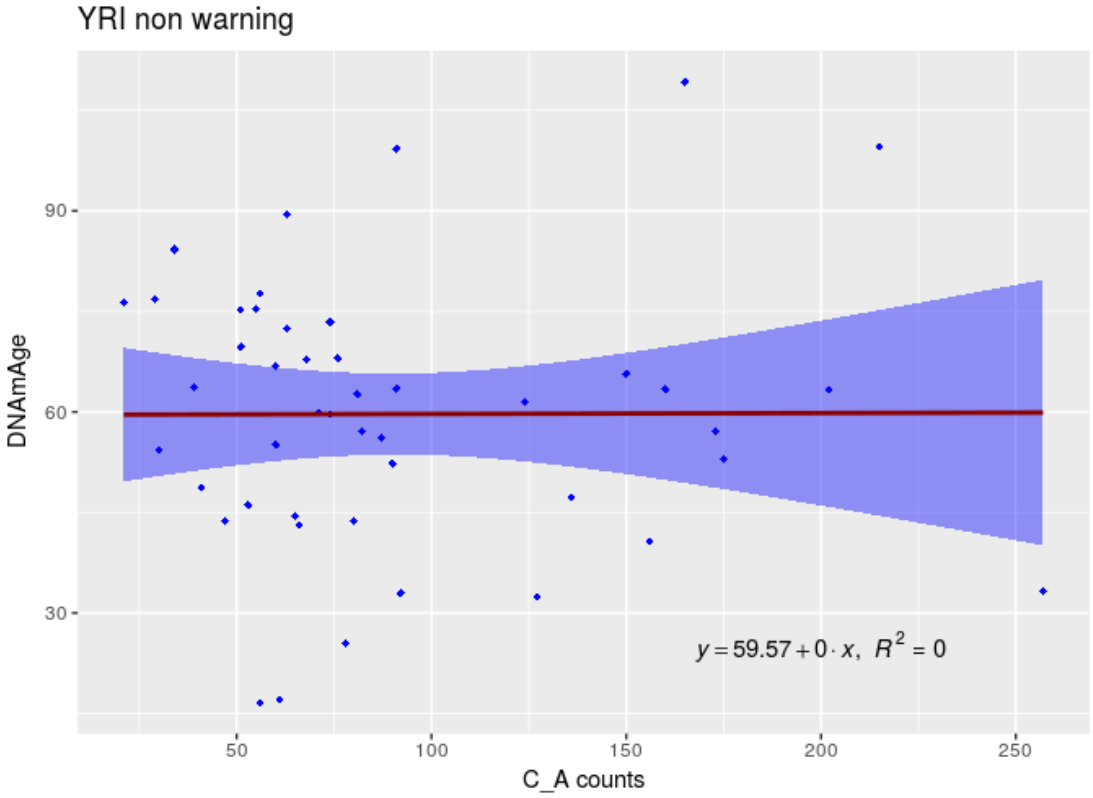

Figure 4: Linear model for the non-warning YRI counts vs DNAm